Mathias Perez, Edouard Rabasse

15 april 2024

INTERMEDIATE REPORT

Protein Cleavage Prediction

Our goal is to predict the cleavage of proteins. We have a dataset of proteins and their cleavage sites. We will use a classification kernel to predict if a chain of amino acids contains a cleavage site and where. We will use a dataset of proteins and their cleavage sites to train our model. We chose to use python to process the data and train the model.

FIRST LOOK AT THE DATASET

We have a dataset of proteins and their cleavage sites. Each protein is repsented over 3 lines :

— the first line contains some informations on the protein (not relevant to our study),
— the second line is the sequence of amino acids,
— the third line is the position of the cleavage site.

Preprocessing

We first had to check the data (add missing spaces...). We then preprocessed the data to extract the amino acid sequence and the cleavage site. We first convert the amino acid sequence to a vector of integers using a one-hot encoding as suggested.

SIMPLE STATISTICAL APPROACH

For the whole dataset, we calculated the frequency of each amino acid at each position (relative to the cleavage site). We were able to determine the q-1 score of a sequence $a_0a_1...a_{p+q}$

defined as $q-1 = \sum_{i=-p}^{q-1} \log(f(a_{i+p})) - \log(g(a_{i+p}))$ where $f(a,i)$ is the frequency of the amino acid a at position i (relative to the cleavage site) and $g(a)$ is the frequency of the amino acid a in the whole dataset.

For example, the most common amino acid over the whole set is L (leucine) and the most common amino acid at position 0 (relative to the cleavage) is A (alanine) with a frequency of 0.21

Our goal with this Statistical approach is to find the best parameters p and q and a threshold to predict if a sequence of length p+q contains a cleavage site at position p. We will then use a sliding window to predict the cleavage site of a protein.

SVM KERNEL APPROACH

We use the scikit-learn library to train our model. To classify the data, we will use two SVM kernel. The first model should be able, for a given sequence of amino acids, to predict if it contains a cleavage site. To get training data for the first model, we will use a sliding window of size n to extract sequences of amino acids and their label : contains a cleavage site or not.

The second model should be able to predict the position of the cleavage site in a sequence of size n of amino acids that contains one. To get training data for the second model, we will take sequences of size n amino acids that contain a cleavage site and extract the position of the cleavage site.

Then we combine the two models to predict if a chain of amino acids contains a cleavage site and where by using a sliding window of size n.

Next steps

We need to choose the right parameters for the SVM kernels and the sliding window. We might as well want to improve our batching process.

CONCLUSION

We have preprocessed the data and are now able to train our model. We have started to implement a simple statistical approach to predict the cleavage site of a protein. We have also started to implement a SVM kernel to predict the cleavage site of a protein. We will now focus on training our model and improving our predictions.