Project 1

Jerry Lucas, Ryan Thimons

Introduction

In this project our goal is to discover information about different secondary structures and amino acids. Particularly we are interested in finding out the preferences of proteins, secondary structures, and amino acids. To do so we will use a dataset of several protein chains contained at <http://dunbrack.fccc.edu/Guoli/pisces_download.php>. The set that was used was cullpdb\_pc30\_res3.0\_R1.0\_d191017\_chains18877.gz however it appears that these sets are updated quite frequently so in order to repeat this experiment one should use a similar dataset such as cullpdb\_pc30\_res3.0\_R1.0\_dXXXXXX\_chainsXXXXX.gz.

Building & Cleaning the Data

The format of this dataset is in the following format one per line.

IDs length Exptl. Resolution R-factor FreeRvalue

12ASA 330 XRAY 2.200 0.16 0.29

The field we are using to build our dataset is the ID section which is the pdbID with the chainID appended to the end. For our data we will be acquiring this pdbID file and acquiring all information pertaining to the chainID. Sections of interest in our pdb file are the HELIX, SHEET, and ATOM fields respectively. Retrieving all of the data can take some time and took on average 3-4 hours to download and process each pdb chain. In order to alleviate retrieving the data multiple times each section is saved to their own respective file with the REMARK field of HELIX, SHEET, or ATOM being replaced with the pdbID and chainID. While we could store all of this in a single file it was decided to be kept separate because not all questions we are answering will need each section which can take awhile to load.

Now that the data has been acquired we will need to clean the data. To do so we analyzed each section and determined the necessary fields to keep. These are the fields kept in order to reduce our dataset size.

HELIX:

pdbIDC, serNum, helixID, initSeqNum, endSeqNum, helixClass, length

SHEET:

pdbIDC, strand, sheetID, numStrands, initSeqNum, endSeqNum

ATOM:

pdbIDC, serial, name, altLoc, resName, resSeq, x, y, z

We also removed every atom that was not part of the backbone, the result is that only atoms of name N, CA, and C remained. This shrunk the size of the dataset considerably which helps to speed up the data processing needed for our analysis. Resulting data is stored in a CSV formatted file for each section of HELIX, SHEET, and ATOM in an effort to reduce processing time so that we do not need to reclean the dataset multiple times.

1. Popularity of Different Helices

We first wanted to discover which types of helices were more prevalent than others and if it coincided with findings of others. To do so we loaded the data from our cleaned data set for helices only. Our helix csv file contains a helixClass column and each row contains a helix making up all of the helices in our data set. The different helix classes and corresponding pdb class number are as follows.

TYPE OF HELIX CLASS NUMBER

Right-handed alpha (default) 1

Right-handed omega 2

Right-handed pi 3

Right-handed gamma 4

Right-handed 3 - 10 5

Left-handed alpha 6

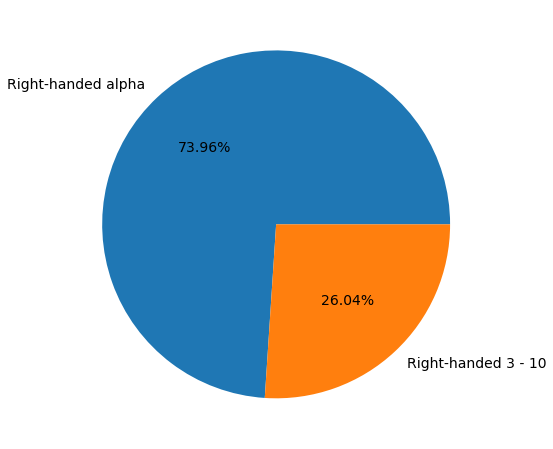
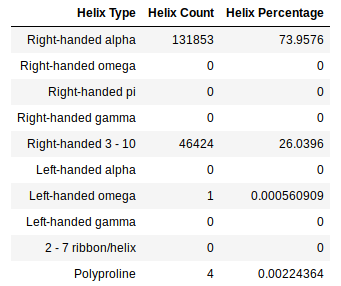
Left-handed omega 7

Left-handed gamma 8

2 - 7 ribbon/helix 9

Polyproline 10

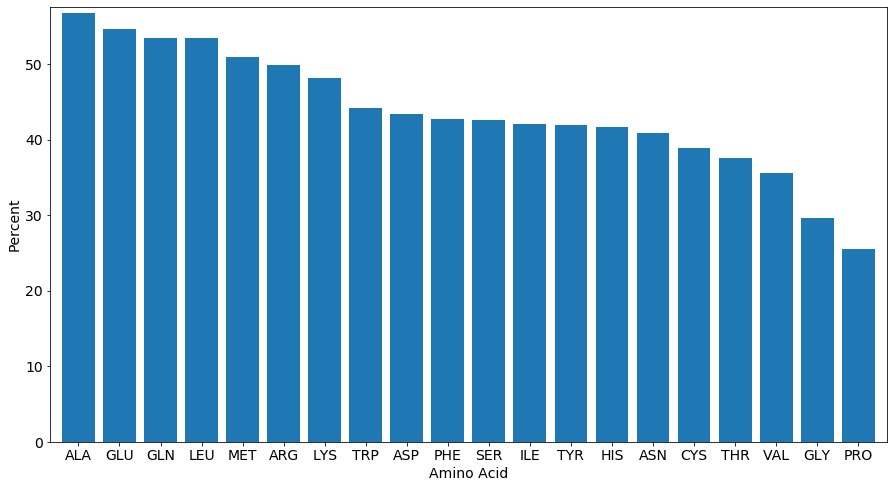
The data set was loaded into a numpy array where we selected the helixClass column and used the np.unique function with the return\_counts option set to True to retrieve all of the types and number each type occurs in the data set. Of the 178282 helices in our data set we discovered the following helix type popularity.



It’s interesting to see that our of Right-handed alpha and Right-handed 3-10 the remaining 8 other types of helices didn’t make up more than % 0.003 of our data.

2A. Amino Acid Helix Preferences

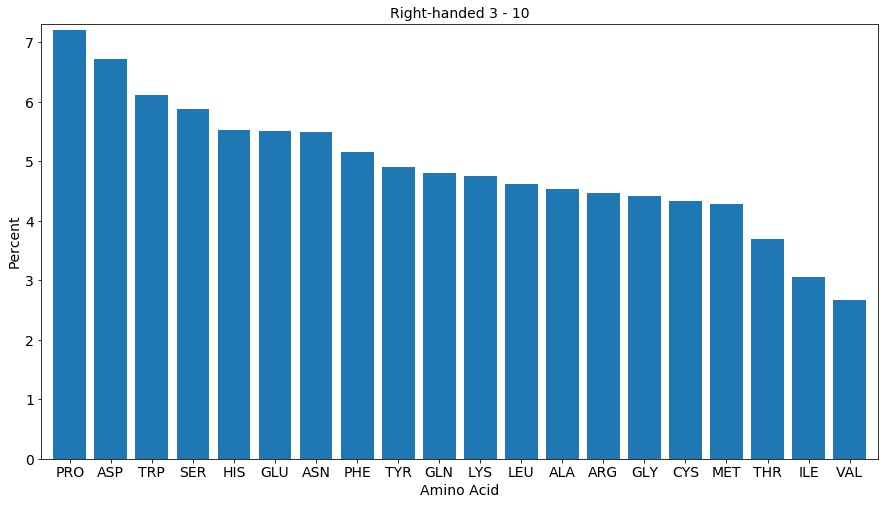
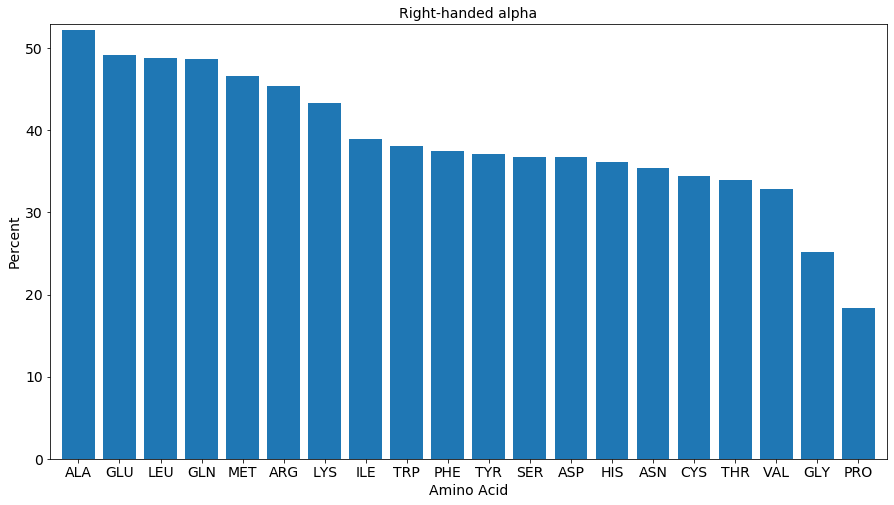
Determined in section 1 was that right-hand alpha and 3 - 10 helices are the most popular helices in our data set. It would also be useful to determine the general composition of helices by seeing the amino acid preference for our data. To do so we loaded our helix and atom data to retrieve all amino acids that occur in helices and get the count of each amino acid. Likewise we retrieve the total count of each amino acid to determine the percentage that each amino acid occurs in a helix resulting our preferences. This is achieved by grabbing the helices and atoms of each pdbIDC and selecting the amino acids that occur within each helix entry using the initSeqNum and endSeqNum of the helix data and the resSeq of the atom data. We then use numpy’s unique method to return the counts of all amino acids that occur in helices and the counts of each amino acid throughout the entire data set to retrieve the preferences. The results are shown below.



We notice that ALA, GLU, GLN, LEU, and MET prefer helices the most. This is an interesting finding and will be explored further in section 5 where we will see which amino acids are preferred by sheets. It is expected to see that helix and sheets have different preferences. Next we will explore the composition of different helix types rather than helices as a whole.

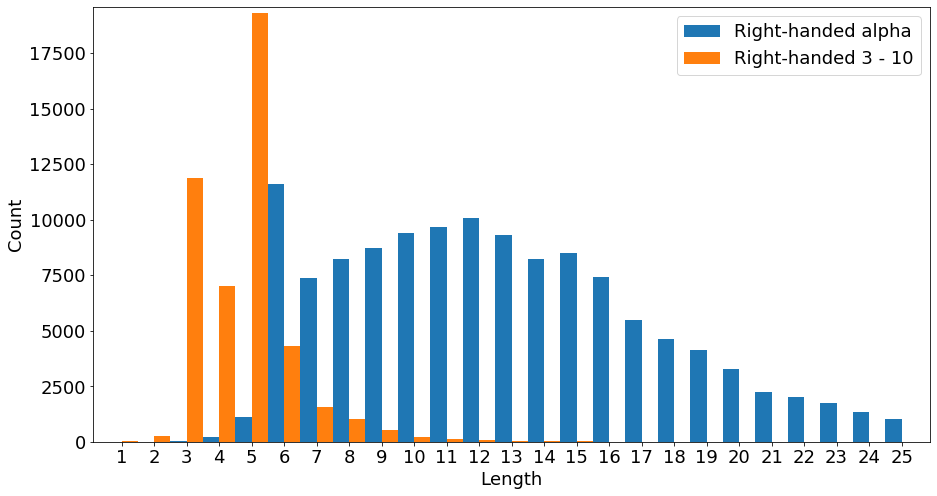
2B. Amino Acid Helix Type Preferences

Analyzing the in the same way except for our main helix types of right-handed alpha and 3 - 10 yields some interesting results. It appears that right-handed alpha follows closely to our results of our analysis in 2A. This is most likely due to the fact that there are many more right-handed alpha helices than right handed 3 - 10 and that is to be expected. However right-handed 3 - 10 yields very different results with PRO, ASP, TRP, SER, and HIS being the amino acids of preference. From this information we can determine that the composition of right-handed alpha and 3 - 10 are composed very differently from one another. It would be interesting if there were enough samples of other helix types to analyze the preferences and see how they compare these helix types.



3. Relations Between Helix Type and Length

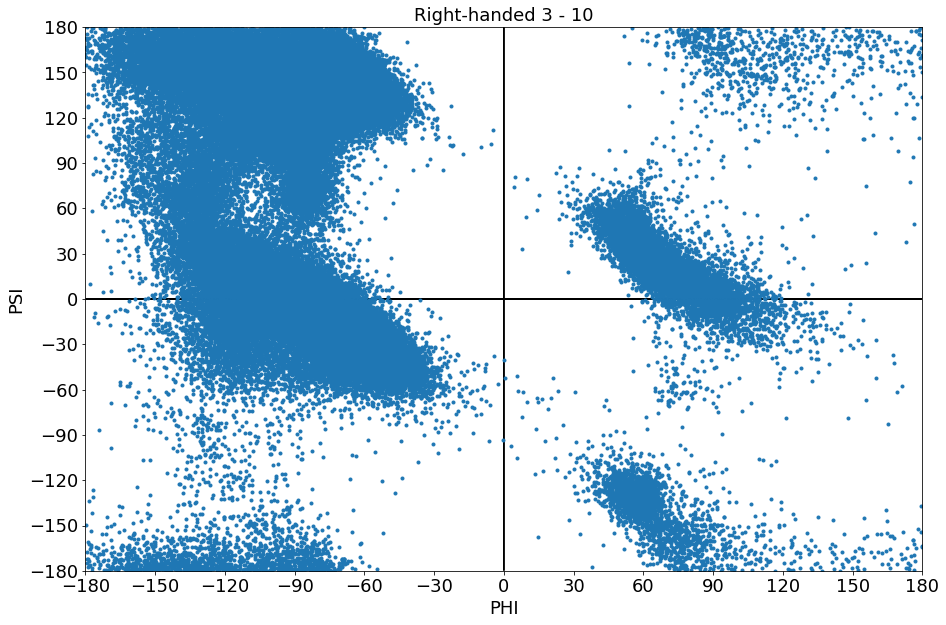
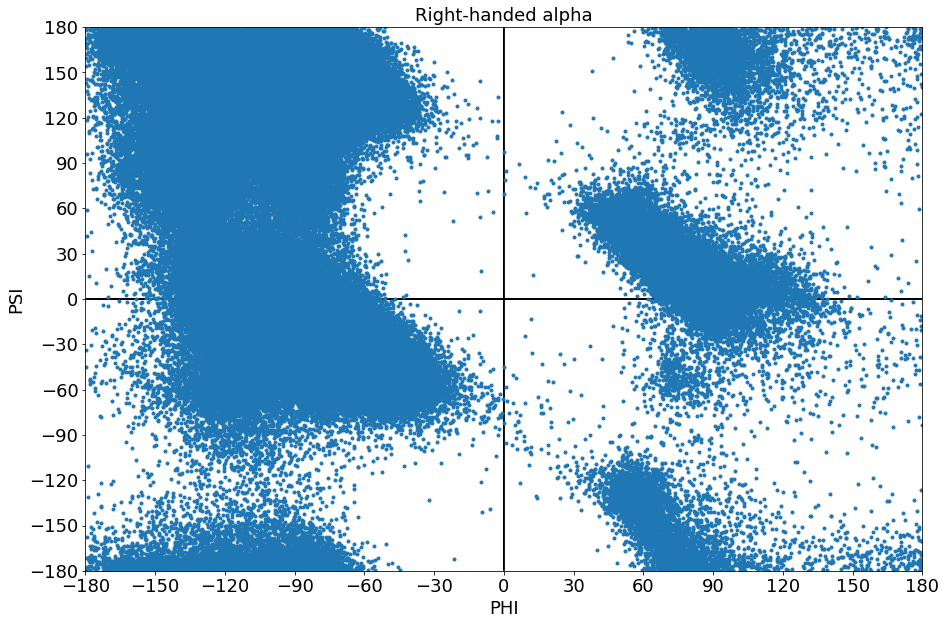
Continuing our exploration of helices we wanted to see if different helix types preferred a certain length. As found within section 1 we noticed that the most prevalent helix types are right-handed alpha and right-handed 3-10 so we analyzed these two helix types as there were not enough samples of other types to come to an accurate conclusion. To begin analysis we used our helix data set contained within a numpy array and selected only our two chosen helix types using the helixClass column, then split them into two arrays for each type respectively. Numpy’s unique function was then used again on each helix type to retrieve lengths from the length column with counts set to True to get the counts for each length. The results were charted below truncating the data to lengths of less than 25 because lengths higher than 25 continued to decrease.



It becomes apparent that most right-handed alpha helices prefer a length of 6 and in general prefer lengths in the range of 5 to 16. Right-handed 3 - 10 prefers a length of 5 and in general sizes in the range of 3 to 5. It is interesting to see that these helices prefer sizes that are right next to one another and it would be interesting to see if other types of helices would also prefer sizes near these if there were enough samples to do so.

4. Torsion Angles of Helix Types

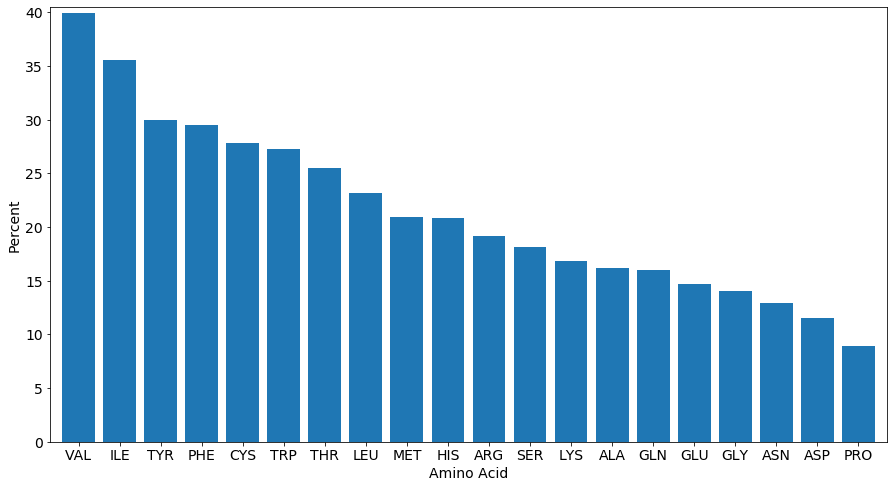
Going further in our analysis between different helices it was decided to build ramachandran like plots of the phi and psi angles of each amino acid contained within the right-hand alpha and right-hand 3 - 10 helices. We want to see if there are many differences between the two with the resulting ramachandran plots. In order to accomplish this we use our numpy arrays for our set of right-handed alpha and 3 - 10 helices. Each atom was pulled from our atom data set which we created during our data preprocessing and stored into their corresponding helix type. The result we have are all the atoms / amino acids contained within right-hand alpha and 3 - 10 types. Each amino acid in each set then has its phi and psi angle calculated for plotting the ramachandran. These are the resulting ramachandran plots.



We can see that the right-handed alpha and 3 - 10 helices are surprisingly similar in their plots. However the right-handed alpha plot appears to be more dense and slightly expanded in some areas as opposed to the right-handed 3 - 10. This discrepancy may be due to number of amino acids in our set of right-handed alpha having 1780618 amino acids and our set of right-handed 3 - 10 containing 216731 amino acids.

5A. Amino Acid Sheet Preferences

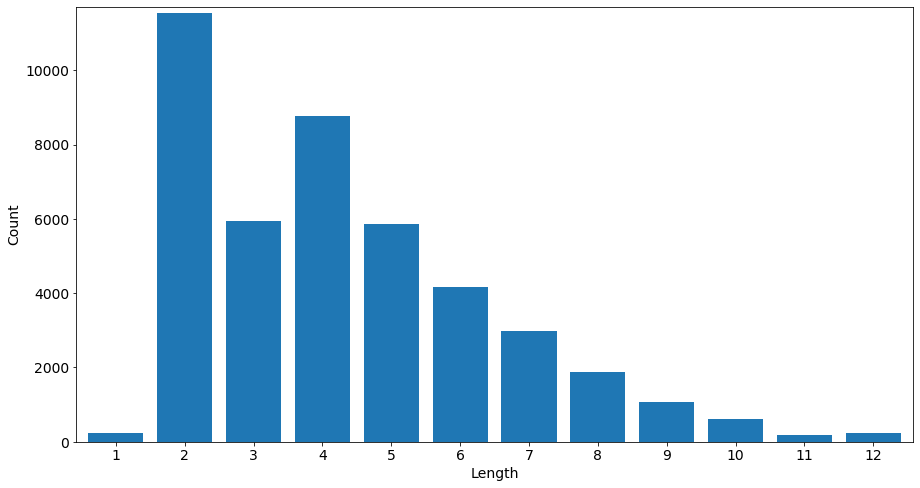
In contrast to the helix preferences explored in section 2 we would like to do the same thing for beta sheets. In order to do so we follow similar steps taken in section 2. We will first load our sheet data and atom data from our cleaned data set. Our atom data is then stripped of all atoms except for our nitrogen atom in order to build a set of amino acids. Then we will retrieve the atoms contained within each strand of every sheet in our data. Sheet entries contain a pdbIDC, initSeqNum field, and endSeqNum which we will use to query our atom data to retrieve each amino acid within the strands resulting in a set of all amino acids that occur in sheets. Finally we utilize numpy’s unique function with return counts to get the counts of every amino acid in sheets and in our data set which we will use to derive the percentage of each amino acid contained in a sheet. The results of our data are as shown below.



From our data set we can see that VAL, ILE, TYR, PHE, and CYS are the 5 amino acids that most prefer sheets which is different from the result from section 2 which is expected and helps to corroborate our results.

5B. Sheet Size Preferences

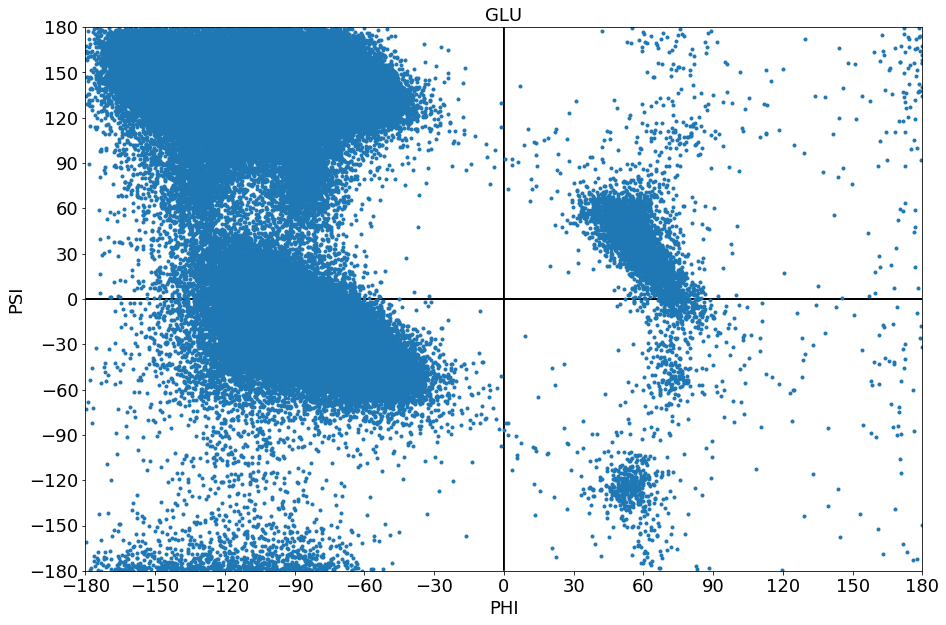
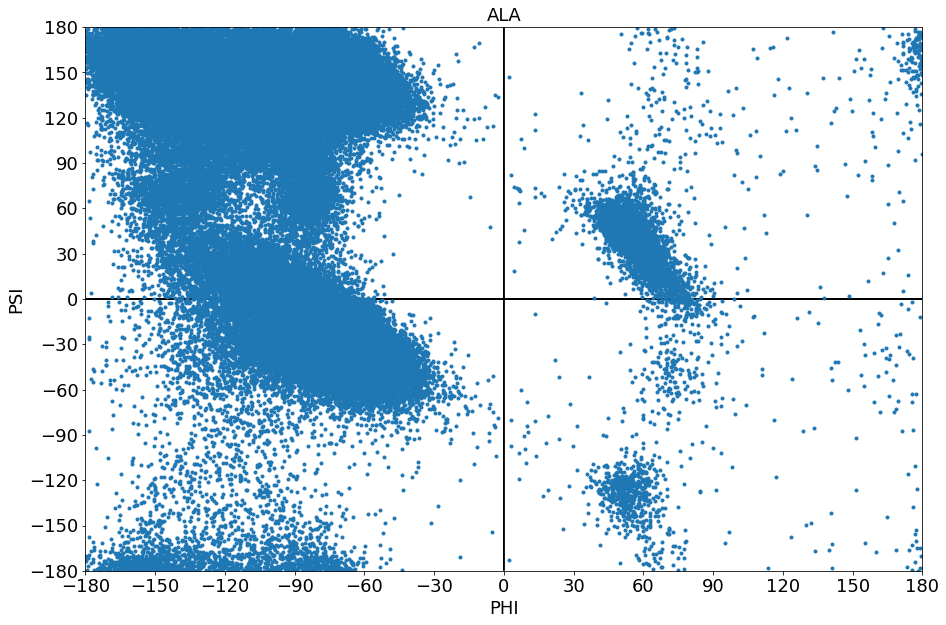
Once again comparing with helices we wanted to see the size preferences of beta sheets. To do so we only need to load our sheet data set. Each sheet in our data starts with the first strand and contains the number of strands contained within the sheet. To make things simple we decide to pull out every sheet’s first strand defined by the strand field and then use numpy unique again to acquire the counts of each sheet length defined by the numStrands field. The resulting data is below.

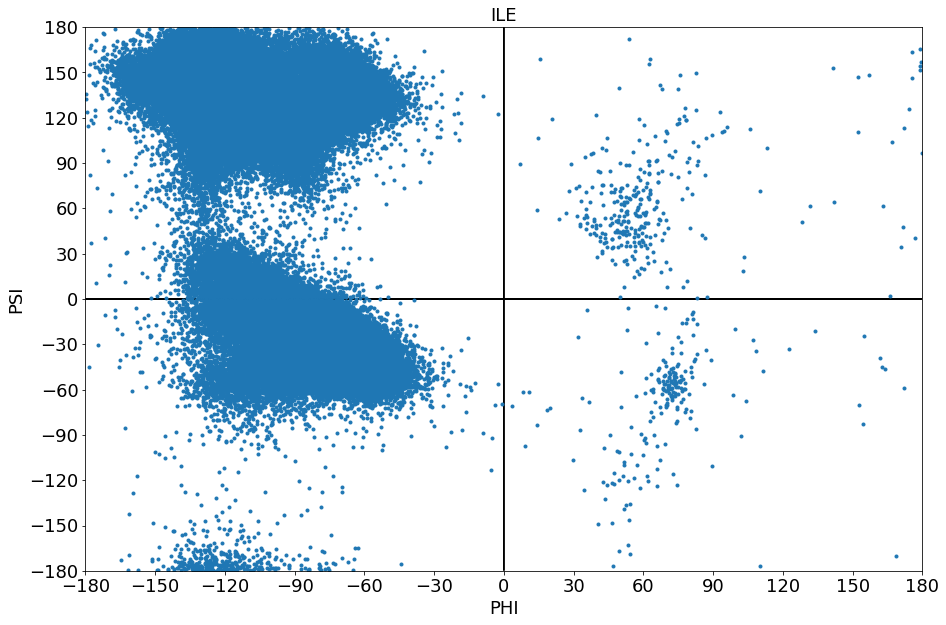
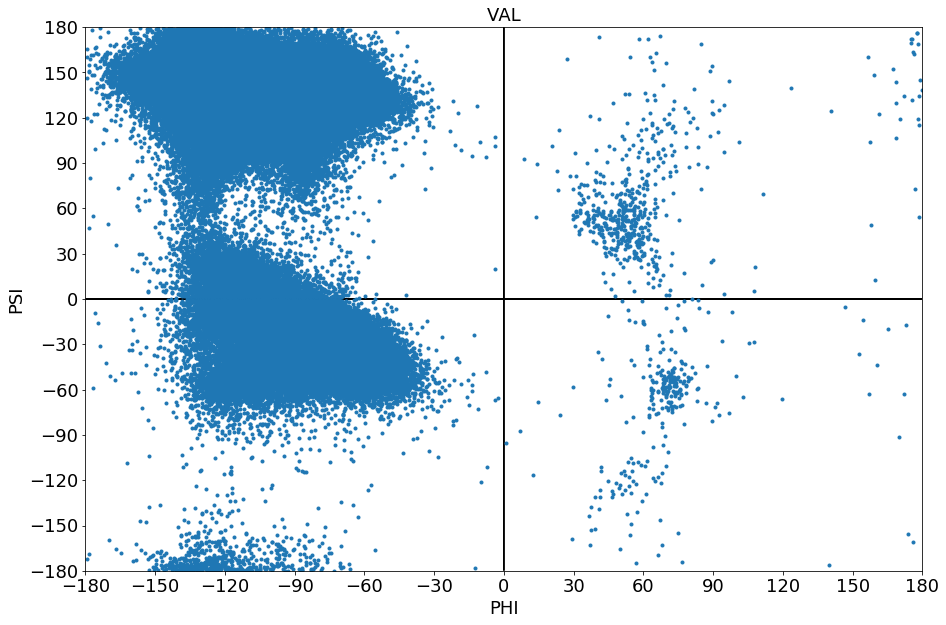


Results are very interesting in that it appears that sheets mostly prefer a length of 2 or 4. In section 3 we discovered that right-handed alpha helices preferred a length of 6 primarily and right-handed 3 - 10 primarily preferred a length of 5. It would be interesting to determine if each different type of helix had a unique primary length preference compared to others. As an example it would be interesting if sheets preferred 2 &4, right-hand alpha preferred 6, right-handed 3 - 10 preferred 5, and the remaining helix types prefer lengths near these but each unique.

6. Ramachandran Comparisons

Since we have been analyzing helices and sheets we thought it would be a good idea to compare the 2 more preferred amino acids from each. The method used is similar to section 4 except it is applied to the set of a given amino acid. Each chosen amino acid has its phi and psi angles calculated from each pdbIDC with the results plotted to build its ramachandran representation. The results are below.

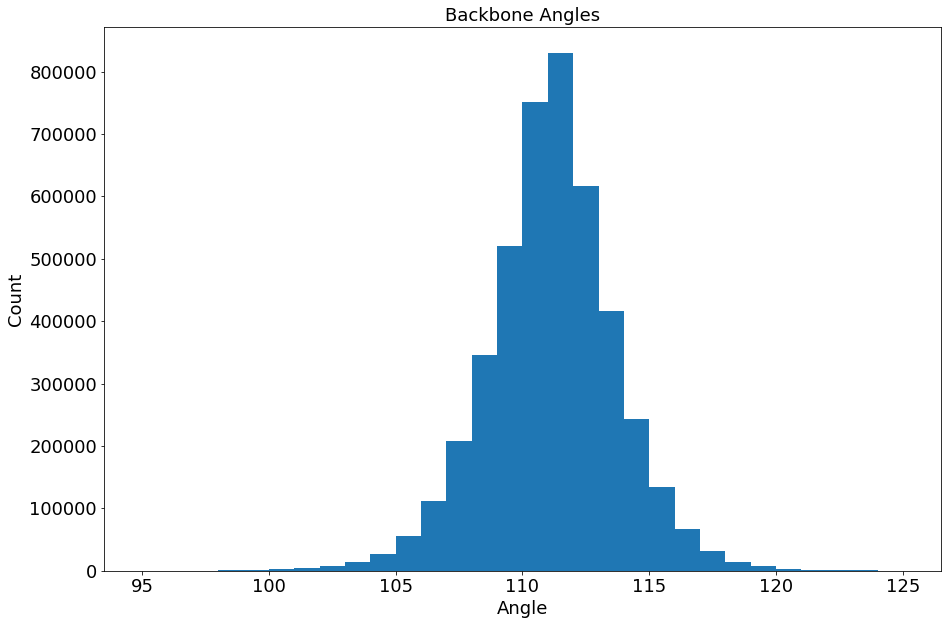
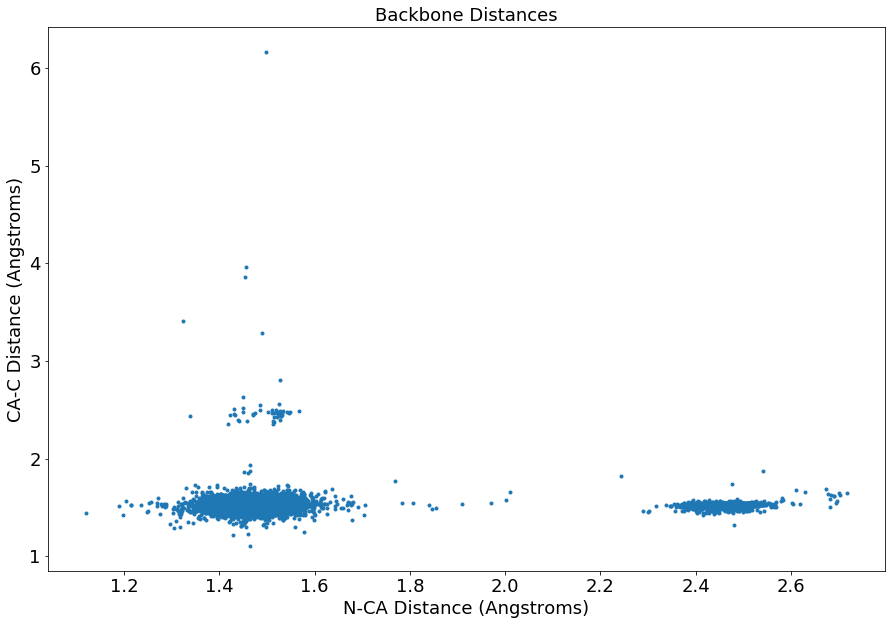




Looking at our plots for ALA, GLU, VAL, and ILE we noticed that there are similarities between them. Generally speaking each of these amino acids share a general shape to them. Paying closer attention to them though we can see that ALA and GLU closely resemble one another and that their both prefer to be in helices. It may be that amino acids to match closely to ALA and GLU are more likely to prefer helices. When looking at our plots for VAL and ILE we also see that they closely resemble one another as well and that they prefer to be in sheets. We believe that it is also likely that amino acids that closely resemble VAL and ILE will also prefer sheets. These plots may also be a good way to classify amino acids and it would be interesting to use machine learning classification algorithms to see if they come up with the same results.

7. Backbone Distances & Angles

While the side chain of amino acids can differ each amino acid contains a backbone consisting of N, CA, and C atoms. To understand more about how these atoms are positioned in the amino acid it was decided to look at the distances between connected atoms as well as the angle between their points. For calculating each of these parameters we iterated over all amino acids contained within each of the pdbIDC’s contained within our dataset that contained the backbone atoms and x y z coordinates to perform the necessary calculations. The results of these calculations are below.



We can notice a few trends once we plot the data. First we notice that almost all of the calculated distance from CA to C is approximately 1.5 angstroms and the distance from N to CA is usually from 1.3 to 1.6 angstroms or from 2.4 to 2.6 angstroms. If we were to further investigate it would be worth knowing which amino acids reside within these two ranges as perhaps certain amino acids will only reside within one of the ranges. Next it can be seen that the angles of the backbone atoms mostly prefers an angle of 112 degrees but in general most angles reside between 107 to 115 degrees. Each of these measurements we have taken helps to understand the amino acids more. It would be interesting to find protein chains that have mutated to see what the measurements of mutated amino acids are compared to an amino acid that was mutated.

8. Conclusion

Throughout our project we have identified certain traits among different secondary structures and the amino acids of different protein chains. It was discovered that the primary helix types were right hand alpha and right hand 3 - 10 where right handed alpha consisted of more than 3 times the amount of right handed 3 - 10 which contributed to the amino acid preferences of helices to favor the right handed alpha types. It was also noted that right handed alpha and right handed 3 - 10 preferred to be of length 6 and 5 respectively which sheet size mostly preferring size 2 but also very much preferred a size 4 leading us to believe that a majority of secondary structures would also prefer a size approximately in this range. A more further in depth study could try to map out the length preferences of the other helix types provided a large enough set of those types is available.

It was also noted that different secondary structure types and sub types appeared to have a different preference into what amino acids are a part of them. It’s very interesting to see that these preferences are so diverse and that they did not share the same more preferred amino acids between them. This leads us to believe that each secondary structure varies greatly between other secondary structures in amino acid composition which could also relate to the purpose of the secondary structure or to the protein.

Lastly we can see how similar some amino acids are by looking at the backbone distances, backbone angles, phi angles, and psi angles. It is amazing how similar each amino acid and the subtle differences between them. It would be interesting if an amino acid was found that did not conform to the general parameter ranges discovered through our experiments.

In the future we could analyze protein similarities using local and global alignments to see what sequences are most commonly among protein chains.