Protein Sequence Analysis

The objective of this project was to design and train a Bidirectional LSTM model on a dataset of human antibody sequences. Generally, the journey from developing an antibody to getting it approved is really long and takes multiple iterations of human trials. The aim was to develop a metric that tells us how likely it is that an antibody will be accepted by a human. For training the model above,we used 900 sequences of antibodies that were collected from a research study at Boston University. These were all real antibody sequences. Out of these 900 sequences, 30 sequences were present in clinically approved drugs.We reviewed the literature and came across a paper that previously used SASA as a metric. The challenge was to create a model that outperformed the model in this paper.

The approach I used was to create a Bidirectional LSTM model training on all 900 sequences. The aim was to be able to generalize this model across a wider set of antibodies in the Observed Antibody Space (OAS) repository. Based on past research, we decided to compute a metric called a solvent accessible surface area (SASA) which measures the solubility of a protein in the human body. For doing this, we used multiple approaches. The first approach was to use every residue of the Antibody. The more practical approach was to use only residues within the CDR regions. A CDR region (Complementarity-determining regions) is the region of the antibody that is part of the variable chain. A challenge that we faced was to find a way to realign the residues so they could be rightly grouped into the CDR regions. Another challenge was to deal with the missing residues within a CDR region. After all the preprocessing, I trained a model on both the heavy and light chains, only on the light chains and finally only on the heavy chains. After training the models, I evaluate the SASA scores for the CDR regions using all three models and compare them to the theoretical values computed from the physical crystal structures of the antibody.

The model I created did have a better performance than the model published in the paper on 4 of the 6 CDR regions. The L3 and H3 CDR regions being the only regions where the performance of our model was worse than the one in the paper. THe L3 and H3 regions are also the most variable CDR regions and hence much harder to train on. A Data Science challenge we faced was to be able to generalize the model. Given only 900 antibody sequences, it is very challenging to be able to generalize the entire observed antibody space.