Task Description

Proteins are large, complex molecules composed of one or more chains of amino acids, known as residues. There are 20 standard types of residues (amino acids) that make up proteins, each with unique chemical properties influencing the protein's structure and function. Proteins can vary in size, containing anywhere from a few dozen to several thousand residues. A typical protein sequence can be represented as a string of single-letter codes for each residue, such as:

Example protein sequence: "`MKTFFVFTTQRYDEQELFFQN`" (`M`, `K`, `T`, … are the protein's residues)

Residue-residue contacts refer to pairs of residues that are spatially close within the 3D structure of the protein. In this task, two residues are considered in contact if the distance between their Cα atoms is below 8 Å (angstroms). Identifying these contacts is crucial for understanding protein folding and interactions, as they provide insights into the overall structure and function of the molecule.

You are tasked with developing a machine learning model for predicting residue-residue contacts within protein sequences. This model will be implemented in Python and trained using an open-source protein database that provides sequence and 3D structural data. The goal is to predict which residues in a protein sequence come into contact, without explicitly predicting the full 3D configuration (i.e. atom coordinates). In your solution you may rely on the fact that similar protein sequences usually have similar 3D structures.

This model should build upon the ESM2 architecture, which already provides a contact prediction head but relies solely on input sequences. The main goal of this task is to re-use ESM2 embeddings and additionally incorporate structural data from proteins with sequences similar to the input sequence, enhancing the prediction performance.

Task Details

Objective: Create a new residue contact prediction model that leverages both the ESM2 embeddings of the input sequence residues and structural data from similar sequences.

Implementation:

Implement your model in Python, modifying or extending the ESM2 model to incorporate additional structural information. The input of the model is a single sequence, the output is a binary contact map (1 - if there is a contact between the residue pair, 0 - if there is no contact).

For the training and evaluation of your model we provide the protein structures data in PDB format. You may download the data from here. When mapping protein residue 3-letter codes to their 1-letter code, you may use the table provided at the end of this document.

Deliverables:

- Submit your code as a ZIP file or upload it to a public GitHub repository.
- Include a concise report that outlines:
    - A precise description of your method and how it extends ESM2.
    - Data pre- and post-processing steps.
    - Hyperparameters and other relevant details of the ML setup.
    - Analysis of the results, including performance metrics and insights.
    - Breakdown of the time (in hours) spent on each aspect of the task.

Evaluation Criteria

- Code quality
- ML training and evaluation setup, data handling
- Model performance
- Results analysis
- Method description

Important notes:

- You are encouraged to research and leverage open-source resources (e.g. for protein sequence alignment if your solution requires it), appropriately citing any external code or libraries used. Ensure originality in your approach and final implementation.
- You should not publish or share the contents of this task.

Protein residue codes mapping

- ALA: A
- ARG: R
- ASP: D
- CYS: C
- CYX: C
- GLN: Q
- GLU: E
- GLY: G
- HIS: H
- HIE: H
- ILE: I
- LEU: L
- LYS: K
- MET: M
- ASN: N
- PHE: F
- PRO: P
- SEC: U
- SER: S
- THR: T
- TRP: W
- TYR: Y
- VAL: V