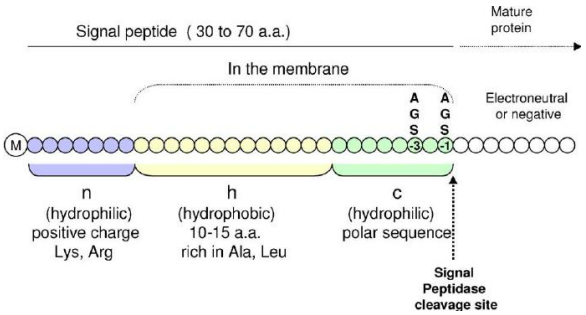
Project plan-Protein structure prediction: Globular and signal peptides

Carolina Savatier-Dupré Bañares, Scilife Master

The process through which a protein reaches its proper final destination, and thus performs its function, is fundamental within cellular life. For this, signals peptides exist. They are short (10 to 60 amino acids) specific sequences of amino acids in the N terminus of proteins that will enter the secretory pathway. The core of the signal peptide is hydrophobic and usually folds as an alpha-helix. It can be flanked by two shorter hydrophilic regions, n and c, referring to N and C terminus respectively.



Taken from Faye et al. (2004)

The main objective of this project is to create a program able to predict whether a given amino acid sequence is a signal peptide or is a globular protein. Following Anfinsen's idea that the amino acid sequence of a protein determines the 3D fold, I will create a predictor by homology, which will find the best match to a database of sequences with known 3D structure. The course lasts around 5 weeks, so I distributed the work load according to the information given in the webpage thereof.

Week 1: I prepared the bash script to create a template project folder, after reading the paper by Stafford Noble (2009), A quick guide to organize computational biology projects. I also went through a Linux tutorial to remember what we had learn in the previous course. Moreover, I got familiar with github and I uploaded there all my directories, folders and documents.

Week 2: During this week I have been working on the script that will parse my data and will give a matrix made of vectors, which is the input I need to perform svm (support vector machine), a supervised machine learning, with SKlearn. I have also been searching literature and I read some of the papers that are attached the course webpage.

Week 3: During next week I will start the SVM training using SKlearn and I will change the window size to see how that affects to the accuracy of the prediction. I will perform cross-

validation by creating different training and testing sets. I would also like to run psi-blast to improve my predictions with evolutionary information.

Week 4: by this time the predictor should give an expected output and it will be time to optimize as much as possible the program. Then I will analyze the results and compare them to previous work. Furthermore, during week 4 I will prepare an oral presentation of a paper and present it in front of a small group of my classmates.

Week 5: I will put the final touches to the predictor but this week will be more focused on writing the final report and then prepare the presentation for the next week.

Week 6: Oral presentations of my work.