Protein Structure Prediction Using Probabilistic Graphical Models

Jacob Berkel, Lee Kim, Andy Nguyen

Department of Computer Science, University of California, Irvine

COMPSCI 179 Introduction to Graphical Models

Professor Alexander Ihler

June 14, 2023

Protein Structure Prediction Using Probabilistic Graphical Models

This report explores the application of probabilistic graphical models for protein secondary structure prediction using amino acid sequences. Six models were examined: the hidden Markov model (HMM), four n-gram models four n-gram models (two bigrams, trigram, quadgram), and the stochastic random model. The aim was to achieve accurate predictions of protein secondary structure using these probabilistic graphical models.

Resources and Methods

Datasets

ss.txt: Sequences and secondary structure information generated using DSSP.
T.txt: Protein names extracted from ss.txt.
X.txt: Amino acid sequence data extracted from ss.txt.
Y.txt: Corresponding secondary structure data for the amino acid sequences in X.txt.

Libraries

pyGMs: Used for building graphical models. (Source: https://github.com/ihler/pyGMs)
scikit-learn: Utilized the train_test_split function to split the dataset. (Source: https://github.com/scikit-learn/scikit-learn)
NumPy: Used for array manipulation and computations.
matplotlib: Used for generating visual representations of experimental results.

Code Implementation

The implementation of the models involved utilizing code provided by Professor Alexander Ihler in the ProteinSS example file. This code encompassed tasks such as data preprocessing, loading, machine learning estimation functions, scoring and prediction functions, and two versions of fully observed HMMs. Our team adapted the provided observed HMM by modifying the emission outputs to consider pairs of amino acids instead of singular ones to capture more information. Additionally, we implemented the bigram models, trigram model, quadgram model, and stochastic random model.

Methods

The HMM utilized probabilistic modeling techniques to consider pairs of amino acids to capture transitions between underlying states and predict secondary structures over 2500 proteins with 10 generations per protein. Four n-gram models were implemented and evaluated on two different sets of train-test datasets: one with an 80:20 training and testing split, trained on 1,000 proteins, tested on 4,000 proteins, 10 generations per protein, and another with a 20:80 split, trained on 4,000 proteins, tested on 1,000 proteins, 10 generations per protein. The amino acid bigram model incorporated one adjacent amino acid to predict the secondary structure, while the secondary structure bigram model incorporated the current secondary structure to predict the subsequent one. The amino acid trigram model extended the amino acid bigram model by considering two adjacent amino acids. Additionally, the amino acid quadgram model considered three adjacent amino acids. Lastly, the stochastic random model randomly predicted secondary structures without considering sequence information or patterns and served as a reference for performance comparisons.

Results

Table 1

HMM

Model	Error Rate (%)	Timing (sec)
Simple HMM	0.00	20.5
Pair HMM	55.3	38.4

Table 2

Amino Acid Bigram Model

Train-Test Ratio	Error Rate (%)	Timing (sec)
80:20	24.7	90.0
20:80	24.7	333

Table 3

Secondary Structure Bigram Model

Train-Test Ratio	Error Rate (%)	Timing (sec)
80:20	22.4	81.0
20:80	22.3	314

Table 4

Amino Acid Trigram Model

Train-Test Ratio	Error Rate (%)	Timing (sec)
80:20	26.7	87.0
20:80	26.5	319

Table 5

Amino Acid Quadgram Model

Train-Test Ratio	Error Rate (%)	Timing (sec)
80:20	28.8	82.0
20:80	28.6	321

Table 6

Stochastic Random Model

Error Rate (%)	Timing (sec)
12.5	235

Discussion

The evaluation techniques used in this research included error rates, timing information, and comparisons to the stochastic random model. These evaluation metrics allowed for a comprehensive assessment of the models' accuracy, computational efficiency, and relative performance in predicting protein secondary structure.

The HMM model performed poorly in this study, exhibiting an error rate of 55.3%. This low performance could be attributed to certain limitations or inadequacies in capturing the complexities of protein structures. While the HMM model utilizes probabilistic modeling techniques to capture transitions between latent states, it may not have effectively captured the complex relationships between amino acids and their corresponding secondary structures. Therefore, incorporating more complex techniques or alternative models may be necessary to enhance prediction accuracy.

Among the n-gram models, the secondary structure bigram model with an 80:20 train-test ratio emerged as the most promising model for protein secondary structure prediction, achieving an error rate of 22.3%. This model's consideration of the current secondary structure to predict the subsequent one appeared to provide valuable information into the sequential dependencies of protein structures. The n-gram models did not exhibit significant improvement with increased training data and became less accurate as more amino acids were considered. This could be due to the models not considering the previous secondary structures.

The stochastic random model played a crucial role in this research as it served as a performance reference for predicting protein secondary structures. Among all the models considered, the stochastic random model demonstrated significantly better prediction results. This model's superior performance suggests that considering sequence information or patterns alone may not be sufficient for accurate protein secondary structure prediction. Further research is needed to investigate potential enhancements.

In summary, the stochastic random model outperformed all other models, suggesting the importance of considering factors beyond sequence information alone for protein secondary structure prediction. Among the other models, the secondary structure bigram model with an 80:20 train-test ratio demonstrated the most promising performance, while the HMM model exhibited lower accuracy. The findings of this study contribute to the understanding of the strengths and limitations of different probabilistic graphical models in protein secondary structure prediction.