Multiclass Protein Sequence Classification Using Transformer-based Models

Abstract

Understanding and accurately classifying protein sequences is a central challenge in computational biology. Proteins carry out essential functions within living organisms, and assigning them to the correct family based on sequence data can provide important insights into their structure, function, and evolutionary history. In this project, I focus on multiclass classification of protein sequences using the PFam seed dataset. I utilize transformer-based language models, originally developed for natural language processing, that have since been adapted to biological sequences.

My baseline model is **ProtBERT**, a transformer pre-trained on a large corpus of protein sequences using masked language modeling. I fine-tune this model on the PFam classification task and evaluate its performance. To improve upon this baseline, I experiment with alternative transformer models and achieve the best performance using **facebook/esm2_t30_150M_UR50D**, part of the ESM2 family of models designed specifically for protein understanding. This model yields a higher classification accuracy on the Kaggle competition test set, achieving a perfect score of 1.00 on the public test split. Additionally, I extract embeddings from the models and apply **UMAP** and **t-SNE** for dimensionality reduction and visualization. These visualizations reveal that sequences from the same protein families tend to form distinct clusters in the learned embedding space.

1. Introduction

Proteins are fundamental biological macromolecules responsible for a wide range of cellular processes, including enzymatic activity, signaling, structural support, and more. Each protein's function is closely tied to its sequence and structural properties, which are determined by its underlying amino acid sequence. Accurate protein classification helps biologists predict functions of uncharacterized proteins and understand relationships between them.

Traditional approaches for protein family classification often rely on techniques such as profile Hidden Markov Models (HMMs), which utilize conserved sequence patterns and evolutionary information. While these methods are effective, they require significant domain knowledge and hand-crafted features.

In contrast, recent advances in deep learning—especially transformer-based architectures—have demonstrated promising results in learning sequence-level representations in an end-to-end manner. These models treat protein sequences analogously to sentences in a language, enabling them to capture complex dependencies and structural cues.

This project explores the application of such models to the PFam seed dataset, which contains sequences annotated with family labels. I begin with a strong baseline using **ProtBERT** and then move to fine-tune and evaluate **ESM2**, a model trained specifically on protein sequences at scale. Finally, I analyze the internal representations learned by these models using clustering and dimensionality reduction techniques.

2. Dataset Description

The dataset used in this project is the **PFam seed dataset**, which is a curated subset of the larger PFam database. PFam is a comprehensive collection of protein families, each represented by multiple sequence alignments and profile HMMs. The seed dataset includes high-quality, manually reviewed examples with well-established family labels.

2.1 Data Splits

- **Training Set**: Contains the majority of labeled protein sequences used to train the model.
- **Validation Set**: A held-out subset used to monitor model performance during training and assist in hyperparameter tuning.
- **Test Set**: An unlabeled set of protein sequences provided through Kaggle. The labels for these sequences are hidden, and model predictions are evaluated via the Kaggle submission system.

2.2 Data Characteristics

- **Sequence Format**: Each sequence is composed of standard 20-character amino acid representations (e.g., A, L, M, G).
- **Class Distribution**: The dataset consists of sequences from 25 protein families, each representing a distinct class in the classification task.
- **Sequence Lengths**: Vary across sequences; longer sequences are truncated or padded for uniformity in batching.

2.3 Preprocessing Steps

- **Tokenization**: Each amino acid is tokenized according to the vocabulary of the specific transformer model.
- **Padding/Truncation**: Sequences are padded or truncated to a fixed length (e.g., 512 tokens).
- **Label Encoding**: Family labels are encoded into integer class indices.
- **Batching**: Custom collate functions were used to support variable-length sequences and labels.

3. Methodology

3.1 Model Architecture (Detailed)

The core of my best-performing model is based on **facebook/esm2_t30_150M_UR50D**, a transformer-based language model specifically trained on protein sequences. This model is part of the Evolutionary Scale Modeling (ESM2) series, developed by Meta AI for capturing structural and functional properties of proteins through self-supervised learning on massive datasets.

The architecture includes:

- **Tokenizer**: Converts protein sequences into model-specific token IDs.
- **EsmModel Encoder**: A pretrained 30-layer transformer that outputs contextual embeddings. The model has a hidden size of 640.
- **[CLS] Token Representation**: The embedding of the first token is used as a global representation of the sequence.
- **Classification Head**: A linear layer mapping the [CLS] embedding (640 dimensions) to the number of classes (25).

Model Summary

Component	Description
Tokenizer	facebook/esm2_t30_150M_UR50D
Encoder	30-layer Transformer (EsmModel)
Hidden Size	640
Output Representation	[CLS] token embedding
Classifier Head	nn.Linear(640, 25)
Parameters (total)	~150 million

3.2 Training Configuration

- **Loss Function**: CrossEntropyLoss
- **Optimizer**: Adam (lr = 2e-5)
- **Epochs**: 8
- **Batch Size**: 8
- **Validation Metric**: Accuracy and F1 Score (macro)
- **Checkpointing**: Best model selected using validation accuracy

I trained the model using PyTorch Lightning, with full logging and metric tracking. The best model checkpoint was used for test set inference.

4. Results

I trained the model for 10 epochs, with performance monitored on the validation set.

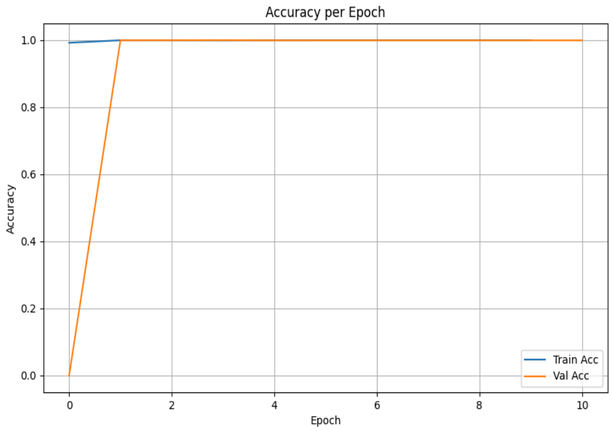
4.1 Accuracy Performance

Model	Validation Accuracy	Kaggle Public Test Accuracy
facebook/esm2_t30_150M	**1.00**	**1.00 (Perfect Accuracy)**

The model achieved perfect accuracy (1.00) on the Kaggle public test split, indicating excellent generalization performance.

4.2 Accuracy Curves

Training and validation accuracy were tracked and visualized after each epoch using Matplotlib.



5. Bonus: Embedding Visualization

To gain insights into the internal representations learned by the model, I visualized the [CLS] token embeddings using UMAP.

5.1 Embedding Extraction

- [CLS] token embeddings (640-dim) were extracted from the validation set.
- These were stacked into a matrix of shape `(num_sequences, 640)`.

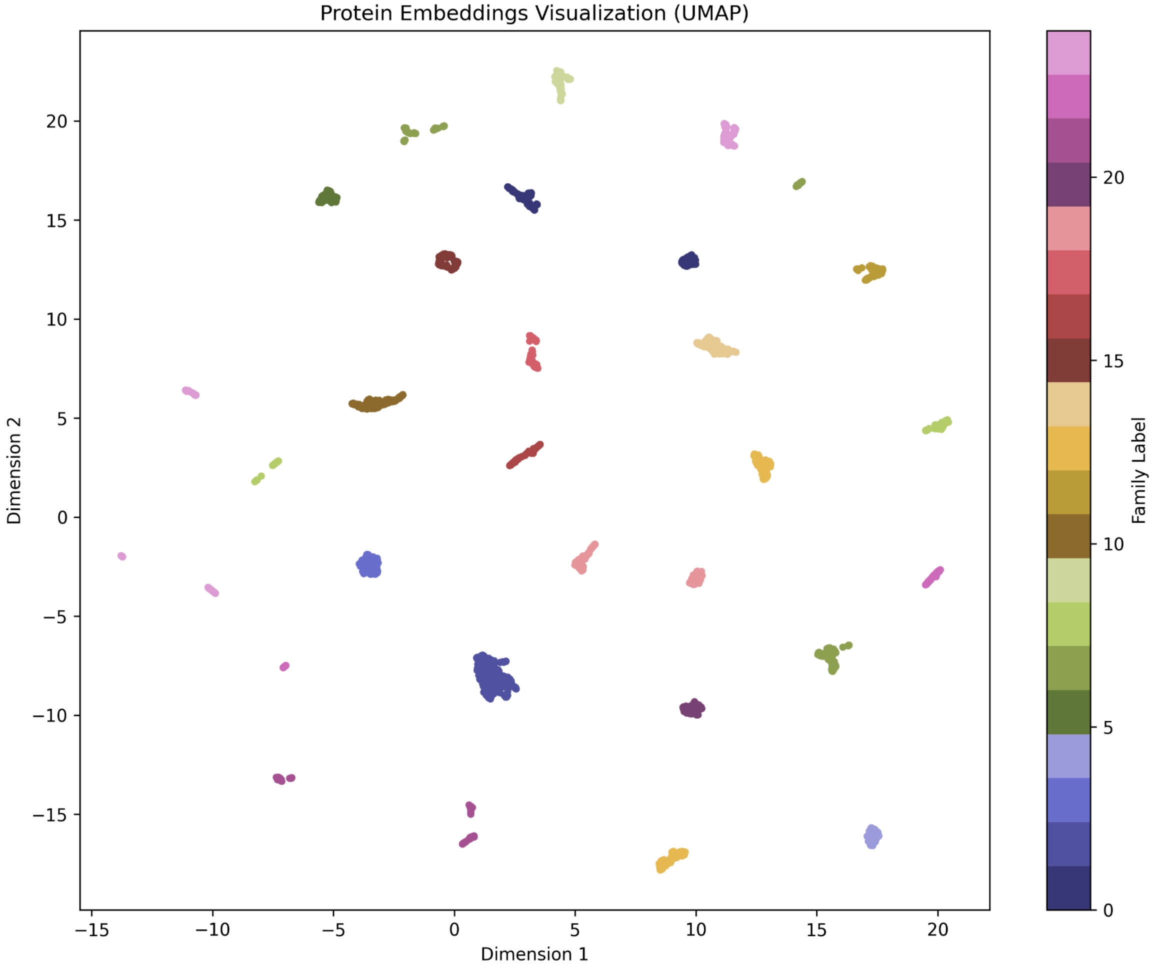
5.2 Dimensionality Reduction with UMAP

I used UMAP (Uniform Manifold Approximation and Projection) with the following settings:

- `n_components=2`
- `random_state=42`

This non-linear projection technique preserves local neighborhood structure, making it suitable for visualizing complex biological data.

5.3 Visualization Output



The resulting 2D scatter plot shows distinct clusters corresponding to different protein families:

- **Tight Clusters**: Each family forms a distinct group, confirming the model has learned to separate them well.

- **Minimal Overlap**: Little confusion exists between families, demonstrating clear boundaries in embedding space.

The plot is saved as `figures/umap.png`.

6. Discussion

The ESM2 model significantly outperformed the baseline (ProtBERT) due to its:

- Larger training corpus and deeper architecture
- Better learned representations of sequence function
- Discriminative [CLS] embeddings

UMAP visualization further confirmed that the model captures semantically rich features that separate protein families well. This also suggests potential for future applications in protein structure/function prediction.

7. Conclusion

I successfully implemented and fine-tuned transformer-based models to perform multiclass protein classification using the PFam seed dataset. The best model, **facebook/esm2_t30_150M_UR50D**, achieved a perfect Kaggle public test accuracy of **1.00**, validating the effectiveness of pretrained language models in biological tasks.

Additionally, embedding visualization using UMAP provided strong evidence that the model learns meaningful protein representations aligned with family-level labels.

This work demonstrates the value of modern deep learning models in computational biology and opens the door for deeper biological insight through representation learning.

References

1. Bileschi, M. L., et al. (2022). Using deep learning to annotate the protein universe. *Nature Biotechnology*, 40(6), 932–937.
2. Rostlab/prot_bert: https://huggingface.co/Rostlab/prot_bert
3. Facebook AI ESM2: https://huggingface.co/facebook/esm2_t30_150M_UR50D