PROTEIN STRUCTURE: UNRAVELED

Ethan Ye, Kelley Tu, Russell Chiu, Sophia Li

INTRODUCTION

Proteins are fundamental biological molecules that perform an array of essential functions like catalysis, structural support, and signal transduction. A protein's function is intimately tied to its shape—this shape is determined by its secondary structure, which folds into a more complex tertiary structure. Protein misfolding is associated with various diseases, such as Alzheimer's and Parkinson's, making accurate structural predictions a valuable tool in drug design and biomarker identification.

We re-implemented a paper that aims to predict the secondary structure of proteins with deep learning techniques, focusing on leveraging raw multiple sequence alignment (rawMSA) data and applying Natural Language Processing methodologies like Long Short-Term Memory (LSTM) to predict secondary structures.

Example amino acid sequence (black) and corresponding secondary structure assignments (color)

DATASET

GATHERING DATA:

- Multi-sequence alignment (MSA) data gathered from **various protein families** with the help of the InterPro database
  - **InterPro Database**: Database operated under EMBL's European Bioinformatics Institute
  - Most of the family domains were selected for being highly conserved across species, and the remainder of domains were chosen randomly
- Data for all of the family domains were split such that:
  - **Training Dataset**: 23 family domains
  - **Validation Dataset:** 5 family domains
  - **Testing Dataset**: 4 family domains

DATA LABELS:

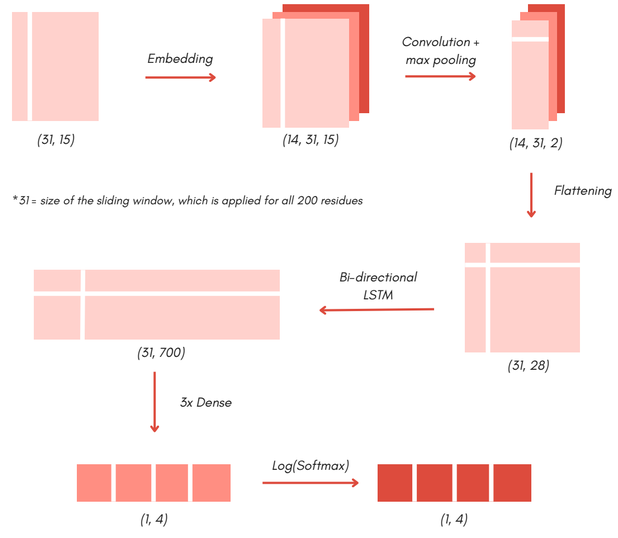
- **25** unique residues:
  - Each residue is a letter that corresponds to an amino acid
- **4** unique secondary structure labels:
  - Coil, Helix, Extended, Other (ex: Bend)

METHODOLOGY

PREPROCESSING:

- Collected data for each family domain converted from a bioinformatics format called "Stockholm"
- Each sequence and its corresponding secondary structure were converted into integers under a consistent mapping:
  - Each amino acid was mapped to a unique integer from **1 to 25** (to represent the 25 different amino acids) while each secondary structure was mapped to a unique integer from **1 to 4** (to represent the 4 possible classifications).

MODEL ARCHITECTURE:



After data collection and preprocessing, the training dataset was of shape (311, 200, 15):

- 311 = number of batches
- 200 = number of amino acids (residues) per sequence
- 15 = number of sequences used to predict the secondary structure of the master sequence

Hence each batch is of the size (1, 200, 15)

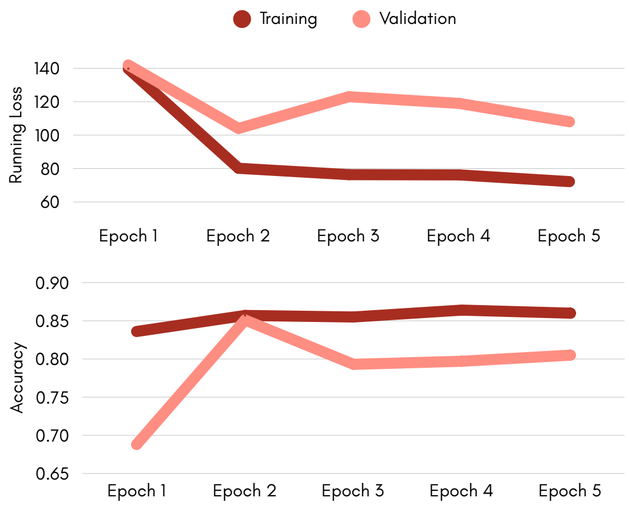
Model Hyperparameters	
Optimizer	Adam
Learning Rate	0.005
Dropout Rate	0.25
Epochs	5

SOURCES

Original paper: "End-to-End Deep Learning Using Raw Multiple Sequence Alignments" by Claudio Mirabello and Björn Wallner. Edited by Yang Zhang. Published: 15 Aug. 2019. journals.plos.org/plosone/article/authors?id=10.1371/journal.pone.0220182.

RESULTS

Below is a table displaying the running loss and accuracy across epochs during training. Under the testing process, the model managed to approach an accuracy of **0.895** with a running loss of **59.42**. Both the training and validation accuracy plateau, indicating that the model is underfitting.



DISCUSSION

CHALLENGES:

- Gathering enough data to train our model and parsing them from "Stockholm" format
  - **Stockholm Format**: Bioinformatics formatting system for protein MSA data
  - Lacks structure compared to other formats (FASTA)
- Understanding how to implement our model with **PyTorch**
  - *PyTorch* differs from *TensorFlow:* all layers must be given an input and output size
  - Caused grief in the form of shape errors
- Avoiding **biases** introduced by parsing master and body sequences separately, which resulted in incomplete utilization of the data during training and testing

FUTURE WORK:

- Implementing the **CMAP** model and evaluating its accuracy in comparison to our SS-RSA predictions.
- Extrapolating our predictions to **real-world applications**, such as:
  - Analyzing the impact of genetic mutations on protein structure to provide insight into genetic diseases