CAP 5510 - Bioinformatics
Project Proposal

Project Title:

Comparative Analysis of Local Sequence Alignment Algorithms: BLAST, FASTA, and Smith-Waterman

Team members:

Name	UFID
Durga Sai Surya Ram Saladi	96891108
Sushanth Reddy Kotha	67664160
Mayur Sai Yaram	29880249

Abstract:

This project aims to conduct a comparative analysis of three widely-used local sequence alignment algorithms: **BLAST**, **Smith-Waterman**, and **FASTA**. Each algorithm has unique strengths in terms of speed, accuracy, and applicability, and this project will evaluate their performance based on common criteria such as alignment accuracy, execution time, computational resources, and biological relevance. By leveraging publicly available sequence datasets from **NCBI GenBank** and **UniProt**, the study will assess the effectiveness of each algorithm in aligning both Protein and DNA sequences. The goal is to provide a comprehensive understanding of how each tool performs under similar conditions and offer insights into their optimal usage in bioinformatics workflows.

Plan Of Action:

What will you implement?

- Install and configure the BLAST, FASTA, and Smith-Waterman tools on a local or cloud-based environment.
- Implement the alignment methods for each algorithm using the same input datasets to ensure consistency.

What methods are you going to compare and how will you get them?

We are comparing the following methods:

1. **BLAST**: A widely used heuristic algorithm for rapid local alignments, optimized for speed.
2. **FASTA:** A heuristic algorithm similar to BLAST, known for its speed and sensitivity in local sequence alignments.
3. **Smith-Waterman:** A dynamic programming-based algorithm that produces optimal local alignments with high precision.

The **BLAST** algorithm can be obtained by downloading the standalone BLAST package from the NCBI website. The **Smith-Waterman** algorithm is available through bioinformatics toolkits such as EMBOSS. The **FASTA** algorithm can be accessed by downloading the software package from the EBI website or other bioinformatics repositories.

Which datasets are you going to use and where will you get them from?

We will utilize the following datasets:

- **NCBI GenBank**: DNA sequences from GenBank for local sequence alignment. (Link: https://www.ncbi.nlm.nih.gov/genbank/)
- **UniProt**: Protein sequence data for evaluating performance on protein alignments. (Link: https://www.uniprot.org/)
- **Ensembl**: The Ensembl dataset will provide high-quality genomic data, including DNA sequences, gene annotations, and protein-coding genes for various species, facilitating comprehensive comparisons. (Link:https://www.ensembl.org/)

Here are the specific datasets we will be using:

1. **Blast:** https://ftp.ncbi.nlm.nih.gov/blast/db/
2. **FASTA**: https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/

What kind of experiment will you run and what will you measure?

- **Alignment Experiments**: We will perform alignment experiments using the selected datasets (NCBI GenBank, UniProt, and Ensembl) with the BLAST, Smith-Waterman, and FASTA algorithms to evaluate their performance in local sequence alignment.
- **Alignment Scores**: We will measure the alignment scores generated by each algorithm, with higher scores indicating better quality alignments based on match/mismatch scores and gap penalties.
- **Execution Times (Time Complexity)** : The time taken by each algorithm to complete the alignments will be recorded to assess computational efficiency, particularly with large datasets.
- **Number of Significant Alignments**: We will count the number of significant alignments that exceed a defined score threshold, indicating meaningful biological relationships.
- **Memory Usage**: Track memory consumption to understand scalability.

Planned workload distribution to team members.

Ram:
Analyze the corresponding BLAST research paper.
Collect and manage datasets relevant to BLAST experiments.
Focus on implementing the BLAST algorithm.
Collaborate on the overall comparison and analysis of all algorithms.

Sushanth:
Analyze the research paper related to FASTA.
Collect and manage datasets relevant to FASTA experiments.
Work on implementing the FASTA algorithm.
Assist with the final comparison and the overall analysis.

Mayur:
Analyze the research paper related to Smith-Waterman.
Collect and manage datasets relevant to Smith-Waterman experiments.
Focus on coding the Smith-Waterman Algorithm.
Collaborate on the overall comparison and analysis of all algorithms.

Collaboration:
Jointly work on the comparison and analysis of all three algorithms (BLAST, FASTA, and Smith-Waterman).
Share responsibilities for writing and preparing the final report.

Research Papers and Resources:

BLAST: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC441573/

BLAST TOOL: https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/

BLAST DB: https://ftp.ncbi.nlm.nih.gov/blast/db/

FASTA: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5072362/

FASTA db: https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/

Smith-Waterman: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9246398/