ROHAN GHOSH DASTIDAR

22CH30028

ASSIGNMENT 1: Serial Ab Initio Protein Folding Using PyRosetta

For this project, I have used 35 AA as per the assignment

INTRODUCTION

This project focuses on the development and implementation of a serial ab initio protein folding pipeline using PyRosetta, inspired by the widely used Rosetta Abinitio Relax protocol. The primary objective is to predict the structure of the 35-residue villin headpiece—a benchmark system in protein folding studies—starting solely from its amino acid sequence.

The pipeline integrates several key steps:

- Generating an idealized starting pose
- Linearizing the structure
- Converting to centroid mode for efficient sampling
- Employing fragment-based Monte Carlo sampling to explore conformational space.

The lowest-energy structure identified during the simulation is then recovered, converted back to full-atom detail, and output for further analysis.

In addition to structure prediction, the project includes analysis and visualization of the NATIVE and PREDICTED structures, such as aligning the predicted structure with the experimentally determined native structure and calculating the root-mean-square deviation (RMSD) using BioPython and py3Dmol.

PIPELINE STAGES

1. Initialization: Set up the PyRosetta environment.
2. Pose Creation: Generate an initial, idealized full-atom pose from the HP35 sequence.
3. Linearization: Set backbone torsion angles (phi, psi, omega) to extended values (-150°, 150°, 180°) to start from an unbiased conformation.
4. Centroid Conversion: Switch the pose representation to centroid mode to simplify the energy landscape and speed up sampling by representing sidechains as single pseudo-atoms.
5. Setup Movers: Load 9-mer and 3-mer fragment libraries and configure ClassicFragmentMovers within a MoveMap allowing full backbone flexibility.
6. Monte Carlo Sampling: Perform extensive conformational sampling using a MonteCarlo object and a TrialMover. The trial move consists of applying repeated fragment insertions (using RepeatMover wrapping the ClassicFragmentMovers,

combined via SequenceMover) followed by Metropolis acceptance/rejection based on the score3 centroid energy function.

7. Decoy Recovery: Identify and recover the lowest-energy centroid conformation encountered during the simulation.

8. Full-Atom Conversion: Convert the recovered low-energy centroid pose back to a full-atom representation (fa_standard).

9. Refinement: Apply the FastRelax protocol to the full-atom pose using the DEFAULT ref2015 score function.

10. Output: Save the final refined full-atom structure to a PDB file (output.pdb) and optionally save the energy trajectory (energy_convergence_data.txt).

11. Analysis: Calculate the RMSD between the predicted model and the provided native structure (native.pdd) using BioPython s Superimposer.

PARAMETERS

PROTEIN_SEQUENCE:
MLSDEDFKAFGMTRSAFANLPLWKQQNLKKEKLLF (35 residues, as specified).

- Fragment Files: aat000_09.frag and aat000_03.frag (Standard filenames for Robetta-generated fragments).
- num_centroid_cycles: 15000. A large number of cycles was chosen for extensive exploration of the conformational space and to achieve the lowest possible energy.
- kT: 1.0. A lower temperature parameter (compared to the example 3.0) biases the Monte Carlo search towards lower-energy states, potentially helping to locate deeper energy minima more effectively, though it may hinder escape from local minima.
- N_long_frag_repeats: 1. One attempt to insert a 9-mer fragment per trial move, focusing on larger conformational changes.
- N_short_frag_repeats: 3. Three attempts to insert 3-mer fragments per trial move, allowing for more local refinement within each Monte Carlo step.
- refinement_cycles: 5. The standard number of cycles for FastRelax, generally sufficient for optimizing the full-atom structure after centroid sampling.
- Score Functions: score3 for centroid sampling and ref2015 for full-atom refinement