Bioinformatics Project Proposal

Najib Ishaq

Project Name

CASP (Critical Assessment of protein Structure Prediction)

This is a biennial competition in Bioinformatics where teams are tasked with predicting the structure of proteins given the amino-acid sequences.

Outside Help

- Course professors for what they know about protein folding.
- My friend Tom Howard, should he have time, because we work very well together.
- Strangers on the internet when I inevitably post on online forums asking questions about various snags I hit.

Team Members

- Me.
- Just me.

Team Leader

- I am the leader by default.

Schedule

- One of the (occasional) downsides of being me is that I am forever saddled with my own company. Sometimes I amuse myself and at other times I frustrate myself. I will be in the lounge on floor 1 of Tyler on Wednesdays from 2pm to 5pm.
- Please note that I too may drop in on any of you to ask questions about relevant topics, talk about life, and possibly share dark chocolate.

Project Plan

I found a paper that is currently in pre-publication. A researcher at Harvard, Mohammad AlQureishi, recently created an end-to-end differentiable model to predict protein tertiary structure. This model takes an amino acid sequence and

a position-specific-scoring-matrix and predicts the tertiary structure. I will start by replicating this model. I plan on improving upon this by using the secondary structure to help guide predictions. I will use the potential energy of the tertiary structure (using ROSETTA) as a scoring metric. I have a suspicion that I can make the model more robust by leaving out the PSSM from the input data. This is what I expect to get done as a bare minimum.

After this, I plan to look for partial structures that are predicted especially well. These can be indicative of templates that the model has learned. If I find any such templates then I will extract them and use them to guide further refinement of the predictions. I can probably get this done by creating a vector of "edit-costs" that is used to penalize editing known templates. I am interested in seeing if this pans out.

I plan on training my model(s) with time reset data i.e. use all data available up to 2012 to predict structures for proteins found after 2012. I already have five relevant data-sets for this purpose. I think the model will perform better when it is trained with more data. I want to find out if I can use my model to augment training data but that might be too much to ask.

I will be extremely happy if I manage to get close to, or surpass, the performance of Google's AlphaFold at CASP13. If I get promising results, I will turn this into a larger research project and perhaps make it into a publication.

Resources Needed

- I will read through several papers in the field. I have already started looked at the paper DeepMind released on AlphaZero. They are yet to publish their paper on AlphaFold. I have already read the paper by Mohammad AlQureishi.
- I have my own GPU to run some serious computations but will likely need access to some better machines on campus. I already have accounts on some servers at URI and will use them when I have a model ready to use the additional computational power.
- I am happy to make use of any other resources that you think will be useful to me.