**GitHub Diary – Project in Molecular Life Science**

**15 February (Start-Introduction)**

Introduction to the project. Read all the documentation regarding Bash, Git and “How to organize your project” to become familiar with the tools to use and the structure of the project.

I wrote a bash script that creates all the different folders that are needed to store the information I will be collecting during the development of the project. For a first try, I created a big directory (*MTLSProject*) inside which I created five other repositories (*Projects, Datasets, Results, Scripts, Bin*). These can be modified later on, when a clearer structure of the project is perceived and more files need to be stored inside the different repositories.

Then, this was uploaded on GitHub inside a repository that was previously created. The link of this repository was sent to the TAs so they can check everything has been done correctly.

Bash script:

*mkdir MTLSProject*

*cd MTLSProject*

*mkdir Projects*

*touch Projects/readme.txt*

*touch Projects/commands.txt*

*mkdir Datasets*

*#touch Datasets/namefile #touch creates a file inside a directory*

*mkdir Results*

*mkdir Scripts*

*mkdir bin*

**16 February (Lab group meeting)**

I attended the lab group meeting at SciLife and listened to the presentations of the group members, who explained what they were exactly working on, as well as the most interesting and relevant scientific findings of their last week.

**19 February (Project help + software carpentry submission)**

Deadline for software carpentry tasks. The link to GitHub was already sent to TAs last week.

**20 February (Project introduction)**

Start of project. Mandatory parts to pass the project:

* *Extract the feature from your dataset*
* *Create cross-validated sets*
* *Train a SVM using single sequence information, using sklearn*
* *Check different window sizes for the inputs*
* *Analyze the results and compare it to previous work*
* *Review the state of art for your predictor*
* *Write a report*

The first step is to look at my dataset and understand what is consists of. It has 42 proteins, each of them with a given ID, its amino acid sequence and the corresponding topology. First of all, I have to read my dataset line by line and create a dictionary to store all the information it contains. The keys will be the IDs and the values will be one list. This list contains two elements: the amino acid sequence and the topology.

**21 February (Project help)**

Since the inputs from SVM must be numbers, it is necessary to convert the protein sequences and topology to a different format. This will then give us the output from which we obtain the predicted topology of the input sequence. In order to do this, the amino acid sequences are encoded into pure binary: every letter is represented by a vector of 20 elements, in which 19 of them are 0 and the other one is a 1. Depending on the position of the 1 in the vector, it will represent one of the 20 amino acids. For example:

["A"]=[1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0], ["C"]=[0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0] and so on.

The topologies also need to be encoded as numbers. In my dataset there are 3 possible topologies: i, M and o. In the new format: i = 0; M = 1; o = 2.

Once both the amino acid sequences and the topologies are transformed into numbers, they are stored in the original dictionary, so that now it has the keys with the sequence ID, and the values with a list containing the amino acid sequences and the topologies in binary format.

**23 February (Project plan + 5 most relevant papers + Journal club + Lab group meeting)**

I submitted a list of the 5 most relevant papers regarding β-barrel prediction that I found online. One of them will be selected for my individual presentation at the end of the course. Also, I also sent in one-page project plan describing how the realization of the project will look like and what steps will be followed to achieve the topology prediction of an input amino acid sequence.

I attended Journal Club at SciLife for a discussion about deep learning and the consequent explanation of the paper “*Visualizing and Understanding Convolutional Networks*” I had read beforehand. Then we had a lab group meeting where we listened to the new findings of the group members.

**26 February (Project help)**

The next step in the project is to build the sliding windows. This is done to take into consideration the amino acids that surround a central one, and consider the topology of such amino acid accounting for the influence of the surrounding ones. For the sliding windows of the first and last amino acids we have to add flanking vectors to the left and the right of the binary amino acid sequence, respectively; this flanking vector contains 20 zeros and the number of times it should be added depends on the size of the sliding window.

**28 February (Project help)**

In order to train the SVM, we need to input the sliding windows of the amino acid sequences and the corresponding topologies for each sliding window, which will be determined by the central amino acid of each window.

Once the model is trained, it is time to predict the topology of an input protein sequence. For this aim, the file testfile.txt needs to be read and its sequences converted into binary. Following the same process as in the main script, these input sequences undergo the sliding window process to give rise to the testsetaa, which will be the input to predict with clf.predict(). The topology is predicted and shown in the command window.

It is still necessary to create all the modules that will make the script look more compact and clear, but this will be done next week!

**1 March (Paper presentation with colleagues)**

We organized groups of 5 students and went over the each of the presentations. Then I wrote a short evaluation of each of the presentations and submitted it.

**2 March (Journal Club + Lab group meeting)**

I attended Journal Club at SciLife for a background presentation about deep transfer and how to make use of datasets when there is not enough data available. Then we listened to a presentation about face recognition using deep learning based on whales. Then we had a lab group meeting where we listened to the new findings of the group members.