Lab Evaluation 1: HIV-2 Mutations and Drug Resistance Analysis

---

1. Novelty

- Most of the existing research works focus mainly on **HIV-1**, whereas this project specifically targets the **HIV-2** variant, which is naturally resistant to many non-nucleoside reverse transcriptase inhibitors and therefore requires a separate protease inhibitor analysis.
- Instead of using simple string matching techniques, this work uses **Global Pairwise Alignment (Needleman–Wunsch)** through BioPython, which helps in accurately identifying mutations even when insertions, deletions, or sequence shifts are present.

---

2. Dataset Retrieval: `retreving_seqences.py` Data Source: NCBI Protein Database.

- **Search Logic (`term`):**
  `("Human immunodeficiency virus 2"[Organism] AND "protease"[All Fields])`
  This Boolean query ensures that only HIV-2 protease protein sequences are retrieved and irrelevant proteins are excluded.
- **Retrieval Scope (`retmax = 1745`):**
  A total of **1,745 raw protein sequences** were successfully retrieved, which exceeds the required minimum of 1,000 sequences.
- **Database Selection (`db = "protein"`):**
  The protein database was chosen to directly obtain amino acid sequences, which are required for mutation and structural analysis.

---

3. Redundancy Removal & Dataset Refining: `refining_of_data.py`

Final Dataset Size:

The dataset was refined from 1,745 raw sequences to **536**

high-confidence sequences, which satisfies and exceeds the 500-sequence requirement.

- **Three-Level Refining Strategy:**
    1. **Level 1 – Length Constraint (90–110 AA):**
       Removes incomplete fragments and large polyprotein sequences to isolate the HIV-2 protease monomer.
    2. **Level 2 – Functional Motif Check (DTG):**
       Verifies the presence of the **Aspartic Acid–Threonine–Glycine (DTG)** catalytic triad. Sequences missing this motif are biologically non-functional.
    3. **Level 3 – Homology Verification (1IVP Reference):**
       Each sequence is compared with the Wild-Type HIV-2 protease reference (1IVP) using **Global Pairwise Alignment**.
- **Global Alignment Justification:**
  Global alignment is used since both the target sequences and the reference are of similar length (~99 amino acids), ensuring that mutations across the entire protease sequence are detected, including terminal regions that may be missed by local alignment.

---

4. Target Specificity & Validation

The specificity of the final 536 sequences to the **HIV-2 Protease** target is validated using the following criteria:

- **Identity Threshold:**
  A strict **80% identity cutoff** with respect to the **1IVP PDB reference** is applied, confirming that all retained sequences belong to HIV-2 protease.
- **Active Site Validation:**
  The presence of the **DTG motif** ensures that all sequences retain the correct active-site architecture required for protease inhibitor binding.
- **Sequence Specificity:**
  An average identity score of approximately **85.5%** indicates strong similarity to the wild-type while still allowing biologically meaningful mutations.

Final Data Summary

- **Initial Sequence Count:** 1,745
- **Final Refined Count:** 536
- **Refining Efficiency:** 16.2%
  (Approximately 83.8% of low-quality or redundant sequences were removed)
- **Final Output File:** `HIV2_PROPER_DATASET.xlsx`
  (Contains Sequence ID, Cleaned Sequence, and Identity Score)

GIT HUB LINKS FOR THE CODES:

nunnajaswant-ship-it/BIO_PROJECT

OUTPUT LINKS:

Link to 2 items