JADAVPUR UNIVERSITY
KOLKATA-700032

GENOMIC DETECTION OF BACTERIAL PROMOTERS AND MICROSATELLITES IN ANCIENT HUMANS AND THE CORONAVIRUS FAMILY.

The thesis is submitted in the partial fulfilment of the requirements of
Masters of Technology in Computer Technology
in the
Department of Computer Science and Engineering
by

Poulami Ghosh.
University Roll Number: 001910504003.
Examination Roll Number: **M6TCT22005.**
Registration Number: 149838 of 2019-20.

Under the supervision of

Dr. Anasua Sarkar, Assistant Professor
Department of Computer Science Engineering
Jadavpur University, Kolkata-700032.

Faculty of Engineering and Technology
Jadavpur University

Certificate of Recommendation

This is to certify that the dissertation entitled "**GENOMIC DETECTION OF BACTERIAL PROMOTERS AND MICROSATELLITES IN ANCIENT HUMANS AND THE CORONAVIRUS FAMILY."** has been conducted by Poulami Ghosh (University Roll No. **001910504003,** Examination Roll Number: **M6TCT22005**, Registration Number: **149838 of 2019-20**), under the guidance and supervision of Dr Anasua Sarkar, Department of Computer science and Technology, Jadavpur University, Kolkata, is being presented for partial fulfilment of the degree of Masters of Technology in Computer Technology during the academic year 2021-22. The research results in the thesis have not been included in any of the papers submitted to award any degree in any other university or institute.

_____

(Signature of thesis supervisor)
Dr. Anasua Sarkar,
Assistant Professor,
Department of Computer Science and Engineering
Jadavpur University, Kolkata- 700032.

Countersigned,

_____

Prof.(Dr.) Anupam Sinha.
Head of the Department,
Department of Computer Science and Engineering.
Jadavpur University, Kolkata-700032.

_____

Prof. Chandan Mazumdar.
Dean,
Faculty of Engineering and Technology
Jadavpur University, Kolkata-700032.

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

Certificate of Approval

(only in case the thesis is approved)

This is to certify that the dissertation entitled "**GENOMIC DETECTION OF BACTERIAL PROMOTERS AND MICROSATELLITES IN ANCIENT HUMANS AND THE CORONAVIRUS FAMILY."** is a bonafide record of the work carried by Poulami Ghosh (University Roll No. **001910504003,** Examination Roll Number: **M6TCT22005**, Registration Number: **149838 of 2019-20**) in partial fulfilment of the requirement for the M.Tech degree in Computer Technology in the Department of Computer Science and Engineering of Jadavpur University during the period of August 2021 to July 2022. By this approval, the undersigned does not necessarily endorse or approve any statement made, opinion expressed, or conclusion drawn therein but supports the thesis only for its submitted purpose.

_____

Signature of the External Examiner
Date:

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

Declaration of Originality and Compliance of Academic Ethics

I now declare that this thesis entitled "**GENOMIC DETECTION OF BACTERIAL PROMOTERS AND MICROSATELLITES IN ANCIENT HUMANS AND THE CORONAVIRUS FAMILY."** contains a literature survey and original research work by the undersigned candidate, as a part of her M.Tech degree in Computer Technology.

All information in this document has been obtained and presented by academic rules and ethical conduct.

I have fully cited and referenced all materials and results that are not original to this work as these rules and conduct requirements.

Name (in block): POULAMI GHOSH.
University Roll No. **001910504003.**
Examination Roll Number: **M6TCT22005**.
Registration Number: **149838 of 2019-20.**

Thesis Title: **"GENOMIC DETECTION OF BACTERIAL PROMOTERS AND MICROSATELLITES IN ANCIENT HUMANS AND THE CORONAVIRUS FAMILY."**

Signature with date:

ACKNOWLEDGEMENT

First and foremost, I would like to express my deepest gratitude and sincere thanks to my respected supervisor, Dr. Anasua Sarkar, SMIEEE, Assistant Professor, Department of Computer Science and Engineering, Jadavpur University, allowing me to do research and for providing valuable guidance throughout. Her dynamism, vision, sincerity and motivation have deeply inspired me. She has taught me the methodology for carrying out the research and presenting it. It was a great privilege and honour. The above words are just a token of my profound respect towards her for all she has done to give my thesis work a present shape.

I am incredibly thankful to Dr. Sucheta Tripathy, Senior Principal Scientist, Structural Biology & Bioinformatics, collaborator for my thesis work from CSIR-IICB, Kolkata. I am not only grateful for her uplifting words but also for the knowledge that she had imparted during my visits to the lab.

I would express my gratitude and thanks to my parents for their unconditional love, support and guidance. Besides, I am also thankful to my seniors, Neelotpal da, Dipayan da, Rudrajit da and Debasmita Di, for their constant guidance and imparting of knowledge that was helpful for my research. And also, my friends, Anindya, Sankha, Aanzil and Debasis, I thank you wholeheartedly for accompanying me patiently on this research journey.

Last but not least, I would thank all my teachers and well-wishers who have always encouraged me to grow to achieve my aim, as your good words and wishes were fuel during my research work journey.

Regards,
Poulami Ghosh.
Department of Computer Science and Engineering,
Jadavpur University,
Kolkata-700032.

Contents.

1 Introduction

Deoxyribonucleic acid, abbreviated as DNA, is the fundamental unit of the block of life. The line between living and non-living is very vague, i.e., the point where some chemical reaction changes into a biological functioning is challenging to decipher. Understanding the basic unit of life is essential to understanding life as a whole, i.e., the gene. Since the birth of science, scientists and thinkers have always contemplated life on earth. History tells us about the several experimentations conducted, hypotheses were proposed on the building blocks of life and the cause of heredity. Several theories were proposed and refuted from Pythagoras to Plato, Doppler to Mendel. Later in the 1900s, a brilliant scientist named Rosalind Franklin discovered the molecular structure of DNA. Unfortunately, her contribution to the discovery remained highly unnoticed. In the later years, the DNA double helical structure was proposed by Watson and Creek.[1]

The advancement of computer science and technology reached a new peak. Its applications are used in different fields of science. Hence, the era of computational biology and bioinformatics was born. Considering different branches of computational biology, computational genomics fascinates me the most. It deals with the application of computational tools to genomics data. [4]

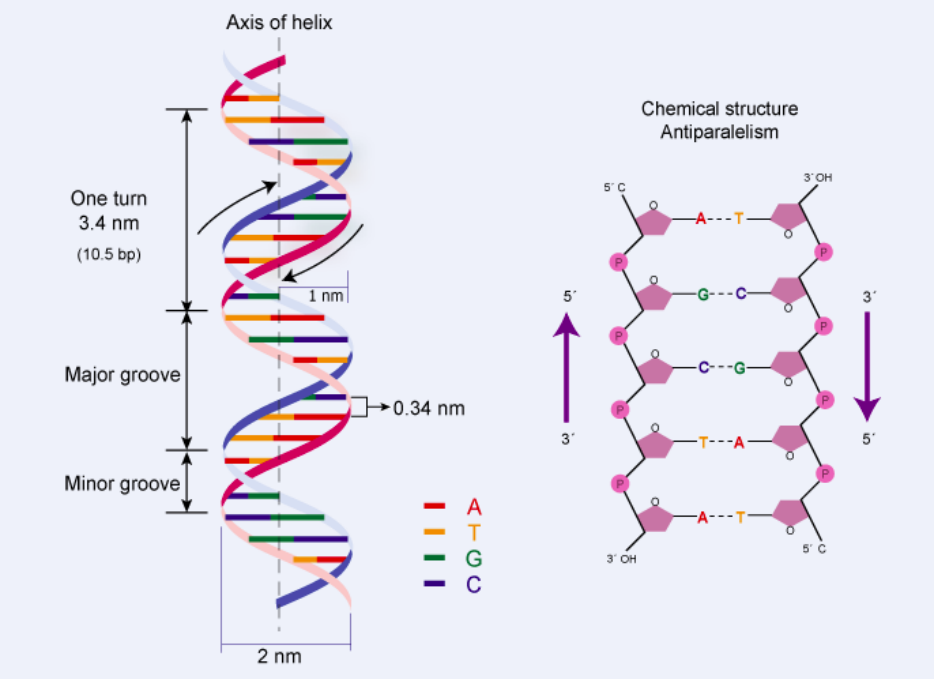

Fig.1 DNA double helix structure.[https://www.toppr.com/ask/question/what-is-nucleosome-draw-diagram-of-double-stranded-polynucleotidechain/]

Mitochondria is rightly known as the cell's powerhouse, as it converts food particles to energy, which can be used for different biochemical functions in our body. Thousands of mitochondria

are present in each cell in our body located inside the nucleus's cytoplasm. We know that DNA is present within the nucleus in the form of chromosomes. A small portion of DNA is also associated with the mitochondria, also known as the mitochondrial DNA, or mtDNA. The human mitochondrial DNA spans about 16,500 base pairs, representing a small fraction of the total DNA strand within the cell. [30][4]

1.1. Human mitochondrial DNA.

The mtDNA is different from the DNA which is present inside the nucleus. It is small in size and circular in shape. It encodes other proteins that are specific for mitochondrial functioning. It is essential to understand that the pathways inside the mitochondria produce energy from the food that we intake. Human mtDNA contains 37 genes, out of which 13 are involved in the process of oxidative phosphorylation, 22 genes encode t-RNA for specific amino acids, and 22 units encode the subunits of 2 ribosomes.[31]

Firstly, we have the control region. The control region contains the signals to control RNA and DNA synthesis in the mitochondria. It is also our fastest evolving DNA sequence. His region of non-coding DNA, also known as the hypervariable region, accumulates mutations at approximately10 times the rate of nuclear DNA. This region can also be called the D-loop. It is also related to a structure formed when the mtDNA replicates.

Complex I gene encode subunits of the protein complex in the mitochondrial membrane. For example, the complex NADH dehydrogenase is an electron transporter involved in the production and storage of energy. Although the genes to make the complex lie within the mtDNA, producing energy requires molecules that are produced outside the mitochondria.

Complex III gene is a part of the electron transport chain of mitochondria, involved in the generation and storage of energy. This mitochondrial gene encodes the cytochrome b protein, a part of the complex in the inner member of a mitochondrion, known as ubiquinol, or cytochrome c oxidoreductase.

Complex IV gene encode protein subunits part of a complex known as the cytochrome c oxidase. This complex acts as an electron transporter while producing and storing energy in the mitochondria.

Complex V gene encode proteins contained within a complex known as ATP synthesis. This protein complex, located in the inner membrane of the mitochondria, acts on electron transportation in the pathway that produces and stores energy.

Ribosomal RNA gene known as 16s and 12s rRNA, encode ribosomal RNAs (rRNAs) used to build ribosomes. Ribosomes assemble and translate the messenger RNA (mRNA)sequence into an assembled protein during protein synthesis. The RNAs produced by these genes build the ‘molecular machinery’ to synthesise proteins encoded by other mtDNA genes.

Transfer RNA gene encodes transfer RNA molecules. Each transfer RNA is associated with a specific amino acid. Transfer RNAs play a vital role in protein synthesis, delivering their amino acids to the ribosomes for incorporation into new proteins.

The human mitochondrial chromosome has vastly reduced with time. Genes for functions that the host could provide, also some genes needed for respiration, were transferred to the nucleus. Millions of years of evolution have resulted in the small mitochondrial chromosomes. During fertilisation, the sperm mitochondria are discarded. The mtDNA is only inherited from the mother within the family, which remains uncombined with the other chromosomes.[30]

1.2. Bacteria

Bacteria are microscopic organisms that exist everywhere on our planet and are essential for our ecosystem. Bacteria, known as extremophiles, also survive in extreme circumstances such as extreme heat and pressure. In the proper biological functioning of the human body, bacteria play a vital role. Most bacteria in our body are harmless, and some are helpful; for instance, in our digestive systems, our gut bacteria are present, which help our body function healthily. However, a minimal class of bacteria causes disease.

Bacteria come in various shapes. They can be in the form of spheres, rods or spirals. The bacteria that cause diseases are known as pathogens. [16]

The whole bacterial genome used as a training dataset in this work is Cyanobacteria.
Cyanobacteria is a phylum that includes photosynthetic bacteria that dwell in aquatic habitats and moist soil. They are known to be the oldest fossils ever found. They were first found at around 3.5 million years old, and they still exist in today's atmosphere. Cyanobacteria can be considered the most senior and most crucial bacterial group on earth. They produce gaseous oxygen as the byproduct of photosynthesis.
Moreover, they are believed to be a part of the great oxygenation event. Some of them fixate nitrogen, and others live singly or in colonies forming filaments or spheres. All living things are methodically arranged into five kingdoms. The cyanobacteria are known to be the Cyanophyta and are one of the kingdom Protista species. Recent discoveries and research have brought light to new changes in the taxonomic positions and led to a unique classification system. Cyanophyta (also known as the blue algae) is now known as cyanobacteria, which falls in the class of bacteria.[16][18]

1.3. Virus

A virus can be considered an infectious microbe consisting of a segment of nucleic acid (DNA or RNA) surrounded by a protein coat. A virus cannot replicate independently. It needs a host cell to infect so it can replicate; in this process, it kills the host cell. Viral infection is prevalent in many organisms, including humans and microorganisms. Understanding how viruses manage themselves while interacting with human cells is vital to understanding human diseases. However, humans get infected by a tiny percentage of viruses on this planet. The viruses that are found in abundance are the ones which infect bacteria. They are known as phages. Genetic

material and proteins are the main ingredients of the inner biology of a virus. They use either DNA, RNA or both. Viral genomes are also found in different shapes and sizes. Their shape s much smaller than the genome of cellular organisms. When a cell gets infected by a virus at the microscopic level, the virus starts replicating itself inside the cell. A viral lifecycle is several steps by which a virus recognises and captures a host cell. Then, it reencodes the cellular DNA with the viral DNA to create its replica from the host cell's resources. Viral infection happens in the following steps:

- *Attachment*. The virus recognises and binds with the host cell.
- *Entry.* The genetic material of the virus is inserted into the cell.
- *Genome replication and gene expression.* The viral DNA reprograms the cellular DNA for replication with the assistance of cellular resources.
- *Assembly.* New viral copies are accumulated from the genome copies and viral proteins.
- *Exit.* Viral particles exit the host cell to infect other cells in the body.

In this thesis, we have worked with the whole genome of a few coronaviruses, which are described in the upcoming paragraphs.

Coronaviruses are a highly diversified family of RNA viruses that are enveloped with positive-sense, single-strands. They infect mammals, including humans and avian species. They belong to the order of *Nedovirales* and the suborder *Coronavirineae* which lies in the class *Coronaviridae.* Further, they are subdivided into subfamily of *Orthocoronavirinae* which consists of four genera: *alpha coronavirus, beta coronavirus, gamma coronavirus* and *delta coronavirus*. The *alpha* and *beta* coronaviruses infect mammals only, whereas *gamma* and *delta* coronaviruses infect the avian class. Coronavirus infection primarily causes respiratory or enteric diseases. [15]

If we look into the biology of a coronavirus virion, it consists of spike proteins (S), envelope (E), membrane (M), and nucleocapsid (N). An encapsulation by N is found in a positive-sense, single-stranded RNA genome (+ssRNA). Bound the strings of RNA are the nucleoproteins; they help give the virus its structure and enable it to replicate. The RNA genome encapsulates the viral envelope of lipids, a waxy barrier containing fat molecules protecting the virus’s precious genetic material. It protects the virus outside the host cell and anchors the structure needed to infect a cell. Envelope proteins embedded in this layer aid the assembly of new virus particles once it has infected the cell. The bulbous projection seen outside the coronavirus is called the spike protein (S). This projection gives the virus its crown-like appearance and the ‘corona’ moniker. They act as gaping hooks, allowing the virus to latch onto the host cells and crack them open for infection.[13]

2. Chapter 1: Characterization of Simple Sequence Repeats: Evolutionary Implications from Ancient Human Mitochondrial Genome.

2.1 Introduction

If we travel back in time, Archaic Homo sapiens originated and spread through the Afro-Asian subcontinent during the Pleistocene period. The different subspecies of Archaic humans are the Neanderthals, Denisovans and the Heidelburgensis. The modern Homo sapien emerged close to 20,000 to 45,000 years ago, assumingly in Africa. We have also implied that the Homo neanderthalensis or Neanderthals emerged in Africa, and Homo sapiens sapiens emerged in Europe and West Asia around the same period. It has been discovered that the Neanderthal-derived DNA can be found in the DNA of possibly all contemporary populations, which vary regionally. The Denisovan ancestry is also called Homo sapiens Altai, which had emerged during the lower and middle Paleolithic periods. The first identification of Denisovan individuals occurs in 2010 based on the mitochondrial DNA extracted from a juvenile female finger bone from the Siberian Desinovan cave in the Altai Mountains. The DNA indicates a close affinity with the Neanderthal DNA sequence[7]. Another sub-species of the Archaic Homo sapiens is the Homo heidelbergensis. It is considered a dynamic species that had evolved from an African form of Homo erectus. It is the most recent common ancestor between modern humans and Neanderthals[8].[31]

This study will consider the mitochondrial DNA genomes of the sub-species mentioned above of the tribe Homo and will discuss their distributions as well as in Simple Sequence Repeats characterization, which may imply their evolutionary tree and for humans to escape from extinction.

2.1.1. The mitochondrial DNA

Mitochondria is an organelle found in all organisms with nuclei releasing energy from food and storing it as ATP (Adenosine triphosphate) molecules. Mitochondria has its DNA. It was once a free-living bacteria that ancestors of eukaryotes engulfed but could not digest. Bacterial enlarged cells established a symbiotic relationship in which the larger cell protected while bacteria produced food. As a result, the mitochondrial chromosomes store many features like bacteria. They are circular molecules. The human mitochondrial chromosome is around 16,546 base pairs long which is approximate to the size of any bacterial plasmid. Unlike nuclear chromosomes, the

mitochondrial chromosome is packed tightly with genes, with a significant region for non-coding DNA between genes. Most mitochondrial genes lack introns which are non-coding information of nuclear genes. Human mtDNA contains 37 genes. and 13 of them are involved in processes of oxidative phosphorylation. Twenty-two genes code for tRNA for specific amino acids, whereas 2 of them regulate the sub-unit of ribosomes. The human mitochondrial chromosome has been vastly reduced. Over time, the functionality of genes that the host could provide was lost. Also, some genes needed for respiration were transferred to the nucleus. As a result, millions of years of evolution have resulted in small mitochondrial chromosomes. During fertilization, when an egg and sperm unite, the sperm mitochondria are discarded. MtDNA is, therefore, only inherited from the mother within a family. Unlike the DNA of the nucleus, the mtDNA does not recombine with other chromosomes. Consequently, we can use them to understand the ancestry and evolution of a species [1][2].

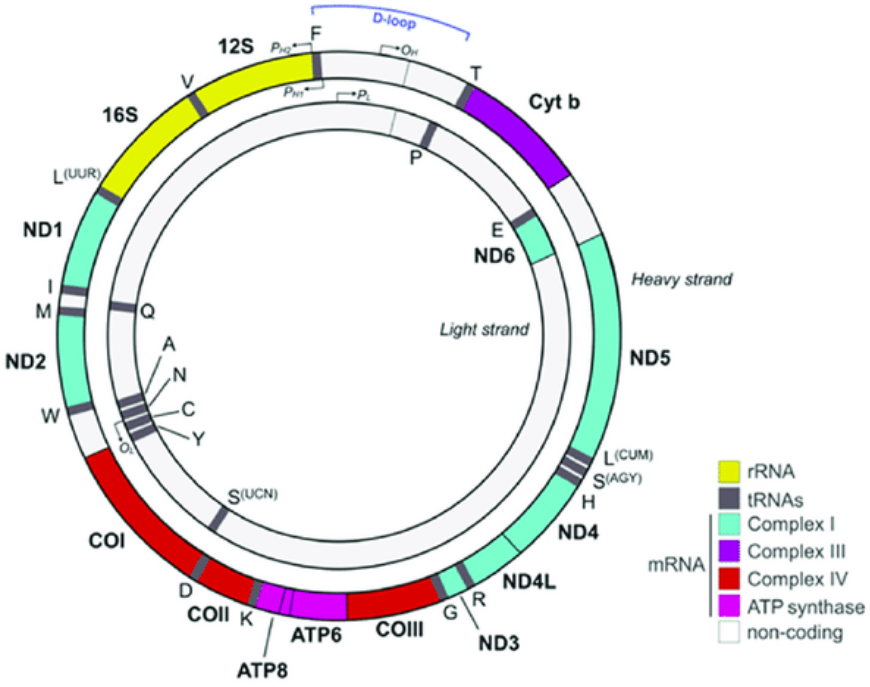

Fig.2. Structure of human mitochondrial DNA.
[https://www.researchgate.net/figure/Human-mitochondrial-DNA-mtDNA-The-mtDNA-consists-of-a-light-inner-and-heavy_fig1_329446024]

2.2. Literature Survey.

The evolution of modern humans is a long and difficult process which started from their first appear- ance and continues to the present day. The study of the genetic origin of populations can help to deter- mine population kinship and to better understand the gradual changes of the gene pool in space and time. Mitochondrial DNA (mtDNA) is a proper tool for the determination of the origin of populations due to its high evolutionary importance. Ancient mito- chondrial DNA

retrieved from museum specimens, archaeological finds and fossil remains can provide direct evidence for population origins and migration processes. Despite the problems with contaminations and authenticity of ancient mitochondrial DNA, there is a developed set of criteria and platforms for obtaining authentic ancient DNA. During the last two decades, the application of different methods and techniques for analysis of ancient mitochondrial DNA gave promising results. Still, the literature is relatively poor with information for the origin of human populations. Using comprehensive phylo- geographic and population analyses we can observe the development and formation of the contemporary populations. The aim of this study was to shed light on human migratory processes and the formation of populations based on available ancient mtDNA data.[30]

The microsatellite search was performed using the IMEx software [31]. Earlier reports on eukaryotes and E. coli genomes focused on assessing microsatel- lites of 12 bp or more (Tóth et al., 2000) but these parameters did not yield any results in carlavirus. The onus for this observation may lie with relatively smaller size of carlavirus genomes. Subsequently, microsatellites from carlavirus genomes were ex- tracted using the 'Advance-Mode' of IMEx using the parameters previously used for HIV (Chen et al., 2012) and potyvirus (Alam et al., 2013) which are as follows: Type of Repeat: perfect; Re- peat Size: all; Minimum Repeat Number: 6, 3, 3, 3, 3, 3; Maxi- mum distance allowed between any two SSRs (dMAX) is 10. Other parameters were used as default. Compound microsatel- lites were not standardized in order to determine real composition.

2.3. Materials and Methods

Whole mitochondrial genome sequences of the early *hominini* subspecies (Neanderthals, Denisovans, *Homo Heidelbergensis* and *Homo sapiens*) were considered. We have downloaded the FASTA files from www.ncbi.nlm.nih.gov. Their reference number is NC_011137.1, NC_012920.1, NC_013993.1, and NC023100.1, respectively. The lengths of those mitochondrial genomes are 16.9kb, 16.9 kb, 16.9 kb, and 16.9 kb, respectively.

2.4. Microsatellites identification and investigation

Simple microsatellites are extracted with the help of a sliding window algorithm. We have computed the program using Python 3.0 programming language with the assistance of the Regular expression package and Biopython modules in the Conda environment in the Jupyter notebook (version: 6.0.3). For finding the GC content, we have used Bio.SeqUtils package from Biopython module. We have also used the standard Sequence Input/Output interface (Bio.SeqIO) for accessing the FASTA files. We have considered repeat type to be Perfect and motif sizes to vary from 6 to 1, i.e., Hexamer, Pentamer, Quadmer, Trimer, Dimer and monomers. The

considered repeat number is 3,3,3,3,3,6 in the same order, and other parameters are set as default. We have also determined the phylogenetic analysis of the given hominini sub-species with the help of the Molecular Evolutionary Genetics Analysis (MEGA) tool. Version 11.0.10.

2.5. Statistical analysis

We have done the statistical analysis and graphical representations with the help of NumPy and matplotlib libraries in Python 3.

2.6. Observations and Results

In Table 1, we see that the number of SSRs in Neanderthal mtDNA is 31, which counts to be the highest among the four sub-species. Therefore they have a higher mutation rate than the other species.

Sub species	Homo sapiens mitochondrion	Homo sapiens neanderthalensis mitochondrion	Homo sapiens Desinova mitochondrion	Homo Sapiens Heidelbergensis mitochondrion
Length of genome	16568	16565	16570	16568
No. of SSR	26	31	25	28
No. of Tri-mer	9	16	9	9
No. of Di-mer	13	11	12	14
No. of Mono-mer	4	4	4	4
GC Content	44.35	44.39	44.3	43.42

Table 1. Overview of the computational results of the whole genome of the four Hominini sub-species.

Relative Abundance (RA) in Fig.3, estimates the biological varieties of a species. RA can be defined as the number of microsatellite repeats divided by genome size in kb. It measures how common or rare a species is comparable to other species in a particular location. According to Fig.1, the Neanderthals had the highest Relative Abundance of the other species in one specific area. We conclude from this context that Neanderthals were the most diverse, and Denisovans were the least varied in a given location among the given species.

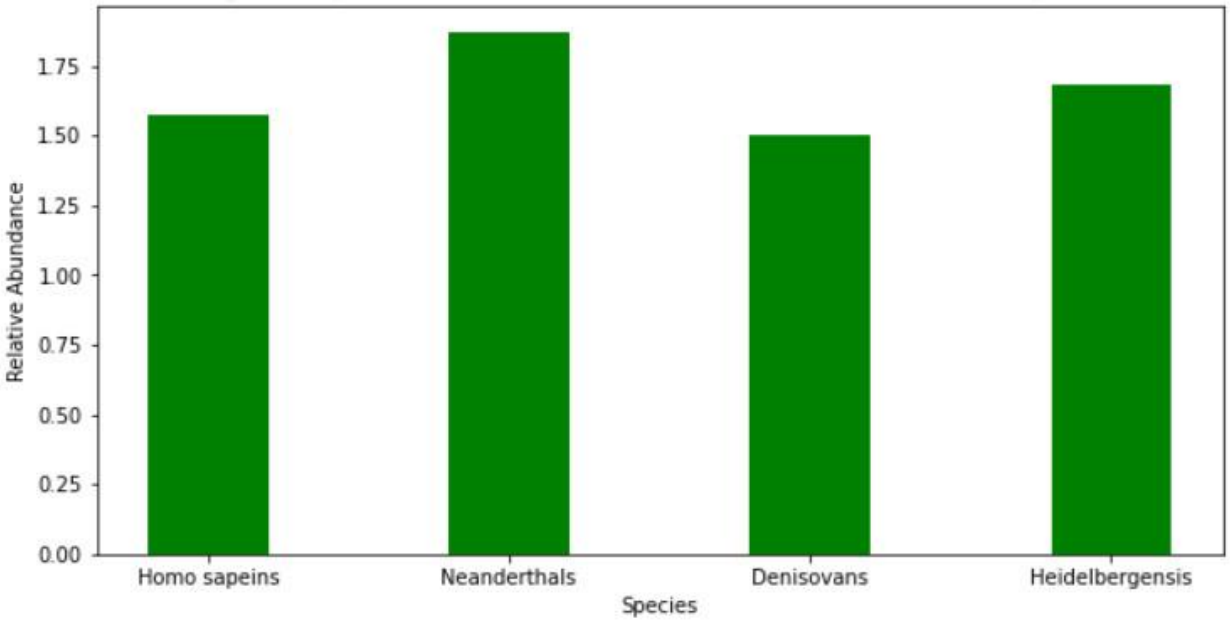

Fig. 3. Comparison of Relative Abundance (RA) of the Hominini sub-species.

Relative Density (RD), shown in Fig.4, measures the total length contributed by each microsatellite upon the length of the genome in kb. It tells us about the number of a given species expressed as a percentage of all species present. Therefore, according to our observation in Fig. _, the relative density of Neanderthals is higher than all the given species, and Denisovans have the lowest relative density among the four species. From this angle, we can conclude that Relative Abundance and Relative Density are directly proportional to each other. Moreover, the percentage of GC content showcases the lifespan of a particular species.

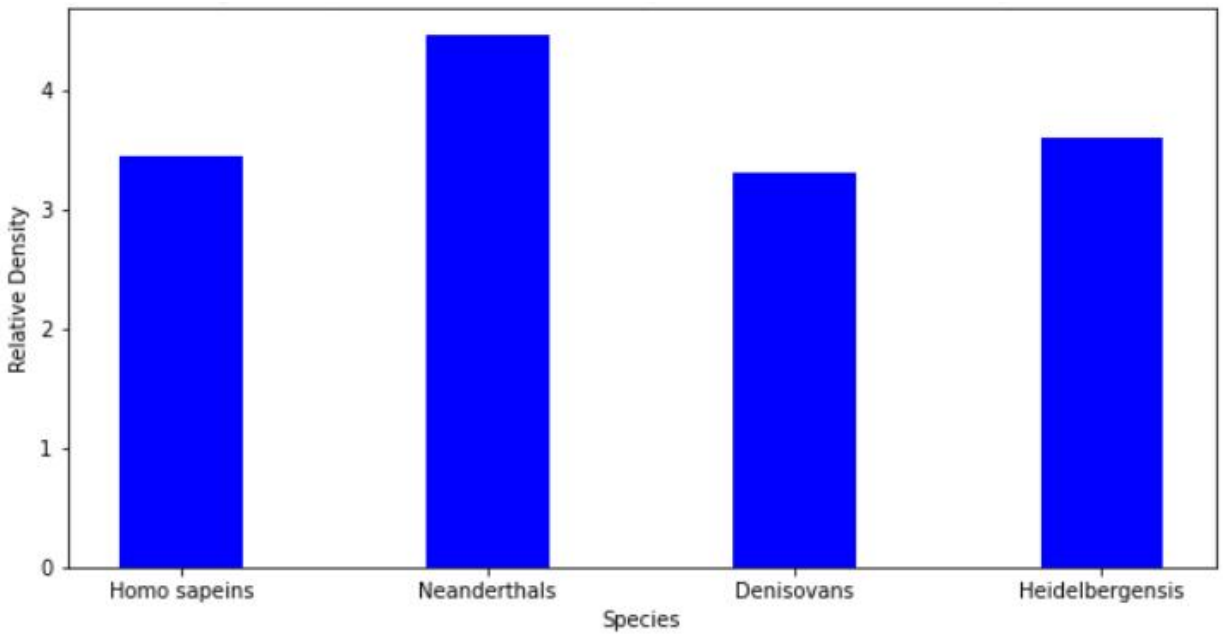

Fig. 4. Comparison of Relative Density (RD) of the Hominini sub-species.

Therefore, according to Table 1, we can conclude that Neanderthals had the longest life span, and Heidelbergensis had the shortest. Fig.4 displays the various simple sequence repeats that show a significant rate of length polymorphism due to mutations of one or more repeat types. The comparison of the common simple sequence repeats in the four species shows that there has been missing and gaining of microsatellites among the four species, which has led to the emergence of one and the extinction of the other.

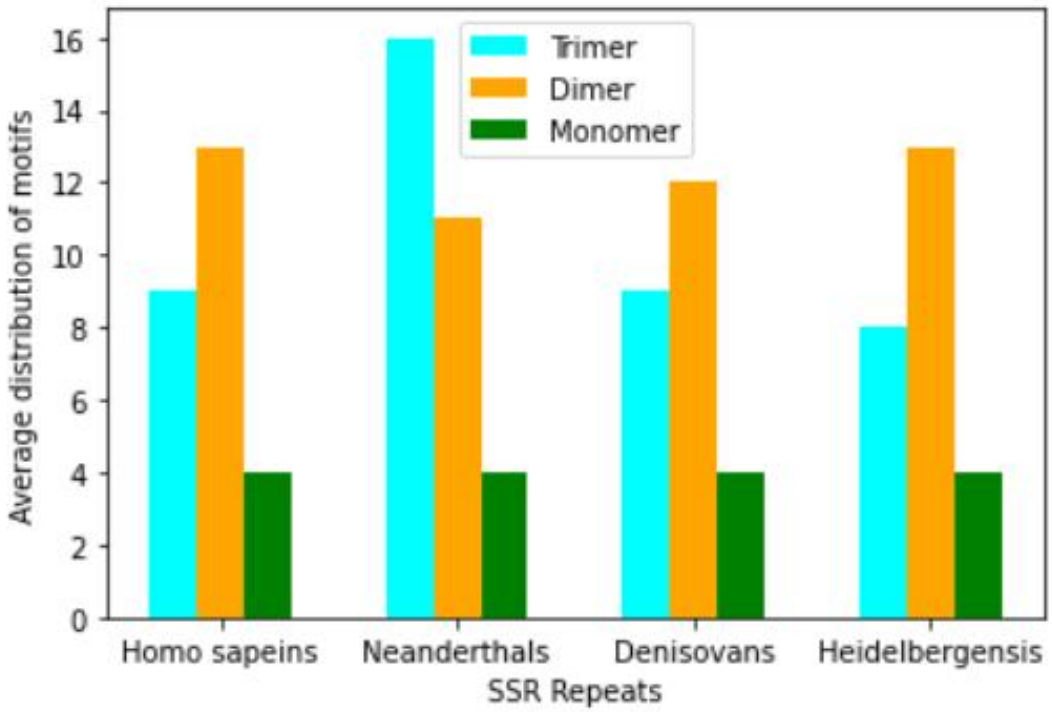

Fig. 5. Observation of SSR distribution in the given Hominini subspecies.

For example, $(ACC)_3$ and $(GGA)_3$ are gained by all the species other than homo sapiens. According to Fig. 5, we can see that the Homo sapiens and the Neanderthals emerged during the same period, and Homo Heidelbergensis is the most recent common ancestor between the two. In contrast, the Homo sapiens Altai (Denisovans) emerged long after their emergence. While performing the phylogenetic analysis, we had to perform sequence alignment of the mentioned sequences. In the process, we observed mutations in the amino acid sequence that had explicitly occurred in the mtDNA sequence of the discussed species, showing some beautiful evolution examples [4].

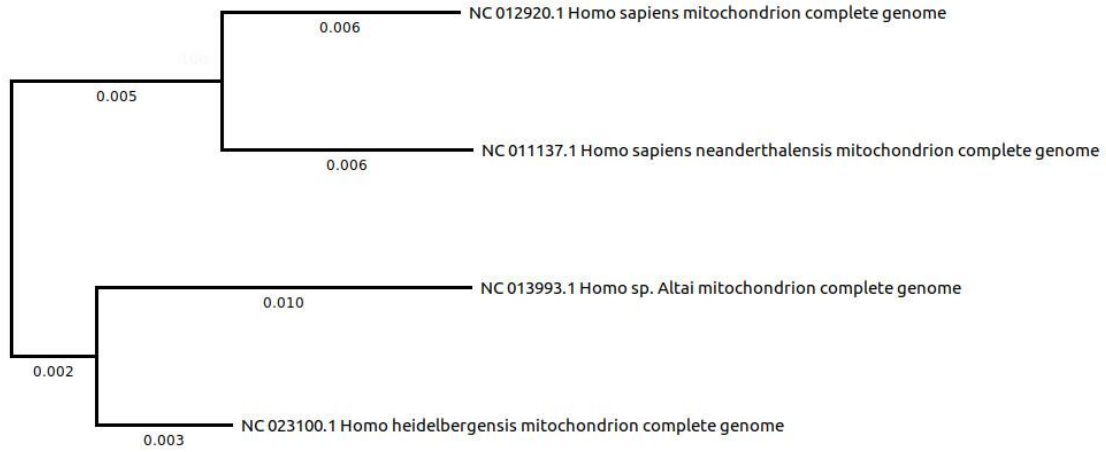

Fig. 6. Observation of the phylogenetic analysis of the hominini sub-species.

Table.1 displays some of those positions and changes in the amino acid. In Fig.7, we observe that the trimer repeat $(ATC)_3$. is gained by Homo sapiens and Denisovans, whereas Neanderthals and Heidelbergensis miss it. We also see that the trimer repeat $(ATA)_3$ is found in Homosapiens, Neanderthals and Heidelbergensis, but it is missing in the mtDNA of Denisovans. The repeats $(CAG)_3$ and $(GCA)_3$ are found in all the subspecies except that of Denisovans. $(CCT)_3$ is found in Homo sapiens and Denisovans, whereas it is missing in Neanderthal and Heidelbergensis. $(ACT)_3$ is found in all the subspecies but Heidelbergensis. Trimer motifs $(GGA)_3$ and $(ACC)_3$ are found in are gained by Neanderthals and Heidelbergensis, whereas missed by the Homo sapiens and Denisovans. $(TTA)_3$ is another repeat that is found in all the subspecies other than the Homo sapiens. $(CAA)_3$, $(GCC)_3$ in Neanderthals and $(CCG)_3$, $(GAG)_3$, $(AAC)_3$ in Denisovans are the repeats exclusively found in their respective mtDNA sequences.

Homo sapiens mitochondrion	A	G	T	T	C	T	T	C	G	T	A	C	A	C
Homo sapiens Neanderthalensis mitochondrion	A	A	T	T	T	T	T	T	A	C	A	T	A	T
Homo sapiens Denisova mitochondrion	G	A	T	T	T	T	C	T	A	T	G	C	A	T
Homo Sapiens Heidelbergensis mitochondrion	A	A	C	C	T	C	T	T	A	T	A	C	G	T
Position:	929	1019	1041	1244	4905	4908	4938	5388	5472	5971	6047	6207	6360	6459

Table 2. Observed changes in protein sequence after sequence alignment.

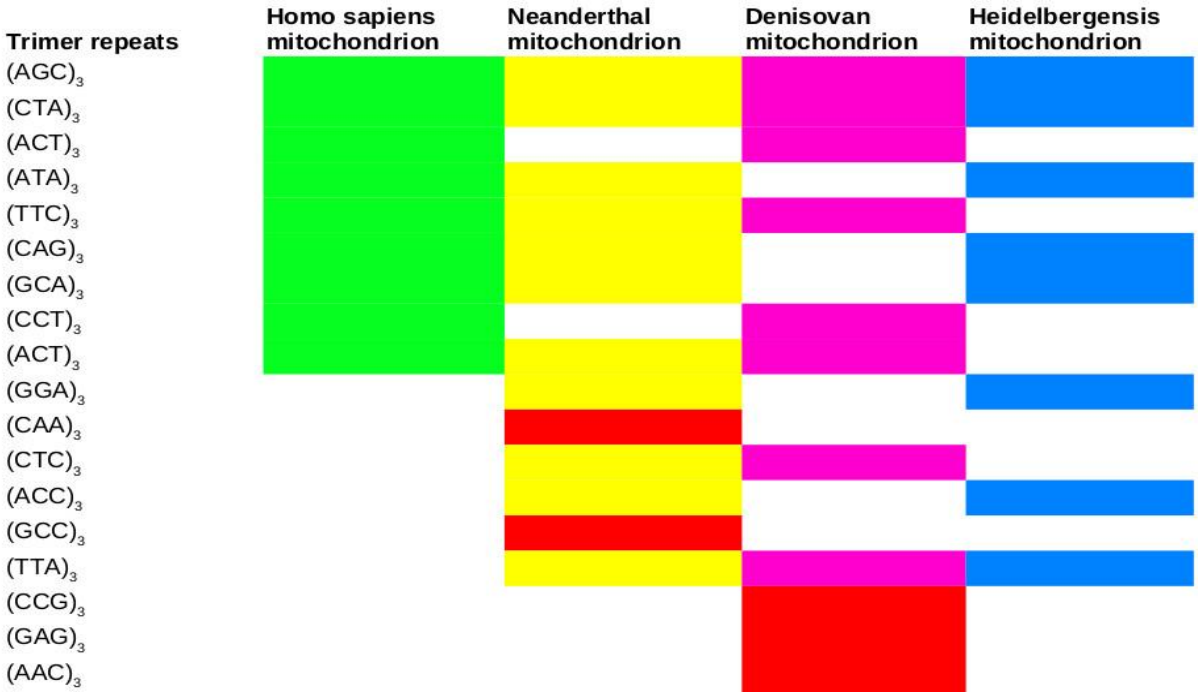

Fig. 7. Representation of the Trimer distribution among the four Hominini sub-species.

Fig.8. represents the changes that have occurred in the SSRs in the subspecies. We observe that most of the Dimers are present in all the mtDNA sequences except $(GA)_2$, which is missing in Neanderthals and Denisovans. Another repeat $(AG)_2$ is present in all the subspecies except the Neanderthals. We have compared our observation with that of the MicroSAtellite identification tool (MISA), and considering the common repeats of all the subspecies, has led us to the conclusion that the repeats $(ATA)_3$, $(ACT)_3$ and $(CCT)_3$ are the SSR markers for Homo sapiens. The presence of $(CCG)_3$, $(CCT)_3$ and $(TTC)_3$ confirm that the mtDNA is of the Denisovan subspecies.

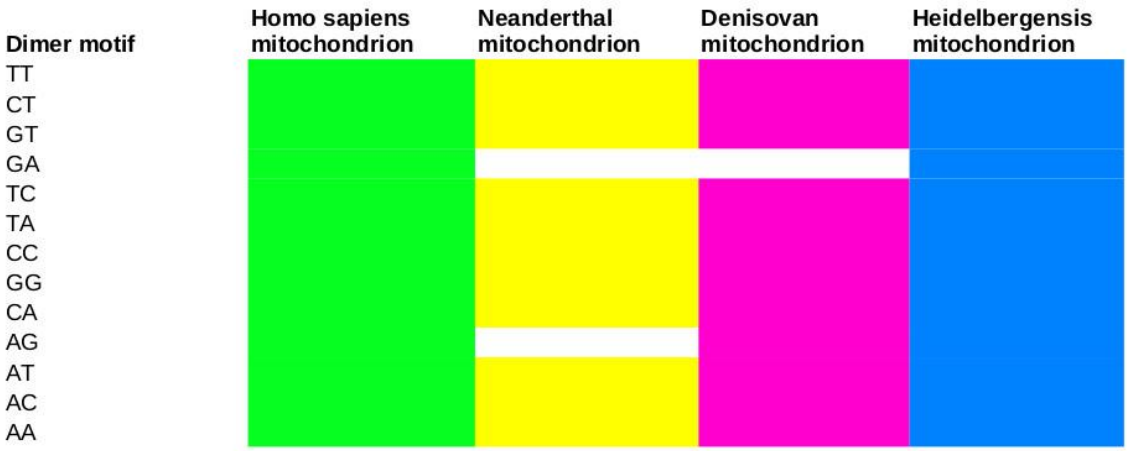

Fig. 8. Representation of the Dimer distribution among the four Hominini sub-species.

The presence of $(CAA)_3$ and $(ACC)_3$ in the mtDNA sequence assures that the mtDNA is of the Neanderthal sub-species, and lastly, the presence of $(GGA)_3$, $(ACC)_3$, and $(AG)_2$ confirms that the mtDNA is of Heidelbergensis.

2.7. Conclusion.

We can identify the different aspects of evolution in the given sub-species of the Hominini tribe by performing the SSRs. From Table. 1., we observe that the SSR count in the mtDNA of the Neanderthals is a maximum of 31, and that of the Desinovans is a minimum. The observation concludes that the Neanderthals have maximum relative density and relative abundance, giving us an idea about the vast diversity of the species across a particular location. The high GC content of the Neanderthals also reveals the fact that the species had the longest life span than the rest. Moreover, by extracting the microsatellite sequences, we have tried to understand the alteration of functionalities in the mtDNA sequences. Working more on these data using advanced computational tools will reveal more information about the prehistoric ancestors of humanity to re-frame the history of evolution in a better way.

3. Chapter 2: Analysis of microsatellite repeated in MERS-CoV, SARS-CoV and SARS-CoV2.

3.1. Introduction.

Viruses are microscopic infectious agents that contain genetic material, either DNA or RNA, and must invade a host in order to multiply. Predominantly, viruses are known for causing disease, as they've triggered widespread outbreaks of illness and death throughout human history. Recent examples of virus-driven outbreaks include the 2014 Ebola outbreak in West Africa, the 2009 swine flu pandemic and the COVID-19 pandemic, which was caused by a coronavirus first identified in late 2019.[17]

When a virus is completely assembled and capable of infection, it is known as a virion. Capsids protect viral nucleic acids from being chewed up and destroyed by special enzymes in the host cell called nucleases. Some viruses have a second protective layer known as the envelope. This layer is usually derived from the cell membrane of a host; little stolen bits that are modified and repurposed for the virus to use. The DNA or RNA found in the inner core constitutes the virus's genome or the sum total of its genetic information. Viral genomes are generally small in size, coding only for essential proteins such as capsid proteins, enzymes and proteins necessary for replication within a host cell.

A virus requires a host cell to replicate, or make more copies of itself, said Jaquelin Dudley, a professor of molecular biosciences at the University of Texas at Austin. "The virus cannot reproduce itself outside the host because it lacks the complicated machinery that a [host] cell possesses," she told Live Science. The host cell's cellular machinery allows viruses to produce RNA from their DNA (a process called transcription) and to build proteins based on the instructions encoded in their RNA (a process called translation).

Therefore, the primary role of a virus is to "deliver its DNA or RNA genome into the host cell so that the genome can be expressed (transcribed and translated) by the host cell," according to "Medical Microbiology." First, viruses break into the host cell, which may be part of a larger organism, in the case of animals and humans. Respiratory passages and open wounds can act as gateways for viruses into the body. Once inside an organism, viruses will then attach themselves to the surface of host cells. They do so by recognizing and binding to cell surface receptors, or proteins that stick off the cell surface; proteins on the viral surface fit onto these receptors like interlocking puzzle pieces. Many different viruses can bind to the same receptor and a single

virus can bind to different cell surface receptors. While viruses use them to their advantage, cell surface receptors are actually designed to serve the cell.

3.1.1. The coronavirus family.

Coronaviruses (CoVs) are a family of viruses that cause respiratory and intestinal illnesses in humans and animals. They usually cause mild colds in people but the emergence of the severe acute respiratory syndrome (SARS) epidemic in China in 2002–2003 and the Middle East respiratory syndrome (MERS) on the Arabian Peninsula in 2012 show they can also cause severe disease.

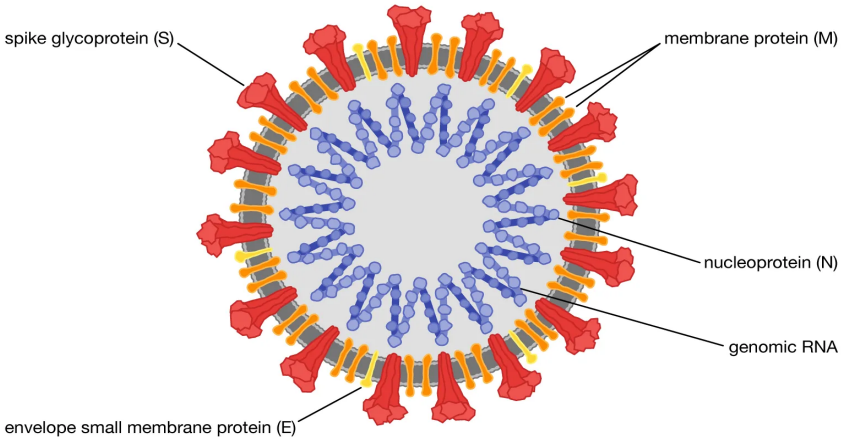

Fig.9. Structure of Coronavirus [https://www.britannica.com/science/Coronavirus-virus-genus]

Since December 2019, the world has been battling another coronavirus. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the virus responsible for the current outbreak of coronavirus disease (COVID-19), which was first identified in Wuhan, China, following reports of serious pneumonia.

3.1.2. Description.

Coronaviruses are relatively simple structures, and their form helps us to understand how they work. They are spherical and coated with spikes of protein. These spikes help the virus bind to and infect healthy cells.

However, the same spikes are also what allow the immune system to 'see' the virus. Bits of the spike can be used in potential coronavirus vaccines to prompt the body to produce antibodies against this new virus.

They are named for the distinctive appearance of their spikes; when seen under a powerful microscope, the spikes look like a crown (corona is the Latin for ‘crown’). Beneath these spikes

is a layer of the membrane. This membrane can be disrupted by detergents and alcohols, which is why soap and water and alcohol hand sanitiser gels are effective against the virus.

Inside the membrane is the virus' genetic material – its genome. Whereas the genomes of some viruses like chickenpox and smallpox are made of DNA like humans, those of coronaviruses are made of closely related RNA. RNA viruses have small genomes which are subject to constant change. These changes, called mutations, help the virus adapt to and infect new host species. It is thought that the new COVID-19 likely originated from bats but it is not yet known whether mutations allowed this jump from animals to humans.

3.1.3. Types of coronaviruses found in humans.

To date, seven human coronaviruses (HCoVs) have been identified (see table below). Four of them are common; less high risk and typically cause only mild respiratory illnesses in healthy human adults. However, they contribute to a third of common cold infections, and, in people with weak immune systems, outbreaks can cause long-term, life-threatening illnesses.

The other three (those causing MERS, SARS and COVID-19 cases) are known to cause more severe illnesses such as shortness of breath and even death. COVID-19 illness tends to be milder than SARS and MERS but more severe than disease caused by the four common coronaviruses.

Human coronavirus names.	Illness.
SARS-CoV-2	COVID-19
SARS-CoV	Severe acute respiratory syndrome (SARS)
MERS-CoV	Middle East respiratory syndrome (MERS)

Table 3. Disease caused by the considered coronavirus family.

3.2. Literature Survey.

An outbreak of atypical pneumonia in Guangdong Province, People's Republic of China, that has continued since November, 2002, is reported to have affected 792 people and caused 31 deaths.1 In adjacent Hong Kong, surveillance of severe atypical pneumonia was heightened in the public hospital network under the Hospital Authority of Hong Kong. [23]By the end of February, 2003, clusters of patients with pneumonia were noted in Hong Kong, along with affected close contacts and health-care workers. The disease did not respond to empirical antimicrobial treatment for acute community-acquired typical or atypical pneumonia.[24] Bacteriological and virological pathogens known to cause pneumonia were not identified. Thus, the new disorder was called severe acute respiratory syndrome (SARS). Subsequently, SARS has spread worldwide to involve patients in North America, Europe, and other Asian countries. We investigated patients

in Hong Kong to try to identify the causal agent.[20] The SARS-CoV was also unusually stable in the environment, more so than other coronaviruses or other respiratory viruses, making infection control in hospitals a challenge. It has been speculated that the enhanced stability of the SARS-CoV at lower temperatures and lower humidity, espe- cially in air-conditioned environments, may help explain the explosive outbreaks that occurred in some regions, compared to others. However, as awareness grew, patients began to be identified and hospitalized earlier in the illness (Leung et al., 2004), and as effective infection control modalities were bet- ter implemented, it became possible to interrupt transmission in the community and in hospitals. Thus, on 5 July 2003, it was possible for the WHO to announce that ‘‘all known chains of hu- man-to-human transmission of the SARS virus now appear to be broken’’. Such an outcome could hardly have been imagined in the dark days of March–April, when, for example, an unprece- dented cluster of around 300 cases emerged over a few days in the Amoy Gardens housing estate in Hong Kong[18].

3.3. Materials and Methods.

3.2.1. Genome sequences.

We have considered several viruses from the coronavirus family. We have provided their details in the following table.

MERS:

S No.	Accession number	Organism	Country of Origin	Collection Date	Genome size (bp)
CoV1	KT861627	Human betacoronavirus 2c Jordon-N3/2012	Jordan	22-Jan-2014	30089
CoV2	KU710264	Middle East Respiratory Syndrome coronavirus (MERS-CoV)	Saudi Arabia.	04-Nov-2014	30096
CoV3	KU851859	Middle East respiratory syndrome-related coronavirus (MERS-CoV)	Saudi Arabia	12-Jul-2015	30076
CoV4	KU851864	Middle East respiratory syndrome-related coronavirus (MERS-CoV)	Saudi Arabia	24-Aug-2015	30096
CoV5	KX034096	Middle East respiratory syndrome-related coronavirus (MERS-CoV)	South Korea	17-Jun-2015	30082
CoV6	KY688120	Middle East respiratory	Saudi Arabia	10-May-2015	30102

		syndrome-related coronavirus (MERS-CoV)			
CoV7	KY581693	Middle East respiratory syndrome-related coronavirus (MERS-CoV)	The United Arab Emirates	20-Apr-2014	30123
CoV8	OL622035	Middle East respiratory syndrome-related coronavirus (MERS-CoV)	Saudi Arabia, Riyadh	03-Jan-2017	29994
CoV9	OL622036	Middle East respiratory syndrome-related coronavirus (MERS-CoV)	Saudi Arabia, Dammam	03-Apr-2019	29994
CoV10	MT387202	Middle East respiratory syndrome-related coronavirus (MERS-CoV)	South Korea	20-May-2015	30108

Table 4. Details of MERS-CoV

SARS:

S No.	Accession number	Organism	Country of Origin	Collection Date	Genome size (bp)
CoV11	AY274119	SARS coronavirus Tor2	Canada, Toronto	Not found	29751
CoV12	AY278487	SARS coronavirus BJ02	Not found	Not found	29745
CoV13	AY278488	SARS coronavirus BJ01	Not found	Not found	29725
CoV14	AY278489	SARS coronavirus GD01	Not found	Not found	29757
CoV15	AY278490	SARS coronavirus BJ03	Not found	Not found	29740
CoV16	AY278491	SARS coronavirus HKU-39849	Not found	Not found	29742
CoV17	AY278554	SARS coronavirus CUHK-W1	Not found	Not found	29736
CoV18	AY279354	SARS coronavirus BJ04	Not found	Not found	29732
CoV19	AY283794	SARS coronavirus Sin2500	Not found	Not found	29711
CoV20	AY268070	SARS coronavirus Hong Kong/03/2003	Guangdong Province, Republic of China.	Not found	646

Table 5. Details of SARS-CoV

SARS2:

S No.	Accession number	Organism	Country of Origin	Collection Date	Genome size (bp)
CoV22	ON366161	Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)	France, Marseille	2022-03-13	29735
CoV21	ON365967	Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)	France, Marseille	2020-10-09	29464
CoV23	ON366175	Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)	France, Marseille	2022-03-07	29795
CoV24	ON366926	Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)	The USA, Virginia	2022-01-06	29766
CoV25	ON366929	Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)	The USA, Virginia	2022-01-20	29752
CoV26	ON367465	Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)	USA, SC, ALLENDALEE	2022-04-12	29720
CoV27	ON369053	Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)	The USA, North Carolina	2022-04-14	29770
CoV28	ON369589	Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)	USA, Virginia	2022-02-08	29752
CoV29	ON372840	Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)	USA, California	2022-04-14	29727
CoV30	ON373029	Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)	USA, Virginia	2022-04-14	29755

Table 6. Details of SARS-CoV2.

Whole genome sequences of several members of the coronavirus family were considered. We have downloaded the FASTA files from www.ncbi.nlm.nih.gov. The details of the sequences are given in Table 4,5,6.

3.3. Microsatellites identification and investigation

Simple microsatellites are extracted with the help of a sliding window algorithm. We have computed the program using Python 3.0 programming language with the assistance of the Regular expression package and Biopython modules in the Conda environment in the Jupyter notebook (version: 6.0.3). For finding the GC content, we have used Bio.SeqUtils package from Biopython module. We have also used the standard Sequence Input/Output interface (Bio.SeqIO) for accessing the FASTA files. We have considered repeat type to be Perfect and motif sizes to vary from 6 to 1, i.e., Hexamer, Pentamer, Quadmer, Trimer, Dimer and monomers. The considered repeat number is 3,3,3,3,3,6 in the same order, and other parameters are set as default. For graphical analysis and representation, we have used MS Excel.

3.4. Statistical analysis

3.4.1. Comparison of the sequence length of the Coronavirus family.

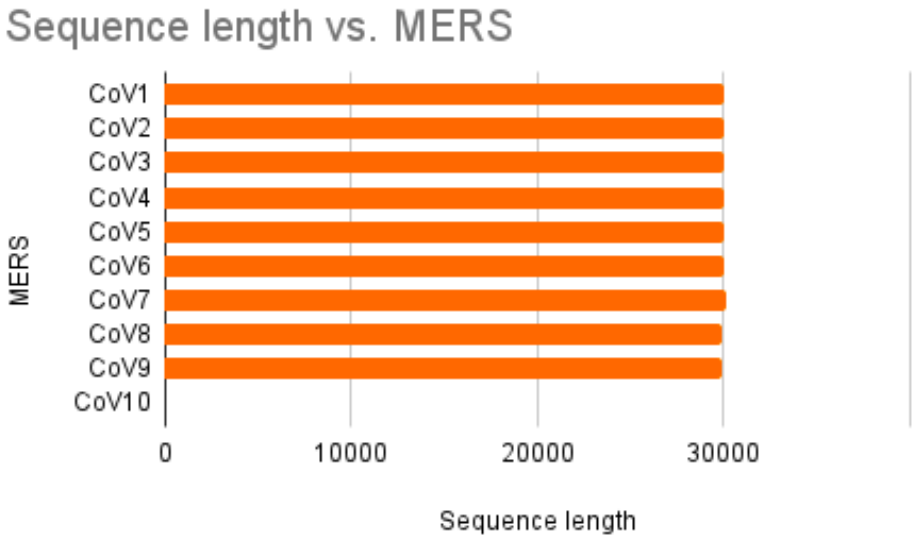

Fig 10. Comparison of the nucleotide sequence length of MERS-CoV.

We have considered the sequence length of 10 MERS-CoV variants whose details can be found in table5. It is observed that the sequence length is around 30,000 bp.

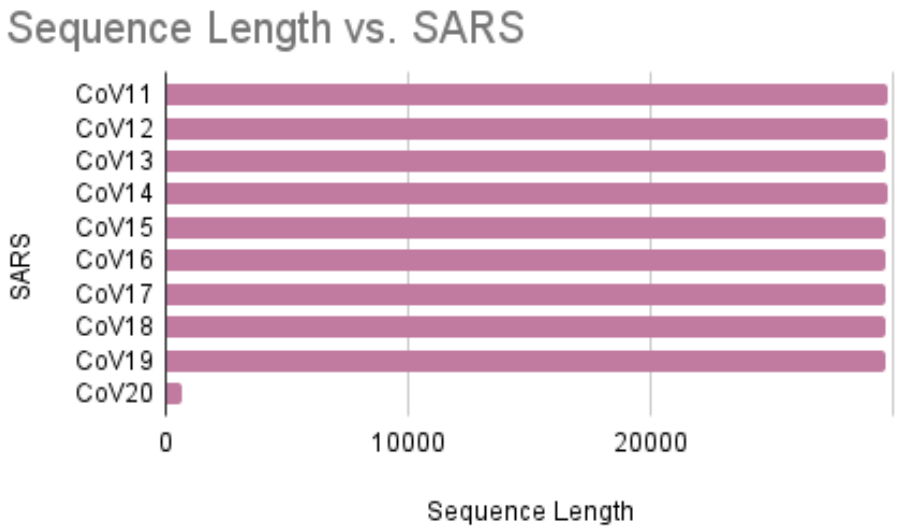

Fig11. Comparison of the nucleotide sequence length of SARS-CoV.

In the case of SARS-CoV, it is observed that the sequence lengths of the variants are shorter if compared with MERS-CoV. They are slightly less than 30,000 bp. We observe that CoV20, the reference of which is given in the table, is exceptionally short if compared with the other sequence length.

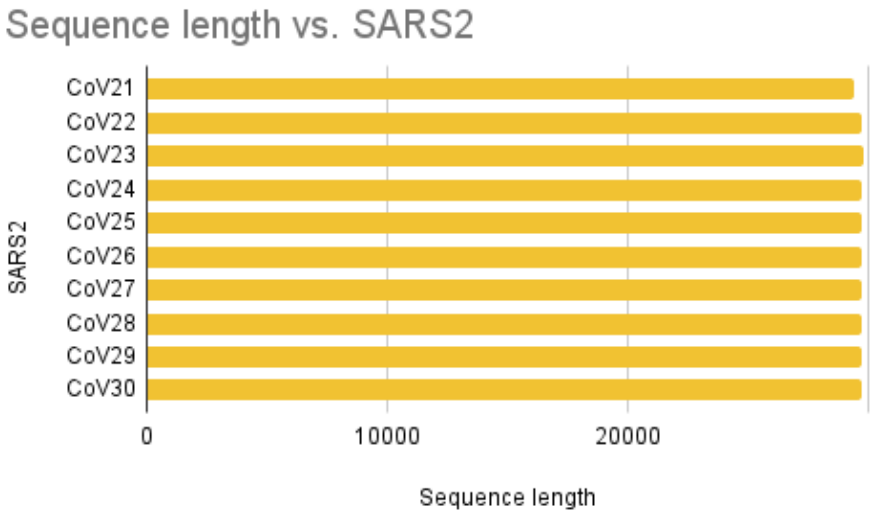

Fig 12. Comparison of the nucleotide sequence length of SARS-CoV2.

We can observe that SARS-CoV and SARS-CoV2 variants have the approximately same length of basepair sequences and they vary largely with that of MERS-CoV. However, the length of the sequence of CoV21 is a bit smaller when compared to other sequences of SARS-CoV2.

3.4.2. Comparison of repeat sequences in the Coronavirus family.

In our computational approach, we have observed that hexamer, pentameter, and quadmer repeats are missing in the sequences. Therefore, we could only make the computational analysis of trimer, dimer and monomer repeats of the Coronavirus variants.

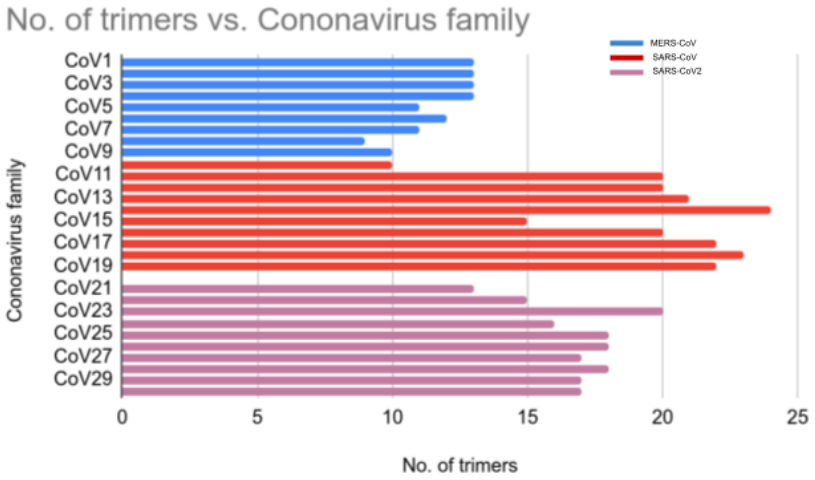

Fig 13. Comparison of trimer repeats in the Coronavirus family.

From Fig.13, we can observe that in the case of the MERS-CoV variant, the trimer repeat is least if compared with that of SARS-CoV and SARS-CoV2. If SARS-CoV and SARS-CoV2 are taken into account, SARS-CoV has the maximum number of trimers. The highest number of trimers are found in CoV14, i.e, 24 and we can also see that CoV20 has no trimers in them. It also belongs to the SARS-CoV variant.[3]

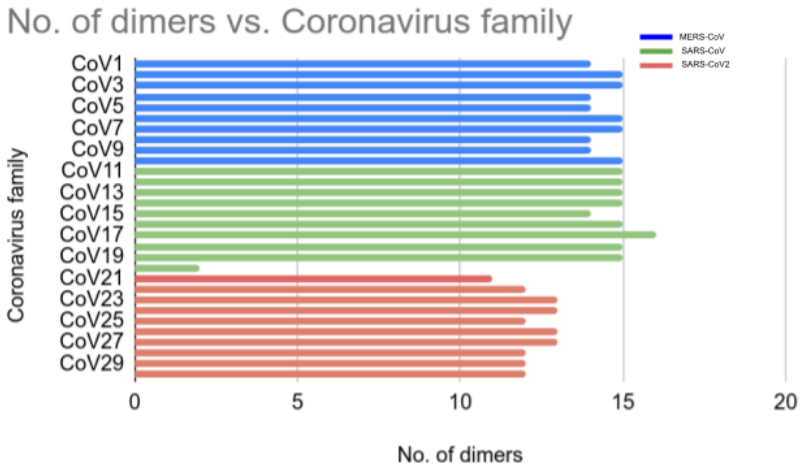

Fig14. Comparison of dimer repeats in the Coronavirus family.

For dimer repeats, we see that SARS-CoV2 has the least number when compared with MERS-CoV and SARS-CoV. MERS-CoV and SARS-CoV have more or less the same number of dimer motifs. CoV 17 has the highest number of dimer motifs 16 and also the least number of 2.[16][20]

In the case of monomer repeats, SARS-CoV ranks the highest. MERS-CoV and SARS-CoV2 have more or less the same number of motifs. The number of monomer repeats found in CoV20 is zero.

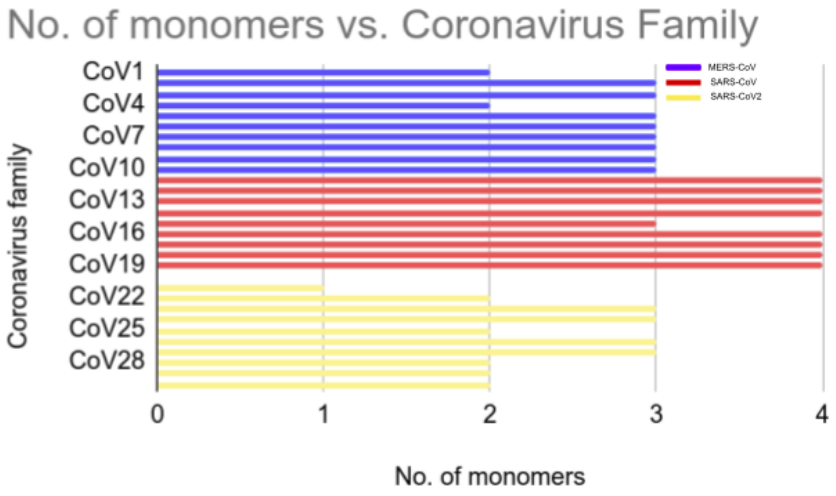

Fig15. Comparison of monomer repeats in the Coronavirus family.

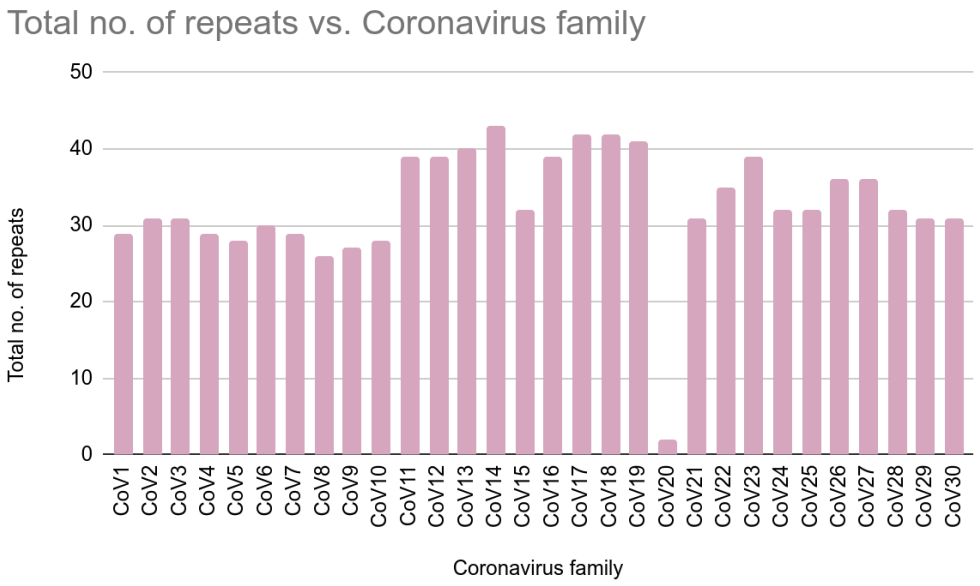

Fig16. An overall observation of short sequence repeats in the considered coronavirus variants.

If we see the scenario for all tandem repeats, we see that SARS-CoV have the most of the counts, and then SARS-CoV2 and MERS have the least of the sequences.[27]

With the help of this information, we can derive the traits of the coronavirus variants which will help us understand their lineage in a better way.

3.4.3. Comparison of Relative Density and Relative Abundance in the Coronavirus family.

The relative abundance of each gene of a simulated metagenome can be defined as the total length of the mapped reads normalised by the gene length. It informs us about the evenness of a particular motif in the entire genome sequence.[9][10]

In the case of MERS-CoV, we see that the relative abundance (RA) of the considered variants is calculated to be around 1.00. There is quite an evenness in the treant other than the drift in the case of CoV8. this brings us to the conclusion that the variant CoV8 has more variance than the other variants of the MERS-CoV family.

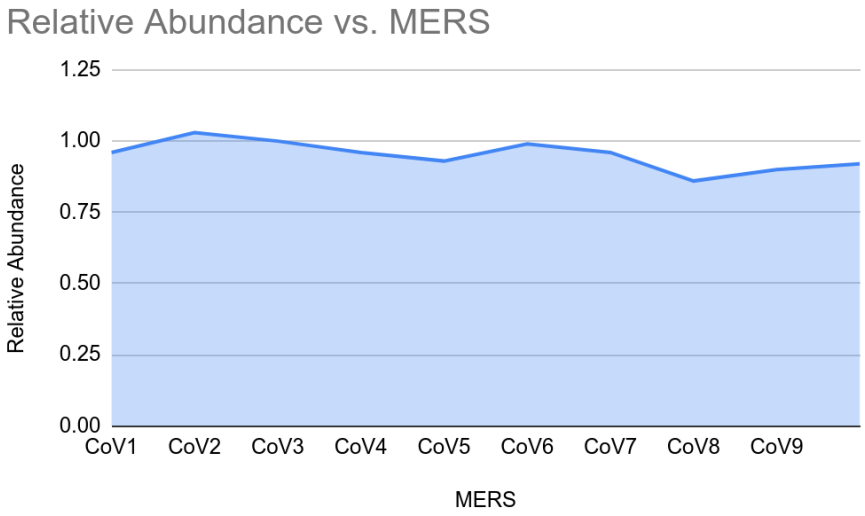

Fig17. Relative abundance graph for MERS-CoV.

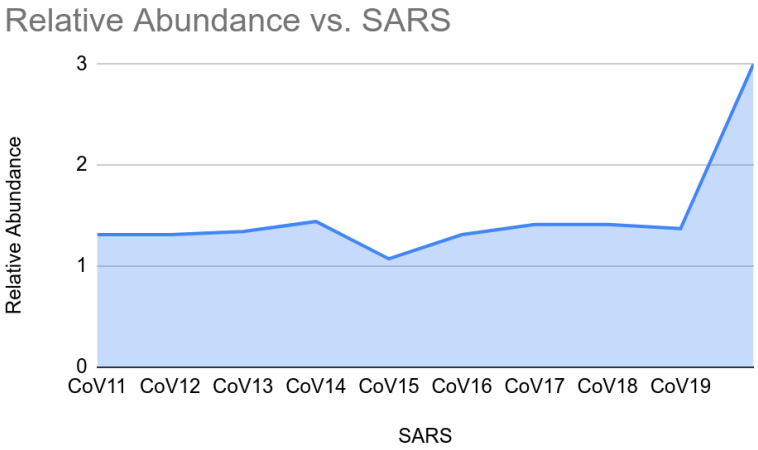

Fig18. Relative abundance graph for MERS-CoV.

In the case of SARS, the RA is not as good as MERS, there is a mild drop at CoV15 and a sharp rise after CoV19. This brings out the fact that the variants are not evenly distributed as MERS-COV.

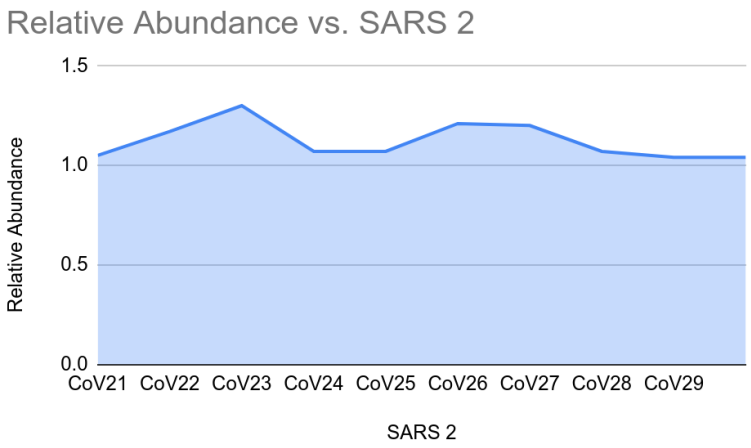

Fig19. Relative abundance graph for SARS-CoV2.

In SARS-CoV2, the RA is seen to be quite uniform, unlike SARS-CoV. Therefore, we can say that the gene distribution in SARS CoV-2 is more even than that of SARS-CoV. Whereas MERS and SARS-COV2 follow quite a similar trend.

The relative density can be defined as the ratio of the number of genes and the total number of base pairs. More precisely it explains how densely populated a particular gene is in the whole genome sequence.[22]

With this inference, and by looking at Fig.20 and 21 we can conclude that the density of the MERS-CoV variant follows a nearly uniform trend. Although they are less dense than SARS-CoV. In the case of SARS-CoV in variant CoV19 we observe a steep rise in the density.[23]

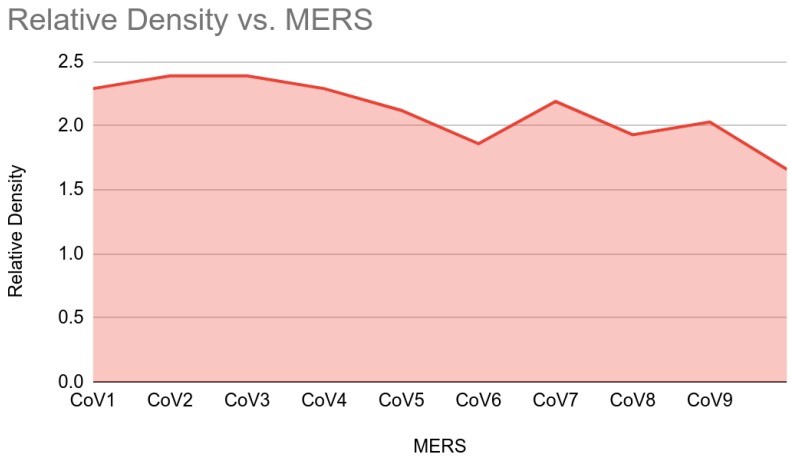

Fig20. Relative density graph for MERS-CoV.

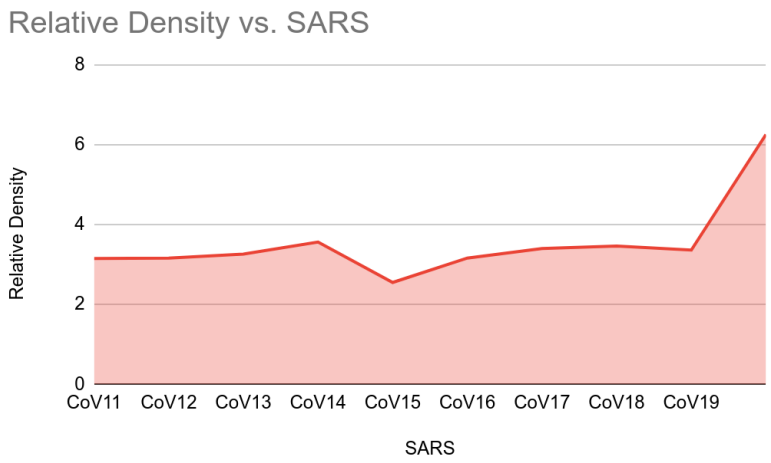

Fig21. Relative density graph for SARS-CoV.

Fig22. shows that in SARS-CoV2, the relative density varies between 2 and 3 making the density quite uniform apart from a slight rise in the case of CoV23.

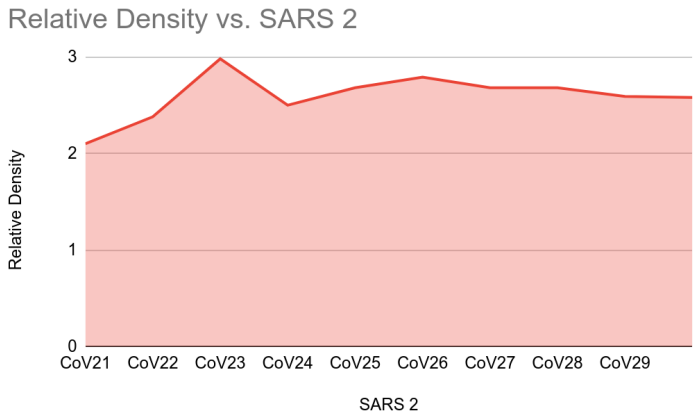

Fig22. Relative density graph for SARS-CoV2.

3.4.4. Discussion on GC content.

In molecular biology and genetics, GC-content (or guanine-cytosine content) is the percentage of nitrogenous bases in a DNA or RNA molecule that are either guanine (G) or cytosine (C). This measure indicates the proportion of G and C bases out of an implied four total bases, also including adenine and thymine in DNA and adenine and uracil in RNA.[26][29]

GC content may be given for a certain fragment of DNA or RNA or for an entire genome. When it refers to a fragment, it may denote the GC-content of an individual gene or section of a gene (domain), a group of genes or gene clusters, a non-coding region, or a synthetic oligonucleotide such as a primer.

Qualitatively, guanine (G) and cytosine (C) undergo a specific hydrogen bonding with each other, whereas adenine (A) binds specifically with thymine (T) in DNA and with uracil (U) in RNA. Quantitatively, each GC base pair is held together by three hydrogen bonds, while AT and AU base pairs are held together by two hydrogen bonds. To emphasize this difference, the base pairings are often represented as "G≡C" versus "A=T" or "A=U".

DNA with low GC-content is less stable than DNA with high GC-content; however, the hydrogen bonds themselves do not have a particularly significant impact on molecular stability, which is instead caused mainly by molecular interactions of base stacking. In spite of the higher thermostability conferred to a nucleic acid with high GC-content, it has been observed that at least some species of bacteria with DNA of high GC-content undergo autolysis more readily, thereby reducing the longevity of the cell *per se*. Because of the thermostability of GC pairs, it was once presumed that high GC content was a necessary adaptation to high temperatures, but this hypothesis was refuted in 2001. Even so, it has been shown that there is a strong correlation between the optimal growth of prokaryotes at higher temperatures and the GC-content of structural RNAs such as ribosomal RNA, transfer RNA, and many other non-coding RNAs. The AU base pairs are less stable than the GC base pairs, making high-GC-content RNA structures more resistant to the effects of high temperatures.

More recently, it has been demonstrated that the most important factor contributing to the thermal stability of double-stranded nucleic acids is actually due to the base stackings of adjacent bases rather than the number of hydrogen bonds between the bases. There is more favourable stacking energy for GC pairs than for AT or AU pairs because of the relative positions of exocyclic groups. Additionally, there is a correlation between the order in which the bases stack and the thermal stability of the molecule as a whole.

GC-content is usually expressed as a percentage value, but sometimes as a ratio (called **G+C ratio** or **GC-ratio**). The GC-content percentage is calculated as:

$$\frac{G+C}{A+T+G+C} \times 100\% \quad \text{.....(1)}$$

Whereas the AT/GC ratio is calculated as:

$$\frac{A+T}{G+C} \quad \text{.....(2)}$$

The GC-content percentages, as well as GC-ratio, can be measured by several means, but one of the simplest methods is to measure the melting temperature of the DNA double helix using spectrophotometry. The absorbance of DNA at a wavelength of 260 nm increases fairly sharply when the double-stranded DNA molecule separates into two single strands when sufficiently heated. The most commonly used protocol for determining GC-ratios uses flow cytometry for

large numbers of samples. The higher the GC content of DNA, the more stable is the double-stranded helical molecule.

The GC content of a gene region can impact its coverage, with regions having 50–60% GC content receiving the highest coverage while regions with high (70–80%) or low (30–40%) GC content having significantly decreased coverage. Yet, adequate coverage of GC-rich regions (which are commonly present in the promoter and first exon of many genes) is necessary for a high analytical sensitivity of a targeted gene panel. [27]

From our research, we see that GC content varing from 40% to 60% is desirable for the stability of the DNA double helix. Whereas in case of GC content < 40% the DNA double helix can lose its stability easily.[15]

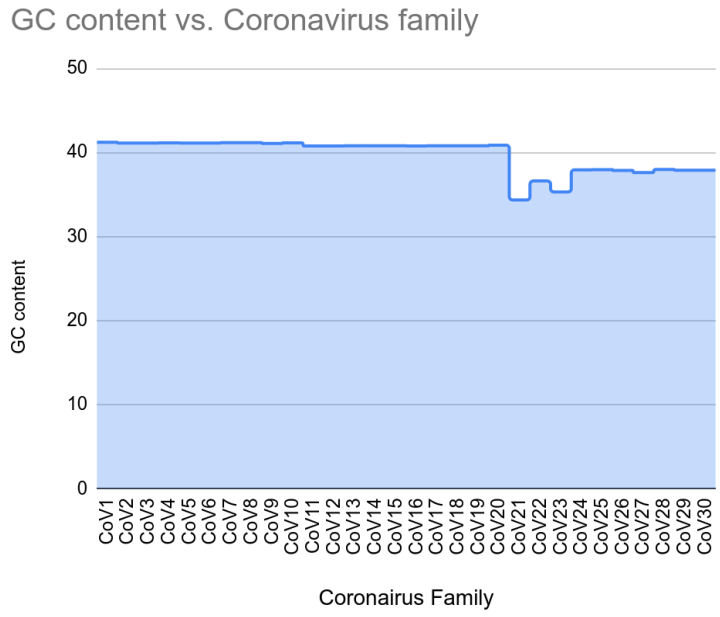

Fig23. Comparison of GC content in the coronavirus family.

Keeping this information in mind, we can conclude from Fig 23. that the GC content ranging from CoV1 to CoV20 is more than 40% which makes the variants more stable. Whereas the variants from CoV21 to CoV30 are less than 40%, CoV21 and CoV24 have even lesser GC content; this shows that they are weakly stable.

3.4.5. Analysis of microsatellites.

In Fig.24, we can see that $(ATC)_3$ and $(CTA)_3$ motifs are repeating elements in MERS-CoV. $(TGA)_3$ and $(GAA)_3$ are SARS-CoV-specific elements. $(AAG)_3$ is common in SARS-CoV and SARS-CoV2. $(TCT)_3$ and $(AGA)_3$ are more dominant in SARS-CoV2. $(TGC)_3$ is mostly

dominant in MERS-CoV, SARS-CoV and SARS-CoV2. Whereas $(TAT)_3$ and $(CCT)_3$ are dominantly present in MERS-CoV.

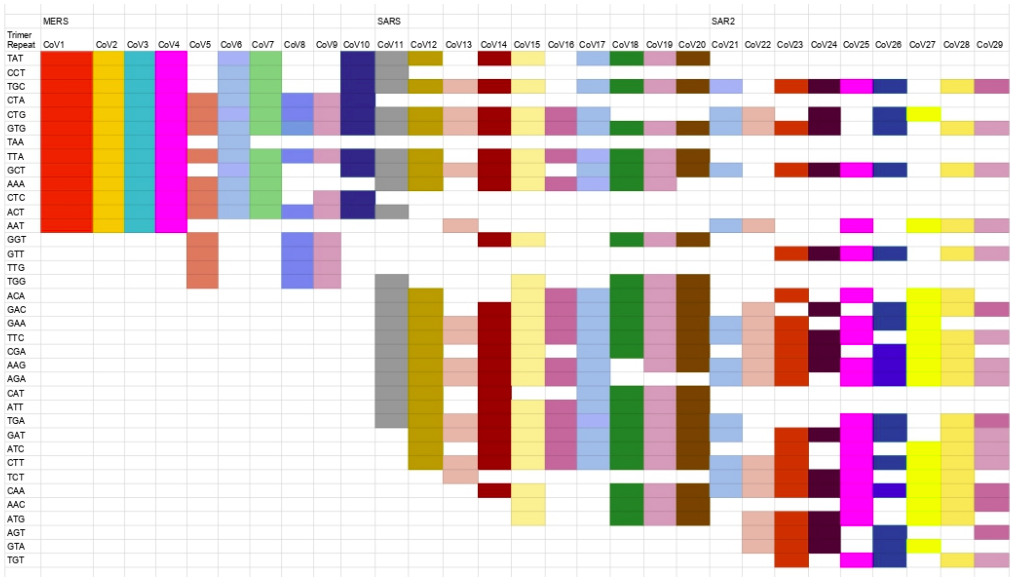

Fig24. Distribution of trimer motifs in the considered coronavirus family.

In Fig.25, $(AT)_2$, $(CT)_2$, $(TA)_2$, $(GA)_2$, $(AG)_2$ and $(AC)_2$ are present in all the three variants. $(TC)_2$ and $(AA)_2$ are present in MERS and SARS-CoV2. $(GG)_2$ is predominantly present in SARS. $(TC)_2$ is scarcely present in SARS but dominantly present in MERS and SARS-CoV2. $(CG)_2$ and $(TG)_2$ are predominantly present in MERS and SARS whereas it is scarcely present in SARS-CoV2. $(GC)_2$ is absent in SARS-CoV2.

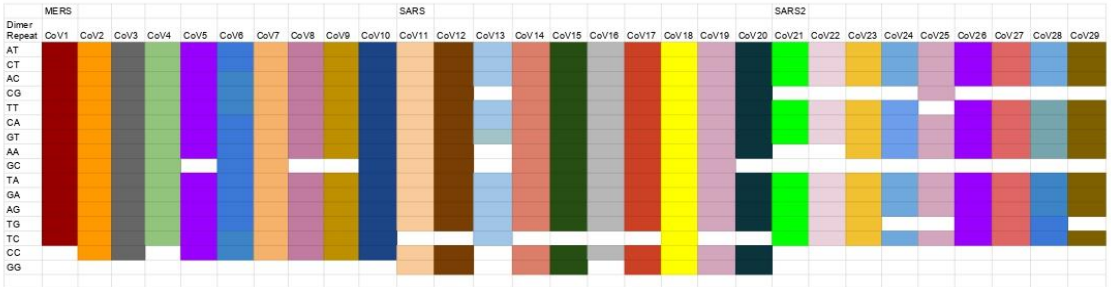

Fig25. Distribution of dimer motifs in the considered coronavirus family.

In Fig26. We can see that motif T is dominant present in MERS, whereas it is rarely present in SARS and SARS2. Therefore T is a marker for MERS-CoV. At the same time, G is dominantly present in SARS-CoV and missing in MERS-CoV and SARS-CoV2.Monomer C is present in both SARS and MERS.

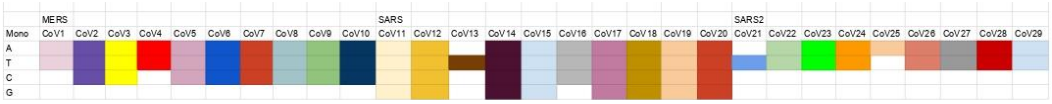

Fig26. Distribution of monomer motifs in the considered coronavirus family.

3.5. Conclusion.

To date, several human coronaviruses have been identified. Three of them are common, less high risk and typically cause only mild respiratory illnesses in healthy human adults. The three those causing MERS, SARS and COVID-19 cases are known to cause more severe illnesses such as shortness of breath and even death. COVID-19 illness tends to be milder than SARS and MERS but more severe than disease caused by the four common coronaviruses. We could not find the information about several variants of the coronavirus family, which we wish to search and investigate in future. And we do hope to find some more observations and derive conclusions from the above data that we have gathered from our computational tools.

4. Chapter 3: Prediction of the promoter region of cyanobacteria using Deep learning models.

4.1. Introduction.

Bacteria are small single-celled organisms. Bacteria are found almost everywhere on Earth and are vital to the planet's ecosystems. Some species can live under extreme conditions of temperature and pressure. The human body is full of bacteria and is estimated to contain more bacterial cells than human cells. Most bacteria in the body are harmless, and some are even helpful. A relatively small number of species cause disease.

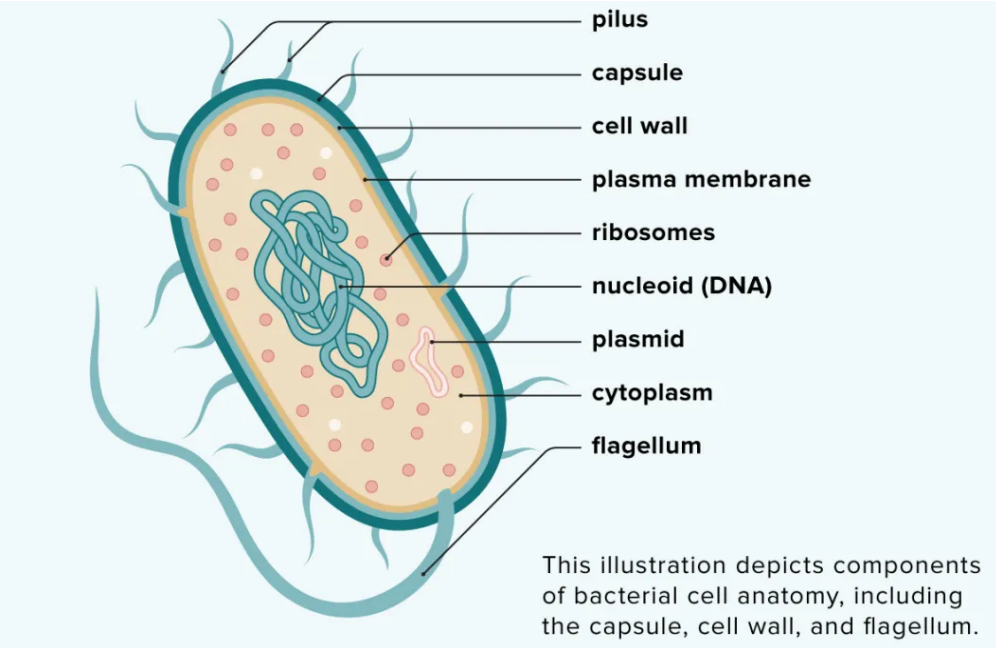

Fig.27. Structure of bacteria.[https://www.healthline.com/health/bacteria]

Bacteria come in various shapes. They can be in the form of spheres, rods or spirals. The bacteria that cause diseases are known as pathogens. The whole bacterial genome used as a training dataset is Cyanobacteria. It is a species that includes photosynthetic bacteria that dwells in aquatic habitats and moist soil. They are known as the oldest fossils. They were first found around 3.5 million years ago. Cyanobacteria can be considered the most senior and crucial bacterial group on earth. They produce gaseous oxygen as the byproduct of photosynthesis.
Moreover, they are believed to be a part of the great oxygenation event. Some of them fixate nitrogen, and others live singly or in colonies forming filaments or spheres. All living things are

methodically arranged into five kingdoms. The cyanobacteria are known to be the Cyanophyta and are one of the kingdom Protista species. Recent discoveries and research have brought light to changes in the taxonomic positions and led to a unique classification system. Cyanophyta (also known as the blue algae) is now known as cyanobacteria, which falls in the class of bacteria.[25]

Cyanobacteria are identified by the presence of pigments such as phycobiliproteins responsible for their bluish-green colour. *Phycobilisomes* are the components present in phycobiliproteins. These pigments are behind the formation of blue-green pigmentation of cyanobacteria, enabling them to synthesise their sugar through photosynthesis. Some of the cyanobacteria lack phycobiliproteins and contain chlorophyll-b instead. Apart from this, they also lack a membrane-bound nucleus but have microcomponents. We can consider a carboxysome a compartmentalised structure surrounded by protein cells. Cyanobacteria use this for concentrating carbon dioxide hence increasing the efficiency of *RuBisCo* (the Co_2 fixing enzyme). Cyanobacteria have separate thylakoids, which are not only responsible for photosynthesis but also for cellular respiration. They are used as electron transport machinery during photosynthesis and photosynthesis during the night. Cyanobacteria reproduce through binary fission.[26][27]

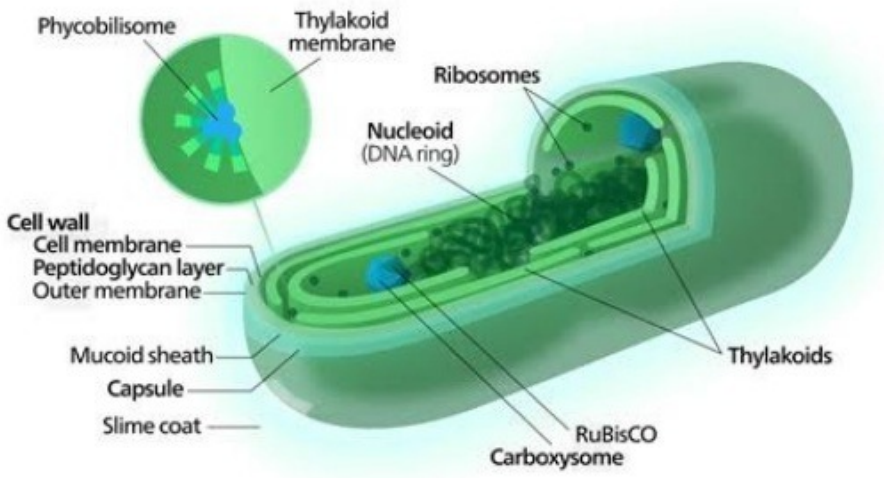

Fig.28. Structure of cyanobacteria.[https://en.wikipedia.org/wiki/Cyanobacteria]

4.2.1 Promoter region in Bacteria.

The region of DNA where transcription of the gene is initiated is called a promoter. They are an essential component of expression vectors because they are responsible for the binding RNA polymerase to DNA. The mRNA is transcribed from DNA with the help of RNA polymerase and

is finally translated into a functional protein. Hence, the promoter region controls the time and position of the expression of the gene of interest.
Promoters in bacteria contain two short DNA sequences located at the -10 (10 bp 5' or upstream) and -35 positions from the transcription start site (TSS). Their equivalent to the eukaryotic TATA box, the Pribnow box (TATAAT) is located at the -10 position and is essential for transcription initiation. The -35 position, simply titled the -35 element, typically consists of the sequence TTGACA and this element controls the rate of transcription. Bacterial cells contain sigma factors which assist the RNA polymerase in binding to the promoter region. Each sigma factor recognises different core promoter sequences.

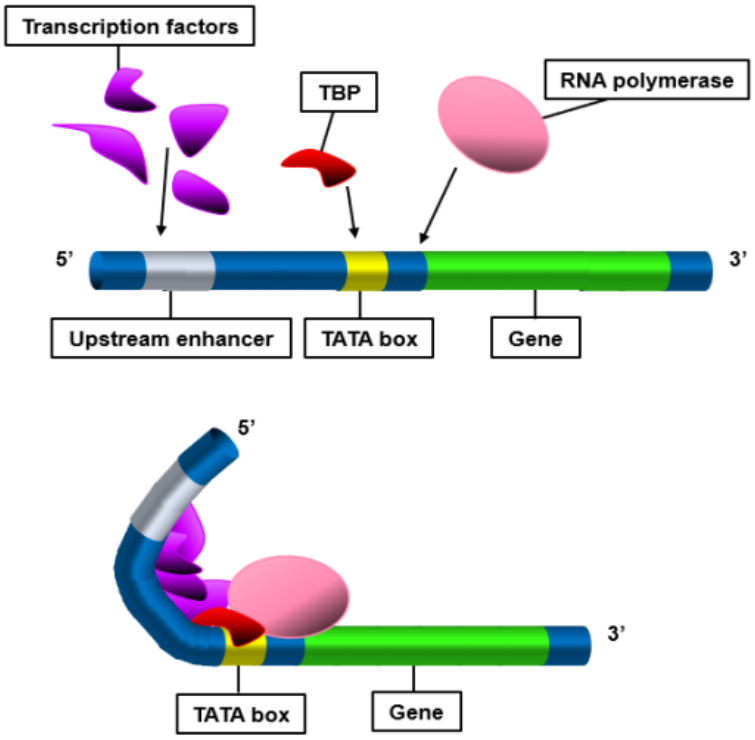

Fig.29. DNA transcription process.[https://www.differencebetween.com/difference-between-tata-and-caat-box/]

Although bacterial transcription is simpler than eukaryotic transcription, bacteria still have complex gene regulation systems, like operons. Operons are a cluster of different genes controlled by a single promoter and operator. Operons are common in prokaryotes, specifically bacteria, but have also been discovered in eukaryotes. Operons consist of a promoter, which is recognized by the RNA polymerase, an operator, a segment of DNA in which a repressor or activator can bind, and the structural genes that are transcribed together.

Operon regulation can be either negative or positive. Negative repressible operons are normally bound by a repressor protein that prevents transcription. When an inducer molecule binds to the

repressor, it changes its conformation, preventing its binding to the operator and thus allowing for transcription. The Lac operon in bacteria is an example of a negatively controlled operon.

A positive repressible operon works oppositely. The operon is normally transcribed until a repressor/corepressor binds to the operator preventing transcription. The trp operon is involved in the production of tryptophan prokaryotes of a positively controlled operon.

Promoter binding is very different in bacteria compared to eukaryotes. In bacteria, the core RNA polymerase requires an associated sigma factor for promoter recognition and binding. On the other hand, the process in eukaryotes is much more complex. Eukaryotes require a minimum of seven transcription factors for RNA polymerase II (a eukaryote-specific RNA polymerase) to bind to a promoter. Transcription is tightly controlled in both bacteria and eukaryotes. Promoters are controlled by various DNA regulatory sequences, including enhancers, boundary elements, insulators, and silencers.

4.2. Literature Review.

The cyanobacteria are photosynthetic prokaryotes found in most, though not all, types of illuminated environment. They are also quantitatively among the most important organisms on Earth. A conservative estimate of their global biomass is 3 × 10 14 g C or a thousand million tonnes (10 15 g) wet biomass [23]. They all synthesize chlorophyll a and typically water is the electron donor during photosynthesis, leading to the evolution of oxygen. Most produce the phycobilin pigment, phycocyanin, which gives the cells a bluish colour when present in suf fi ciently high concentration, and is responsible for the popular name, blue-green algae; in some cases the red accessory pigment, phycoerythrin, is formed as well. A few genera, however,produce neither, but form other accessory pigments. These include some ecologically very important members of the ocean plankton.[23]

A tool called IMEx (Imperfect Microsatellite Extractor). IMEx uses simple string-matching algorithm with sliding window approach is frequently used to screen DNA sequences for microsatellites and reports the motif, copy number, genomic location, nearby genes, mutational events and many other features useful for in-depth studies. IMEx is more sensitive, efficient and useful than the available widely used tools. IMEx is available in the form of a stand-alone program as well as in the form of a web-server. [33]
Microsatellites or simple sequence repeats (SSRs) are the nucleotide sequences arising out of tandem repeating of short sequence motifs of the size 1–6 bp. Microsatellites have been found in all the known genomes so far and are widely distributed both in coding and non-coding regions [33] Increase/decrease of repeat copy numbers in microsatellites in coding regions often lead to shifts in reading frames thereby causing changes in protein products [34] and in non-coding regions, known to effect the gene regulation [33]. Mutations occurring at microsatellite loci within or near certain genes have been implicated to be responsible for some human

neurodegenerative diseases. Furthermore, microsatellite instability has also been implicated in the induction of cancer [30]. Owing to their high mutability, it is thought that the microsatellites are one of the sources of genetic diversity [29]. In the recent times, efforts have also been made to study the possible functional roles of microsatellites in giving rise to certain amount of plasticity and also in the evolution of genomes[33].

4.3. Data collection and preprocessing.

4.3.2. Data preprocessing.

Two datasets are collected from the CSIR-IICB computational genomics lab. We have collected two separate datasets. one was a mix of 3 species of Nostoc, and the other was from a single species *N. sphaeroides Kutzing En.* This is done for testing the CNN model. Nostoc nuclear acid sequences are fed to another CNN model obtained from Berkeley Drosophila Genome Project (prokaryotic predictor). It was used to predict 55 bases of sequences which are positive/ promoter regions. These positive promoter regions are taken as our CNN model's promoter or positive dataset. Random 55 base sequences were used as the negative dataset.

The positive cyanobacterial dataset is the 100 bases upstream region of genes present in genomes. The genomic files were generated in-house using Prokka (please read up on this software). The negative dataset was the genic region between upstream -200 to -100 of isolated genes (no genes 300 bases upstream or downstream of that particular gene). After collecting the data, CSV files of both data frames were created for the model training purpose.

The data is categorized between “DATASET” and “LABEL”. Then we concatenate the integer 0 to the length of the string to every column of the data frame. After this, the dataset is one-hot encoded. The length of each sequence is 100 bp and the number of sequences is 600.

4.4 Deep Learning Models and methods used for comparative analysis.

4.4.1. Recurrent Neural Network Architecture

RNNs are powerful and robust neural networks and belong to the most promising algorithms because they are the only ones with internal memory. Like many other deep learning algorithms, recurrent neural networks are relatively old. They were initially created in the 1980s, but we have only seen their true potential in recent years. An increase in computational power, the

massive amounts of data we now have to work with, and the invention of long short-term memory (LSTM) in the 1990s brought RNNs to the foreground. Because of their internal memory, RNNs can remember important things about the input they received, which allows them to be very precise in predicting what is coming next. As a result, they are the preferred algorithm for sequential data like time series, speech, text, financial data, audio, video, weather, etc. In addition, recurrent neural networks can form a much deeper understanding of a sequence and its context than other algorithms.

Recurrent neural networks (RNN) are a class of neural networks that are helpful in modelling sequence data. Derived from feedforward networks, RNNs behave similarly to how human brains function. Recurrent neural networks produce predictive results in sequential data that other algorithms cannot.[22][27]

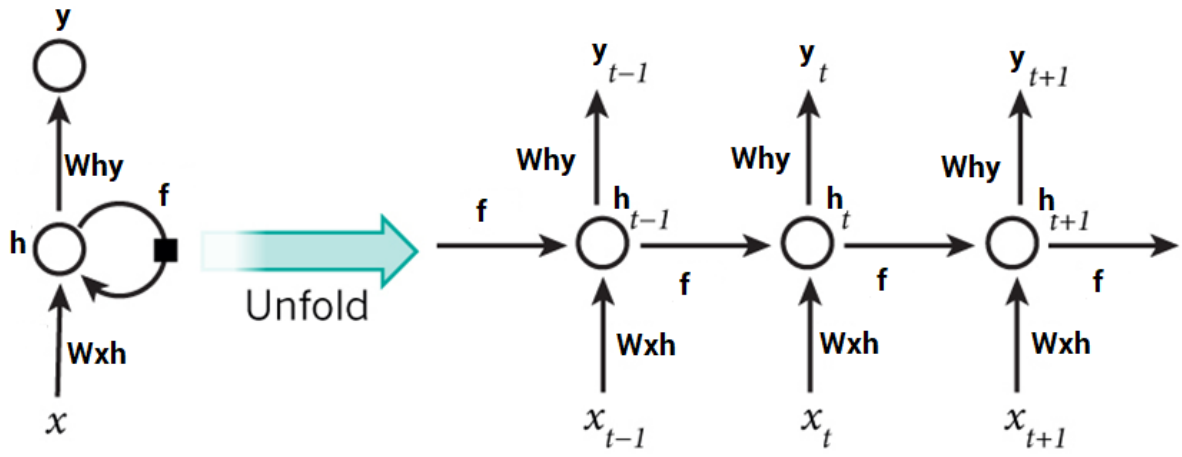

Fig.30. Workflow of the mechanism of RNN[https://www.analyticsvidhya.com/blog/2017/12/introduction-to-recurrent-neural-networks/]

Based on available runtime hardware and constraints, the RNN layer will choose different implementations (cuDNN-based or pure-TensorFlow) to maximise the performance. If a GPU is available and all the layers' arguments meet the requirement of the cuDNN kernel, the layer will use a fast cuDNN implementation. The requirements to use the cuDNN implementation are: activation=tanh,recurrent_activation=sigmoid,use_bias=True,kernel_initializer='glorot_uniform', recurrent_initializer='orthogonal',unit_forget_bias=True,kernel_regularizer=None,recurrent_regularizer=None,bias_regularizer=None,activity_regularizer=None,kernel_constraint=None,recurrent_constraint=None,kernel_constraint=None,recurrent_constraint=None,bias_constraint=None, dropout=0.0,recurrent_dropout=0.0,implementation=1,return_sequences=True,return_state=False, go_backwards=False, stateful=False, unroll=False, input_shape=(x_units, y_units).

The arguments are discussed below:

- units: Positive integer, the dimensionality of the output space.
- activation: Activation function to use. Default: hyperbolic tangent (tan h). If you pass None, no activation is applied (i.e, "linear" activation: a(x) = x).
- recurrent_activation: Activation function to use for the recurrent step. Default: sigmoid (sigmoid). If you pass None, no activation is applied (ie. "linear" activation: a(x) = x).
- use_bias: Boolean (default True), whether the layer uses a bias vector.
- Kernel_initializer: Initializer for the kernel weights matrix, used for the linear transformation of the inputs. Default: glorot_uniform.
- recurrent_initializer: Initializer for the recurrent_kernel weights matrix, used for the linear transformation of the recurrent state. Default: orthogonal.
- bias_initializer: Initializer for the bias vector. Default: zeros.
- unit_forget_bias: Boolean (default True). If True, add 1 to the bias of the forget gate at initialisation. Setting it to true will also force bias_initializer="zeros". kernel_regularizer: Regularizer function applied to the kernel weights matrix. Default: None.
- recurrent_regularizer: Regularizer function applied to the recurrent_kernel weights matrix. Default: None.
- bias_regularizer: Regularizer function applied to the bias vector. Default: None.
- activity_regularizer: Regularizer function applied to the layer's output (its "activation"). Default: None.
- kernel_constraint: Constraint function applied to the kernel weights matrix. Default: None.
- recurrent_constraint: Constraint function applied to the recurrent_kernel weights matrix. Default: None.
- bias_constraint: Constraint function applied to the bias vector. Default: None.
- Dropout: Float between 0 and 1. Fraction of the units to drop for the linear transformation of the inputs. Default: 0.
- recurrent_dropout: Float between 0 and 1. Fraction of the units to drop for the linear transformation of the recurrent state. Default: 0.
- return_sequences: Boolean. Whether to return the final output. In the output sequence or the complete sequence. Default: False.
- return_state: Boolean. Whether to return the last state in addition to the output. Default: False.
- go_backwards: Boolean (default False). If True, process the input sequence backwards and return the reversed sequence.
- stateful: Boolean (default False). If True, the last state for each sample at index i in a batch will be used as the initial state for the sample of index i in the following batch.
- time_major: The shape format of the inputs and outputs tensors. If True, the inputs and outputs will be in shape [timesteps, batch, feature], whereas in the False case, [batch, timesteps, feature]. Using time_major = True is more efficient because it avoids transposes at the beginning and end of the RNN calculation. However, most TensorFlow

data is batch-major, so by default, this function accepts input and emits output in batch-major form.

- unroll Boolean (default False). If True, the network will be unrolled; else, a symbolic loop will be used. Unrolling can speed up an RNN, although it tends to be more memory-intensive. Therefore, unrolling is only suitable for short sequences.

4.4.2. Logistic regression model

Logistic regression is a statistical method used to build machine learning models where the dependent variable is dichotomous: i.e. binary. Logistic regression describes data and the relationship between one dependent variable and one or more independent variables. The independent variables can be nominal, ordinal, or interval type. The name "logistic regression" is derived from the concept of the logistic function that it uses. The logistic function is also known as the sigmoid function. The value of this logistic function lies between zero and one.

4.4.2(a). Mathematics behind logistic regression

Probability always ranges between 0 (does not happen) and 1 (happens). Using our Covid-19 example, in the case of binary classification, the probability of testing positive and not testing positive will sum up to 1. We use the logistic function or sigmoid function to calculate probability in logistic regression. The logistic function is a simple S-shaped curve used to convert data into a value between 0 and 1.

$$h\theta(x) = 1/1 + e^{-(b_o + b_1 x)}$$

Where,

$h\theta(x)$ is the output of the logistic function, where 0<= $h\theta(x)$>=1

b_o is the slope

b_1 is the y-intercept

x is the independent variable.

$(b_o + b_1 x)$ is derived from the equation of a line y (predicted) = $(b_o + b_1 x)$ + *Error value.*

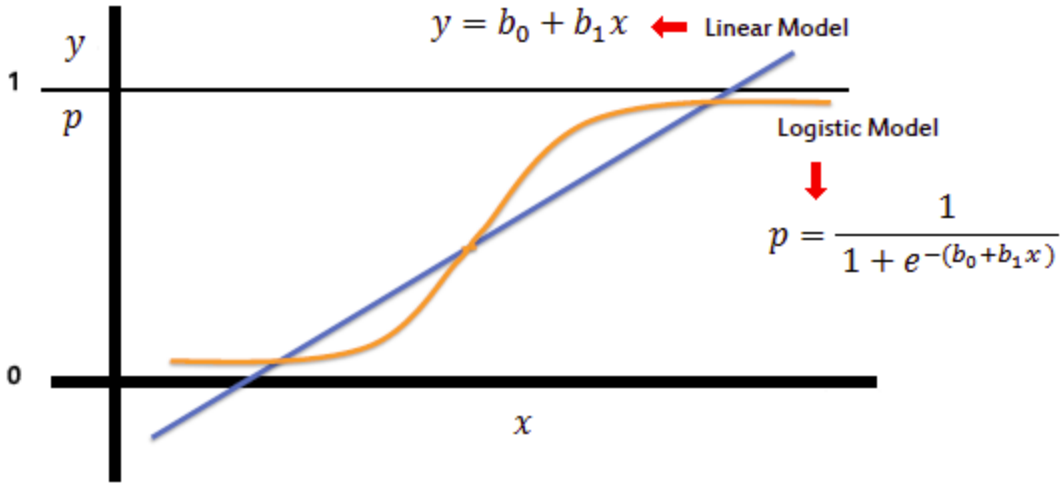

Fig.31. The logistic regression curve.[https://www.saedsayad.com/logistic_regression.htm]

Logistic regression is used for classification problems when the output or dependent variable is dichotomous or categorical. There are some assumptions to keep in mind while implementing logistic regressions, such as the different types of logistic regression and the different types of independent variables and the training data available.

4.4.3. Support vector machine as a classifier.

SVC is a supervised machine learning algorithm that helps in classification problems**.** It aims to find an optimal boundary between the possible outputs. Simply put, SVM does complex data transformations depending on the selected kernel function and based on that transformations, it tries to maximise the separation boundaries between our data points depending on the labels or classes we have defined.

In the base form, linear separation, SVM tries to find a line that maximises the separation between a two-class data set of 2-dimensional space points. To generalise, the objective is to find a hyperplane that maximizes the separation of the data points to their potential classes in N-dimensional space. The data points with the minimum distance to the hyperplane (closest points) are called *Support Vectors*.

In the image below, the Support Vectors are the 3 points (2 blue and 1 green) laying on the scattered lines, and the separation hyperplane is the solid red line:

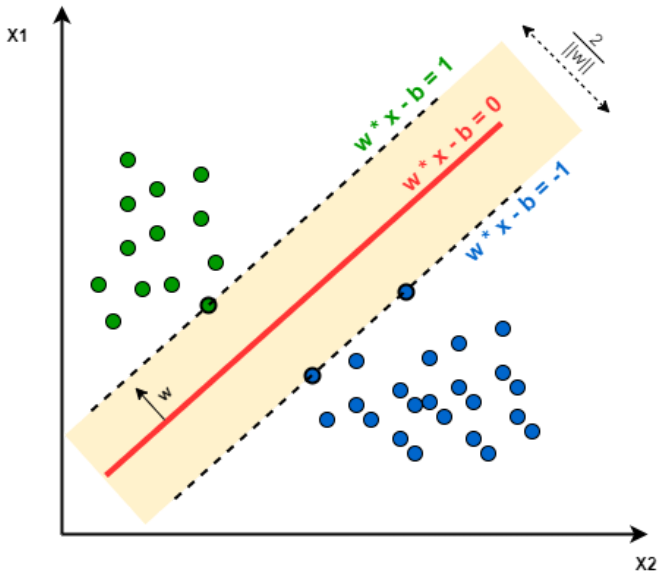

Fig.32. Support Vector machine used as Classifier.[https://towardsdatascience.com/so-why-the-heck-are-they-called-support-vector-machines-52fc72c990a1]

The computations of data point separation depend on a kernel function. There are different kernel functions: Linear, Polynomial, Gaussian, Radial Basis Function (RBF), and Sigmoid. Simply put, these functions determine the smoothness and efficiency of class separation, and playing around with their hyperparameters may lead to overfitting or underfitting.

4.4.4. Deepinsight

There are several numbers of data such as genomic, transcriptomic, methylation, mutation, text, and spoken words in non-image form. CNN cannot be used on these data as CNN needs image data as input. However, it is possible to convert non-image data to a well-organised image format, which the CNN model can use for higher classification performance. To improve the detection rate, we have integrated element arrangement, feature extraction and classification in a proposed method called Deepinsight. Deepinsight is used to construct images by placing similar elements or features together and dissimilar ones apart, enabling the collective use of the neighbourhood elements. This collective approach to element arrangement is vital for decoding mechanisms (e.g. pathways) or understanding the relationship between a set of features like texts or vowels. Hence, image conversion by applying similar features as clusters is more valuable than dealing with individual features. Proper arrangement of elements is the key to disclosing information. Moreover, it helps in understanding the importance of features towards an outcome. Further, Deepinsight also helps in feature extraction and classification via the image classification model.

The function of Deepinsight is first to transform non-image data into image format and then use some classifier for classification or feature extraction purposes. In Fig33. We can see feature

vector x consisting of gene expression values is transformed to feature matrix M which transforms T. the position of the features in the cartesian coordinates depends on the similarity of features. For instance, features g1, g3, g6 and gd are closer to each other in Fig.33a. Once we determine the location of the features in the feature matrix, the feature values or expression values are mapped. As a result, a unique image for each sample will be generated. N samples of d features will provide N samples of $m \times n$ feature matrices. This 2D matrix form will possess all the d features. After that, this set of N feature matrices is passed to the CNN architecture for learning the model and providing prediction.

If the data dimensionality is enormous and tedious to handle due to hardware limitations, then the dimensionality reduction technique (DRT) can be considered before applying DeepInsight. The DRT can be either feature selection or feature extraction, depending upon the problem. [33]

4.4.4(i). Deepinsight pipeline.

The training set is used in finding the feature locations. Let us consider the training set consists of n samples, defined as $\chi = \{x\ 1, x\ 2, \ldots, x\ n\}$ where a feature vector has d features or $x \in d$, then a gene or feature set is considered as $G = \{g\ 1, g\ 2, \ldots, g\ d\}$ where $g \in n$; i.e., a feature g_j has n training samples. G is obtained by transposing χ. We used this feature set G and applied dimensionality reduction techniques like t-SNE or kernel principal component analysis (kPCA) to obtain a 2D plane. They are non-linear dimensionality reduction techniques. The points in this Cartesian plane are the features or genes. These points only point out the location of features, not the feature itself. Once the location of features is identified, the convex hull algorithm is applied to find the smallest rectangle containing all the points. Since images are framed in a horizontal or vertical form for the CNN architecture, a rotation is performed. After that, the Cartesian coordinates are converted to pixels. The conversion from Cartesian coordinates to pixel frames is done by averaging some features as the image size has a pixel limitation. The pixel frame will, therefore, consist of the positions of features for a sample x_j (for $j = 1, 2, \ldots, n$). Once the location is identified, the next step is to map the feature values to the pixel locations. If more than one feature acquired the same location in the pixel frame, then the features will be averaged and placed in the same location during the mapping of the features. Therefore, if the image's resolution or grid size is minimal, many features overlap, and the image presentation may not be very accurate. An appropriate resolution should be selected given the hardware capacity and the number of features required to process. Alternatively, dimensionality reduction may be deductible.[33]

4.4.4(ii). Feature Normalisation.

The single layer of the image has 256 shades which are normalised in the range of [0, 1]. Therefore, feature values are normalised before the application of image transformation. In this work, we performed two types of normalizations: (1) each feature is assumed independent and

therefore normalized by its minimum and maximum, and (2) the topology of mutual features are retained up to some extent by normalizing it with the one maximum value from the entire training set. DeepInsight evaluates validation set performance on both the types of normalisations and accepts the one with the lowest validation error.[33]

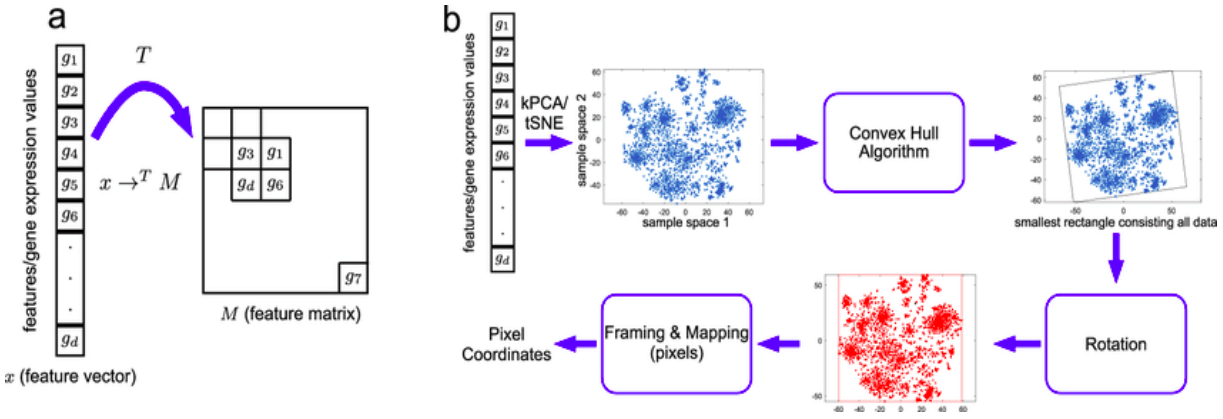

Fig.33. Workflow of the Deepinsight method[https://www.nature.com/articles/s41598-019-47765-6].

SimCLR

Learning visual representation effectively without human supervision is a real-life problem. The standard approaches are either generative or discriminative. Productive methods learn ways to generate or model pixels into the input space. However, pixel-level generation is quite computationally expensive and may be unnecessary for representation learning. Whereas discriminative approaches use objective functions to learn representations similar to those used for supervised learning. But from unlabeled datasets, they derive the inputs and labels used to train pretrained networks to perform pretext tasks. Several such approaches are dependent on heuristics to design pretext tasks which limit the learned representations' generality. Discriminative methods are grounded on contrastive learning in the latent space, which has shown promising results. The simCLR framework learns representations by maximising agreement between the augmented views of the same data. It is done by contrastive loss in the latent space. The simCLR framework has the following major components.[9]

- A stochastic data augmentation module that transforms any given data example randomly resulting in two correlated views of the same example, denoted a_i and a_j, which is considered a positive pair. SimCLR sequentially applies three simple augmentations: random cropping followed by resizing back to the original size, random colour distortions, and random Gaussian blur. The authors find random crop and colour distortion is crucial to achieving good performance.

- A neural network base encoder $f_o(.)$ that extracts representation vectors from augmented data examples. Various choices of network architecture are allowed in the simCLR framework without any constraints. We use the ResNet Maximize agreement to maintain simplicity and to obtain $h_i = f_o(a_i) = ResNET(a_i)$ where $h_i \in R^d$ is the output of the average pooling layer.

- A small neural network projection head $g_o(.)$ that maps representations to the space where contrastive loss is applied. One hidden layer is used with MLP to obtain,

$$z_i = g_o(h_i) = W^{(2)}\sigma(W^{(1)}h_i)$$

Where σ is the RELU non-linearity. It is essential to define the contrastive loss on z_i rather than h_i.

- For the contrastive prediction task, a contrastive loss function is defined. For instance, a given set $\{a_k\}$ including a positive pair of example a_i and a_j, are contrastive prediction task which aims to identify a_j in $\{a_k\}_{k \neq i}$ for a given a_i.

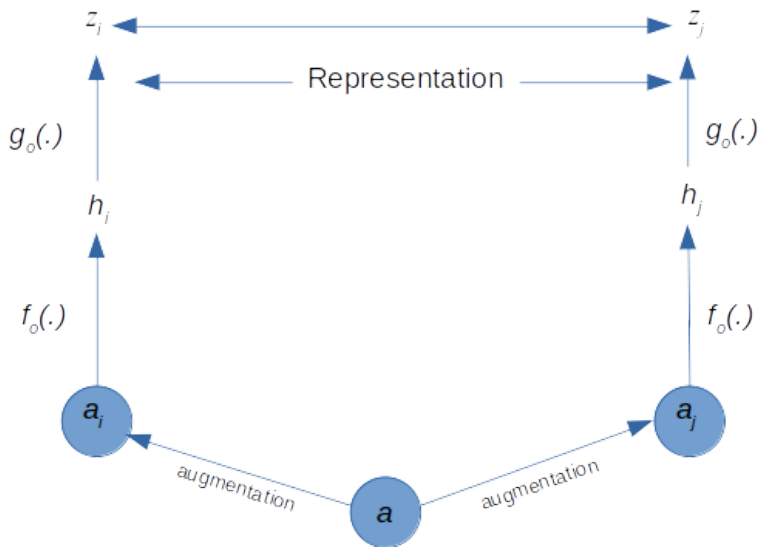

Fig.34. Workflow of simCLR framework.

We first sample a minibatch of N and then define the contrastive prediction task on pairs of augmented samples that are derived from the minibatch, thus resulting in 2N data points. The negative examples are not sampled but given a positive pair. The other 2(N-1) augmented examples within a minibatch are treated as negative examples. [9]

4.5. Comparative analysis.

We have derived a comparative analysis of all the above-discussed supervised models. This is mentioned in the table below.

Name of the Model	Testing accuracy.	Type of Model
Support Vector Classifier	75%	Supervised
Logistic Regression	91.2%	Supervised
Recurrent Neural Network	94.75%	Supervised
simCLR Framework	100%	Unsupervised

Table7. Comparison of testing accuracy of the considered machine learning models.

Supervised learning, as the name indicates, has the presence of a supervisor as a teacher. Basically supervised learning is when we teach or train the machine using well-labelled data. This means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labelled data. Supervised learning deals with or learns with “labelled” data. This implies that some data is already tagged with the correct answer.

Unsupervised learning is the training of a machine using information that is neither classified nor labelled and allowing the algorithm to act on that information without guidance. Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training in data.

Unlike supervised learning, no teacher is provided, which means no training will be given to the machine. Therefore the machine is restricted to finding the hidden structure in unlabeled data by itself.

4.6. Feature extraction methodology.

For obtaining the location feature matrix of the training dataset shown in Fig.36, we have used the t-SNE method which stands for t-Distributed Stochastic Neighbor Embedding. The algorithm works well even for large datasets. The main goal of t-SNE is to project multi-dimensional points to 2- or 3-dimensional plots so that if two points were close in the initial high-dimensional space,

they stay close in the resulting projection. If the points were far from each other, they should stay far in the target low-dimensional space too. To do that, t-SNE first creates a probability distribution that captures these mutual distance relationships between the points in the initial high-dimensional space. After this, the algorithm tries to create a low-dimensional space that has similar relations between the points. We can describe the classification networks on a high level as follows. They have a backbone that extracts valuable information, or features, from the image. It also has a classifier applied right after the backbone. The classifier makes a final decision based on the information extracted by the backbone. We have used the t-SNE model with specification: n_components=2, metric='cosine', perplexity=perplexity, n_iter=1000, method='exact', random_state=rand_seed, n_jobs=-1.[33][5][9]

In Fig.36, we find the location matrix with perplexity 50 of the extracted feature of the training dataset. We in infer from the figure that the vectors of the matrix scatter largely between -0.5 to 2.5 along the y-axis and -1.0 and 2.0 along the x-axis.

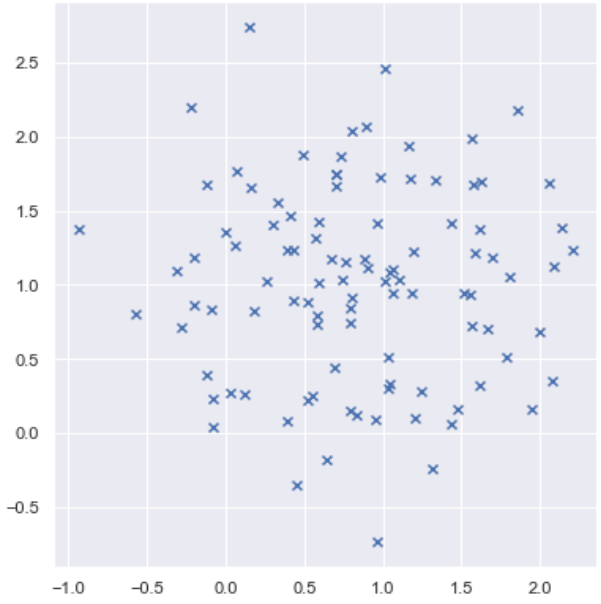

Fig35. Feature location matrix of the training dataset.

The data distribution shown in the graph in Fig.37, is also extracted with the help of the t-SNE method. The feature density matrix shows that the density of extracted features with resolution 101x101 from the training set is quite less.

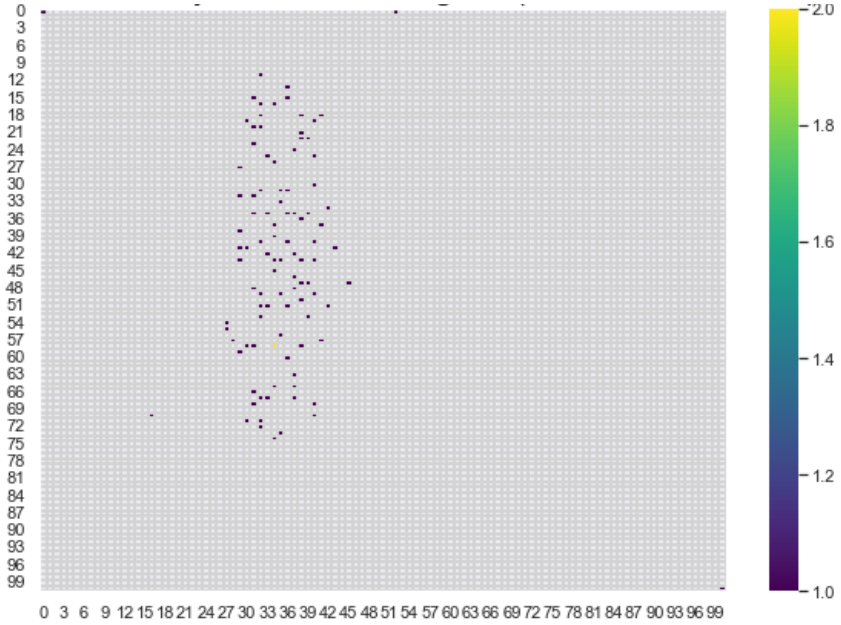

Fig36. Feature density matrix of the Training dataset.

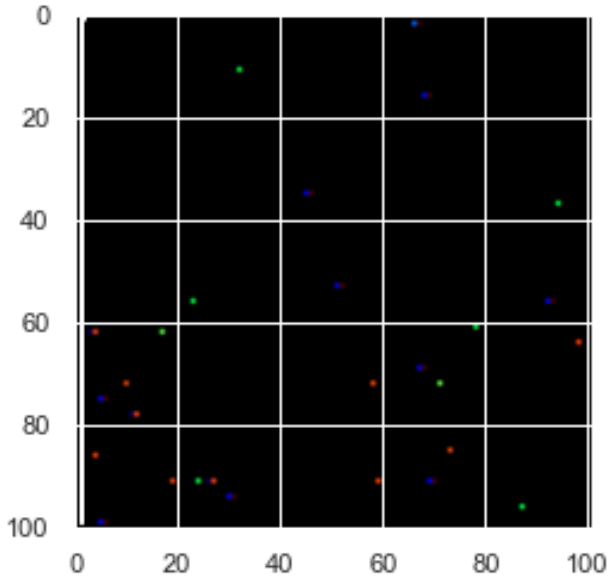

Fig.37. Feature overlapping count.

The overlapping count method generates planarised overlapping features from the input features. The count of overlapping features is written to the output features. It is shown in Fig38. The different colours represent the different extracted overlapping features.

4.7. Conclusion.

From our work on classifying promoter sequences from non-promoter sequences using deep learning models mentioned in work, we could attend good accuracy of the models we have created. We have also used a dimensionality reduction method such as t-SNE and derived the feature location matrix and the feature. From the feature density matrix, we conclude that the density of the extracted features is low. We can also see the overlapping feature matrix in which three features are detected and the clusters are represented in different colours.

As future work we can find the location of the promoter regions that are present in the DNA sequences and make other biological inferences from it.

3. Conclusion.

In chapter 1, 2 and 3 we have used several computational methods such as microsatellite extraction using sliding window algorithm for the observation of repeated sequences in mitochondrial DNA sequences of the hominini subspecies and the coronavirus family. We have also done the phylogenetic analysis of the *homo sapiens*. We have also performed a detailed analysis of the presence or absence of different motifs in the DNA sequences that can be considered markers for the specific species. Moreover, the relative density, relative abundance and the GC content of the DNA sequences are also derived, and conclusions were drawn from the results.
Apart from this, we have also classified between promoters and non-promoter regions of DNA sequences of bacteria using deep learning models such as logistic regression, support vector classifier Recurrent neural network and the simCLR framework for contrastive learning. We have used dimension reduction methods and clustering algorithms for the feature extraction of the training datasets. We have used the Deepinsight algorithm to convert our dataset to image format as the input for the simCLR framework.

Last but not least, we will also try to extract more information from the computational outputs we have calculated, adding to the information required about these life forms for further studies.

4. References.

[1] H. Horiuchi, “Molecular structure of nuclei,” *Eur. Phys. J. A*, vol. 15, no. 1–2, pp. 131–133, 2002, doi: 10.1140/epja/i2001-10240-x.

[2] C. M. Alam, A. K. Singh, C. Sharfuddin, and S. Ali, “Genome-wide scan for analysis of simple and imperfect microsatellites in diverse carlaviruses,” *Infect. Genet. Evol.*, vol. 21, pp. 287–294, 2014, doi: 10.1016/j.meegid.2013.11.018.

[3] J. S. M. Peiris *et al.*, “Coronavirus as a possible cause of severe acute respiratory syndrome,” *Lancet*, vol. 361, no. 9366, pp. 1319–1325, 2003, doi: 10.1016/S0140-6736(03)13077-2.

[4] NDSU, “Deoxyribonucleotide Structure,” *Pubweb*, 2017, [Online]. Available: https://www.ndsu.edu/pubweb/~mcclean/plsc731/Genome-sequencing-PMG-overheads.pdf

[5] D. Luo *et al.*, “Unsupervised Document Embedding via Contrastive Augmentation,” 2021, [Online]. Available: http://arxiv.org/abs/2103.14542

[6] M. Kim *et al.*, “An infectious cDNA clone of a growth attenuated Korean isolate of MERS coronavirus KNIH002 in clade B,” *Emerg. Microbes Infect.*, vol. 9, no. 1, pp. 2714–2726, 2020, doi: 10.1080/22221751.2020.1861914.

[7] A. Field, “Logistic regression Logistic regression Logistic regression,” *Discov. Stat. Using SPSS*, pp. 731–735, 2012.

[8] S. Dreiseitl and L. Ohno-Machado, “Logistic regression and artificial neural network classification models: A methodology review,” *J. Biomed. Inform.*, vol. 35, no. 5–6, pp. 352–359, 2002, doi: 10.1016/S1532-0464(03)00034-0.

[9] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” *37th Int. Conf. Mach. Learn. ICML 2020*, vol. PartF16814, no. Figure 1, pp. 1575–1585, 2020.

[10] C. M. Alam, A. K. Singh, C. Sharfuddin, and S. Ali, “In-silico analysis of simple and imperfect microsatellites in diverse tobamovirus genomes,” *Gene*, vol. 530, no. 2, pp. 193–200, 2013, doi: 10.1016/j.gene.2013.08.046.

[11] C. M. Alam, A. K. Singh, C. Sharfuddin, and S. Ali, “Incidence, complexity and diversity of simple sequence repeats across potexvirus genomes,” *Gene*, vol. 537, no. 2, pp. 189–196, 2014, doi: 10.1016/j.gene.2014.01.007.

[12] M. Naghibzadeh, H. Savari, A. Savadi, N. Saadati, and E. Mehrazin, “Developing an ultra-efficient microsatellite discoverer to find structural differences between SARS-CoV-1 and

Covid-19," *Informatics Med. Unlocked*, vol. 19, p. 100356, 2020, doi: 10.1016/j.imu.2020.100356.

[13] A. Merkel and N. Gemmell, "Detecting short tandem repeats from genome data: Opening the software black box," *Brief. Bioinform.*, vol. 9, no. 5, pp. 355–366, 2008, doi: 10.1093/bib/bbn028.

[14] M. A. M. Atia, G. H. Osman, and W. H. Elmenofy, "Genome-wide in silico analysis, characterization and identification of microsatellites in spodoptera littoralis multiple nucleopolyhedrovirus (SpliMNPV)," *Sci. Rep.*, vol. 6, no. February, pp. 1–9, 2016, doi: 10.1038/srep33741.

[15] A. G. Wrobel *et al.*, "SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects," *Nat. Struct. Mol. Biol.*, vol. 27, no. 8, pp. 763–767, 2020, doi: 10.1038/s41594-020-0468-7.

[16] M. J. Pajuelo *et al.*, "Identification and Characterization of Microsatellite Markers Derived from the Whole Genome Analysis of Taenia solium," *PLoS Negl. Trop. Dis.*, vol. 9, no. 12, pp. 1–15, 2015, doi: 10.1371/journal.pntd.0004316.

[17] M. Rashighi and J. E. Harris, "乳鼠心肌提取 HHS Public Access," *Physiol. Behav.*, vol. 176, no. 3, pp. 139–148, 2017, doi: 10.1145/3233547.3233604.ULTRA.

[18] R. Hilgenfeld and M. Peiris, "From SARS to MERS: 10 years of research on highly pathogenic human coronaviruses," *Antiviral Res.*, vol. 100, no. 1, pp. 286–295, 2013, doi: 10.1016/j.antiviral.2013.08.015.

[19] C. M. Alam, A. K. Singh, C. Sharfuddin, and S. Ali, "In- silico exploration of thirty alphavirus genomes for analysis of the simple sequence repeats," *Meta Gene*, vol. 2, pp. 694–705, 2014, doi: 10.1016/j.mgene.2014.09.005.

[20] R. Siddiqe and A. Ghosh, "Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19 . The COVID-19 resource centre is hosted on Elsevier Connect , the company ' s public news and information ," no. January, 2020.

[21] C. M. Alam, A. Iqbal, A. Sharma, A. H. Schulman, and S. Ali, "Microsatellite diversity, complexity, and host range of mycobacteriophage genomes of the Siphoviridae family," *Front. Genet.*, vol. 10, no. MAR, pp. 1–11, 2019, doi: 10.3389/fgene.2019.00207.

[22] B. C. Faircloth, "MSATCOMMANDER: Detection of microsatellite repeat arrays and automated, locus-specific primer design," *Mol. Ecol. Resour.*, vol. 8, no. 1, pp. 92–94, 2008, doi: 10.1111/j.1471-8286.2007.01884.x.

[23] K. Cwiklinski, K. Allen, J. LaCourse, D. J. Williams, S. Paterson, and J. E. Hodgkinson, "Characterisation of a novel panel of polymorphic microsatellite loci for the liver fluke, Fasciola hepatica, using a next generation sequencing approach," *Infect. Genet. Evol.*, vol. 32, pp. 298–304, 2015, doi: 10.1016/j.meegid.2015.03.014.

[24] A. K. Avvaru, D. T. Sowpati, and R. K. Mishra, “PERF: An exhaustive algorithm for ultra-fast and efficient identification of microsatellites from large DNA sequences,” *Bioinformatics*, vol. 34, no. 6, pp. 943–948, 2018, doi: 10.1093/bioinformatics/btx721.

[25] M. L. Ledenyova, G. A. Tkachenko, and I. M. Shpak, “Imperfect and Compound Microsatellites in the Genomes of Burkholderia pseudomallei Strains,” *Mol. Biol.*, vol. 53, no. 1, pp. 127–137, 2019, doi: 10.1134/S0026893319010084.

[26] B. Hu, H. Guo, P. Zhou, and Z. L. Shi, “Characteristics of SARS-CoV-2 and COVID-19,” *Nat. Rev. Microbiol.*, vol. 19, no. 3, pp. 141–154, 2021, doi: 10.1038/s41579-020-00459-7.

[27] A. Sherstinsky, “Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network,” *Phys. D Nonlinear Phenom.*, vol. 404, no. March, pp. 1–43, 2020, doi: 10.1016/j.physd.2019.132306.

[28] S. S. Keerthi, O. Chapelle, and D. DeCoste, “Building support vector machines with reduced classifier complexity,” *J. Mach. Learn. Res.*, vol. 7, pp. 1493–1515, 2006.

[29] V. Proutski and E. C. Holmes, “SWAN: Sliding window analysis of nucleotide sequence variability,” *Bioinformatics*, vol. 14, no. 5, pp. 467–468, 1998, doi: 10.1093/bioinformatics/14.5.467.

[30] D. Mishmar *et al.*, “Natural selection shaped regional mtDNA variation in humans,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 1, pp. 171–176, 2003, doi: 10.1073/pnas.0136972100.

[31] D. V. Nesheva, “Aspects of ancient mitochondrial dna analysis in different populations for understanding human evolution,” *Balk. J. Med. Genet.*, vol. 17, no. 1, pp. 5–14, 2014, doi: 10.2478/bjmg-2014-0019.

[32] S. B. Mudunuri and H. A. Nagarajaram, “IMEx: Imperfect microsatellite extractor,” *Bioinformatics*, vol. 23, no. 10, pp. 1181–1187, 2007, doi: 10.1093/bioinformatics/btm097.

[33] A. Sharma, E. Vans, D. Shigemizu, K. A. Boroevich, and T. Tsunoda, “DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture,” *Sci. Rep.*, vol. 9, no. 1, pp. 1–7, 2019, doi: 10.1038/s41598-019-47765-6.