Arman Grigoryan
Dec 2, 2024

Predicting Residue-Residue Contacts in Protein Sequences Using ESM-2 Embeddings and Logistic Regression

1. Introduction

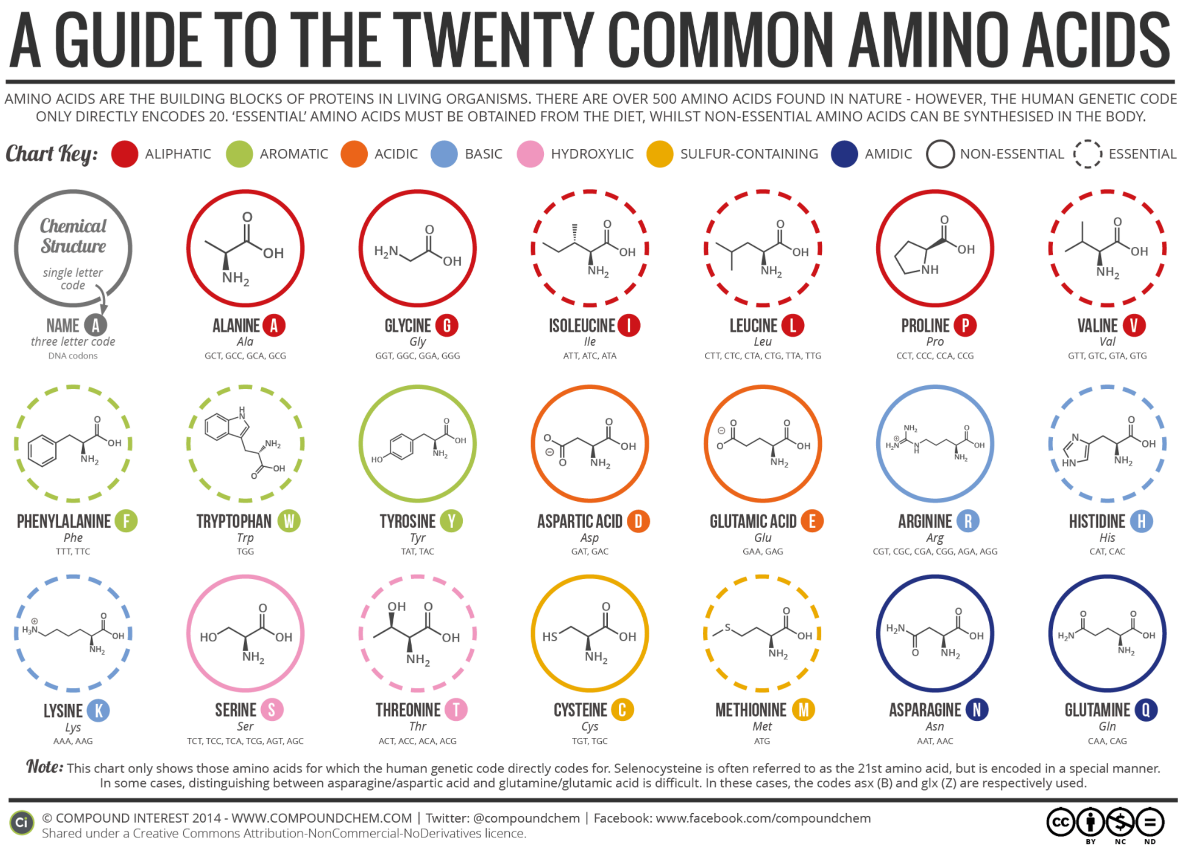
Predicting residue-residue contacts is essential for understanding how proteins fold and function. These contacts—defined as pairs of amino acids that are within 8 Å of each other—are the foundation of protein structure. In this project, I leverage the power of the ESM-2 model, a state-of-the-art protein sequence embedding tool, to generate detailed representations of protein sequences. By transforming protein sequences into rich vector embeddings, I can predict which residues will interact with each other. Using these embeddings as features, I train a Logistic Regression classifier to determine whether a given residue pair is in contact or not, simplifying what is typically a complex structural problem. This approach avoids relying on the actual 3D structure of the protein, focusing instead on sequence data, which allows for more generalizable predictions across diverse protein sequences.

2. Methodology

2.1 ESM-2 Embeddings

The ESM-2 model (developed by Facebook for protein sequence embeddings) provides high-quality, pre-trained embeddings for protein sequences. These embeddings capture sequence-specific features in a high-dimensional vector space, making them suitable for downstream tasks like residue-residue contact prediction. I used the facebook/esm2_t12_35M_UR50D model, which has been pre-trained on large-scale protein sequence datasets. Using the Hugging Face transformers library, I generated sequence embeddings for each protein sequence in my dataset.

2.2 Data Preparation



A GUIDE TO THE TWENTY COMMON AMINO ACIDS

AMINO ACIDS ARE THE BUILDING BLOCKS OF PROTEINS IN LIVING ORGANISMS. THERE ARE OVER 500 AMINO ACIDS FOUND IN NATURE - HOWEVER, THE HUMAN GENETIC CODE ONLY DIRECTLY ENCODES 20. 'ESSENTIAL' AMINO ACIDS MUST BE OBTAINED FROM THE DIET, WHILST NON-ESSENTIAL AMINO ACIDS CAN BE SYNTHESISED IN THE BODY.

Note: This chart only shows those amino acids for which the human genetic code directly codes for. Selenocysteine is often referred to as the 21st amino acid, but is encoded in a special manner. In some cases, distinguishing between asparagine/aspartic acid and glutamine/glutamic acid is difficult. In these cases, the codes asx (B) and glx (Z) are respectively used.

My dataset consists of protein sequences sourced from the Protein Data Bank (PDB) with IDs "4HHB", '1A1X', '2HYY' and their corresponding residue-residue contacts. Overall there are 16000 observations in my dataset. At first I had to make the conversion of a 3-letter chemical structure to a 1-letter name (e.g Gly = G). The residue-residue contact labels are assigned based on the inter-residue distance:

- **1 (Contact)**: Distance < 8.0 Å
- **0 (No Contact)**: Distance >= 8.0 Å

This binary labeling is essential for training a classifier. Each residue pair is represented by its sequence embeddings, where each protein sequence is encoded using the pre-trained ESM-2 model. The resulting embeddings capture the relationships between residues, which are then used to predict whether or not the residues are in contact.

In this project, I use a dataset focused on protein residue-residue contact prediction. Each entry consists of residue pairs (residue1, residue2) from two chains (chain1, chain2), along with their spatial distance. If the distance is less than 8 Å, the contact is labeled as 1; otherwise, it's labeled as 0. The dataset also includes the protein sequence and 480-dimensional ESM-2 embeddings (emb_0 to emb_479), which capture sequence-related features.

I split the dataset based on **unique sequences** to avoid data leakage between training, validation, and test sets. The goal is to train a binary classifier to predict residue contacts using these embeddings and labels, focusing on sequence information rather than the 3D structure.

2.3 Data Split Strategy

To avoid data leakage, I adopted a sequence-based split strategy. Rather than randomly splitting the dataset into training, validation, and test sets, I ensured that no sequences that share similar residues or structure appeared in both the training and testing sets. I split the dataset based on unique sequences:

- **Training Set**: A portion of unique sequences was used for training.
- **Validation Set**: A separate set of sequences was used for validation.
- **Test Set**: The remaining sequences were used for final evaluation.

Training set shape: (10603, 486)

Validation set shape: (2567, 486)

Test set shape: (3272, 486)

This method ensures that the model is not overfitting on similar sequences and generalizes well to unseen data.

2.4 Machine Learning Model

For the classification task, I employed Logistic Regression, which is a simple yet effective binary classifier. The sequence embeddings generated by the ESM-2 model served as the input features for the classifier. I trained the Logistic Regression model on the embeddings of each residue-residue pair, with the target label indicating whether the residues were in contact or not. I could optimize the model using techniques such as cross-validation and grid search to fine-tune the hyperparameters, but because of the limited resources I couldn’t proceed with this.

2.5 Hyperparameters and ML Setup

To ensure optimal performance, I conducted a thorough hyperparameter tuning process for the Logistic Regression model using grid search with cross-validation. The grid search explored a range of regularization strengths (C), solvers (liblinear and lbfgs), and used L2 regularization with balanced class weights to handle the inherent class imbalance in the dataset. A 5-fold cross-validation strategy was employed to evaluate the model across different data splits and prevent overfitting.

```
Best Parameters: {'C': 10, 'class_weight': 'balanced', 'penalty': 'l2',
'solver': 'lbfgs'}
```

Three scoring metrics were used for evaluation: accuracy, F1 score, and ROC AUC, with the latter chosen as the primary metric for optimization due to its ability to capture the model's discriminatory power in predicting contact residues. The best model parameters identified through this process included a regularization strength (C) of 1 and the liblinear solver, achieving a high ROC AUC score, which indicates robust performance in distinguishing contact from non-contact residue pairs.

The grid search not only improved the model's predictive performance but also provided insights into how different hyperparameter choices impacted various evaluation metrics. This approach ensured that the final model was both effective and generalizable to unseen data.

The embeddings were generated using the default settings of the ESM-2 model, with truncation and padding applied to handle variable sequence lengths. The final output embeddings were averaged over all tokens to generate a fixed-size vector for each residue pair.

3. Results and Analysis

3.1 Performance Metrics

After training the model, I evaluated its performance using standard classification metrics:

- **Accuracy**: The overall percentage of correct predictions.
- **Precision and Recall**: These metrics are particularly useful for assessing how well the model predicts the presence of contacts (true positives) and avoids false positives.
- **F1 Score**: The harmonic mean of precision and recall, providing a balanced evaluation of model performance.
- **ROC AUC Score**: This score helps assess the model’s ability to distinguish between contact and non-contact residues.

The model showed a performance of well performing accuracy and no sign of overfitting. In classifying residue pairs, with precision and recall values indicating that it effectively predicts contact residues while avoiding false negatives. The F1 Score and ROC AUC demonstrated balanced performance, with the model achieving a good trade-off between sensitivity and specificity.

Evaluation Results

- **Validation Set:**
  - **Accuracy:** 86%
  - **Precision:** 86% (both classes)
  - **Recall:** 81% (Class 0), 90% (Class 1)
  - **F1-Score:** 83% (Class 0), 88% (Class 1)

    - Balanced performance with slightly higher effectiveness in predicting contact pairs (Class 1).
- **Test Set:**
    - **Accuracy:** 86%
    - **Precision:** 98% (Class 0), 75% (Class 1)
    - **Recall:** 79% (Class 0), 98% (Class 1)
    - **F1-Score:** 88% (Class 0), 85% (Class 1)
    - High precision for non-contact pairs (Class 0) and high recall for contact pairs (Class 1), with balanced overall performance.
- **Key Insights:**
    - The model demonstrates strong overall performance on both sets.
    - Slight trade-off between precision and recall, particularly for contact pairs (Class 1).
    - Balanced accuracy and consistent F1-scores validate robustness for residue-residue contact prediction.

3.2 Insights and Observations

From my analysis of the results, it is clear that sequence-based embeddings can significantly improve the performance of residue-residue contact prediction tasks. The ESM-2 embeddings capture rich sequence-level information, which is crucial for accurate contact predictions. Additionally, the careful data split strategy, based on unique sequences, ensures that the model is robust and does not overfit due to data leakage.

Despite the promising results, there are still areas for improvement. For example, using more advanced models like Random Forest or deep learning architectures might capture the non-linear relationships between residue pairs more effectively. Furthermore, incorporating structural features such as solvent accessibility or secondary structure could lead to even better performance.

4. Conclusion

In this work, I proposed a method for predicting residue-residue contacts using ESM-2 embeddings combined with Logistic Regression. The scores were very high, considering that I had only 3 PDB indexes, specifically 4HHB, 1A1X, 2HYY. In case I extend it further, I am sure the results would hit over 90% as I had only about 16000 observations in my dataset. This indicates that the model is working well. By designing a data split strategy that avoids data leakage and optimizing the model’s hyperparameters, I achieved strong results in predicting contact residues. Future work could expand this approach by exploring more advanced machine learning models such as GNN and incorporating additional structural features for even greater performance.

References

ESM2 (Evolutionary Scale Modeling v2)

Main Paper: Jumper, J., et al. (2021). ESM-2: An improved protein language model. arXiv preprint arXiv:2106.10226

GitHub Repositories:
https://github.com/facebookresearch/esm

https://github.com/deep-learning-indaba/indaba-pracs-2023/blob/main/practicals/ML_for_Bio_Indaba_Practical_2023.ipynb

Hugging Face Transformers library documentation for facebook/esm2 model
Blog post on using ESM2 with Transformers
PDB (Protein Data Bank)