Protein Thermostability Prediction with Modeled Structures

Presenters: Or Levy, Maya Epstein, Nir Feldman | **Supervisor:** Dr. Jerome Tubiana

1. Abstract

Protein stability is critical for maintaining structure and function, especially at high temperatures, in various fields like biology and industry. We propose a new machine learning method to forecast protein stability using sequences. We utilize the Kaggle Novozyme Enzyme Stability Challenge dataset and ESMFold-predicted structures to extract features and employ a Random Forest Classifier for prediction. Our model shows promising results in cross-validation, with interpretable feature analysis. This study enhances our understanding of protein stability and offers insights into improving protein performance across different environments. Our model shows the importance of making predictions about how proteins react to temperature changes. It hints at the potential of using computers to understand and manipulate proteins in industries like biotechnology and beyond.

2. Introduction

Protein thermostability, the ability of proteins to withstand high temperatures, is crucial for optimal performance, especially in challenging industrial conditions. Many proteins become less stable and lose their structure beyond ~40 degrees Celsius, limiting their usefulness and reducing cellular production. Predicting protein thermostability is important for enhancing protein engineering, making proteins function better in high-temperature environments, and improving the efficiency of cellular protein production. Developing smart computer tools for accurate predictions is essential for these advancements. This article emphasizes the significance of understanding and predicting protein thermostability, offering insights into how proteins behave in various conditions.

In this article, we're talking about how we use computer models (using machine learning techniques) to guess the temperatures where proteins stay stable. We're using a dataset from Kaggle, which comes from a competition called novozymes-enzyme-stability-prediction. We're comparing our model with a simpler one to see which is better at guessing temperatures.

We're also seeing if our model can tell the difference between proteins from humans and those from thermophilic organisms. This helps us figure out if our method works for different kinds of proteins.

Our goal is to make it easier to understand how we predict protein temperatures and how it can help in different industries and biology.

3. Methods

In our study, we used advanced methods to predict protein thermostability. We started by collecting a dataset from Kaggle, focusing on the Novozymes-enzyme-stability-prediction competition. With over 30,000 sequences, we carefully partitioned the data and identified sequence pairs with MMSeq2. Using SciPy, we grouped similar sequences into cohesive units. Next, we divided the dataset for training, testing, and validation. We created a protein feature map with the ESM algorithm and trained a Random Forest Regressor model. Comparing our model with a basic feature map highlighted its effectiveness. Through these methods, we aimed to enhance protein research and industrial applications.

The methods:

1. **Dataset Collection:**

We gathered our dataset from Kaggle, a platform known for data science competitions, specifically focusing on the Novozymes-enzyme-stability-prediction contest. The dataset includes sequence identifiers (seq_ids), protein sequences, pH values, data sources, and melting temperatures (tm), offering a complete picture of protein behavior.

The initial dataset holds over 30,000 sequences, each with unique sequence identifiers (seq_ids).

2. **Preprocessing:**

We organized our data for analysis, considering practical aspects like RAM availability in Google Colab and memory constraints. Initially, we filtered the dataset to keep unique protein sequences under 400 amino acids, a crucial step to manage memory efficiently and streamline subsequent analyses.

MMSeqs2 Algorithm:

For uncovering meaningful patterns, we employed the MMSeqs2 algorithm. This tool diligently searched and aligned sequences, revealing pairs with over 70% identity. The output of this algorithm is a curated list of pairs, where each pair exhibits more than 70% identity. This list played a pivotal role in the following stages of our analysis.

Sparse Matrix Construction:

Using SciPy, we transformed the pair list into a sparse matrix, a powerful computational tool. This matrix facilitated discovering connected components by grouping sequences with 70% or more identity together. This strategic step provided a structured foundation for further analyses.

3. Data Partition:

In the final data preparation step, we carefully divided the dataset into training, testing, and validation sets. The division allocated 60% for training, 20% for testing, and an additional 20% for validation. Additionally, we grouped all vertices within each connected component, ensuring our model captures cohesive patterns during training.

4. Creating Protein Feature Map:

We used the ESM (Evolutionary Scale Modeling) algorithm, known as ESM of Meta, to convert protein info into a computer-friendly format. This process generated a feature map with 320 embeddings for each protein sequence, allowing us to better understand and study proteins.

5. Training the Predictive Model:

To predict protein thermostability, we relied on the Random Forest Regressor. Trained with 30 decision-making experts ($n\_estimators$=30), this model became the heart of our predictions. The ensemble nature of Random Forests, coupled with expert decision-makers, ensured a powerful tool for capturing the nuances of protein behavior and stability.

6. Comparison with Basic Feature Map:

In a comparison, we looked at our advanced feature map model and a simpler one using a basic 20-dimensional vector for amino acid frequencies. This step revealed the effectiveness of our feature-rich protein language models.

4. Results

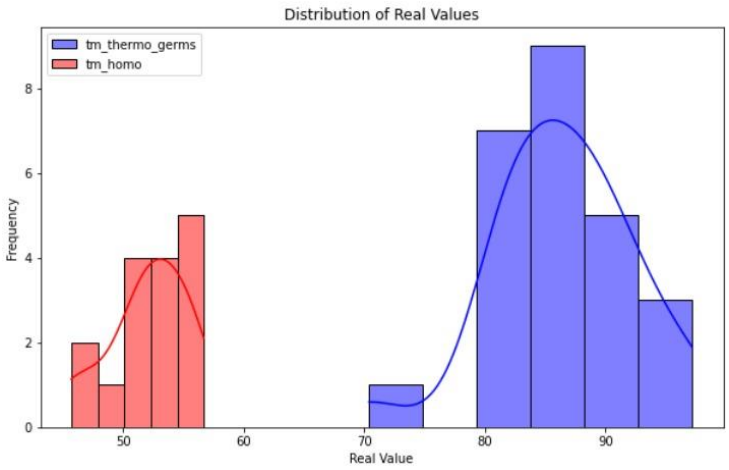
Performance Comparison:

In our study, we compared the predictive performance of two models: our pLM and the amino acid content model. Here are the key findings:

- Our pLM exhibited an average deviation of 8.432 in predicting protein thermostability temperatures, while the amino acid content model showed an average deviation of 8.610. This suggests that our pLM generally outperforms the simpler amino acid content model in terms of accuracy.
- The Pearson correlation coefficient for our model is 0.629, slightly higher than the coefficient of 0.619 for the other model. While the differences seem marginal, they become more pronounced when visualized through histograms, as discussed below.

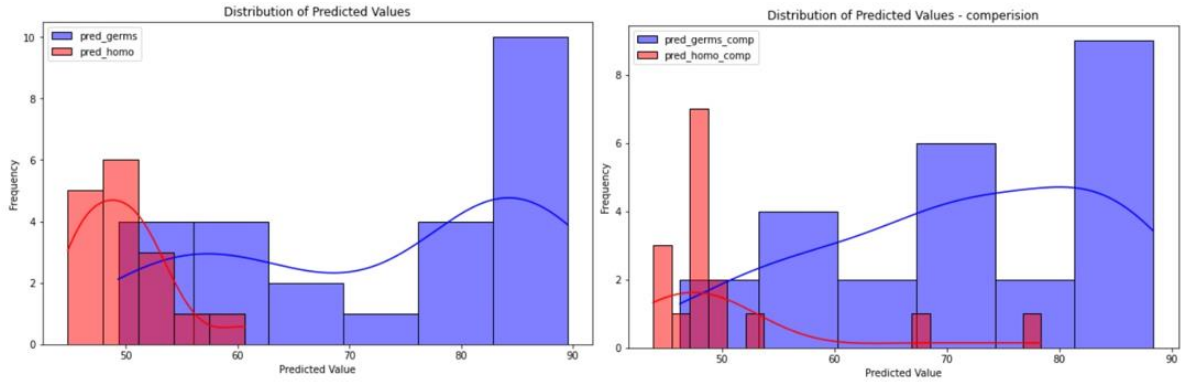
Human vs. Thermophilic Sequence Predictions:

Figure 1



In Figure 1, we present the actual values extracted directly from our dataset. This figure comprises 16 sequences from human organisms and 25 sequences from thermophilic organisms. Notably, there's a distinct separation evident between the two groups depicted by the red and blue segments. Human sequences are primarily clustered within the range of 45 to nearly 60 degrees Celsius, while thermophilic sequences span from approximately 80 to nearly 100 degrees Celsius. This clear distinction underscores the unique thermal preferences of proteins derived from different organisms.

Figure 2



Examining the histograms in Figure 2, we observed that both models exhibit alignment with the actual temperature values, particularly in the red group. However, our pLM demonstrates a smaller deviation, indicating a closer proximity to the true values.

We observed that our model tends to perform better in predicting temperatures for sequences resembling those found in humans compared to those from thermophilic organisms. This can be attributed to the abundance of sequences with temperatures around 50 degrees Celsius in our dataset (close to 8000 sequences), which the model learned more effectively than the thermophilic sequences.

An intriguing observation is the close proximity of the global maximum point in the red graph of our model to the global maximum point in the true values graph. This alignment underscores the precision of our model in predicting temperatures for sequences resembling those found in humans.

Furthermore, our model outperforms the amino acid content model in predicting temperatures for sequences with higher values (75 degrees Celsius and above), despite their scarcity in the dataset. Specifically, our model identifies 14 such sequences, compared to the 11 identified by the amino acid content model.

Additionally, our model's superiority can be attributed to its more comprehensive feature map, encompassing 320 features compared to the 20 features in the amino acid content model. This broader feature set contributes to greater accuracy in temperature predictions.

5. Discussion:

In conclusion, the implementation of machine learning models for predicting protein thermostability from primary sequences offers promising insights into enhancing protein stability under harsh conditions. The model constructed through this study represents a significant step towards understanding the factors influencing protein stability and provides a framework for developing more robust computational approaches in the future.

The construction of a Random Forest Classifier and the evaluation of its performance provided valuable insights into the predictive capabilities of the model.

The exploration of follow-up tasks, including the training of models based on protein language models and the generation of proteome-wide predictions, presents exciting avenues for further research and application. By expanding the scope of analysis to include a broader range of organisms and incorporating advanced machine learning techniques, we can refine our understanding of protein thermostability prediction and its implications across various biological and industrial contexts.

What could have been done with more time and resources:

1. **Inclusion of more thermophilic organism sequences:** Expanding the training dataset to include sequences from a diverse range of thermophilic organisms could enhance the model's predictive accuracy and generalizability. Incorporating data from organisms known for their extreme thermostability could provide valuable insights into the underlying mechanisms governing protein stability under high temperatures.

2. **Utilization of BLAST for data partitioning:** Integration of BLAST (Basic Local Alignment Search Tool) for data partitioning could improve the quality of the training and testing datasets by ensuring a more representative distribution of sequences across different protein families and structural motifs. This approach could help mitigate biases and enhance the robustness of the predictive model.

3. **Exploration of More Complex ESM Models:** The exploration of more complex ESM (Evolutionary Scale Modeling) models with larger feature maps could capture a broader range of structural and evolutionary information, potentially improving the model's ability to discriminate between stable and unstable proteins. Investing time in

optimizing the selection and integration of structural features derived from advanced ESM models could lead to more accurate predictions of protein thermostability.

In summary, while the current study represents a significant advancement in the field of protein thermostability prediction, there remains ample room for further exploration and refinement. By addressing the above and leveraging additional resources and methodologies, future research endeavors hold the potential to unlock new insights into the fundamental principles governing protein stability and facilitate the development of innovative strategies for enhancing protein performance in diverse applications.

6. Bibliography

1. Thermostability:
    a. Wikipedia https://en.wikipedia.org/wiki/Thermostability
    b. Lecture on energetics and stability (Chapter 4, Kessel & Ben Tal): https://www.bentalab.com/protein-book
2. ESMFold: https://www.biorxiv.org/content/10.1101/2022.07.20.500902v2
3. Language models: https://arxiv.org/pdf/2007.06225.pdf
4. MMseqs2: https://github.com/soedinglab/MMseqs2
5. SciPy: https://docs.scipy.org/doc/scipy/reference/sparse.html
6. Random Forest Regressor: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html