Overview

This project forms a component of the Laboratory of Bioinformatics I at the University of Bologna. The primary objective is to develop a Profile Hidden Markov Model (HMM) for the Kunitz-type protease inhibitor domain. These domains are functional regions of proteins that inhibit proteolytic enzymes, including notable examples such as aprotinin (BPTI), Alzheimer's amyloid precursor protein (APP), and tissue factor pathway inhibitor (TFPI). The focus of the project is to employ the HMM to identify the optimal e-value threshold for binary classification derived from HMM search outcomes.

Project Description

Main Objective

The principal aim is to construct a Profile HMM for the Kunitz-type protease inhibitor domain, utilizing available structural data, and to use it to ascertain the optimal e-value threshold for binary classification in HMM searches.

Specific Objectives

- Develop a model for the Kunitz domain using structural data.
- Identify the optimal e-value threshold for binary classification based on HMM search outcomes.

Procedure:

Step 1: Data Collection

1. **Access the PDB Database**:
    - Query using CATH ID or PFAM ID.
    - Set data collection resolution to ≤ 3 Å.
    - Specify size between 50-80 residues.
    - Exclude mutant proteins.

2. **Handle Redundancy**:
    - Cluster proteins based on similarity, selecting one representative from each cluster.
    - Use PDB search: return -> polymer entities, group by -> S.I. 50% or above.
    - Alternatively, perform pairwise alignments and cluster according to similarity level.

    - Obtain the ID, sequence, and chain (Auth Asym ID), and download in JSON or CSV format.
    - Clean up results by removing quotes (`tr -d ""`).
    - Extract values and disregard lines without entry IDs (`awk -F "," 'if ($1!="") { print $1 $2 $3}'`).

3. **Perform Pairwise Alignments**:
    - Use `blastclust` for all-against-all pairwise alignments.
    - Set score and coverage thresholds (`blastclust -i <fastafile> -o <outfile> -S 80 -L 0.8`).
    - Each line in the output file represents a cluster.
    - For a limited number of sequences, consider global alignment.

4. **Alternative Method**:
    - Select one representative and use PDBeFold for pairwise alignment against the PDB to find similar structures, noting possible functional differences.

Step 2: Multiple Structure Alignment

- Upload the list of PDB codes to PDBeFold.
- Address alignment issues where positions contain mostly gaps, which can hinder HMMER's ability to calculate probabilities.
- Format alignments with one whole sequence per line (`awk '{if (substr($0,1,1)==">") {print "\n"$1} else { printf "%s",$1 } }'`).

Step 3: Generate HMM

- Use `hmmbuild` to create the HMM.
- Ensure it skips the first and last positions with too many gaps.
- Verify the length of the model.