MARINA’S DIARY

Friday, February 16th

- **Creation of a bash script that when run creates a template project folder structure.**

The specified file structure is the following:

```
projects/MTLS project
./scripts/ – directory for all the python/R/perl scripts
./bash/ – driver scripts that call all other scripts and execute pipelines
./bash/runall.sh – the main driver script
./bash/filter*sh – scripts that I use usually only once to filter input directories and create soft links
./input/ – input directories
./output/ – output directories
./logs – stdout and stderr of runall.sh scripts
readme.txt – description of files and scripts in this project folder
commands.txt – commands that I run in this directory (could be a mess)
```

```
datasets (directory for storing all datasets)
CASP9
CASP10
CASP11
Human reference genome
etc…
```

```
general_scripts (directory for storing scripts that are use in many projects)
pdb_scripts
templates
etc..
```

```
bin (all small scripts that I want to be always in PATH, remember to att this
to your PATH-variable)
fastalen
svm_to_txt
txt_to_svm
echo_both
etc…
```

**The blue entries are the ones that I specified in the bash script as examples. The rest of them will be added gradually when needed.

- **Signing up on github**

Also, I signed up on github. (E-mail used: marina.martinez.hernandez@stud.ki.se) and I also created a new repo on github. The previous bash script was uploaded to my github and sent to John.

Tuesday, February 20th

- **Literature searching and reading for my paper presentation**

Searching and reading of the 5 papers/protocols relevant to my project (beta barrel 2 state). The most relevant ones are:

- Prediction of Transmembrane Regions of β-Barrel Proteins Using ANN- and SVM-Based Methods. Navjyot K. Natt, Haroreet Kaur, and G.P.S. Raghava. Institute of Microbial Technology, Chandigarh, India. PROTEINS: Structure, Function, and Bioinformatics 56:11–18 (2004). Link: http://onlinelibrary.wiley.com/doi/10.1002/prot.20092/pdf

- PRED-TMBB2: improved topology prediction and detection of beta-barrel outer membrane proteins. Konstantinos D. Tsirigos, Arne Elofsson and Pantelis G.Bagos. Bioinformatics. 2016 Sep 1;32(17): i665-i671.Doi: 10.1093/bioinformatics/btw444. Link: https://www.ncbi.nlm.nih.gov/pubmed/27587687

- BOCTOPUS: improved topology prediction of transmembrane β barrel proteins. Sikander Hayat and Arne Elofsson. Structural bioinformatics. Vol. 28 no. 4 2012, pages 516–522 doi:10.1093/bioinformatics/btr710. Link: https://academic.oup.com/bioinformatics/article/28/4/516/213408

- Beta barrel trans-membrane proteins: Enhanced prediction using a Bayesian approach. Paul D. Taylor,Christopher P. Toseland, Teresa K. Attwood and Darren R. Flower. Bioinformation. 2006; 1(6): 231–233. Published online 2006 Oct 7. PMCID: PMC1891693. Link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1891693/

- Predicting Beta Barrel Transmembrane Proteins Using HMMs. Tsaousis G.N., Hamodrakas S.J., Bagos P.G. (2017). In: Westhead D., Vijayabaskar M. (eds) Hidden Markov Models pp 43-61. Methods in Molecular Biology, vol 1552. Humana Press, New York, NY. Link: https://link.springer.com/protocol/10.1007%2F978-1-4939-6753-7_4

Extra paper, although I prefer the above ones:
- Predicting transmembrane beta-barrels in proteomes. Henry R. Bigelow, Donald S. Petrey, Jinfeng Liu, Dariusz Przybylski and Burkhard Rost. Nucleic Acids Research, 2004, Vol. 32, No. 8. DOI: 10.1093/nar/gkh580. Link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC419468/pdf/gkh580.pdf

Papers sent to Arne by e-mail.

My project: beta barrel 2 state. 2 state means that my prediction should be able to distinguish between two different states: g (globular) or B (beta barrel).

Start programming in python for my project. Make a python file from the given txt. Start reading the lines (line by line) and understanding the content. First aim to start with the project: handle my dataset as a dictionary. The ID will be the key entries of the dictionary, and the values will correspond to a 2-element-list containing the aminoacid sequence (list[0]) and the topology (list[1]).

Wednesday, February 21st

- **Theoretical background research / analysis of the project's goal**

Main aim: build up a predictor (SVM to do so). Protein sequence as input → the trained predictor processes the input → topology is predicted.

How does a SVM work? It is supervised, used for classifying my sequence into g (globular) or B (beta). We need an input so that the SVM gives you an output: input (protein sequence) → SVM → output (topology). SVM only works with numbers. Therefore, the sequences that are used as inputs need to be transformed into binary identifiers. OneHotEnconder can be used to convert aminoacid residues into binary identifiers automatically.

Part of the input is the known topology because SVM is a supervised method, so the topology has also to be in numbers.

When designing the input vectors that we are going to introduce in the SVM, we must bear in mind that topology depends on the environment, not only on the one aminoacid residue to which the topology corresponds. Thus, the training set derived from the residues should contain more than one single aminoacid residue. The question that comes now is: which is the optimal length of this input sequence associated to the topology? In other words: how many aminoacid residues define the topology? We need to find out the answer to this question by trial-and-error, using different window sizes. Also, when trying out different sliding sizes we must assess different parameters: speed, performance, cross validation, etc.

Example of a window size of 5:
Sequence: ADERRM
Topology: BB gg B g
Sliding window of 5, third residue: AD-E-RR. Topology: g
Sliding window of 5, fourth residue: DE-R-RM: Topology: g

With this approach we will miss Window_Size/2 residues (the ones at the edges), but we still need them. A procedure that can be used to retain those flanking residues is to add something floating. Example: add virtual aminoacids that are only zeros at both ends. NOTE: Do not mix the flanking "fake" residues with the actual aminoacids. If sliding window is 5: 5/2=2,5 → add 2 aminoacids that are zeros at both the beginning and the ending of the protein (flanking). As seen, this depends on the window size (int(Window_Size/2)). Although the environment for the first and last residues is inexistent, but we still need to add something.

Further considerations for SVM. SVM expects a 2D array as an input, not a 3D. When doing the sliding windows, we should consume only 1dimension. We cannot do this [[..],[..],[,..]….], because otherwise we will end up with a 3D array that cannot be used as input. Instead we should use: […. ….. …..], which has a length of the (sliding window)*(number of 0,1 needed for each aminoacid) = 5*20=100. Thus, each sliding window is a 1D list of length 20*(window size). The SVM will complain if we do not do this.

After this, cross validation sets need to be done. I will worry about this once I have everything done as stated above.

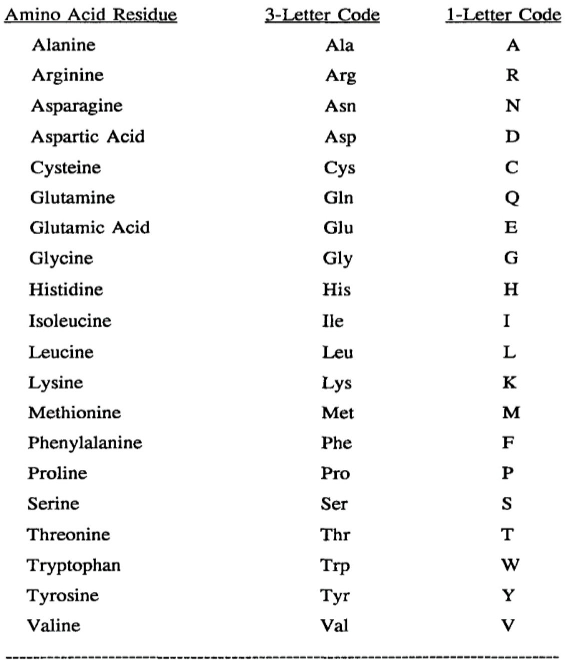
Summary: input format for the SVM. Sequences translated to binary, and then picked up by a sliding window of predefined size. This has to be in one dimension. Topology encoded as a number, not in binary as the output of a SVM has to be a number.

- **Interiorize all the background needed**
- **Analysis of OneHotEncoder's functioning**

Thursday, February 22nd

- **Translation of sequences into a binary code**

I cannot make OneHotEncoder properly work, so I created a dictionary with all the amino acid residues and their corresponding binary code. I specified them in the same order as in the following list:



Amino Acid Residue	3-Letter Code	1-Letter Code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic Acid	Asp	D
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamic Acid	Glu	E
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

- **Paper reading for the journal club: Visualizing and Understanding Convolutional Networks**

- **Writing down the project plan**

Planning of the timelines in the project. One-page project plan and submission to elofsson.arne.su@analys.urkund.se in a PDF format.

Friday, February 23rd

- **Journal club: Visualizing and Understanding Convolutional Networks**

10-11 am. Gamma Lunch Room, Scilifelab.

- **Elofsson Group meeting**

1:30-3 pm. Gamma Lunch Room, Scilifelab.

Monday, February 26th

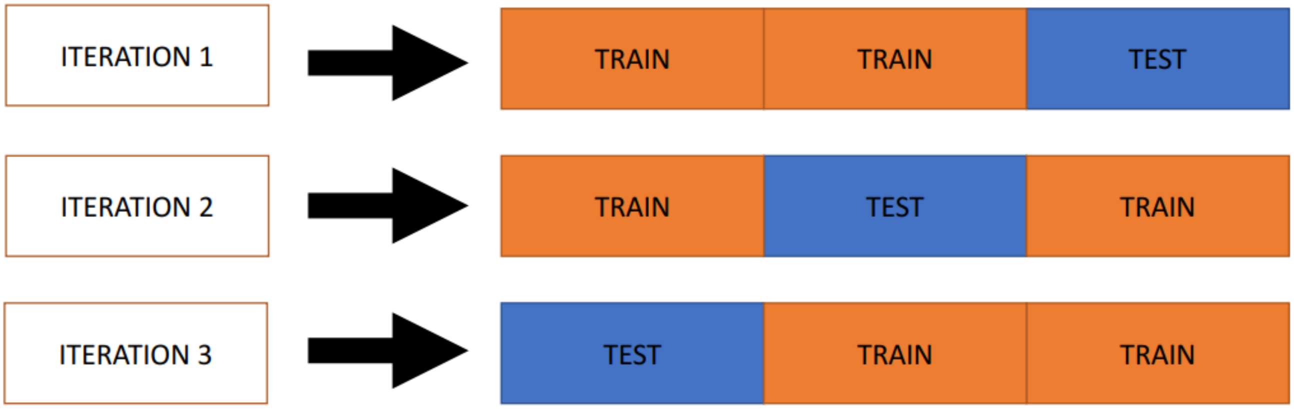
- **Create input to SKlearn**

Create two input vectors for the SVM. The first one is composed by segments of sequences defined/picked up by the sliding window of a specified size, all in binary and 1D. The second input is the vector with the corresponding features, specified in numbers (no letters).

Tuesday, February 27th

- **Create cross-validated sets**

With the input lists created yesterday, create 3 different sets for training and testing in the following manner:



ITERATION 1	TRAIN	TRAIN	TEST
ITERATION 2	TRAIN	TEST	TRAIN
ITERATION 3	TEST	TRAIN	TRAIN

Reading: https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6

- **Reading the paper for my presentation in depth**

Paper: BOCTOPUS: improved topology prediction of transmembrane β barrel proteins. Sikander Hayat and Arne Elofsson. Structural bioinformatics. Vol. 28 no. 4 2012, pages 516–522 doi:10.1093/bioinformatics/btr710.
Link: https://academic.oup.com/bioinformatics/article/28/4/516/213408

Wednesday, February 28th

- **Prepare paper presentation**

PowerPoint presentation, for an estimate time of 12 minutes. Paper: BOCTOPUS: improved topology prediction of transmembrane β barrel proteins. Sikander Hayat and Arne Elofsson. Structural bioinformatics. Vol. 28 no. 4 2012, pages 516–522 doi:10.1093/bioinformatics/btr710.
Link: https://academic.oup.com/bioinformatics/article/28/4/516/213408

Thursday, March 1st

- **Paper presentation in front of my group**

PowerPoint presentation, for an estimate time of 12 minutes. Paper: BOCTOPUS: improved topology prediction of transmembrane β barrel proteins. Sikander Hayat and Arne Elofsson. Structural bioinformatics. Vol. 28 no. 4 2012, pages 516–522 doi:10.1093/bioinformatics/btr710.
Link: https://academic.oup.com/bioinformatics/article/28/4/516/213408

- **Evaluation of the presentations**

Receive evaluation from the colleagues in my group and also evaluate their presentations.

- **Train a SVM using single sequence information, using sklearn and the input I created on Monday**

Upload file on github.

- **Blog reading for the journal club: Which whale is it, anyway? Face recognition for right whales using deep learning**

Friday, March 2nd

- **Journal club: Which whale is it, anyway? Face recognition for right whales using deep learning**

10-11 am. Gamma Lunch Room, Scilifelab.

- **Elofsson Group meeting**

1:30-3 pm. Gamma Lunch Room, Scilifelab.

- **Submit the evaluation of the presentations.**

Submit the evaluations that I did yesterday about the 4 presentations that my colleagues presented.

Monday, March 5th

- **Decode the predictor's output**

The output from the predictor was in numbers. Translation of the results (decoding) to the 'g' and 'B' topologies. There was no specific mention to this last week…

- **Optimization of the predictor**

Trying different sliding windows. Let it run overnight.

Tuesday, March 6th

- **Literature searching**

Fully understand the function of C (parameter for the soft margin cost function) and gamma (for non linear functions). Also, different kernels.

- **Optimization of the predictor**

Trying different kernel types and gamma values. Let it run overnight.

- **Enhancing my paper presentation**

Include the changes that my classmates proposed me last week when practicing the presentation.

Wednesday, March 7th

- **Optimization of the predictor**

Trying different kernel C values. Let it run overnight.

- **Paper presentation**

Practice paper presentation for tomorrow

Thursday, March 8th

- **Paper presentation**

Paper: BOCTOPUS: improved topology prediction of transmembrane β barrel proteins. Sikander Hayat and Arne Elofsson. Structural bioinformatics. Vol. 28 no. 4 2012, pages 516–522 doi:10.1093/bioinformatics/btr710.
Link: https://academic.oup.com/bioinformatics/article/28/4/516/213408

- **Paper reading for the journal club: DeepLoc: prediction of protein subcellular localization using deep learning**

Friday, March 9th

- **Journal club: DeepLoc: prediction of protein subcellular localization using deep learning**

10-11 am. Gamma Lunch Room, Scilifelab.

- **Elofsson Group meeting**

1:30-3 pm. Gamma Lunch Room, Scilifelab.

- **Self-evaluation of the presentation (yesterday).**

One page self-evaluation of your presentation including the written comments by your peers submitted to elofsson.arne.su@analys.urkund.se
Write it down and send it by mail.

Monday, March 12nd

- **PSI-BLAST using Uniref90**

Remember: format your database using formatdb and then specify the formatted database as input to blast.

Create bash script for running it with Uniref90. Save the outputs (align and PSSM). Finally, understand the structure of the PSSM file so that I know how to parse it into the input for a SVM (tomorrow's aim).

- **Random Forest**

Reading of Random Forest. Playing around with different parameters. Sliding windows: (3,36,2), n_estimators (100,400,50) and min_samples_split (2,11) and all the possible combinations. Let it run overnight.

Tuesday, March 13rd

- **Predictor training using multiple sequence information**

Create parsers, training and predicting scripts able to work with multiple sequence information (from the PSSMs obtained yesterday).

- **Random Forest**

Data collection. I have a lot of data (more than 200 pages of doc when the results have been pasted). I need to start analyzing the data ASAP, cause I am accumulating it.

- **Simple Decision Tree**

Reading of Simple Decision Tree. Playing around with different parameters. Sliding windows: (3,36,2), min_samples_split (2,11) and all the possible combinations. Let it run and collect the data.

- **State of the art reading**

Collecting info from previous work related to my project.

Wednesday, March 14th

- **Predictor using multiple sequence information**

Final predictor trained on multiple sequence information.

- **State of the art reading**

Collecting info from previous work related to my project

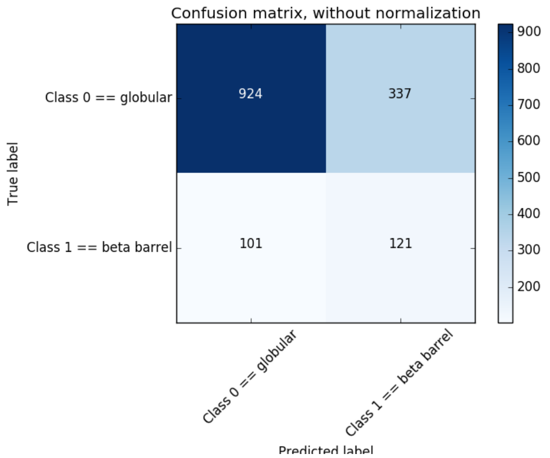
- **Extract the data from 50 other proteins**

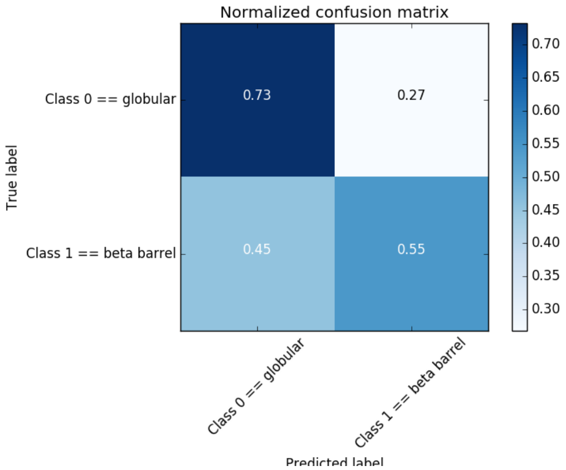
New extracted dataset of TMBs. Some parsing has been needed to match the data with the topologies that can be predicted by my predictor (2 states).

Thursday, March 15th

- **Test the performance of my predictors (both the trained on PSSM and on residue information).**

Performance tested. Example output (confusion matrix), among other parameters tested.





- **State of the art reading**

Collecting info from previous work related to my project

- **Data analysis**

Summarize and analyze all the data collected throughout this project.

Friday, March 16th

- **Journal club: Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model, Wang et al 2017.**

10-11 am. Gamma Lunch Room, Scilifelab. http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005324

- **Elofsson Group meeting**

1:30-3 pm. Gamma Lunch Room, Scilifelab.

- **Data analysis**

Summarize and analyze all the data collected throughout this project.

Saturday & Sunday, March 17th and 18th

- **Data analysis**

Summarize and analyze all the data collected throughout this project.

- **Report**

Writing the report