DEEP LEARNING FRAMEWORK FOR AB INITIO PROTEIN SECONDARY STRUCTURE PREDICTION

Christopher Sun & Bowen Zheng
Department of Computer Science
Stanford University
{chrisun, bowen07}stanford@edu

1 Introduction

Protein structure prediction is one of the key goals and challenges in the field of computational biology. Since the discovery of the 3D folding of proteins and how structures are tightly related to their functions, there have been efforts to use computational power to predict the 3D structure of proteins, from initial models like homology modeling to more advanced approaches tackling ab initio structure prediction.

Secondary structures of proteins refer to localized folding patterns driven by hydrogen bonding – structures such as alpha-helices and beta-sheets. These structures are critical for protein stability and function, and their prediction can provide insights into the biological functions of certain uncharacterized proteins.

Recent advancements in machine learning and neural networks have revolutionized protein structure prediction. For example, models like AlphaFold and Evolutionary Scale Modeling (ESM) use deep learning techniques to predict 3D protein structures and have achieved great accuracy [1][2].

In this project, we use the pre-trained protein language model, ESM, to generate embeddings that represent the evolutionary and structural properties of protein sequences. We design a computationally lightweight deep learning architecture that directly predicts protein secondary structure, notably without multiple sequence alignment or homology modeling. The model is trained and evaluated on JPred, a dataset of protein sequences with known secondary structure annotations [3]. The model's performance is assessed using metrics like categorical classification accuracy and macro F1 score.

In this report, we outline the methodology for developing and evaluating our protein secondary structure prediction model. We describe the data preprocessing steps, model architecture, training procedure, and evaluation results. Finally, we suggest potential avenues for future work to improve accuracy and generalizability.

2 Background

Secondary structures of proteins are of vital importance in that they are the first level of folding. Correctly predicting secondary structures is important for the downstream prediction of 3D structure, for example. ESM, using evolutionary information embedded with protein sequences, has shown great promise in generating embeddings that capture both evolutionary and structural features of proteins.

Evolutionary Scale Models for Embedding Generation The core idea behind ESM is large-scale, multiple sequence alignment data that captures evolutionary patterns. The architecture of ESM models is based on transformer networks. Similar to how transformer-based models process language, ESM uses transformers to generating sequence embeddings that encode information about the protein's structure, function, and evolutionary history. These embeddings are highly informative and can be used as features in downstream tasks like secondary structure prediction, fold recognition, and functional annotation.

The main contribution of ESM models is their ability to generate high-quality embeddings from primary protein sequences. These embeddings are vector representations that capture a protein's evolutionary relationships and encodes the essential features. For instance, it takes into account how conserved regions of the sequence are likely to fold or function. This is especially useful for understanding proteins with few or no homologous structures available in the Protein Data Bank, as the model can still provide useful predictions based on evolutionary constraints.

Each protein sequence is processed by the model, and the output is a tensor of dimension sequence length by embedding dimension, where ESM's embedding dimension is 1,280.

3 Methods

3.1 Data Preprocessing

For the training and evaluation of our protein secondary structure prediction model, we used protein sequences and their secondary structure annotations from the JPred/JNet v.2.3.1 dataset [3]. This dataset consists of 1,348 samples and 149 blind-test samples.

The sequences in this dataset were derived from the UniRef90 database released in July of 2014. The original JPred/JNet v.2.3.1 model was trained using amino-acid-level secondary structure annotations derived from the Dictionary of Secondary Structure of Proteins.

Each protein sequence was processed in FASTA format, and passed into the pretrained ESM2 model with 33 layers and 650 million parameters for embedding generation. Secondary structure annotations were transformed from string form to one-hot encodings indicating one of three possible secondary structures at each residue: 1) alpha helix, 2) beta sheet, or 3) another structure.

3.2 Embedding Protein Sequences Using ESM

After data preparation, we convert raw protein sequences into tokenized representations suitable for input into the ESM model. We embed the protein sequences using the 650 million parameter model, pre-trained on large-scale protein sequence data to generate sequence embeddings.

The embeddings are extracted from the 33rd layer of the model and stored as the sequence's representation. We extract all tokens except the first and last, because they represent start and end tokens.

Critically, not all protein sequences are of the same length. Hence, we apply padding and masking techniques to assist model training. Each batch of training data is padded with zeros to the maximum sequence length of a protein within the batch, and corresponding binary attention masks are used to prevent the model from learning on such padded zeros, and to correctly calculate the loss and accuracy during training.

3.3 Model Definition

We design a deep learning model whose architecture is depicted in Figure 1. We make use of a convolutional motif consisting of a 1D Convolutional layer, the Gaussian Error Linear Unit (GELU) activation function, and 1D Batch Normalization.

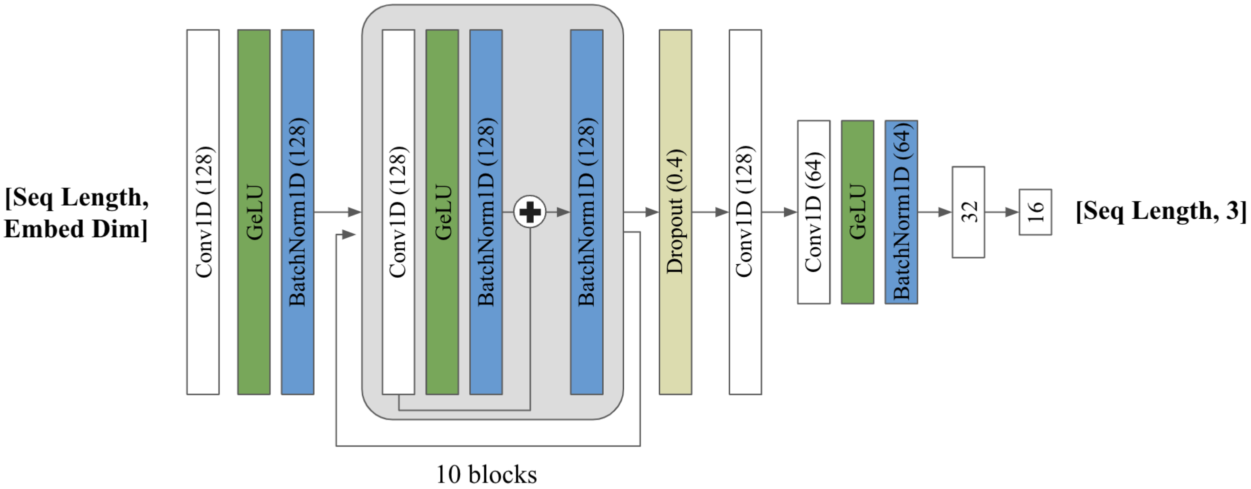

Figure 1: Convolution-Based Secondary Structure Prediction Network

The embeddings generated by ESM are passed as inputs to this architecture, where they are first processed through one convolutional motif that condenses the embedding dimension of ESM – 1,280 – down to 128. This representation is then passed through ten sequential convolutional motifs, where the representation that is used as input to the motif is added to the representation that results from the motif before the next round of processing. We implement these residual blocks because they are useful in learning deeper representations, allowing the model to better capture complex patterns in the data.

After this, we apply dropout regularization with a dropout probability of 0.4, followed by several 1D convolutional layer that project from 128 channels to 64, 32, 16, and eventually 3, which is the output dimension of the data. This output, which is a one-hot encoded tensor of dimension sequence length by output dimension, is the model's prediction of protein secondary structure for each position in the original 1D protein sequence.

Overall, the ESM embeddings are expected to capture complex relationships between sequence and structure, while the convolutional architecture aims to effectively learn and generalize from these embeddings. In total, the deep learning model contains 366,947 trainable parameters.

We use a 70%-30% train-test split, a batch size of 32 sequences, a learning rate of 5×10^{-4}, the Adam optimizer and the Cross Entropy Loss.

4 Results and Analysis

We see peak performance on the test data set after training for roughly 10 epochs. Throughout training, we monitor the train and test loss, accuracy, and macro F1 scores for each epoch. With light hyperparameter tuning, we produce a model that achieves 86.18% categorical accuracy and 0.8547 macro F1 score on the test set. There is a small amount of overfitting that can further be optimized to improve test set loss and metrics, but at this stage the model is already indicating effective learning.

We aim to find more granularity in the model's performance, so we examine accuracy for each ground-truth category of secondary structure. The model achieves 89.67% accuracy for alpha helices, 74.40% for beta sheets, and 85.38% for other structures on the test set of 405 sequences. These results suggest that perhaps it is easier for the model to recognize helices, but harder to recognize sheets. Biologically, this could make sense because an alpha helix is considered a simpler structure than a beta sheet.

To further improve the model, we would address the discrepancy between the training and validation loss, as the model began to overfit after the fourth or fifth epoch. Future work could involve regularization techniques or an increased amount of training data to address this.

Also, the beta-sheet regions showed slightly poorer prediction performance, suggesting that the model may struggle with certain structural motifs that are less prominent in the training data. Future work could involve improving the model's performance on beta-sheet regions by collecting a more balanced dataset that includes a higher proportion of sequences with beta-sheet structures.

Finally, we can also experiment with more complex network architectures, such as transformer-based layers containing attention mechanisms, to capture long-range dependencies and specific cross-amino-acid dependencies in the protein sequences.

Contributions

Both authors contributed to the design, implementation, and writing of this project.

References

[1] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[2] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[3] Alexey Drozdetskiy, Christian Cole, James Procter, and Geoffrey J Barton. Jpred4: a protein secondary structure prediction server. *Nucleic acids research*, 43(W1):W389–W394, 2015.