2023
5TH EDITION

GEN(E)IUS

www.sastra.edu/combigs

The COMBIGS Magazine

FOREWORD

Welcome to the 5th edition of GEN(E)IUS, a captivating showcase of the brilliance and innovation brought forth by the 2024 batch of Bioinformatics B.Tech students at SASTRA University. Unveiled during Prathyarth 2023, an annual symposium-like event, this magazine reflects the dynamic spirit of our Bioinformatics community.

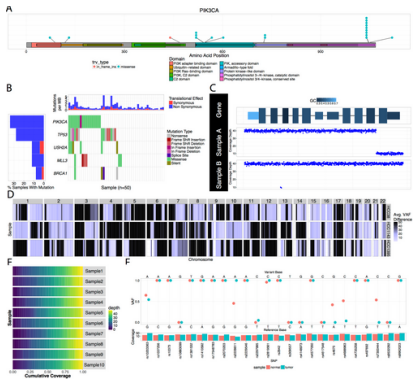
COMBIGS, the student committee driving the Department of Bioinformatics, has been at the forefront of fostering awareness and knowledge exchange since 2005. Through symposiums, conferences, and bi-semester magazines like GEN(E)IUS, COMBIGS has empowered students to explore the latest in Bioinformatics.

In these pages, you'll discover a diverse range of articles, each a testament to the dedication and brilliance of our students. From exploring DNA as a storage device to unraveling the wonders of the CRISPR-Cas system, this edition encapsulates the multidimensional nature of Bioinformatics.

Extending appreciation to all contributors and editors for making this edition a reality. May GEN(E)IUS serve as a source of inspiration and ignite the curiosity of its readers.

Bioinformatics Pioneering the Battle Against Neurodegenerative Diseases: A Focus on Multiple Sclerosis

The progressive loss of neurons in the central nervous system is a hallmark of neurodegenerative diseases like multiple sclerosis (MS) and Parkinson's disease. These disorders place a heavy burden on the healthcare system and have a great impact on the quality of life of affected patients. Many diseases remain incurable after decades of research, and the medicines currently in use can only partially halt the disease's course. However, bioinformatics has emerged as a powerful tool in the fight against neurodegenerative diseases. Bioinformatics is an interdisciplinary field that amalgamates computer science, statistics, and biology to answer biological questions and interpret biological data. In recent years, bioinformatics has evolved in value in the investigation of neurodegenerative disorders, accelerating the study of these diseases by giving researchers strong tools for analyzing and interpreting biological data. It has also opened up new directions for research and the creation of new treatments.



Genome data visualisation

How Bioinformatics Helps Fight Against Multiple Sclerosis:

Identification of Genetic Risk Factors:

Bioinformatics has shown great promise in identifying genetic risk factors for neurodegenerative diseases, including Multiple Sclerosis. For instance, Multiple Sclerosis is a long-term autoimmune disorder of the central nervous system that leads to progressive loss of myelin and nerve cells. It comes with an array of symptoms such as muscle weakness, tremors, and cognitive impairment. Numerous genes and environmental factors are known to contribute to the development of Multiple Sclerosis. Multiple Sclerosis is a complicated disease with a strong hereditary component. Researchers have discovered specific genes and genetic variants linked to an elevated risk of getting Multiple Sclerosis by examining big genomic datasets. For example, variations in the Human Leukocyte Antigen (HLA) genes increase the risk of developing Multiple Sclerosis.

Studying the Molecular Pathways Involved in the Disease:

Bioinformatics has been employed to study the underlying molecular pathways of neurodegenerative diseases. For example, researchers have used bioinformatics to examine changes in gene expression and protein levels in the brains of individuals with MS. This has led to the identification of specific biological pathways disrupted in the disease, such as the regulation of immune response, inflammation, and oxidative stress.

Identification of Therapeutic Targets:

Identification of potential therapeutic targets for Multiple Sclerosis is an important application of bioinformatics. Researchers have identified individual proteins and other molecular targets that may be important in the onset and

Bioinformatics Pioneering the Battle Against Neurodegenerative Diseases: A Focus on Multiple Sclerosis

progression of Multiple Sclerosis by examining massive biological data sets. Bioinformatics has been utilized, for instance, to identify possible targets associated with the regulation of immunological response, neuroprotection, and myelin repair.

Identification of Genetic Risk Factors:

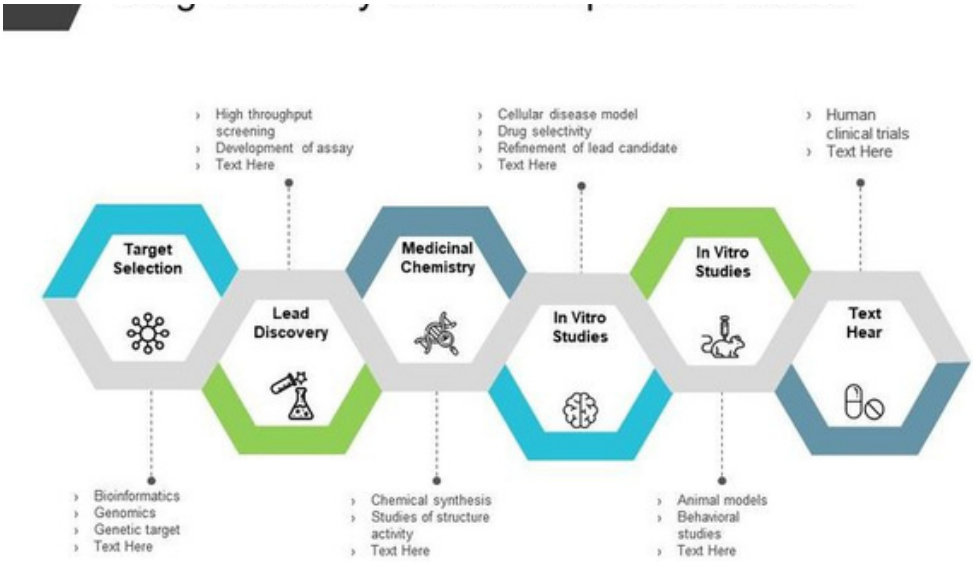
Potential Multiple Sclerosis biomarkers have also been found using bioinformatics. Biomarkers are markers that can be used to determine if a disease is present or is progressing. Researchers have discovered particular biomarkers that may be helpful in the diagnosis of Multiple Sclerosis, detecting disease progression, and tracking the response to treatment by examining vast datasets of biological data.

Drug Discovery:

One of the most important applications of bioinformatics is drug discovery. Drug discovery is a complex and time-consuming process that involves the identification, design, and development of novel drugs for treatment of specific diseases. At the beginning of the drug discovery process, researchers identify potential drug targets, which are proteins or other molecules involved in a specific disease. Bioinformatics tools can be used to analyze the structure of these targets, as well as their function and interactions with other molecules. This analysis allows researchers to identify potential drug candidates that can interact with the target molecule and treat the disease.

In conclusion, bioinformatics has fundamentally changed how we understand neurodegenerative diseases like Multiple Sclerosis and created new opportunities for investigation and the creation of novel therapies. The application of bioinformatics has enabled the identification of genetic risk factors, molecular mechanisms, prospective therapeutic targets, and biomarkers, and this field will surely continue to play a crucial role in the struggle against neurodegenerative diseases in the future. Although the use of bioinformatics in the treatment of neurodegenerative illnesses is still in the preliminary stage, it holds great promise for enhancing the lives of patients with these diseases with continuous research and innovation.

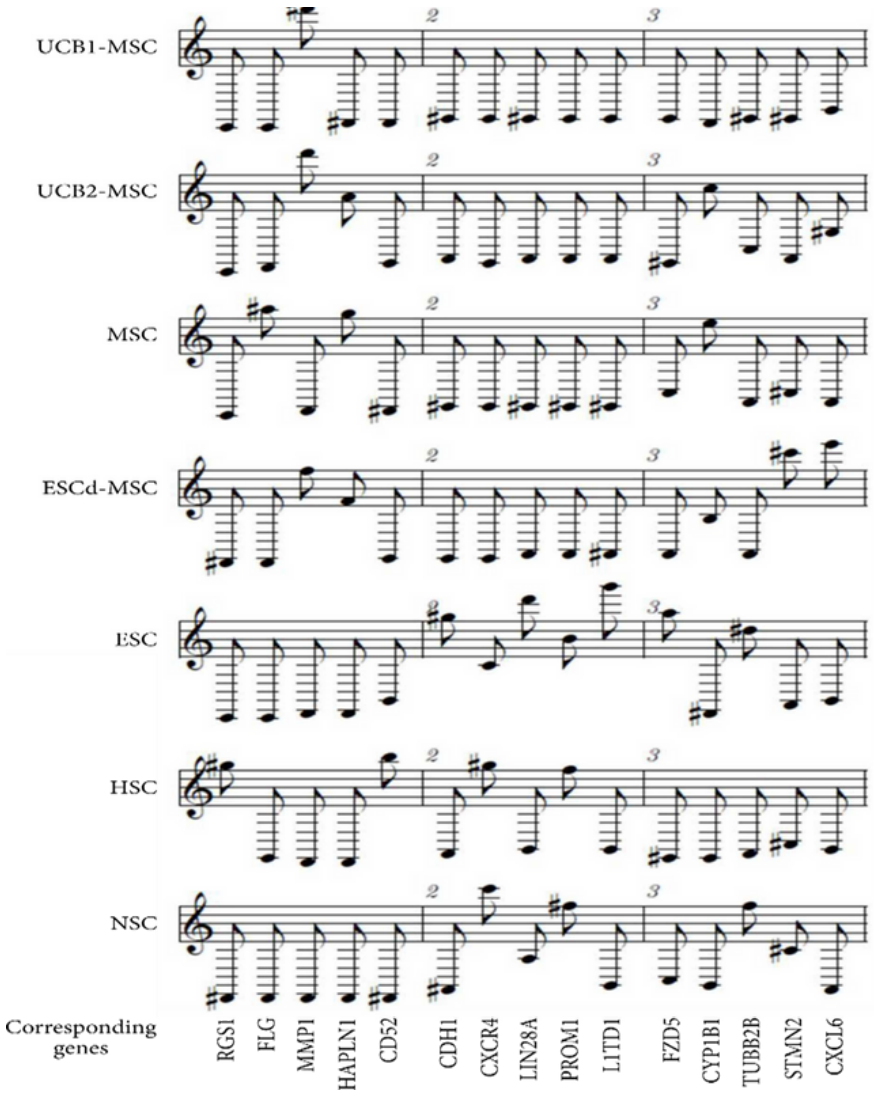
S. Helina Hilda
(Class of '24)



Geno-Music: Harmonizing Genetic Data into Melodies for Cancer Exploration

Gene expression music, also known as Geno-music or biocomputational music, is a genre of music generated from biological data, such as DNA sequences or gene expression data. The concept involves transforming data into musical notes and rhythms, offering a unique auditory representation of biological information. Various methods exist for creating gene expression music, with one common approach mapping the four DNA nucleotides to musical notes or chords. For instance, adenine may be mapped to the note C, guanine to G, cytosine to A, and thymine to D.

The expression levels of genes can then be used to determine the duration and intensity of the notes, creating a melody that represents the gene expression pattern. The resulting music can provide a new perspective on the biological data and help researchers identify patterns that might not be immediately apparent from looking at graphs or charts. It can also be a fun and engaging way to communicate science to the general public. However, it's important to note that gene expression music is a creative interpretation of scientific data and should not be used as a replacement for rigorous analysis or as a primary source of scientific information.

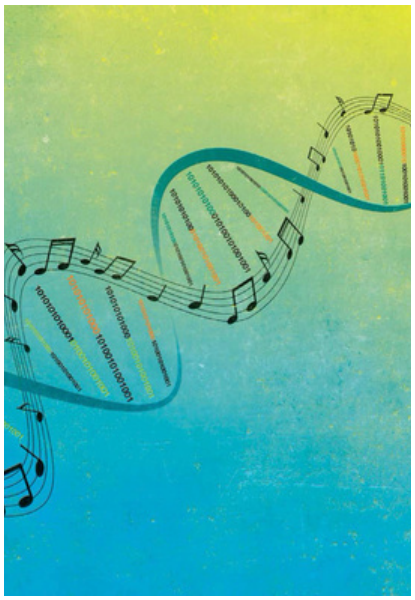


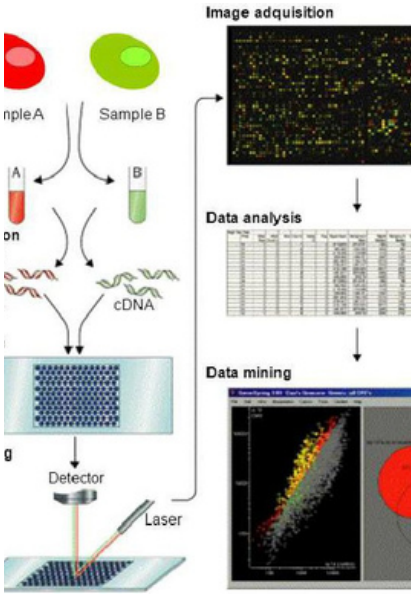
Characterization of the Ewing Sarcoma Stem Cell Signature based on Gene Expression Music Algorithm

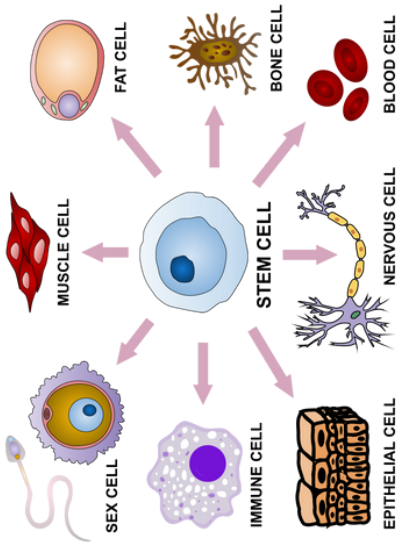
Ewing sarcoma is a type of bone cancer that affects children and young adults. The exact cause of Ewing sarcoma is unknown, but it is thought to be related to changes in specific genes. Researchers have used music to explore the genetic patterns of Ewing sarcoma. In recent years, researchers have started to use a technique called "sonification" to turn genetic data into music. This process involves assigning musical notes to different genetic markers and then creating a piece of music that reflects the patterns of gene expression in a particular sample.

One study published in the journal Genome Biology used sonification to analyze the genetic patterns of Ewing sarcoma. The researchers created a musical composition that highlighted the different genetic changes that occur in Ewing sarcoma tumors. This technique can help researchers visualize and explore complex genetic data in a new way.

Geno-Music: Harmonizing Genetic Data into Melodies for Cancer Exploration







Additionally, it may make it easier for non-scientists to understand and appreciate the complexities of cancer genetics.

A technique called Gene Expression Music Algorithm (GEMusicA) is used for conversion of DNA microarray data into melodies that are used for identification of differentially expressed genes. This technique can be used for distinguishing between various stem cell types by comparing the gene expression profiles of these stem cells.

Methods:

1.Microarray datasets collection

DNA microarray data for Ewing sarcoma cell lines and biopsies from 10 studies were downloaded from GEO. DNA microarray cel files from the following stem cells were RMA (Robust Multiarray Average) normalized and used for the selection of probe sets which exhibited differential expression between different stem cell types :

- Embryonic stems cells (ESC)
- Neuronal stem cells (NSC)
- Hematopoietic stem cells (HSC)
- Different subtypes of mesenchymal cells (MSC)

These cel files were processed using the GEMusicAR script

2.GEMusicA Analysis

GEMusicA algorithm filters the microarray probe sets with high variance of the signal intensities. This method was used to characterize tumour specific gene expression profiles. GEMusicA algorithm generates melodies from the combined DNA microarray data of all the stem cells.

Hence, GEMusicA is useful for characterization of stem cell-type specific gene expression signatures.

References:

Staege MS. Gene Expression Music Algorithm-Based Characterization of the Ewing Sarcoma Stem Cell Signature. Stem Cells Int. 2016;2016:7674824. doi:10.1155/2016/767482

Rindhya
(Class of '24)

The Benjamin Button of Jellyfish: Decoding the Genomic Secrets of Immortality

All organisms born on Earth grow, age, and die. We, humans, do too, with an average life span of about 70 years. But, can you imagine that an advanced species like us, the smartest and most evolved being, was beaten in age by a small jellyfish, barely the size of half a thumbnail?

Turritopsis dohrnii : The Immortal Jellyfish

Meet Turritopsis dohrnii, one of the only three cnidarian species in the world to have the ability to regenerate after sexual reproduction. Unlike its cousins, T. dohrnii is the only one that maintains its high rejuvenation potential (~ 100%) in post-reproductive stages. It has cracked the secret to immortality, confusing scientists and defying the understanding of cellular senescence or genomic instability with the same genomic structures and genes as bilaterians.

The Mystery Unveiled: Genomic Analysis

We're all wondering the same thing, right? How? How does such a tiny creature defy the laws of nature? Maria Pascual-Tornera and her team set out to solve the mystery. They sequenced the genomes of both T. dohrnii and T. rubra, a closely related species with no post-reproductive rejuvenation, and used comparative genomic analyses to identify and understand molecular mechanisms of almost 1,000 genes related to aging and DNA repair between both

species, as well as between them and the cnidarians Hydra vulgaris, Clytia hemisphaerica, and Aurelia aurita.

Genome Size and Amplification: A Closer Look

First things first, they studied the genome sizes of the two Turritopsis species. Using MAKER, the team predicted a set of 17,468 genes in T. dohrnii and 9,324 genes in T. rubra. Repetitive elements represented around half of the genome, and 28 copy number variations and 10 variants unique to the species were identified. Naturally, the larger the size, the more gene amplification, right? Absolutely true, but what was rather interesting was the amount of amplification observed in T. dohrnii, which was 3.77 times higher than the representative gene set of the WHOLE genome in this species. The reason was likely due to positive or neutral selection processes, rather than genome expansion.

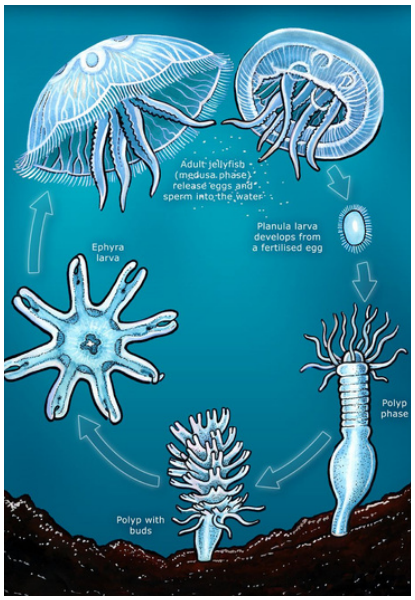
Factors Contributing to Aging: A Comprehensive Approach

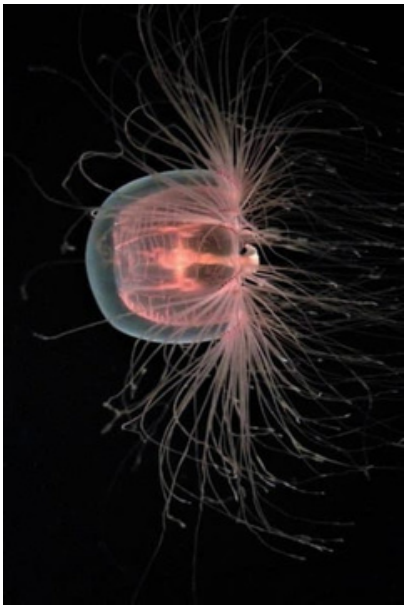
The team focused on 5 main factors that contributed to aging - DNA replication and repair, response to oxidative stress, telomeric lengths, cell communication pathways, and transcription regulation.

DNA Replication and Repair Mechanisms:

Aging and DNA replication and repair go hand in hand. It was found that T. dohrnii may have more efficient and enhanced replicative mechanisms and repair systems than the other cnidarians. For instance, the amplifications of four copies of POLD1 and two copies of POLA2, both encoding DNA Polymerase genes in T. dohrnii, were found compared to one copy in T. rubra. Duplication of replication factors (RFC3) and topoisomerases(TOP3B) was also detected.

The Benjamin Button of Jellyfish: Decoding the Genomic Secrets of Immortality





Response to Oxidative Stress:

Response to oxidative stress dictates genomic stability in a cell, as it prevents internal damage and brings about homeostasis. In this regard, enhanced copies of thioredoxin(TXN), and glutathione reductase (GSR) were present in T. dohrnii. Overexpression of these genes resulted in maintaining the redox environment of the cell, thereby increasing lifespan.

Telomeric Stability:

Telomeric degradation is one of the major indicators of cellular aging. Several variations in the telomerase and shelterin complexes were observed. Two amplified copies of ribonucleoprotein homolog (GAR1) in the genome of T. dohrnii, while other cnidarians only have one copy of this gene, suggesting a finer regulation of telomerase activity.

Cell Communication Pathways:

Genes related to apoptosis pathways, such as BMP7, had eight copies in T. dohrnii (five of them found to be active during LCR), in contrast to five copies in T. rubra and three copies in H. vulgaris. Fascinatingly, none of the other genes were amplified. Instead, overexpression was observed in caspase (CASP3), B-cell lymphoma (BCL2), and apoptosis regulator (BAX) during LCR.

Transcription Regulation:

Deficient transcriptional regulation leads to the loss of proteostasis, another major hallmark of aging. Increased copy number variations affecting genes that act as chromatin-binding modulators (MORC3), involved in transcriptional and posttranscriptional regulation were detected. Gene amplifications in microtubules and cytoplasmic elements also lead to increased neuronal function and cell plasticity.

Cell Reprogramming Mechanisms:

Finally, the team found two of the main mechanisms for cell reprogramming, the first being the silencing of the Polycomb Repressive Complex 2 (PRC2) target, which is a key chromatin modifier involved in the maintenance of transcriptional silencing, and the second being the activation of pluripotency targets, thereby increasing the ability of a cell to differentiate and form any type cell in the organism.

In conclusion, the jellyfish had both amplification and overexpression of genes related to DNA repair and replication, telomeric stability, cellular signaling, and transcription regulation, which is how it had cracked the secret to everlasting youth, effectively cheating senescence and living up to its name of 'The Immortal Jellyfish!

V Aruna
(Class of '24)

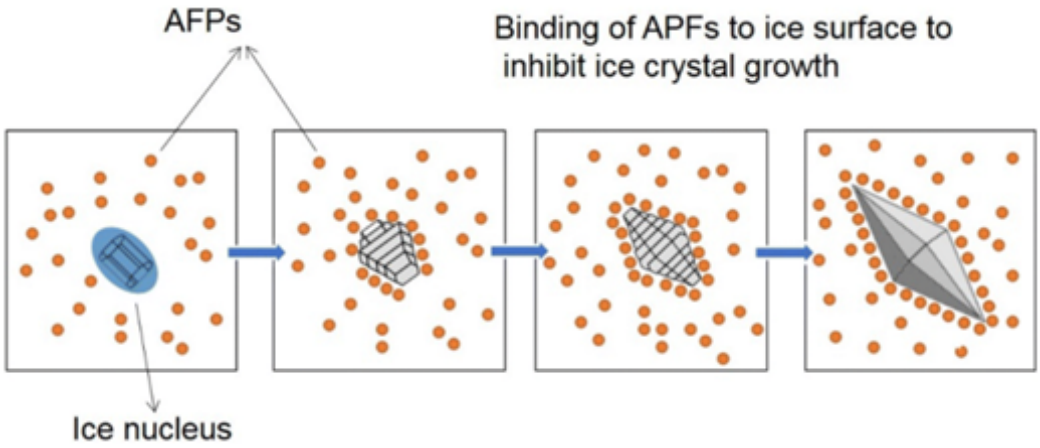
Marvel of Antifreeze Proteins: Exploring Their Significance

Structure of the Tenebrio molitor Beta-Helical Antifreeze Protein:

Antifreeze proteins (AFPs), also known as ice-binding proteins, possess the unique ability to bind to ice crystals and prevent their growth, thereby safeguarding the organism's tissues from freezing damage. The primary role of AFPs is to protect organisms in sub-zero temperatures by inhibiting the growth of ice crystals both outside and inside the cells. This is crucial as the formation of ice crystals within cells can lead to cell rupture and death. AFPs play a vital role in the survival of organisms in extremely cold environments, such as the Arctic and Antarctic.

AFPs have independently evolved in various organisms, including fish, insects, plants, and bacteria. Despite their diverse origins, AFPs share common characteristics, including a high content of specific amino acids like Alanine and Threonine, essential for binding to ice crystals.

Mechanism of action of AFP



AFPs are intriguing proteins that have evolved in response to extreme environmental conditions, thriving in some of the coldest regions globally. Their ability to bind to ice crystals and prevent their growth presents numerous potential applications, forming an active area of ongoing research.

Mechanism of AFP:

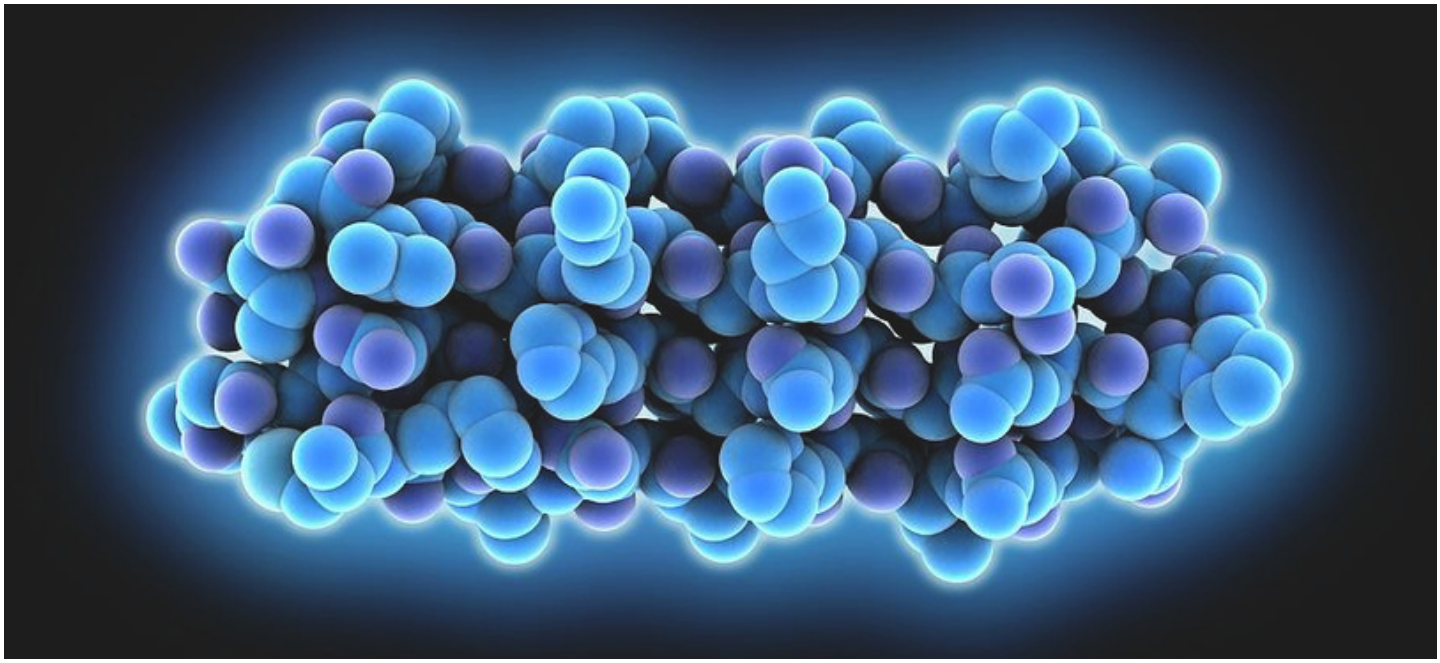
Antifreeze proteins (AFPs) function by binding to the surfaces of ice crystals and inhibiting their growth, thereby preventing the formation of larger ice crystals. This is achieved through the unique molecular structure of AFPs, which contain specific amino acid sequences that allow them to bind to the surface of ice crystals.
When ice crystals begin to form, they do so at specific sites, called nucleation sites. AFPs bind to these nucleation sites, preventing further ice crystal growth. The specific mechanism by which AFPs bind to ice crystals is still the subject of ongoing research, but it is thought to involve the formation of hydrogen bonds between the AFPs and the ice crystal surface.
AFPs work by lowering the freezing point of water, which is the temperature at which water freezes into ice. Normally, the freezing point of pure water is 0 degrees Celsius. However, when AFPs are present, they can lower the freezing point of water by several degrees, depending on the concentration of the AFPs.

Applications of AFP:

Antifreeze proteins (AFPs) hold broad potential applications across various fields, including the food industry, medical field, and biotechnology.

- **Food Industry**

In the food industry, AFPs have been used as cryoprotectants to prevent ice crystal formation in frozen foods, which can damage their texture and flavor. They have also been investigated for their potential to extend the shelf life of perishable foods, such as fruits and vegetables, by protecting them from freezing damage during storage and transportation.



- **Medical field**

  In the medical field, AFPs have been studied for their potential to improve the cryopreservation of cells, tissues, and organs. Cryopreservation is the process of freezing and storing biological material at very low temperatures, which can damage the cells and tissues due to ice crystal formation. By using AFPs as cryoprotectants, it may be possible to reduce ice crystal formation and increase the survival rate of cells and tissues during the freezing and thawing process. AFPs may also have potential applications in improving the preservation of blood and other biological fluids during storage and transportation.

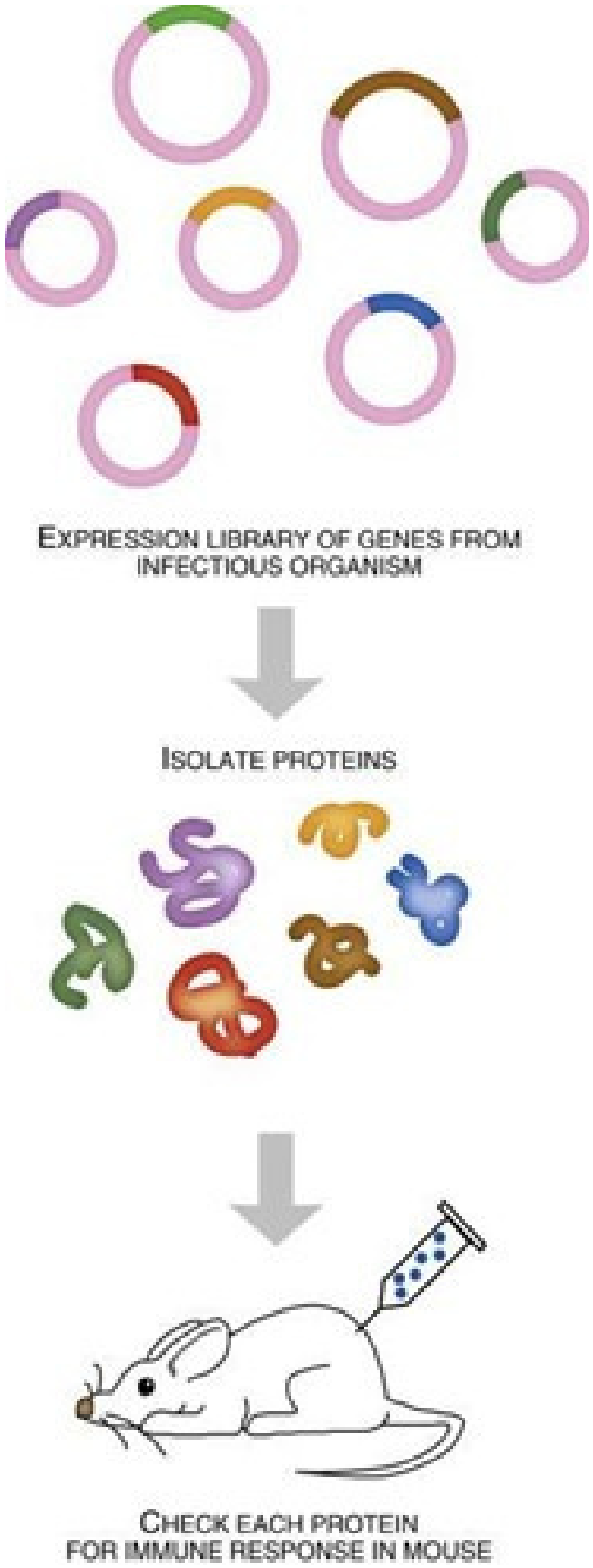
- **Biotechnology:**

In biotechnology, AFPs have been used as tools for protein purification, as well as in the development of biosensors and diagnostic assays. They have also been investigated for their potential to improve the performance of industrial enzymes and to protect cells and microorganisms from freezing damage during bioprocessing.

Overall, AFPs are a group of interesting proteins that have a diverse range of potential applications, and their unique ice-binding properties make them valuable tools for protecting biological material from freezing damage in a variety of fields.

Ananya S
(Class of '25)

Reverse vaccinology



Vaccines are the pharmaceutical products that offer the best cost-benefit ratio in the prevention or treatment of diseases. Vaccine development and production are costly and it takes years for this to be accomplished. Several approaches have been applied to reduce the time and cost of vaccine development, mainly focusing on the selection of appropriate antigens or antigenic structures, carriers, and adjuvants. One of these approaches is the incorporation of bioinformatics methods and analyses into vaccine development.

The success of vaccination is reflected in its worldwide impact by improving human and veterinary health and life expectancy. In addition to the invaluable role of traditional vaccines to prevent diseases, the society has observed remarkable scientific and technological progress since the last century in the improvement of these vaccines and the generation of new ones. The application of bioinformatics strategies in vaccine design and development shows some successful examples of vaccines in which bioinformatics has furnished a cutting edge in their development.

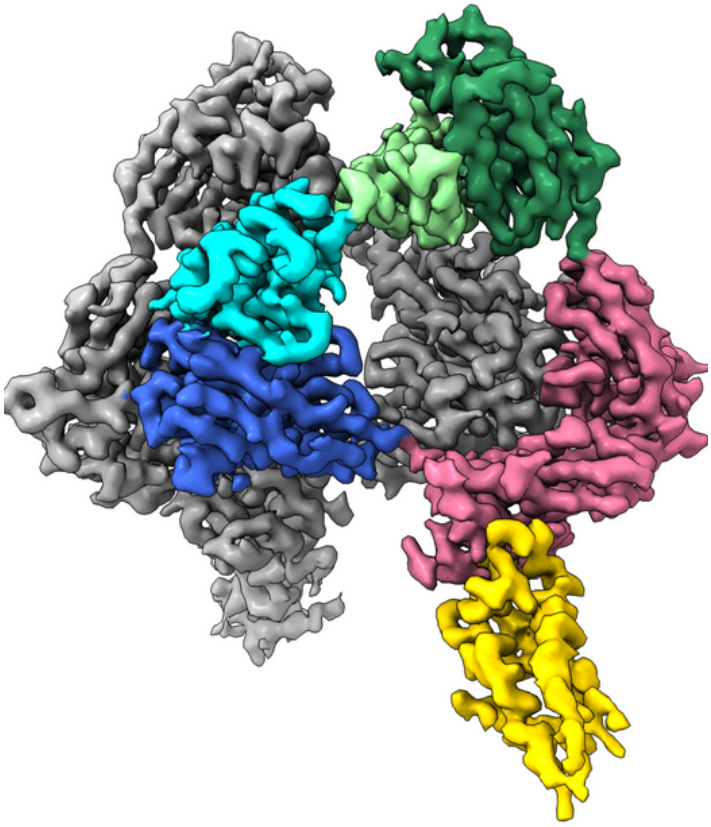
Reverse vaccinology:

Reverse vaccinology is a methodology that uses bioinformatics tools for the identification of structures of bacteria, virus, parasites, cancer cells, or allergens that could induce an immune response. The characteristics of each protein in the proteome under study should be analyzed, employing different bioinformatics approaches to select the proteins with the best properties for testing through in vitro and in vivo assays, in order to demonstrate its safety and immunogenicity.

With the best vaccine candidates, different types ofvaccines can be designed and developed, For example, subunit, recombinant, and nucleic acid vaccines . These technologies have been used to study pathogenic agents including eukaryotic organisms and those involved in diseases transmitted by vectors. Hence, this helps us to design and obtain vaccines not only for humans but also for animals.

Structural Vaccinology:

Structural vaccinology focuses on the conformational features of macromolecules, mainly proteins that make them good candidate antigens. This approach to vaccine design has been used mainly to select or design peptide-based vaccines or cross-reactive antigens with the capability of generating immunity against different antigenically divergent pathogens. There are many bioinformatics programs that predict protein epitopes.



The approach that has been employed to develop vaccines is to perform several bioinformatics analyses at both the sequence and structure levels. Bioinformatics analyses have been done to improve the functionality of antibodies. One premise of bioinformatics is to detect epitopes that can be recognized by antibodies, but modeling antibody-antigen complexes has been difficult because of the mobility of protein loops in the Fab region of antibodies.

Vaccines against infectious and non-infectious diseases:

Infectious diseases: Tuberculosis

Tuberculosis is an infectious disease caused by Mycobacterium tuberculosis, which is the most virulent and transmissible bacteria. One strategy applied for vaccine design is to identify the structures that are present only in M.tuberculosis and absent in Mycobacterium Bovis BCG. Several candidates and epitopes have been found with different softwares. Some of these have been expressed and proven in vitro and in vivo, demonstrating their immunogenicity and protective effect.

Non-infectious diseases: Cancer

Since T cells present in the thymus do not recognize mutated antigens expressed in cancer cells, there is no negative selection, and these neoantigens are ideal targets for therapeutic vaccination; furthermore, they are not present in healthy tissue. The complexity of some experimental tools such as mass spectrometry hampers its usefulness in the selection of targets in a clinical setting where personalized therapy is needed. It is not possible to analyze all of the mutations, bioinformatics addresses this problem and has become important in the selection of targets.

A drug or vaccine against these deadly diseases can be developed using Bioinformatics approaches, as it provides critical insight into the genetic makeup of the pathogens.

VARSHEENEE A
(Class of '25)

CARBON FOOTPRINT IN BIOINFORMATICS

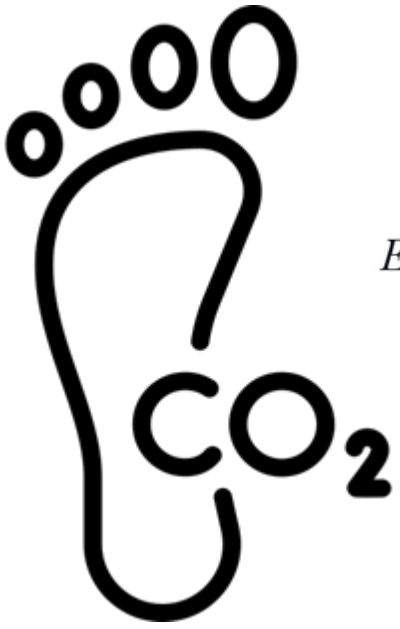
A carbon footprint is **the total amount of greenhouse gases (including carbon dioxide and methane) that are generated by our actions.** Now a question will arise that how carbon footprint and bioinformatics were relatable. Yes, it is!

The role of bioinformatics in biological research requires the use of large-scale computational resources to analyze large and complex data sets. Greenhouse gas (GHG) emissions caused by the usage of computers is harmful for human health. Air pollution is one of the world's leading risk factors for death, attributed to millions of deaths each year. Air pollution is attributed to **11.65%** of deaths globally.

The carbon footprint of bioinformatics tools and analyses can be estimated by using the Green Algorithms tool which is available online. The Green Algorithms project **aims at promoting more environmentally sustainable computational science.** It regroups calculators that researchers can use to estimate the carbon footprint of their projects.

Sequence alignment, gene prediction, genome assembly, genome-wide metagenomics, RNA sequencing, and phylogenetics analysis are some of the bioinformatics analyses that use various softwares like Pymol and MegaX which has a considerable effect on carbon footprint.

ESTIMATING THE CARBON FOOTPRINT:



The carbon footprint is measured in kilograms of CO2-equivalent (CO2e), which is the amount of carbon dioxide with an equivalent global warming impact as a mix of GHGs.

C = E ×CI

Here C is the carbon footprint, E is the energy needed, and CI is the carbon intensity.

The energy needs of an algorithm are measured based on running time, processing cores used, memory deployed, and efficiency of the data center:

$$E = (n_c \times P_c \times u_c + n_m \times P_m) \times t \times \mathrm{PUE} \times 0.001$$

Here, nc is the number of computing cores, pc is the power used by "one" core(Watt), uc is the
core usage factor, nm is the memory available for computing (GB), pm is the energy used by memory (Watt), t is the running time(h), PUE is the power usage effectiveness of the data center.

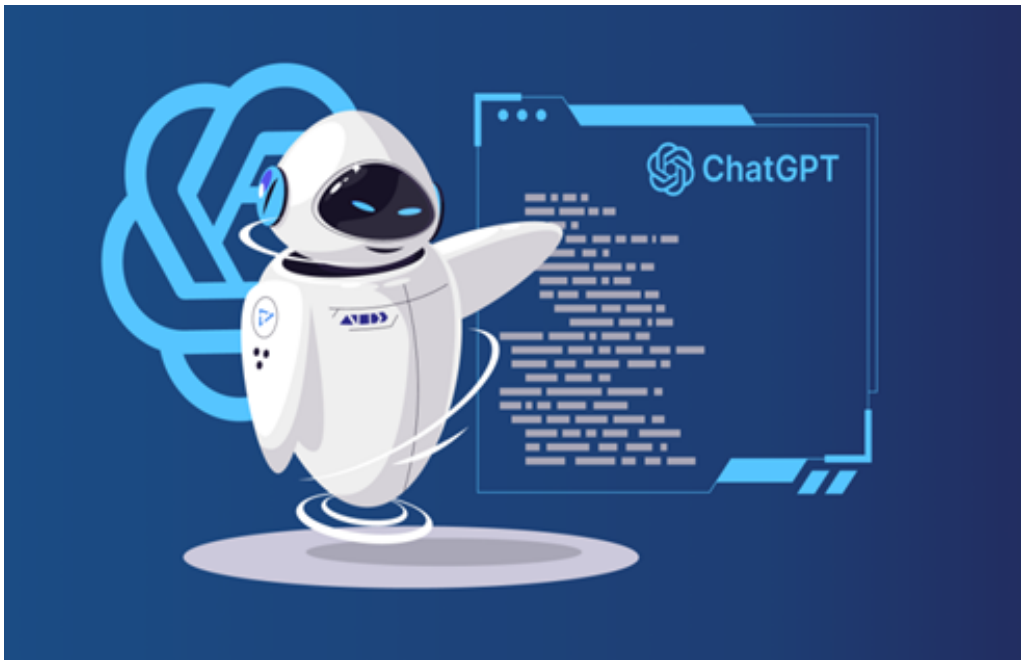
NOORUL AEIN
(Class of '25)

ChatGPT: A Promising Solution for Complex Biological Data Analysis in Bioinformatics

What is ChatGPT?

Artificial intelligence (AI) is a rapidly growing field that focuses on developing machine learning models that can learn, solve problems, and make decisions with the aid of human intelligence. One of the most intriguing applications of AI is natural language processing (NLP), which involves developing machines that can understand and generate human language. The potential of NLP in AI is best demonstrated by ChatGPT, a vast language model created by OpenAI.

ChatGPT has shown immense potential in numerous fields, including bioinformatics, the field that analyzes and interprets biological data using computational techniques. The volume of biological data being produced has increased exponentially since the development of high-throughput technologies, offering a considerable challenge to traditional analytical techniques.



Applications of ChatGPT in Bioinformatics

One of the main applications of ChatGPT in bioinformatics is the analysis of genomic data. The genomic data is enormous and analyzing it using traditional techniques can be time-consuming and computationally intensive. Certain analytical tasks like identifying relevant genes or predicting the functions of unknown genes can be automated using ChatGPT. Moreover, ChatGPT can be used to extract data from the enormous volume of genomics literature, facilitating quick access to relevant information.

Another area where ChatGPT is useful in bioinformatics is in protein structure prediction. Protein structure prediction is a challenging task, and current approaches have some accuracy issues. By examining the amino acid sequence and predicting the secondary structure, ChatGPT can be utilized to predict not only the protein structure but also to predict the protein-protein interactions which are crucial in understanding the biological pathways.

In order to classify the biological data for determining disease subtypes or for predicting drug targets, ChatGPT can be employed. ChatGPT can identify patterns and correlations in huge datasets that might not be evident to human researchers. This strategy can aid in the identification of new drug targets and the creation of patient-specific therapies and treatments.

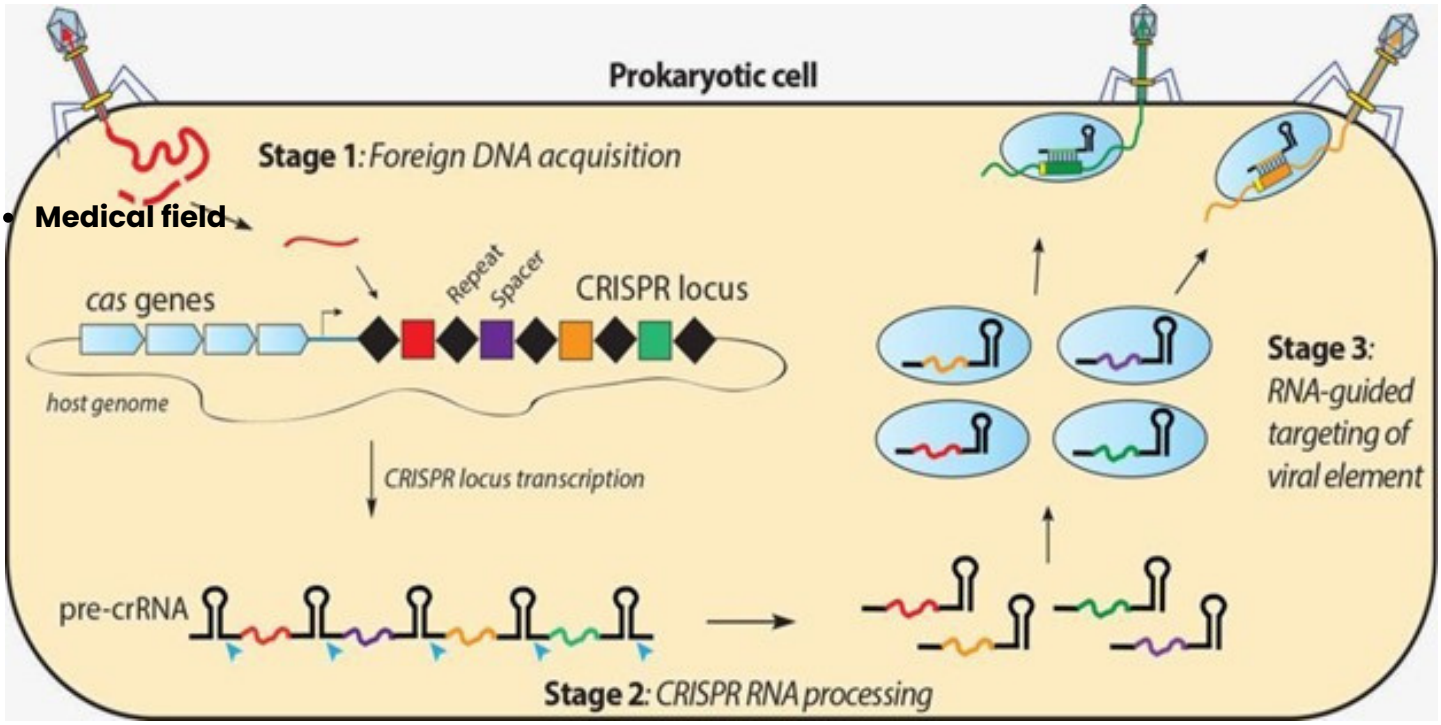
ChatGPT's ability to learn from large datasets is one of the added advantages. By training the model on enormous amounts of biological data, ChatGPT can develop a deep understanding of the relationships between different biological entities. This can result in more accurate predictions and faster analysis.

In conclusion, ChatGPT is a promising solution for complex biological data analysis in bioinformatics. It has the potential to revolutionize the field and overcome the challenges posed by large volumes of data. Its ability to process complex data sets accurately and efficiently will undoubtedly play a crucial role in advancing research and development in bioinformatics.

SHRICHARAN S
(Class of '25)

Editing a manuscript may be boring sometimes, but genome editing is always interesting. They say "prevention is better than cure" while both prevention and cure are genetically possible by CRISPR technology by editing the genome. CRISPR (Clustered Regularly Interspaced Palindromic Repeats) is the bacterial or archaeal adaptive immune system found against viruses or bacteriophages. Yoshizumi Ishino was the first person to describe the architecture of the CRISPR arrays. By bioinformatics analyses, he disclosed that the system was found in many archaeal and bacterial genomes. Later, bioinformatics studies revealed that CRISPR-cas systems target DNA rather than RNA.

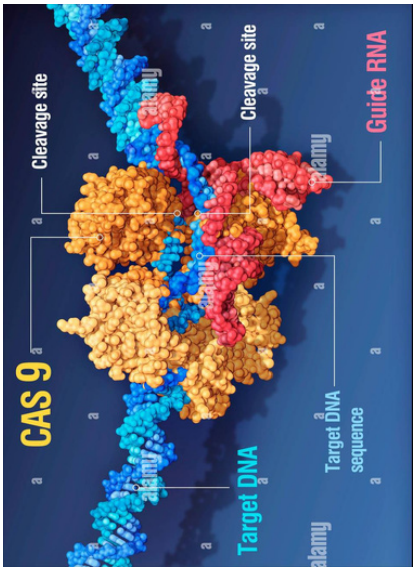
An insight into CRISPR-cas system

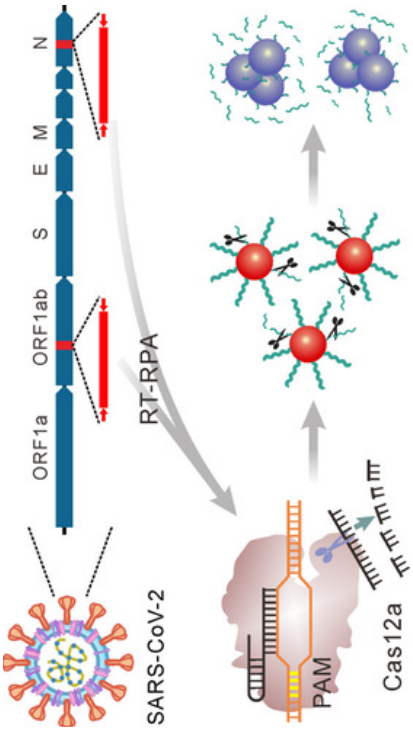


- **Medical field**

CRISPR-cas systems consist of an array of repeats separated by spacers, leader sequence and cas proteins that are required to process the information within the CRISPR array. Cas9 which belongs to class 2 type II CRIPSR system is the most used genome editing tool. Streptococcus pyogenes Cas9 was the first CRISPR system to be reprogrammed for the genome editing of eukaryotes. Cas12a also has been reprogrammed for gene editing in human cells.

CRISPR Cas system : The saviour of the future





Applications of CRISPR-cas system

The CRISPR cas system is mostly useful in point mutations which can lead to genetic disorders. There is much research going on about the cas proteins to use the CRISPR-Cas systems as genetic scissors. Nowadays, people are interested in using these systems in gene-edited babies where the features of the zygote are designed and created.

Researches are using the CRISPR cas system to destroy the HIV proviruses which is prevalent nowadays.Also, the SARS-COV-2 virus which was responsible for the COVID-19 pandemic, is detected by CRISPR-Cas12a assay. Viral diseases are difficult to treat while CRISPR-Cas systems may provide an outbreak for it. WHO has declared that Antimicrobial Resistance is one of the top 10 global health threats against humans. The CRISPR-Cas system provides an option for the development of the next-generation antimicrobials to fight against the infectious diseases caused by AMR pathogens.

Bioinformatics and CRISPR-cas system

Application of the technology depends on the selection of CRISPR-cas system, gRNA(guide RNA) design, transfection and screening. This selection is made one step easier with the help of the bioinformatics tools and computational techniques. CRISPRCasFinder is one of the most popular tools that help in the CRISPR array identification. CRISPRdisco predicts the CRISPR array along with the identification of the cas proteins. Even for the metagenomics studies, CRASS and metaCRISPR were developed for identifying the arrays for unassembled sequences. There are many databases developed for CRISPR-Cas proteins like CRISPRdb, CRISPRone and for Anti-CRISPRdb for anti-CRISPR proteins.

Thus, we have a long way to go in this field and establish better solutions for the available problems using this technology.

DHARANI G
(Class of '25)

DNA AS A STORAGE DEVICE TO STORE DIGITAL INFORMATIONS

EDNA is nature's oldest storage device which stores biological information in sequences of four bases of nucleic acid. But what if DNA can store digital information whatever we want likes photos, documents, etc.

Need for DNA as a storage device:

There are several reasons why DNA can be a potential storage device for digital information.

1. **High storage density**, DNA is a very dense storage medium with the capacity to store a wide range of information in a very small space. 1gm of DNA can store billions of terabytes of data. This makes it an attractive option for the long-term archiving of large datasets.
2. **Long-term stability**, as DNA is a very stable molecule it can survive for thousands of years under the right conditions.
3. **Energy efficiency**, unlike traditional storage devices like hard drives, and flash memory, DNA does not require any electricity or constant power to maintain storage capacity.

The data storage density of silicon chips is limited, and magnetic tapes need high maintenance and also begin to degrade within 20 years. Silicon has less storage ability and also has more disadvantages like environmental pollution and human health hazards. So DNA is a good option for storage due to its endurance, durability, and a higher degree of compaction.

The process of encoding and decoding DNA is still in its early stages, and many of the tools and techniques used in DNA synthesis and sequencing are already well known. This indicates that the infrastructure and expertise needed to work with DNA are already in place, making it a potentially practical and cost-effective solution.

Steps for storing digital information in DNA

1. **Encode digital information into a DNA sequence:**

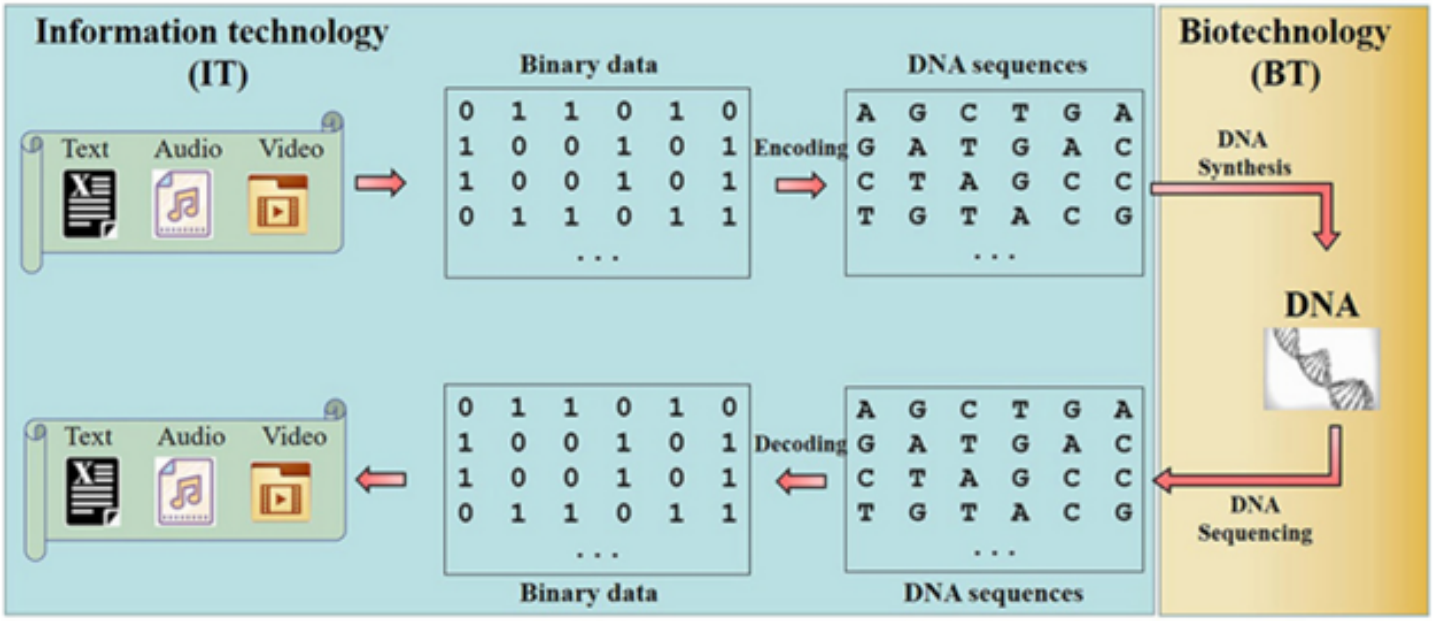
DNA is made up of four nucleotide bases like adenine, guanine , thymine, and cytosine . It is important to modify the digital information (which is in 0s and 1s) into these four nucleotide bases. Once each chunk has been assigned to a DNA base, concatenate them together to form a DNA sequence. This sequence can be synthesized in the laboratory and stored in Bacterial cells.

2. Synthesize the DNA in the lab and insert the DNA into Bacterial cells:

Synthesizing DNA in the lab involves chemical reactions to create a sequence of nucleotides that form a specific DNA sequence. This process can be done using various techniques, including polymerase chain reaction (PCR) and gene synthesis. Once the DNA is synthesized, it can be inserted into Bacterial cells using a process called transformation.

3. Amplify the DNA periodically:

Periodic amplification of DNA can be achieved by repeating the PCR cycles multiple times until the desired amount of DNA has been produced. Care should be taken to optimize the PCR conditions, including the primer concentration, annealing temperature, and extension time, to ensure efficient and specific amplification of the target DNA sequence.



4. Extract the DNA and decode the information back into digital format when needed:

The DNA is sequenced to decode the information back into digital format. This involves breaking down the DNA into its individual nucleotides and reading the sequence using a sequencing technology such as Next-Generation Sequencing (NGS). Once the DNA sequence has been read, specialized software can be used to translate the sequence back into the original digital information. This process can be performed using a computer or other digital device.

Challenges in using DNA as storage system:

There are several challenges to using DNA for digital storage.

1. First, synthesizing and reading DNA sequences is still relatively expensive and time-consuming.
2. Second, there are still technical challenges to overcome in terms of encoding and decoding digital information into DNA sequences.

Storing digital information in DNA is currently a slow and expensive process, and it requires specialized equipment and expertisewhich makes it less accessible and less reliable than other digital storage methods.

Another potential limitation is the risk of errors during the encoding and decoding process, which could result in data loss or corruption. And also there are ethical and regulatory concerns surrounding the use of DNA as a data storage medium, particularly in terms of privacy and security.

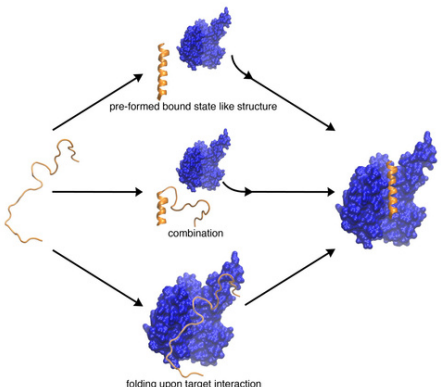
Overall the introduction of DNA as a storage device represents an innovative approach to store digital data for long term that could have a wide range of applications in many fields. This emerging field of DNA as means of data storage has the potential to transform science fiction into reality.

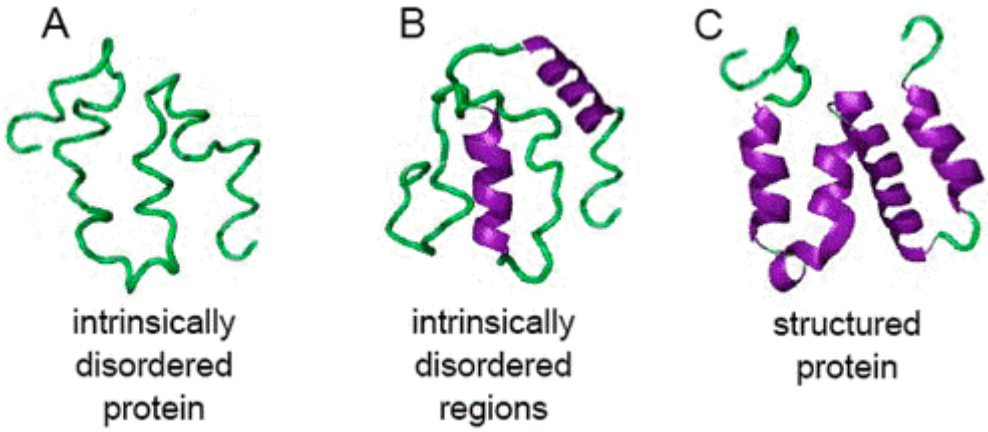
KEERTHANALAKSHMI R
(Class of '25)

Intrinsically Disordered Proteins: Unleashing the Power of Disorder

Proteins that do not follow the traditional framework of rules or the basic lock and key model are called intrinsically disordered proteins. They are also called as intrinsically unstructured proteins. They are characterized by biased amino acids and low sequence complexity. The intrinsically disordered proteins have a low proportion of bulky hydrophobic groups and high proportions of charged and hydrophilic groups.

In some proteins only the regions are disordered and these are called as intrinsically disordered regions. The intrinsically disordered proteins do not have a predefined or well defined 3 dimensional structure. These proteins are highly flexible and adaptable. To classify a region or protein as intrinsically disordered they must at least contain 40 residues





Their high flexibility and structural instability is due to the amino acid sequence that is present. There are 2 types of amino acids in these sequences :

- Order promoting residues like Trp , Tyr , phe Ile, Val , Cys
- Disorder promoting residues like Ala , Arg , Gly , Gln , Pro , Glu , Lys

The intrinsically disordered proteins do not entirely rely on the spatial pockets of the tertiary structure but on the target bound conformation of the conformational selection of the protein. The intrinsically disordered proteins can undergo post translational modification and also regulate conformational ensembles that make it interesting for the researchers. The post translational modifications also called PTM act as switches controlling the activities of the protein .

The disorder in intrinsically disordered proteins highly helps in cell signaling as there is a high potential to bind to multiple sites or partners and they use different structures. The disordered regions are highly accessible in comparison as they have multiple binding motifs and the sites for post translational modifications help in the control of signaling pathways.

The intrinsically disordered proteins or intrinsically disorder regions have remarkable conformational flexibility and structural plasticity that break multiple rules such as the lock and key model , the structure , folding and functionality of protein. They play a very important role in cell signaling , protein-protein interaction and gene regulation networks . There are many experimental , computational and bioinformatics analysis and methods that combine to identify and characterize the disordered regions of proteins and thus know the biological process that is present in it and understand them.

Hence, Bioinformatics methods and approaches are highly useful in identification of intrinsically disordered proteins.

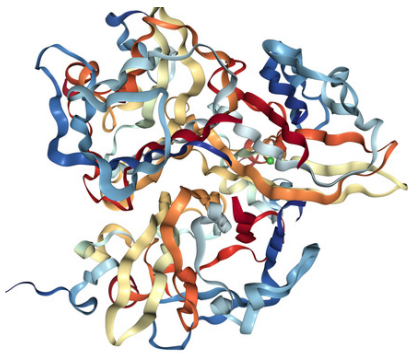
PRIYANKA R
(Class of '25)

Protein : The lifeline of living organisms

Proteins are the building blocks of life. Every cells in human body contain proteins. Proteins can be created from a set of 20 amino acids. Protein is essential for the maintenance and building of body tissues and muscle. DNA is transcribed into RNA, then RNA is translated into protein. Proteins catalyze chemical reactions, provide structural support, regulate substances, protect against disease, and coordinate signalling pathways. We use Ramachandran plot to validate protein structure.

What are Proteins ?

Proteins are biological macromolecules made up of different combinations of amino acids. Functional properties of the proteins depend on the three-dimension structure. Understanding the three-dimensional structure of a protein from its amino acid sequence is a long-standing goal. Proteins are composed of a polypeptide chain of amino acids that have specific functions and extend from the amino (N) to the carboxyl (C) terminus. Proteins are composed of polypeptide chains arranged in long strands and globular shapes.



Protein sequence analysis:

Protein sequences provide information about the preference of amino acid residues and their distribution for understanding secondary and tertiary structures of proteins. Identify similar motifs in protein sequences can help predict important regions. Single amino acid properties can be used to identify similar amino acids with similar properties.

Protein structure analysis:

Protein structures provide information about the stability of proteins, interactions between amino acids , residues present in the interior or surface of the protein. Secondary structure information of protein is used to predict three-dimensional structures, protein stability, and binding site residues.

Protein structure prediction:

The prediction of protein secondary structures is an intermediate goal for determining its tertiary structure. Several methods have been proposed to predict the structural class, secondary structure content, location of secondary structures, and modeling tertiary structures. The concept of "amino acid composition" plays a major role in predicting the structural class of globular proteins, and the success rate is limited.

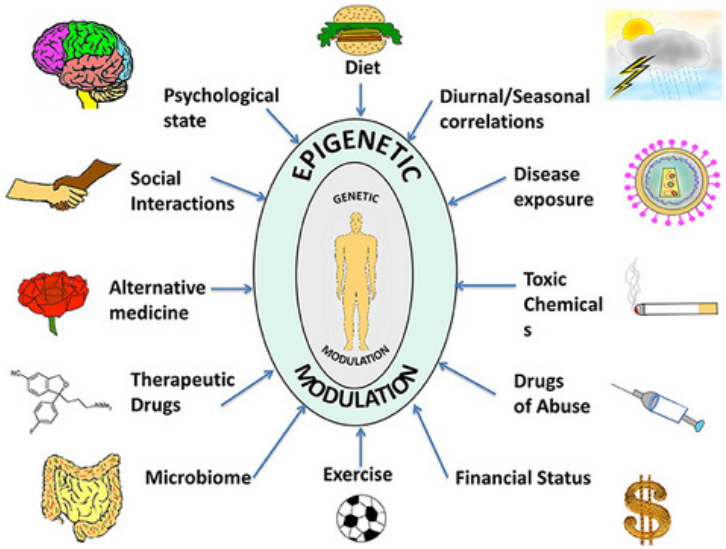
Proteins are composed of chains of amino acids that fold into three-dimensional shapes, forming stable structures that are well-adapted for their functions. Hence, proper functioning of proteins is necessary for all the biological processes in our body.

MONICA S
(Class of '25)

Epigenetics, which means literally "around" the gene allows us to see how environmental factors alter our gene expression in a specific place within each cell. As a result, we now know that when we take active control of these factors, we can learn to keep ourselves healthy.

So, what exactly is epigenetics and how does Bioinformatics help us in understanding it?
Epigenetics is the study of heritable genes that regulate the physiology of cells without changing the DNA sequence that underlies within them. It doesn't change the genetic code but it changes how the code is read. Genes are like blueprints and epigenetics is the contractor. It changes the assembly, and the structure.

The epigenetic factors that affect gene regulation are, histone protein modification, DNA methylation, chromatin modelling, and RNA-mediated silencing. They have a significant impact on how the DNA structure is changed. Epigenetic changes can be brought about naturally or artificially by illnesses and the environment.



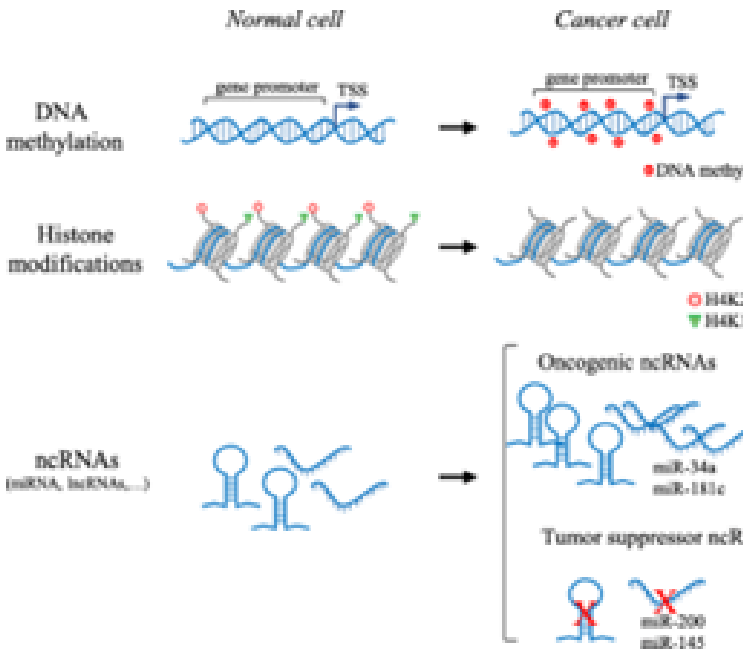
Bioinformatics has become one of the most useful approaches for understanding epigenetic mechanisms and analysing the epigenetic data. This data is generated using the following experimental techniques

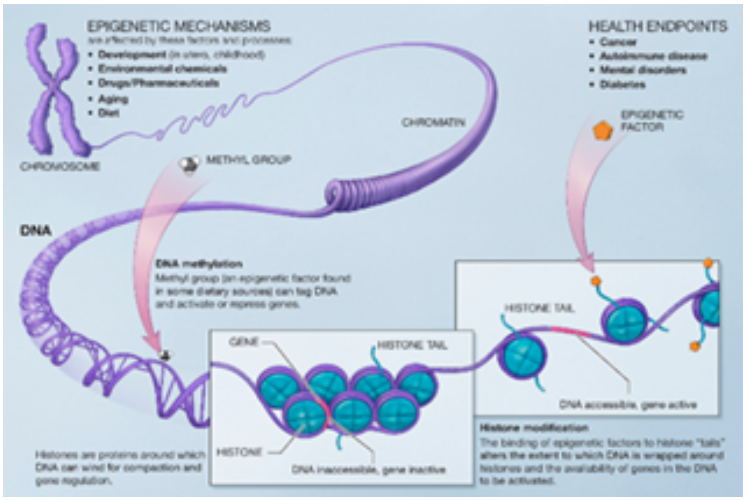
- **ChIP-sequencing** can be analysed using tools like Aclust (Algorithm used to detect sets of neighbouring CpG sites)
- **BLAST** (tool for searching homology)
- **CoSBI** (used to check chromatin modification patterns in human genome)
- **Epidaurus** (analysis of epigenetic data to provide a deeper understanding of epigenome)
- **PANTHER** (gene analysis tools)

Bioinformatics plays an important part in the integration of epigenetic data with other forms of biological data in addition to creating tools for the analysis of epigenetic data. To determine which genes are differently expressed in response to epigenetic alterations, researchers can utilize bioinformatics methods to integrate gene expression data with epigenetic data. Epigenetic data integrated with genomic data can help to identify genetic variants that are associated with epigenetic changes.

The following databases helps to predict the sites of epigenetic modifications :

- ENCODE (Encyclopaedia of DNA elements)
- ROADMAP Epigenetics (offers maps of histone modifications, chromatin assembly, DNA methylation and mRNA expression across 100s of human cell types and tissues
- IHEC data Portal (epigenomes related to health and diseases)
- EWAS Atlas (a knowledgebase of epigenome-wide association studies), 4D Genome (chromatin interactions)
- NonCode (Database of all kinds of noncoding RNA)





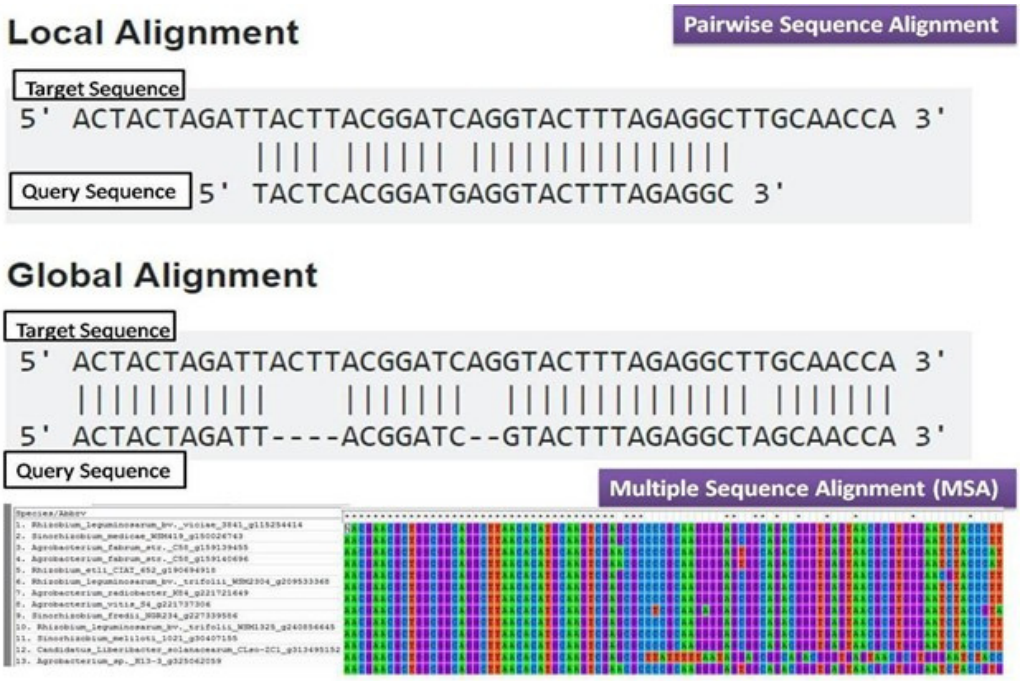
Let's look an example database containing epigenetic data. dbEM (database of Epigenetic Modifiers) has been created by Jagpreet Singh Nanda, Rahul Kumar and Gajendra P.S. Raghava to maintain the genomic information about 167 epigenetic modifiers/ proteins which are considered as potential cancer targets. This database helps in locating the altered epigenetic proteins that might contribute to oncogenesis and may be studied as potential therapeutic targets.

Bioinformatics has greatly helped in the development of drugs that target epigenetic modifications. For example, it has aided in the identification of small molecules that target specific epigenetic regulators, such as histone deacetylases (HDACs) and DNA methyltransferases (DNMTs). These have been shown to have great potential in the treatment of cancer and other diseases.

In conclusion, the analysis and interpretation of the complicated and enormous volumes of data produced by epigenetic research, identification of small compounds that target epigenetic alterations, and integration of epigenetic data with other types of biological data. Thus, Bioinformatics is an integral part of epigenetic research and will keep playing a crucial role in expanding our knowledge of epigenetics and its function in health and illness.

SRINIVASAN GAYATHRI
(Class of '25)

What is Sequence Alignment?



Sequence alignment is a method of aligning DNA or Protein sequences to derive useful insights from the set of sequences. The further, functional, structural, and evolutionary relationships can be found by identifying the similarity regions.

Global and Local Alignment:

Global alignments attempt to align every residue in every sequence, and are most helpful when the sequences in the query set are comparable and about equal in size. Using dynamic programming as its foundation, the Needleman-Wunsch algorithm is a comprehensive global alignment method. On the other hand, local alignment aims to align only a region of the sequences. Hence, they are more advantageous for divergent sequences that may contain similar sequence motifs or patches of commonality within their larger context. A general local alignment technique based on dynamic programming is the Smith-Waterman algorithm.

Dynamic programming:

Global alignments are possible with the Needleman-Wunsch algorithm, and local alignments are possible with the Smith-Waterman algorithm. Protein alignments frequently use a substitution matrix to provide points for matches and mismatches in amino acids as well as a gap penalty for matching an amino acid from one sequence to a gap in the other. There are two distinct gap penalties for opening and extending a gap. The former is typically substantially greater than the latter, for instance, -10 for gap open and -2 for gap extension. As a result, gaps in alignments are typically kept to a minimum, and residues and gaps are frequently kept together, which makes biological sense.

Multiple sequence Alignment:

Multiple sequences in a query set are aligned using multiple alignment algorithms. Multiple alignments are frequently used to find conserved sequence areas across a group of sequences that are thought to be related evolutionarily. The catalytic active sites of enzymes can be found using conserved sequence patterns in combination with structural and mechanistic information, it is used to build phylogenetic trees. it is used to build phylogenetic trees. Multiple sequence alignments are computationally challenging, and the majority of issue formulations result in NP-complete combinatorial optimization problems. Nonetheless, because of the importance of these alignments in bioinformatics, a variety of approaches for aligning three or more sequences have been developed.

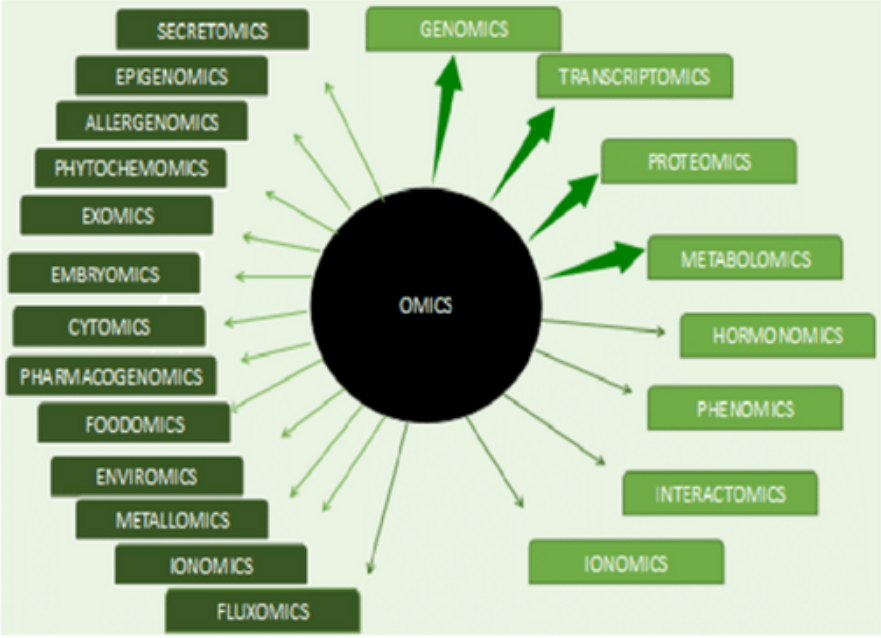
Thus, Sequence alignment is one of the major applications of Bioinformatics. We can easily find out if two organisms have a common ancestor or not just by aligning their sequences!

GAYAN SWEETY A
(Class of '25)

Multi-omics : Amalgam of heterogenous data

The term "**omics**" refers to a broad category of scientific fields, which includes **genomics, proteomics, transcriptomics,** etc. To study the complexity of biological processes systematically, it is vital to take an integrative approach that combines multi-omics data to estimate the inter-relationships of the biomolecules and their functions. Bioinformatics tools and methods adopt integrative approach to analyse multi-omics data and also helps in for disease sub-typing, biomarker prediction, and deriving insights into the data.

Recent research has shown that integrating omics information results in a greater understanding of the system being studied. For instance, Yu Zhang et al, 2013 addressed the role of integrating proteomics data with genomic and transcriptomic data which resulted in the identification of driver genes in colon and rectal tumours. Their findings demonstrated that the highest overall alterations in messenger RNA (mRNA) and protein levels were related to the chromosome 20q amplicon. Potential 20q candidates were found by integrating proteomics data. These candidates were HNF4A (hepatocyte nuclear factor 4, alpha), TOMM34 (translocase of outer mitochondrial membrane 34), and SRC (SRC proto-oncogene, nonreceptor tyrosine kinase).



Insights into the flow of biological information at various levels, can be obtained from the multi-omics data from the same set of samples and this can aid in the unravelling the mechanisms driving various biological questions. The use of the multi-omics approach has led to the creation of a number of tools, methods, and platforms that aids in the analysis, visualisation, and interpretation of multi-omics data. Tools such as GENEASE79, CGDV80, and SLIDE81, make it simple to visualize and comprehend massive biological data sets. Nevertheless, these tools make it easier in analysis and visualization of single omic data at a time.

There are a number of challenges involved in integrating multi-omics data to generate a comprehensive understanding of biological processes and diseases. This is a difficult undertaking due to the underlying heterogeneity in single omics data, huge datasets requiring computationally intensive analysis and lack of research that aid in prioritising the varied collection of tools.

Multi-omics data is a powerful strategy to understand the systematic details of the information flow in a cell. Nowadays, there are a number of tools and methods available in the public domain to integrate multi-omics data sets to produce insightful results. As the tools and methods are largely isolated, there is a need to have an uniform framework that can effectively process and analyse multi-omics data in an end-to-end manner easily.

Subhiksha.M
(Class of '24)