Lab Notebook

# Project summary:

Xxxx

## 10-12-2017 – Project initialisation

**Event:** Initial python coding for this project has been done. The initial code aims to parsing raw data (obtained from course Canvas) into python data types. The following step is performed on the raw data to transform into Machine learning ready data.

1. Open and read each data file
2. Parse Fasta format data into python dictionaries
3. Use Count vectoriser to vectorise raw sequence strings into vectors
4. Split into train/test set, initial split ration is 0.8/0.2

Using the train and test set, 2 Naïve Bayes models and logistic regression is trained.

The result is stored under /results/10-12-2017

**Errors:**

1. The fasta parser from Bio package does not filter out commenting lines, this results an error in the training data. The solution was build our own fasta reader.

**Lessons learned:**

1. Double checking the training data is important, invalid input data will produce incorrect results.

**What’s next?**

1. Separate and clean existing python into several modules
2. Improve documentation and comments
3. Write code to visualise outcomes
4. Write automated scripts that train data and generates results
5. Choose two proteomes and write scripts/code to predict on the proteomes