# Self Directed Programming Projects
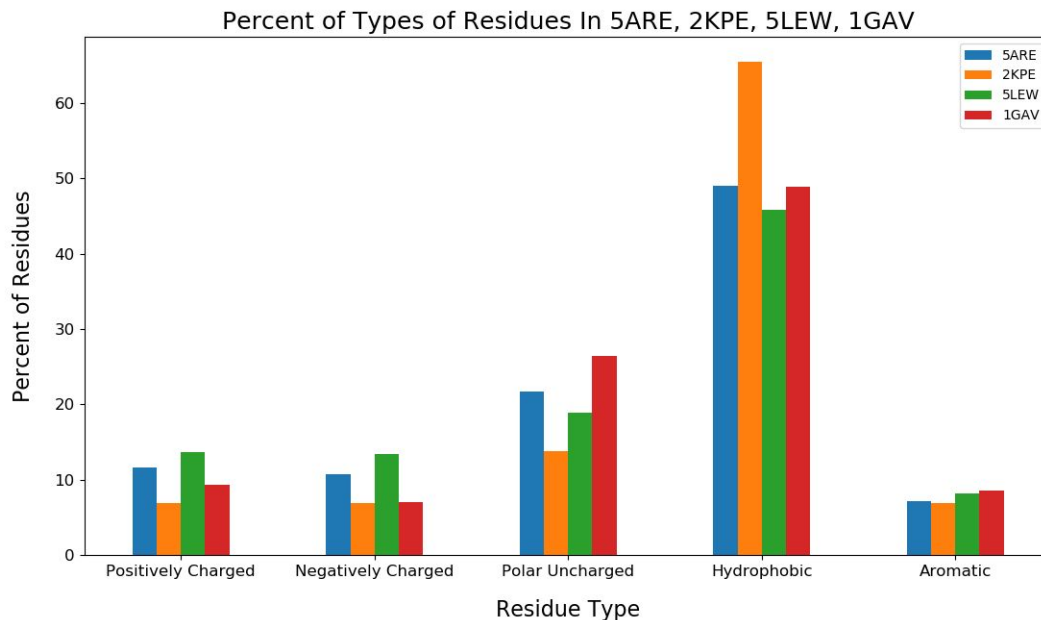
Project 1 : Protein Comparison Project

Description: Program allows a user to either graph: the percent of atoms, percent of each amino acid or percent of each type of amino acid based on chemical properties (e.g. hydrophobic) of several different proteins. The number of proteins that can be graphed is based on the type of graphing done whether it be atoms or amino acids. Proteins are fetched from the protein data bank using the Python library prody. The proteins are fetched by their four letter codes. Graphs are done using matplotlib, examples of the graphs produced are shown below and in the examples folder and code is included. Another program also allows the user to compare these different characteristics of the proteins but for the sum of all the proteins in the whole database. So the user can see, for example, what the average percent of negatively charged amino acids is for all the proteins in the database. This is under the "FullStats" folder.

Example 1: Graph of the percent of each atom (carbon, hydrogen, oxygen, sulfur, and zinc) in four different proteins (5ARE, 2KPE, 5LEW, 1GAV). So for example we can see that the protein (5ARE) has a higher percentage of carbon atoms than the protein 2KPE.
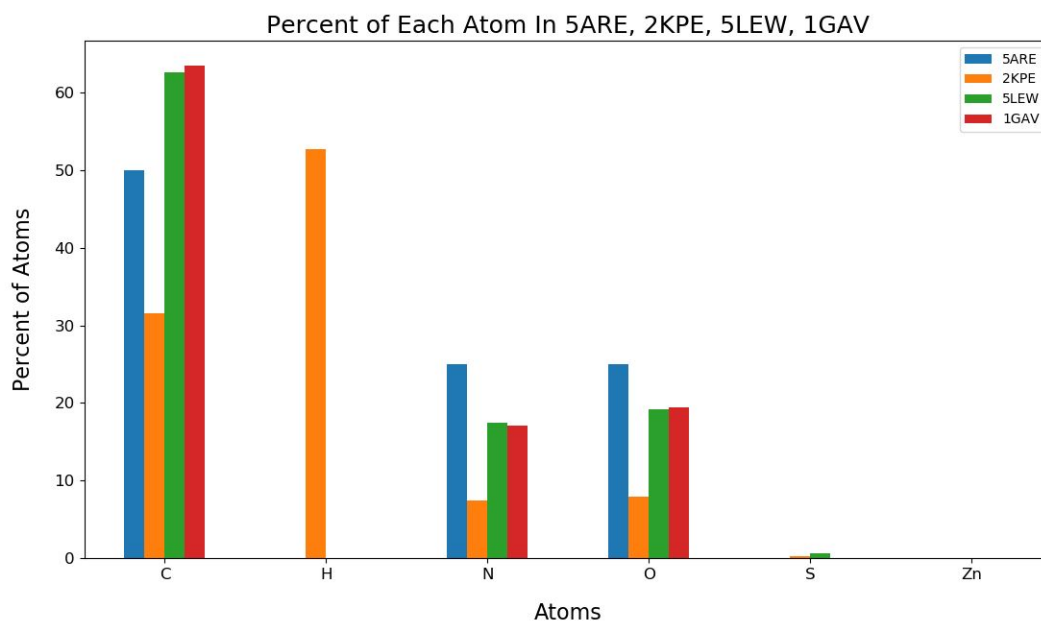
Example 2: Graph of the percent of different types of amino acids (positively charged, negatively charged, polar uncharged, hydrophobic, aromatic) in four different proteins (5ARE, 2KPE, 5LEW, 1GAV). So for example we can see that the protein (5LEW) has a higher percentage of negatively charged amino acids than the protein 2KPE.


Project 2 : Soccer Teams Comparison Project

Description: Allows a user to graph different statistics of multiple soccer teams over time. So we could specify a time range, a statistic and then pick soccer teams from almost any top league and then it would output that graph with time as x-axis and the statistic as y-axis. For example we could graph: Goals scored per season from 2004-2016 for Barcelona, Real Madrid and Sevilla. In this way a user is able to actually graphically compare statistics from multiple different teams instead of just seeing statistics of one team on a website. And so one can easily investigate questions like "are the top spanish league teams winning more games per season than the top english league teams." The program grabs and properly organizes this data from the website: fbref.com, and utilizes some methods to prevent from being detected and banned from the website.

Project 3 : Machine Learning Protein Project

Description: Created a program that trains a machine learning model to recognize from a protein image, whether it is globular (spherical in shape) or fibrous (elongated in shape). So the end result of the program is a model that can be given images of proteins and guess whether the protein is globular or fibrous. A Python program used commands from pymol (molecular visualization application) to automate a download of images of different globular and fibrous proteins from the 3d protein database. These images were for the model to be trained on and validate its training. Another Python program used these images to train a machine learning model to recognize images of proteins and identify whether globular or fibrous. This model was then saved to a file. The model was able to reach an accuracy of 100% in some runs but the data size was small and not so diverse. I used the PyTorch library for the machine learning part.

The following image is an example of a fibrous protein that was used in the program for validation:
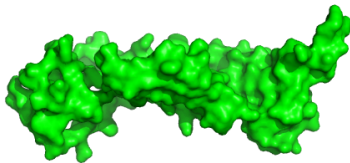
The following image is an example of a globular protein that was used in the program for validation:
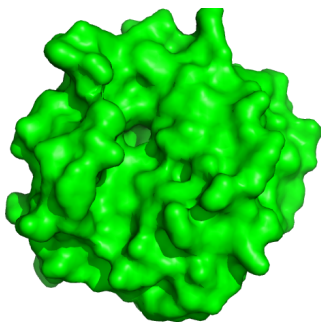
Notes

Project 1 and 2 were completely self-directed. Currently project 2 is not working because of changes in the websites format. Project 3 was with the help of a mentor.