**To do**

* ~~Preprocessing of the data files~~
* Account for redundancy
* Building & training the model (ANN)
  + Grid searching
  + GANs
* Nested-CV
* SMM model
* Comparative plotting, sequence logo plotting
* +++++
* Report writing

**Project description**

In our project we are going to employ the full HLA-A\*0201 peptide data set with the purpose of implementing and fine tuning an ANN for the identification of MHC-I binding peptides and finally performing a comparative analysis against methodologies like SMM. In this regard, approaches like grid search and data redundancy techniques are going to be explored, while a nested cross validation will be defined for validating the performance of the models. +++++++

**GANs**

https://www.wikiwand.com/en/Generative\_adversarial\_network

<https://www.biorxiv.org/content/10.1101/730143v1.full>

<https://github.com/psipred/FFPredGAN/tree/master/src>

<http://bioinfadmin.cs.ucl.ac.uk/downloads/FFPredGAN/RealTrainingData/GO0000003_Real_Training_Negative.txt>

**3d modelling**

Create HLA models and superimposed averaged out peptide Gaussian-like surfaces based on the predictions that would be generated of the best performing model.

**Example**

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0010353>

**Use of validating tools**

…like NetMHCIIpan etc

<http://www.cbs.dtu.dk/services/NetMHCIIpan/>

**Proposal on workload allocation**

|  |  |
| --- | --- |
| **Marius** | Redundancy + seq logo plotting+ Report writing |
| **Hannah** | GANs + Report writing |
| **Frederik** | SMM + Comparative plotting+ Report writing |
| **Konstantinos** | ANN + Grid searching+ Report writing |
| **Thanos** | Preprocessing + Nested CV+3D modelling + Report writing |