Requirements Document

Najib Ishaq

Introduction

Protein Tertiary Structure Prediction is an open problem. Proteins dictate every biological function in our bodies. Understanding protein structure will provide an incredible understanding of how the proteins work, allowing us to modify its functionality. This can give us an unprecedented level of control over our bodies, allowing us to neutralize many diseases and improve our own bodies to the point of bringing to life the dreams of many science-fiction writers. I aim to build an end-to-end differentiable deep-learning model to predict the tertiary structure, and incidentally the secondary structure, of a protein given the primary structure and a PSSM from PSIBLAST. I build upon the current state-of-the-art models in terms of model architecture and techniques.

Milestones and Schedule

Division of labor is a non-problem for a team of one. My tentative schedule for milestones is as follows:

- Acquire training data: (Done)
  - I am using time-reset data as was available for the biennial iterations of the CASP competition. My dataset includes the solved structures of some 87,000 proteins. I have the primary, secondary, and tertiary structures along with the PSSMs for each of these proteins.
- Clean training data: (Apr 8)
  - The secondary structure data comes from a different source from the other pieces of data. Some os it is mis-aligned. I have a working implementation of the Needleman-Wunsch algorithm and am using it to merge my data-sets. The global sequence alignment is somewhat slow so this is taking some time.
- Create the model architecture: (Done)
  - The model consists of three parts. Each will be trained separately at first, and then combined into one pipeline. The details follow later in this document.
- Train the model: (Early May)
  - Once the cleaning and pre-processing steps are completed, I will implement the model in Keras and TensorFlow.
  - I have my own GPU to use for initial training. I will run a tentative search over the hyper-parameters for my model and will then offload longer training session to the servers at URI.

  - I expect the training and fine-tuning to take up quite some computational time.
- Compare against state-of-the-art prediction methods: (Mid May)
  - Predictions from a trained deep-learning models only take milliseconds to produce. I will use the time-reset data that I have to benchmark the prediction metrics for my model against the leading models that are publicly available. I am most interested in how my work compares to that done by Google's DeepMind for CASP '18. They have not yet published their paper or model and so I wait in eagerness.
- Publishing possibilities: (??)
  - If my model performs well, which it should, I will look to expand on the work I am doing for this course and get it published.

Front-End and User Interface

I plan on packaging my work as modules and creating a public GitHub Repository. Users will be able to use import statements in python and remain agnostic to the inner workings of my modules. The user will have to use the primary structure and a PSSM, perhaps from PSIBLAST, as inputs. The model will then predict the secondary and tertiary structures and return them as outputs.

I do not plan on writing PDB files. That format is outdated and I am a firm believer in that we need to improve to a modern and intuitively readable format for storing protein structures. I hope that one of the externalities of my work will be to help move the field in this direction.

Model Architecture and Specifications

The three parts of my model are very similar to each other. They take their inputs and use convolutional layers for feature extraction. The features are fed into LSTM layers that build an internal representation of the predicted structure. These representations are then extracted and converted into the structures that are compared to known structures during training.

The first part takes the primary structure and the PSSM as inputs and produces the secondary and tertiary structures as outputs. The second part takes the primary structure, PSSM, and the secondary structure to produce the tertiary structure. The third part takes the primary structure, PSSM, and the tertiary structure as inputs to produce the secondary structure.

I will I will train each of these three parts separately, using supervised learning, with the training data that I have. I will then use the first part to seed the inputs the second and third parts. I will then wrap them up in a reinforcement learning agent to feed the predictions back into the individual parts to continuously

improve the predictions. I will use MATT, and perhaps other tools, to score the prediction quality and guide the agent.

The final result will be a single pipeline that takes the primary structure and the PSSM as inputs and predicts the secondary and tertiary structure as outputs.

Future Work and Improvements

I am using time-reset data wo simulate what future improvements to my work might look like. I expect the model to perform better when it has more training data to work with. I look forward to testing this once training begins. Any trend I might find is likely to continue as more structures are solved using traditional methods line NRM and X-ray Crystallography.

I expect computational power to continue to increase according to Moore's Law. Better hardware in the future can be used to further train the model. Theoretical improvements in the field of Deep Learning are impossible for me to predict. I expect future generations of researchers to be able to build upon my work, as I am building now, and continue to make strides in this field. *Nanos gigantum humeris insidentes.*

My model predicts the structure of the backbone of the protein. Adding the residues and side-chains to a given backbone is essentially a solved task. I, but preferably someone else, will have to add this piece to the model. This can be treated as a post-processing step but will likely require some drudgery that I am not looking forward to.

Glossary

- **Protein:** any of a class of nitrogenous organic compounds that consist of large molecules composed of one or more long chains of amino acids and are an essential part of all living organisms, especially as structural components of body tissues such as muscle, hair, collagen, etc., and as enzymes and antibodies.
- **Amino Acid:** A relatively simple organic compound. There are twenty different amino acids present in nature. For my purposes, each amino acid consists of a C-alpha atom, a C atom and a N atom. These three atoms constitute a section of the backbone of the protein. Each amino acid also has a side-chain. The characteristics of these side-chains account for the differences between the twenty amino-acids.
- **Primary Structure:** the amino acid sequence of a protein. The analogy of beads on a string is helpful.
- **Secondary Structure:** the local structure of the atoms. This is broken down into alpha-helices, beta-sheets, and 'other'.

- **Tertiary Structure:** The full three-dimensional arrangement of the major atoms in the protein.
- **PSSM:** Position Specific Scoring Matrix. These are constructed using multiple sequence alignments in PSIBLAST. Each column in the matrix corresponds to an amino acid in the protein and represents the probability that the amino acid will 'morph' into another amino acid.
- **Protein Backbone:** The sequence of the C-alpha, C, and N atoms from each amino acid. The orientation of this backbone essentially defines the tertiary structure.
- **CASP:** Critical Assessment of protein Structure Prediction. This is a biennial competition that is used to stimulate work in this field and to compare the work done by various research groups. The last iteration of this was in 2018 and the next iteration will be in 2020.
- **Time-Reset Data:** The solved protein structures as they were available for each iteration of CASP. I am starting with data-sets from 2008 and onwards because that was when several aspects of CASP and its surrounding data were standardized.
- **BLAST:** A publicly method used to search for protein sequences that are similar to query sequences.
- **PSIBLAST:** BLAST with PSSMs.
- **MATT:** Multiple Alignment with Translations and Twists. This is like BLAST except that is aligns the tertiary structure instead of the primary structure.
- **Deep Learning:** A subset of Machine Learning in which Neural Networks with several layers (creating the depth) are used to approximate the functions that produce the required outputs from given inputs.
- **Machine Learning:** A subset of Artificial Intelligence in which the inputs and corresponding outputs to some real-life function are known and models are created that learn approximate that function from repeated examples of input-output pairs.
- **End-to-end Differentiable Model:** A model that requires no human intervention for the intermediate steps in the prediction process.
- **Supervised Learning:** Machine Learning with known input-output pairs as training data.
- **Reinforcement Learning:** Machine Learning methods in which a 'world' is created and an Agent is trained to act in this world by using some scoring function that depends on the state of the world and that of the agent.
- **GitHub:** And online version of the version control system known as git. This is highly useful to keep records of work done on computer science projects.
- **GitHub Repository:** An online location where all of the files and meta-information for a project is stored.
- **Moore's Law:** the principle that the speed and capability of computers can be expected to double every two years, as a result of increases in the number of transistors a microchip can contain.