Protein Domain Boundary prediction based on Deep Neural Networks

Kevin Bryson

2392886

# Project Outline

## Motivation

Predicting the structure of a protein from its amino acid sequence helps better understand the structure of a protein which provides a better understanding of a protein’s function. A novel method has been developed by Deepmind, called AlphaFold, but its task is to predict the entire 3D structure of a protein and does not include predicting the protein boundaries. It is argued that simpler and less time consuming methods can provide comparable results to novel approaches such as AlphaFold.

## Aims

This project will explore different Deep Neural Network architectures to predict protein domain boundaries from the amino acid sequence. The architectures include Fully Convolutional Deep Neural Networks and Recurrent Neural Networks (GRUs, LSTMs etc.) Afterwards, the predicted 3D structures provided by AlphaFold will be processed to extract the domain boundaries from that. Finally, the models’ performances will be compared to AlphaFold’s performance for this particular task.

# Progress

* Language and main libraries chosen:
  + Python
  + Sword2 for predicting AlphaPhold’s protein domains from its predicted structure.
  + Biopython for processing PDB files.
  + Pytorch for building and training the Neural Networks.
* A prototype pipeline for evaluating the performance of AlphaFold and of the models to be trained has been built.
* The architecture of the models to be built has been discussed and agreed on: fully Convolutional Deep Neural Networks and Recurrent Neural Networks (GRUs, LSTMs etc.)
* The input data and output format has been discussed and agreed on:
  + The input data will be one-hot encodings of length *n* of the amino acids.
  + The output data will be a vector of length *n* and which will be the probability of an amino acid being the boundary of a domain.
* Some future work has also been discussed: Utilising contact maps in order to also predict the protein domains for both continuous and discontinuous proteins.

# Problems and Risks

## Problems

* AlphaFold does not directly predict the domain boundaries. Instead, SWORD2 will be utilised to predict the domain each amino acid belongs to and that will be further processed to mark the domain boundaries. Taking this into consideration, the ground truth (i.e. the PDB files from RCSB) will also be fed to SWORD2 in order to find the best score that can be achieved.

## Risks

* The split between one, two, three or more domain boundaries is unbalanced. Most Proteins have less than 3 domains, therefore, there may be some bias during training. To mitigate this, an even split of the proteins can be chosen for training but that may leave us with a small number of proteins to train on.
* Predicting the domains is much harder than predicting the boundaries. That is because proteins can also have discontinuous domains. Therefore, the boundaries will be predicted instead.

# Plan

## Semester 2

* Week 1-2: Finish preprocessing and preparing data.
  + Deliverable: Have data ready to be fed into the models for training. This includes picking out a good ratio of proteins with one, two, three or more domains, extracting the sequence, performing one hot encoding and also converting AlphaFold’s predictions into predicted boundaries.
* Week 3-4: Develop initial models and Finish the evaluation pipeline.
  + Deliverable: Implement the DNNs to be trained using Pytorch.
  + Deliverable: Finish and fully automate the prototype that has already been built that can take the predictions in and display the results from 1) the trained models, AlphaFold and the ground truth (PDB files).
* Week 5-6: Fine tune models.
  + Deliverable: Increase the accuracy of the models as much as possible by fine tuning.
  + This period also leaves space for potential setbacks.
* Week 7-8: Finish evaluation.
  + Deliverable: Evaluate the models and AlphaFold’s performance and have results ready to present in the dissertation.
* Week 9-10:
  + Deliverable: first draft submitted to supervisor two weeks before final deadline.