Lab Notebook for Predicting Signal Peptides

**Feride Eren**

Objective:

It is needed to build, train and test a signal peptide classifier with data obtained by Applied Bioinformatics professor. 4 different models are chosen to be used which are Bernoulli Naïve Bayes, Multinomial Naïve Bayes, Logistic Regression and SVM.

## **10-12-2017 – Project initialisation**

**Event/findings:**

Machine learning is a new concept for me so my first aim is to understand key elements of machine learning. My group mate is previously experienced in machine learning so it is a great chance for me to learn with him and decide how to construct the project. First, we have to decide how to open and read each file and parse the FASTA files to python dictionaries. Data given to us must be transformed to a data for machine learning. In order to do that I learned to Vectorise raw sequences into vectors to via Count Vectoriser. Validation sets are decided. Train/test set ratio is chosen as 0.8/0.2. Each result is stored under Result folder.

Me and my partner opened a github account to have a workflow on the project and track the changes we’ve done.

**The next steps must include:** Also adding SVM as an algorithm and implementing a code to visualize results.

**12-12-2017 – Understanding and implementing Support Vector Machine**

Since we are trying to classify signal peptides from other types of proteins Support Vector Machines are suitable because they are mostly used in classification problems.

**16-12-2017 – Running control tests on the script**

My project partner added several statistical tests to measuring the performances of the algorithms that is been used in the project. Today my aim is to understand the how script is constructed with precision and recalls and also run the script several times to see their differences. Also, the different results coming from F-test is observed. Matplotlib is used to make algorithm comparison.

**17-12-2017 – Starting the work on the report**

I started to write the introduction part for the report. I made literature search for the previous studies on predicting signal peptides.

Next Steps will be included write a code to run the script from command line. Preferably argparse will be used.

**22-12-2017 – Adding argparse module**

More commands are added and argparse module is implemented.

**23-12-2017 – Running prediction on chosen proteome data**

2 proteomes are downloaded from Ensembl’s BioMart service which are Drosophila Melanogaster and Mus Musculus and ran logistic regression on those data.

**03-01-2018 –Project Report**

I have checked the README file and gave feedback to my project partner while I added the Results and Discussion part to the report. I’ve run the script many times and tried to compare it with previous studies.

**12-01-2018 – Final changes**

Last discussion on the report and the code is done. Also Sequence logo is created for Drosophila melanogaster.