MCB 198: Final Project

Submitted by: Ilia Gelfat

Background and Problem Statement

Proteins are responsible for a broad range of biological functions, including catalysis of biochemical reactions, structural support, cell signaling, transport of biomolecules and many more. In principle, the amino acid sequence of a protein contains all the information required to determine its function, three-dimensional structure and affinity to various molecules. However, in practice, deducing protein properties from its sequence remains an incredibly difficult computational challenge.

One common approach to predict protein properties is utilizing sequence homology. By comparing the amino acid sequence of an unknown protein to an existing database, function and structure of various domains within the protein can be inferred. While this strategy can be extremely useful in many cases, it cannot identify or predict the chemical or functional properties of novel or unique protein domains. Another approach involves molecular dynamics simulations. This method can make predictions for arbitrary amino acid sequences, regardless of their origin, while taking into account the rich chemical information found within each amino acid. However, despite notable advances in the field[(1)], such simulations are extremely demanding in terms of computational resources, time and expertise. In this project, I propose and develop a new computational approach for predicting protein function from its sequence. More specifically, I will focus on determining binding of the protein to various small molecule ligands.

Rationale and Approach

Protein interactions with small molecules can provide valuable information regarding overall function. Armed with such knowledge about an unfamiliar protein, one could infer possible substrates for enzymatic reactions, identify potential drug targets based on interactions with known drug molecules, and predict additional functions based on binding to ATP, nucleic acids, metal ions etc.

The approach proposed here involves training a neural network on an existing database of protein-ligand interactions. Upon training, the network will learn to classify a sequence of amino acids into several categories – each associated with binding a specific small-molecule ligand.

Importantly, this approach aims to take into consideration some of the chemical information found in the protein sequence, while still representing it in a compact way to facilitate computation. Therefore, each amino acid will be represented by three of its chemical properties: hydrophobicity, charge and molecular weight. These three properties are important in understanding protein structure: hydrophobic residues tend to be hidden within the protein core, while hydrophilic ones are more likely to be exposed to the solvent; charge can dictate the repulsion or attraction between residues; large, bulky amino acids can cause steric hinderance, while those will lower molecular weight confer more flexibility. They are far from the only relevant properties, but should nevertheless be a reasonable starting point for the implementation of the proposed computational strategy.

Implementation

Database and Ligand Selection

Training data was obtained from the BioLiP database[2]. This database contains a semi-curated listing of proteins, categorized by ligand binding. From the wide variety of ligands found in the database, three were selected: heme, glucose and nicotinamide adenine dinucleotide (NAD). These molecules were chosen for three main reasons. First, they are quite chemically distinct, making it reasonable to assume that there would be detectable differences between the proteins that interact with them. Second, they are well-defined, individual substrates, as opposed to a broad class of biomolecules (e.g. peptides or nucleic acids). Although the database contained a sizeable listing of nucleic acid binding proteins, the interaction would generally depend on the specific nucleic acid sequence, thereby potentially adding a confounding factor into the analysis. Lastly, the database contained a relatively large set of proteins binding to each of these three ligands. After removing redundant sequences, as well as sequences over 1000 amino acids in length, the total number of proteins added up to 1765 (733 heme, 443 glucose, 589 NAD).

Representation of Chemical Properties

As mentioned above, three chemical characteristics were chosen to represent each of the twenty amino acids: hydrophobicity, charge and molecular weight. There are several scales in use to quantify hydrophobicity. The one chosen here is the Kyte-Doolittle hydrophobicity scale[3], which ranges from a minimal hydrophobicity of -4.5 (arginine) up to the most hydrophobic value of 4.5 (isoleucine).

To represent the variation in charge, it is important to remember that the charge of an amino acid depends on the pH of its environment. For this reason, this property can be best captured by the isoelectric point (pI) – the pH value at which the molecule is neutral. For the twenty amino acids, the pI ranges from 2.87 (aspartic acid) to 10.76 (arginine)[4]. Lastly, the molecular weight of each amino acid

can be easily derived from its chemical composition, ranging from 75.1 Da (glycine) to 204.2 Da (tryptophan)[5]. To prevent any effect of the different scales on the overall result, all three properties were normalized such that they span the range [-1, 1]. The distribution of amino acids within this range is illustrated in Figure 1.

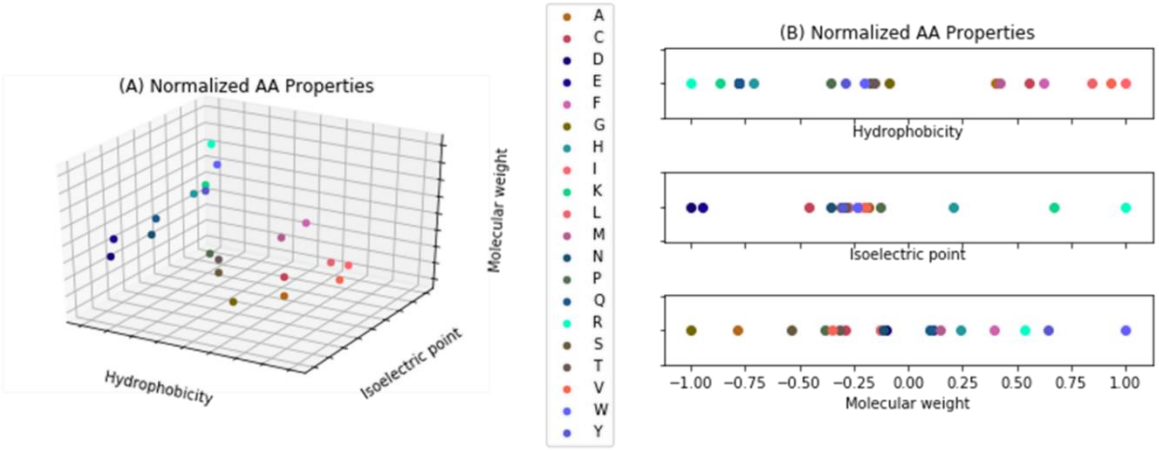

Figure 1: Normalized amino acid properties. (A) Displayed in three-dimensional space. (B) Each property is displayed individually.

It should be noted that these properties are not entirely independent. For example, a charged molecule – whether positive or negative – would be less hydrophobic than an uncharged one. Nevertheless, the correlations do not appear to be strong or straightforward enough to make any of the variables redundant. This can be seen qualitatively in both Figure 1 and Figure 2.

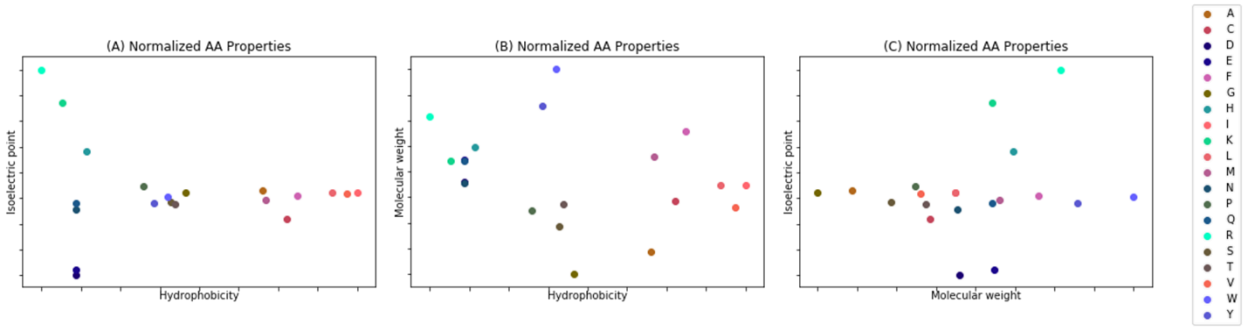

Figure 2: Normalized amino acid properties, projected onto the following planes: (A) hydrophobicity vs. isoelectric point, (B) hydrophobicity vs. molecular weight, (C) molecular weight vs. isoelectric point

Neural Network Implementation

Two machine learning approaches were implemented and compared. The first is a "conventional" fully-connected feed-forward network. This would serve as a naïve first attempt and a benchmark for comparison. The second is a one-dimensional convolutional neural network (CNN).

This latter approach is motivated by the anticipated structure within the data. Proteins often contain motifs – short sequences of about 10 amino acids, which act together to serve a functional or structural purpose. Generally, it is not initially obvious where along the sequence a relevant motif might appear. Since convolutional networks are translationally invariant, they should be well-suited to detect such features regardless of their position. Convolutional neural networks are not often used from protein sequence analysis and are best known for their applications in computer vision. In this context, it may be useful to draw an analogy to common CNN applications: the three chemical properties of each amino acid can be thought of as "channels," analogous to the RGB pixel values of an image. To use another example (which is perhaps more fitting to this one-dimensional case), the three channels are analogous to the x, y and z components of an accelerometer signal over time.

Proteins vary in length and complexity, and this dataset is no exception. However, the neural networks used here require a fixed input size. To overcome this technical issue, the sequence length was chosen to be 1000, and all proteins of shorter length were "padded" with zeros to fit this input size. In principle, if any of the protein groups have a distinct size compared to the others (and therefore would have more zeros at the end), this could be detected by the neural network in addition to any chemical information. Here, though, this is not expected to be the case, as all three categories have broad length distributions. The mean and standard deviation were calculated and are displayed below:

Heme: 305 ± 180 aa; glucose: 475 ± 210 aa; NAD: 367 ± 117 aa.

Hyperparameter selection was carried out empirically, as is common practice in current machine learning applications. However, a few aspects of the network structure were rationally chosen, motivated by the nature of the problem at hand. As mentioned earlier, a reasonable estimate for the length of a motif is 10 amino acids[6]. Therefore, 10 was selected as the width of the convolutional window of the first hidden layer. As for the number of motifs one can expect to find within a group of proteins – this would be difficult to estimate *a prioi*. Some estimates suggest approximately 100,000 motifs exist overall[6], which would require a large number of features to be detected computationally. However, it is reasonable to assume many of these motifs would be either unique or irrelevant to ligand binding, and therefore using a smaller number of features can be justified. Ultimately, the number of convolutional filters in the first hidden layer of the CNN was chosen to be 1000.

For the CNN, the network structure was chosen to contain 2 hidden layers, both of which were one-dimensional convolutional layers. The first was comprised of 1000 features with a convolutional filter of length 10. The second contained 100 features with a convolutional window of length 3, which was then fully connected to the 3-neuron output layer. This network contained 331,403 parameters.

The structure of the conventional feed-forward network was chosen to be similar to that of the CNN, so as to make the two comparable. It was therefore also comprised of 2 hidden layers, with the first containing 300 neurons, and the second – 50. This network contained 345,503 parameters.

Results

The data was shuffled and split, with 20% of the sequences going towards validation, and the remaining 80% used for training the models. The two models were trained over 20 epochs and evaluated by accuracy and loss (i.e. categorical cross-entropy). The results are summarized below in Figure 3:

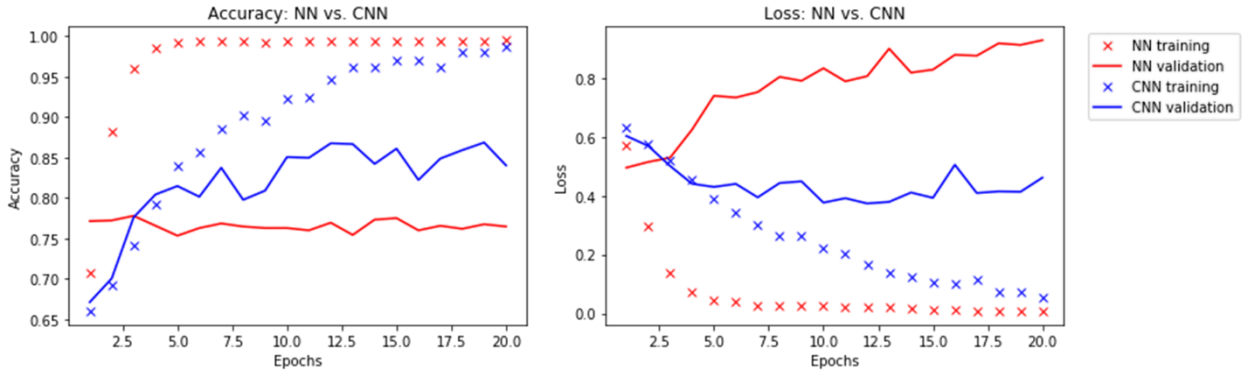

Figure 3: Comparison of neural network performance over 20 epochs – conventional (NN) vs. convolutional neural network (CNN). Left: accuracy, right: loss

These results suggest that the CNN outperforms the conventional approach. On the training data, both networks ultimately reach near perfect accuracy (~99%) and low loss (< 0.1). The difference becomes apparent when examining the validation data. While the conventional network plateaus around 75% accuracy, the CNN manages to surpass it and reach approximately 85%. A similar trend is present in the validation loss curves.

The discrepancy between the training and validation outcomes is likely the result of overfitting. Numerous attempts were made to minimize the extent of overfitting by adjusting the network hyperparameters, while maintaining relative structural simplicity, though it is possible that further optimization could lead to even better results.

Discussion

The results described here demonstrate the potential of a more biologically informed approach to protein sequence analysis. With additional development and a larger dataset, this method could be expanded to include a broader range of ligands, including drug molecules. Furthermore, a similar approach could be applied to other structural and functional features of proteins, such as prediction of three-dimensional folding, sequence specificity of DNA and RNA binding proteins, enzymatic reaction rates and many more. The method could be further fine-tuned to specific applications by the selection of additional relevant amino acid properties (e.g. Ramachandran angles or energies associated with hydrogen bond formation). Despite the many opportunities it presents, this method is not without its limitations. In particular, the lack of a systematic way to optimize hyperparameters and network structures renders it susceptible to suboptimal performance. It is unclear, for instance, how many convolutional layers are required to reliably detect interactions between domains or residues that are not adjacent along the protein sequence. Hopefully, as the field evolves, these challenges would become more tractable.

References

(1) Huang, Po-Ssu, Scott E. Boyken, and David Baker. "The coming of age of de novo protein design." *Nature* 537.7620 (2016): 320.

(2) Yang, Jianyi, Ambrish Roy, and Yang Zhang. "BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions." *Nucleic acids research* 41.D1 (2012): D1096-D1103.

(3) Kyte, Jack, and Russell F. Doolittle. "A simple method for displaying the hydropathic character of a protein." *Journal of molecular biology* 157.1 (1982): 105-132.

(4) Chemistry LibreTexts. *Soult, Allison*. "Chem 103 – Chemistry for Allied Health." *University of Kentucky* (2019): 13.1. [online] Available at: https://chem.libretexts.org/Courses/University_of_Kentucky/UK%3A_CHE_103_-_Chemistry_for_Allied_Health_(Soult)/Chapters/Chapter_13%3A_Amino_Acids_and_Proteins/13.1%3A_Amino_Acids [Accessed 15 Apr. 2019].

(5) ThermoFisher Scientific. [online] Available at: https://www.thermofisher.com/us/en/home/references/ambion-tech-support/rna-tools-and-calculators/proteins-and-amino-acids.html [Accessed 15 Apr. 2019].

(6) Tompa, Peter, et al. "A million peptide motifs for the molecular biologist." *Molecular cell* 55.2 (2014): 161-169.