Protein Function Classification

Sean Whalen, Kirsten Winter, Boyang Zhang

Problem Formulation

There are many different types of proteins that exist, each with a different responsibilities for maintaining the human body. Proteins of the same type exhibit similarities in their amino acid sequences. Given a set of proteins' amino acid sequences labeled with their corresponding functions, we want to learn the sequential properties of proteins for a given function. We will compare two different approaches in protein function classification.

Dataset

We will be using data from the Research Collaboratory for Structural Bioinformatics' (RCSB) Protein Data Bank (PDB). The database has over 400,000 protein sequences labeled with protein function. We plan to choose 15 classes of protein sequences and use different sizes of training and testing sets to examine the scalability of the algorithms. We will also test the effect of larger class size's effect on the performance if time is permitted.

Technical Approach

Hidden Markov Model

Hidden Markov Modelling sequences as Markov chains is known to have success in predicting potential future states of a sequence based on a previous state. For proteins with similar functions, the sequence structure is constructed through similar process. Using a set of protein sequences from the same function, we can calculate the parameters of the hidden markov model and find the most probable path through the hidden states for this class. Then, for a given sequence, we can derive the likelihood of constructing such sequence through the process learned with training data.

RNN (LSTM)

Recurrent neural network has had numerous successes in sequential data classification problems, especially the long-short term memory algorithm. RNN has the advantage to take inputs of different lengths, and learn from the sequential data directly without feature engineering. We will use a softmax layer after the LSTM layers to obtain an output indicates the probabilities of the input sequence belongs to each function group.

Scoring

The two metrics we will use to evaluate the performance are running time and top 3 categorical accuracy, i.e. success when the target class is in the top 3 predictions provided.

References

Deep Recurrent Neural Network for Protein Function Prediction from Sequence
https://arxiv.org/abs/1701.08318