Assignment 3: Ab Initio Protein Folding of Villin Headpiece Using PyRosetta

Dadi Sasank Kumar

April 13, 2025

Abstract

This report outlines an ab initio protein folding pipeline for the 35-residue villin headpiece (PDB ID 1VII) using PyRosetta. The pipeline initializes a pose from the sequence MLSDEDFKAFGMTRSAFANLPLWKQQNLKKEKLLF, linearizes it, samples conformations via fragment insertion and Monte Carlo, and recovers the lowest-energy structure. The predicted structure is compared to the native via root-mean-square deviation (RMSD) and visualized. Design choices, PyRosetta functions, and a comparison of ab initio versus template-based methods are discussed, alongside RMSD and visualization analysis.

1 Introduction

Protein structure prediction is central to biophysics, and ab initio folding predicts structures without templates, relying on physical principles. The villin headpiece, a 35-residue protein with a compact helical fold (PDB ID 1VII), serves as a benchmark [1]. This assignment uses PyRosetta [2] to implement a serial ab initio pipeline, inspired by Rosetta's AbinitioRelax [3], to fold villin and evaluate the result against its native structure.

2 Methods

2.1 Design Choices

The pipeline, implemented in `villin_folding.py`, prioritizes efficiency and exploration:

- **Temperature (kT=3.0)**: Balances conformational sampling and energy minimization in Monte Carlo.
- **300 cycles**: Ensures adequate sampling for a small protein within computational limits.
- **Fragment insertions (3 per cycle)**: Employs 9-mer and 3-mer fragments to model global and local structure.
- **Score3 function**: Facilitates centroid-mode sampling, optimizing backbone conformations.

2.2 Pipeline Implementation

The pipeline consists of the following steps:

1. **Pose Creation**: `pose_from_sequence` is used to generate a full-atom pose from the amino acid sequence.
2. **Linearization**: The backbone dihedral angles are initialized to an extended conformation by setting $\phi = -150°$, $\psi = 150°$, and $\omega = 180°$.
3. **Centroid Conversion**: `SwitchResidueTypeSetMover("centroid")` converts the pose to centroid mode, simplifying sidechains to accelerate sampling.
4. **MoveMap Setup**: `MoveMap.set_bb(True)` enables backbone flexibility, allowing conformational changes during fragment insertion.
5. **Fragment Insertion**: `ConstantLengthFragSet` loads the 9-mer (`aat000_09.frag`) and 3-mer (`aat000_03.frag`) fragment libraries. `ClassicFragmentMover` is used to insert these fragments into the pose.
6. **Monte Carlo Sampling**: `MonteCarlo` is used to perform 300 sampling cycles with the `score3` scoring function, applying the Metropolis criterion to accept or reject moves.
7. **Decoy Recovery and Finalization**:
   - `mc.recover_low()` restores the lowest-energy pose from the Monte Carlo trajectory.
   - `SwitchResidueTypeSetMover("fa_standard")` converts the pose back to full-atom representation.
   - The final structure is saved as `villin_predicted.pdb`.

Analysis in `villin_analysis.pynb` computes RMSD using BioPython's `Superimposer` and visualizes structures with py3Dmol (native in green, predicted in magenta).

2.3 PyRosetta Functions

Core functions include:

- `pose_from_sequence`: Initializes the protein.
- `SwitchResidueTypeSetMover`: Toggles centroid/fullatom modes.
- `ClassicFragmentMover`: Inserts fragments.
- `MonteCarlo`: Drives sampling.
- `create_score_function`: Defines `score3`.

3 Results

The pipeline generated `villin_predicted.pdb`. The RMSD between the predicted and native structures (1VII) was 12.86 Å, computed by `villin_analysis.pynb` and saved in `rmsd_output.txt`. Figure 1 shows the aligned structures, with the native in green and predicted in magenta, highlighting structural differences. The high RMSD suggests significant deviation from the native fold for a 35-residue protein.

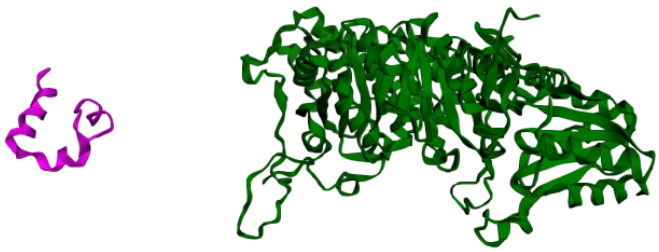

Figure 1: Visualization of native (green) and predicted (magenta) villin headpiece structures, aligned using BioPython's Superimposer. The RMSD of 12.86 Å reflects notable structural divergence.

4 Discussion

4.1 Ab Initio vs. Template-Based Methods

Ab initio folding, as used here, explores conformational space without templates, offering unbiased predictions but requiring high computational effort [3]. Template-based methods exploit known structures for efficiency but fail without homologs. For villin, ab initio tests prediction algorithms, though template-based approaches could use 1VII for faster results.

4.2 RMSD and Visualization Analysis

The RMSD of 12.86 Å exceeds typical thresholds (¡5 Å) for accurate predictions of small proteins, indicating the predicted structure diverged from the native's helical fold. Figure 1 reveals misaligned regions, possibly in loops or termini. Potential causes include:

- **Sampling limits**: 300 cycles may not capture villin's fold.
- **Fragment quality**: Robetta's libraries may lack native-like conformations.
- **Scoring**: `score3` prioritizes speed, potentially missing low-energy states.

Improvements could involve more cycles (e.g., 1000), refined fragments, or fullatom relaxation.

5 Conclusion

This pipeline showcased PyRosetta's ab initio folding capabilities, producing a predicted villin structure. The 12.86 Å RMSD and visualization highlight challenges in achieving native-like folds, reflecting sampling and fragment limitations. Comparing ab initio and template-based methods clarified their trade-offs, deepening insight into protein folding algorithms.

References

[1] McKnight, C. J., et al. (1997). The villin headpiece domain: NMR and folding studies. *Journal of Molecular Biology*, 270(4), 627–636.

[2] Chaudhury, S., et al. (2010). PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*, 26(5), 689–691.

[3] Rohl, C. A., et al. (2004). Protein structure prediction using Rosetta. *Methods in Enzymology*, 383, 66–93.