B.Tech 2020-24 CSE- Project Phase 1

Proposal

I. Group No: B2

Project Title: A STUDY OF GENOME VARIATIONS OF HUMAN PATHOGENS USING PANGENOME GRAPH

Team members :

AM.EN.U4CSE20131 GUTTAPATI RUSHIKESH REDDY

AM.EN.U4CSE20137 K.V.K SATHVIK

AM.EN.U4CSE20140 KOLLIPARA CHARAN SUNEEL

AM.EN.U4CSE20173 GUNTURU VENKATA SAI KOUSIK

II. Abstract:

The project aims to study genome variations of human pathogens, such as bacteria or viruses, using a pangenome graph approach. This involves combining multiple genomes of a species into a single graph structure to capture genetic diversity within a population. Understanding pathogen genome variations is crucial for comprehending their evolutionary patterns, disease transmission, virulence, and drug resistance. By analyzing these variations, researchers can gain insights into adaptive strategies, improve diagnostics, and develop targeted interventions for disease control and prevention. The motivation arises from the importance of uncovering critical insights into pathogen behavior and developing effective strategies to combat infectious diseases. However, persisting challenges include identifying structural variations in the DNA and pinpointing specific locations where mutations are more prevalent.

III. Background Study

Title &year	Problem	Contributions	Limitations	Open problems/Future work
1) Alam, Intikhab, et al. "Functional pangenome analysis provides insights into the origin, function and pathways to therapy of SARS-CoV-2 coronavirus." *bioRxiv* (2020): 2020-02.	This approach can help identify the core genome of *Betacoronaviruses* and extract accessory genomic features shared by a subset of these viruses or unique to SARS-CoV-2.	1) Functional pangenome analysis of SARS-CoV-2 can reveal gene cluster origins and functions related to pathogenicity. 2) Structural analysis predicts host cell locations uncharacterized accessory gene clusters. 3)Phylogenetic tree comparisons withBetacoronaviruses identify dissimilar regions. Utilizing 4)DeepGOPlus, uncharacterized gene clusters are explored further to gain insights.	Analysis may focus primarily on viral genes and pathways but might not fully account for the complex interactions between the virus and the host's immune system, which can significantly influence the disease outcome and potential therapeutic approaches.	Comparing this virus with similar ones can provide more insights. Analyzing how the virus affects our cells and testing existing drugs for treating COVID-19 are essential. Understanding how the virus changes over time can help us find better ways to control it and develop effective treatments

2) Chen, Hongxin, et al. "Combined pangenomics and transcriptomics reveals core and redundant virulence processes in a rapidly evolving fungal plant pathogen." *BMC biology* 21.1 (2023): 24.	To find the genes that are responsible for the rapid evolution in plant fungi like Zymoseptoria tritici.	1) The integration of pangenomics and transcriptomics is a potent approach to study Z. tritici's genetic diversity and gene expression patterns. 2) It sheds light on population structure, evolution, and interactions with wheat. Genetic sequence analysis and transcriptomics identify infection and virulence-related genes. 3) This knowledge informs crop protection and advances biological and medical research	Z. tritici's expansive accessory genome facilitates swift adaptation to various conditions, yet comprehending its functional significance and pathogenicity involvement remains challenging. Pangenomics and transcriptomics provide insights, demanding advanced bioinformatics for data analysis.	Studying Z. tritici's rapid evolution is vital for predicting its adaptation to different environments and hosts. Analyzing transposable elements, recombination, and selection pressures can yield valuable insights. Understanding interactions with wheat aids disease management and informs breeding for resistant varieties.
3) Haseeb, Muhammad, Afreenish Amir, and Hamza Irshad. "Pangenome analysis of SARS-CoV2 strains to Identify Potential vaccine targets by	The paper addresses urgent vaccine design needs against SARS-CoV2, using computational immunoinformatics to predict antigenic epitopes and create a multiepitope	1) The study's computational predictions need experimental validation. The vaccine's efficacy and safety require in vivo and in vitro testing. 2) Considering structural variations in mutations is	The paper stresses the importance of considering structural variations in mutations for vaccine design against emerging viruses.	Future work involves validating the multiepitope subunit vaccine's efficacy while accounting for these variations. Advanced immunoinformatics tools will be explored to improve design

Reverse Vaccinology." *bioRxiv* (2022): 2022-07.	subunit vaccine while considering emerging virus variants and structural variations due to mutations. 1) The paper proposes a multiepitope subunit vaccine for SARS-CoV2 using computational immunoinformatics. 2) It identifies B and T cell epitopes, considers structural variations in mutations, and designs a non-allergenic, antigenic, and non-toxic vaccine with significant HLA binding alleles, providing global population coverage of 84.38%.	important, but further research is needed to fully assess the impact on vaccine effectiveness.		and prediction accuracy amid evolving virus variants

4) Eizenga, Jordan M., et al. "Pangenome graphs." *Annual review of genomics and human genetics* 21 (2020): 139-162.	The paper addresses limitations of single linear reference genomes in genomics due to increasing genomic variation data. It introduces bioinformatic methods using graphical pangenomic reference systems to consider genetic diversity at every analysis stage.	1) The paper introduces pangenomic models representing the complete genomic elements in a species, providing an alternative to linear reference genomes. 2) It emphasizes graphical pangenomic models, explaining their construction from sequencing data or assembled genomes. The paper also surveys index data structures facilitating efficient interactions with pangenomes, enabling novel bioinformatic tasks like read alignment and visualization.	1) Limited evaluation of the practical applicability of graphical pangenomic methods in real-world genomic analyses. 2) Lack of in-depth comparison with existing linear reference genome-based methods to demonstrate the advantages of using pangenomic reference systems. 3) The need for further validation and benchmarking of the proposed bioinformatic methods on diverse datasets to assess their accuracy and performance.	1) Future research can explore standardized pangenomic reference models and their harmonious relationship with linear genomic models. 2) Advancements in graphical pangenomic methods will be vital in a future where genomes are easily sequenced and assembled. The paper encourages the development of bioinformatic methods to effectively construct, query, and operate on pangenomic reference systems
5) Mathur, Garima, Anjana Pandey, and Sachin Goyal. "A comprehensive tool for rapid and accurate prediction of disease using	The paper addresses the need for earlier disease identification, especially during pandemics like COVID-19, to prevent outbreaks and aid drug design.	The paper introduces a comprehensive tool for disease prediction using a DNA sequence classifier. It proposes a novel approach with a machine learning-based classifier	The study acknowledges limitations due to the scarcity of disease patterns, potentially affecting classifier performance. The reliance on	The paper's findings open avenues for disease prediction and DNA sequence classification research. Accuracy can improve by incorporating

DNA sequence classifier." *Journal of Ambient Intelligence and Humanized Computing* (2022): 1-17.	Shortage of disease patterns makes identification challenging. The study proposes DNA sequence classification to identify diseases using NCBI GenBank samples.	and a hot vector matrix for feature extraction. The method achieves an accuracy of 93.9%, outperforming other classifiers.	training data quality and availability is another concern. Further validation with real-world experiments and larger datasets is required.	diverse diseases and viruses in the dataset. Advanced feature extraction and machine learning methods can expedite and enhance disease identification. Integrating structural variations and more genomic data may boost predictive capabilities.

IV. Challenges

1. Practical Implementation: The practical integration of graphical pangenomic methods into existing genomic analyses and workflows may pose challenges due to the need for specialized expertise in graph-theoretic concepts.
2. Comparison and Validation: Thorough comparison and validation of graphical pangenomic methods against linear reference genome-based methods are essential to demonstrate their superiority in real-world applications.
3. Data Complexity: Dealing with large-scale pangenomic data and representing genetic diversity in a concise yet informative manner remains a challenge.
4. Computational Efficiency: Ensuring computational efficiency for indexing, querying, and analyzing pangenomes is crucial to handle the increasing volume of genomic data.
5. Standardization: Establishing standardized pangenomic reference models and formats to enable seamless integration with existing genomic databases and tools requires careful consideration and community efforts.
6. Interoperability: Achieving interoperability between different pangenomic models and linear reference genomes is essential for cross-referencing and data comparison.
7. Generalization: Extending graphical pangenomic methods to handle diverse species and datasets, beyond well-studied organisms, presents generalization challenges.
8. Biological Interpretability: Integrating pangenomic information with biological knowledge and interpreting complex variations to extract meaningful insights is an ongoing challenge.

V. Deliverables of Phase I

1. Pangenome Graphs: Complete pangenome graphs representing the full set of genomic elements for the selected species or clade, incorporating large-scale genomic variations and genetic diversity.
2. Dataset Collection: A well-curated and comprehensive dataset comprising whole genome assemblies from multiple organisms, enabling the construction of accurate pangenome graphs.
3. Pangenome Analysis Pipeline: A robust and efficient bioinformatic pipeline for constructing pangenome graphs, indexing pangenomes, and performing various analyses on the graphs.
4. Visualization Tools: User-friendly visualization tools to explore and interpret the pangenome graphs, enabling researchers to gain insights into the genetic variations and relationships among genomes.
5. Analysis Results: Detailed analysis results, including functional genomics, association studies, variant calling, and genotype information, obtained by leveraging the additional information provided by the pangenome graphs.
6. Documentation: Comprehensive documentation of the methods, algorithms, and tools developed during the project to facilitate reproducibility and future research.

VI. Assumptions/Declarations:

Assumptions :

1. Data Quality and Completeness: One assumption is that the dataset collected from NCBI or other sources is of high quality, reliable, and comprehensive. Any limitations or errors in the dataset may impact the accuracy and validity of the pangenome graph construction and subsequent analysis.
2. Computational Constraints: The project assumes that the available computational resources are sufficient for handling the large-scale genomic data and executing computationally intensive tasks. Any limitations in computational power may lead to challenges in processing and analyzing the data efficiently.
3. Applicability of Pangenome Graphs: An assumption is that pangenome graphs are a suitable representation of genetic diversity in the given dataset. Potential biases or inaccuracies in the pangenome graph model may affect downstream analyses and interpretations.

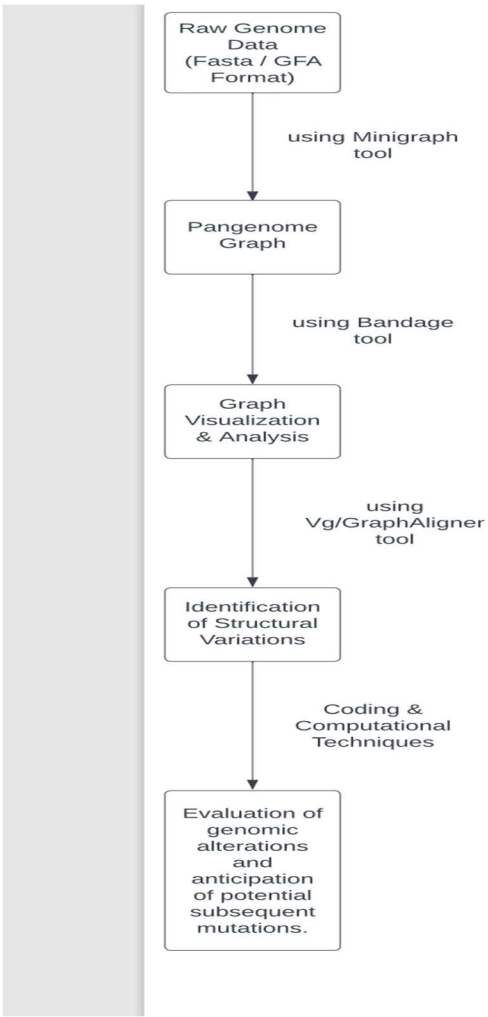
Declarations :

1. Utilization of NCBI Dataset: The project declares the use of the dataset collected from NCBI as the primary data source. However, the quality and integrity of the dataset will be continuously assessed, and any necessary data cleaning or preprocessing steps will be implemented.
2. Focus on Pangenome Graph Analysis: The project emphasizes a focus on constructing pangenome graphs and conducting in-depth analysis using these graphs. It aims to explore and evaluate the impact of pangenome graphs on various bioinformatic tasks.

VII. Tools to be used

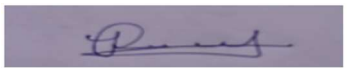
Software/Hardware Tools	**Specifications**
Minigraph	Pan-Genome Graph Construction
GFA format or fasta format	Pangenome Data Representation
Using Bandage	Pangenome Data Visualization
Vg or GraphAligner	Structural variations in pangenome graphs

VIII. High Level Design

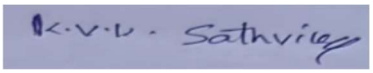


Students' Name and Signature:

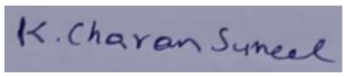
RUSHIKESH REDDY



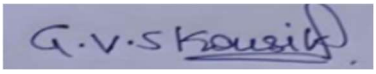
SATHVIK KESAVA



CHARAN SUNEEL



KOUSIK



Guide's Signature: