Project plan

Prediction of residue burial status in membrane proteins

The goal of the project is to write a predictor for the residue burial status for membrane proteins. This will involve performing the following steps:

1. Extracting the feature from the dataset

   The current format of the dataset is a text file with three lines per element: ID, AA-sequence and feature. It should be parsed to a dictionary, where keys are IDs and values are lists of two items (AA-sequence and feature). The dictionary and window size can then be used as variables in a function which creates two arrays, a suitable format for SVM analysis: X, which holds the training samples and Y, which contains the labels. One-hot encoding for amino acids and binary encodings for labels will be used in these arrays.

2. Creating cross-validated sets

   The dataset will be split into three equal parts and each part restructured into a suitable format as described in the step above so that the training can be performed on two sets and tested on the third in three different iterations.

3. Training an SVM using single sequence information, using sklearn

   The training will be performed by using the two constructed arrays in the form of X = [[0, 0], [1, 1]] and Y = [0, 1] for the svm package from sklearn. The trained model can then be used to predict the feature for the test dataset.

4. Check different window sizes for the inputs

   As the window size will be used as a variable for the function creating the input arrays for SVM analysis, this step will involve running the code with different window sizes and comparing how it changes the accuracy and sensitivity on the prediction when the model is trained and tested with different window sizes.

5. Analyze the results and compare it to previous work

   The results will be evaluated by calculating the accuracy and sensitivity of the prediction and comparisons will be made by looking at the already developed methods as described in literature.

6. Review the state of art for your predictor

   Previous publications on the developed prediction methods will be reviewed to compare with the methodology used in my predictor.

7. Write a report

   The report will be 5-10 pages long, will be structured as a scientific paper and will include the description of how the project was performed, as well as analysis and discussion of the results.