PROPOSAL : Leveraging Big Data Generated by AlphaFold for Protein Structure Analysis

(Team Member: Alexander Ratzan- asr655, Eshaan Raj Sharma- es6146)

PROBLEM STATEMENT:

Proteins are the building blocks of life, essential to virtually every biological process. Their functions are intrinsically linked to their three-dimensional structures. However, unveiling the structure of proteins through their amino acid sequences, known as the protein folding problem, has been a formidable challenge. AlphaFold, developed by Google DeepMind, represents a paradigm shift, predicting protein structures with the groundbreaking accuracy. With over 200 million protein structures estimated and a growing dataset of .pdb files, there is a pressing need for scalable, efficient tools to parse, analyze, and derive meaningful insights from this vast repository of data.

WHY IT IS BIG DATA:

The AlphaFold dataset exemplifies the four V's of Big Data:

- **Volume:** The sheer quantity of data, with over 200 million protein structures and a current accessible subset of 20,000 structures in .pdb format, totalling approximately 5 Gb, represents a significant volume challenge.

- **Velocity:** The rate at which new protein structures are predicted and added to the database requires tools that can quickly adapt and process incoming data efficiently.

- **Variety:** The data, stored in .pdb format, contains a wide range of information on protein structures, necessitating sophisticated parsing and analysis methods to extract relevant insights.

- **Veracity:** Given the critical role of these structures in biological research and potential therapeutic applications, the accuracy and reliability of data analysis are paramount.

AIM OF THE PROJECT:

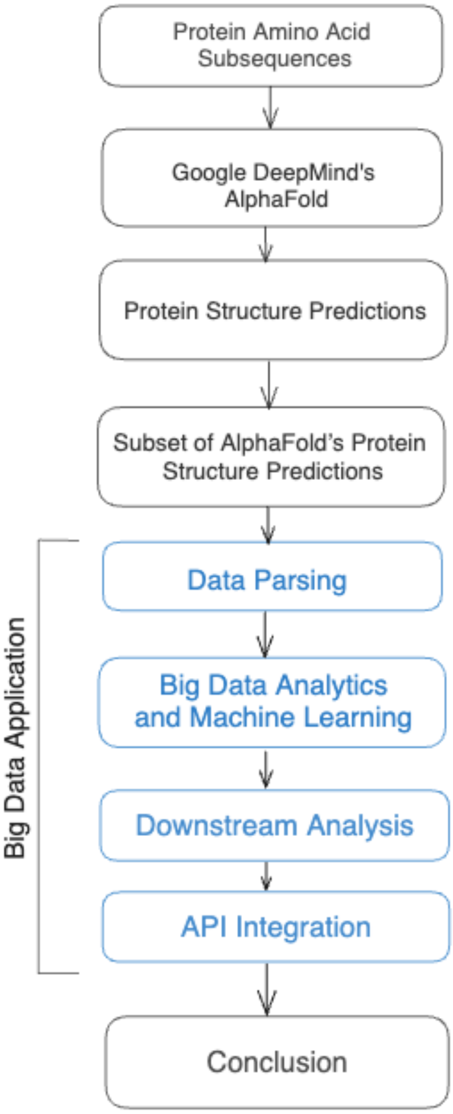
This project aims to work with a publicly available subset of AlphaFold's protein structure predictions. This database contains approximately 20,000 predicted protein structures of the over 200 million total protein structures predicted by the model. As more protein structures are released to the research community there will be substantial need for downstream applications that can scale with the volume of data. The current dataset contains approximately 5 Gb of .pdb (Program database) files.

In summary, this project proposes to leverage AlphaFold's predictions to conduct comprehensive analysis and exploration of protein structures in an efficient and scalable manner. By combining big data techniques with deep learning insights, this project may offer lightweight tools to support research in protein biology.

POTENTIAL TECHNOLOGIES:

PySpark and ML libraries, Vector Database, Web APIs.

WORK FLOW:


Protein Amino Acid Subsequences
Google DeepMind's AlphaFold
Protein Structure Predictions
Subset of AlphaFold's Protein Structure Predictions
Big Data Application
Data Parsing
Big Data Analytics and Machine Learning
Downstream Analysis
API Integration
Conclusion