Y3890114

Using Ramachandran Plots to Determine the Stereochemistry of a Large Dataset of Proteins

1. Introduction

Developed in 1963 by Viswanathan Sasisekharan, C. Ramakrishnan and Gopalasamudram Narayana Ramachandran, the Ramachandran plot shows the Phi (ϕ) and Psi (ψ) bond angles of Amino Acids in protein chains (called dihedral angles). This can allow us to find stereochemical structures such as α-helices and β-sheets. α-helices tend to have a backbone which wraps around in a sharp, elongated spiral shape. This is shown in the plots as small dihedral angles. β-sheets on the other hand, tend to be planar. This results in large dihedral angles.

The Ramachandran plot is useful as it allows us to confirm the 3D structure of proteins. It has applications in computational chemistry and biology and is used extensively in drug discovery to study interactions between proteins and receptors.

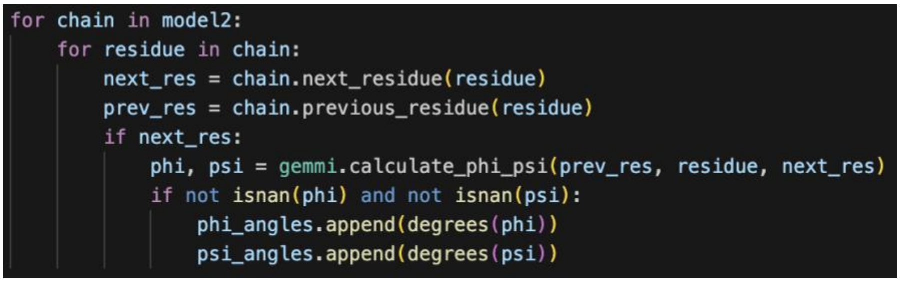
There are several limitations to Ramachandran plots, however. The main one being that it attempts to display 3D structure in a 2D space which can miss some more subtle structures, leading to unexpected interactions between proteins. It is also subject to several inaccuracies such as resolution of measuring equipment, deformations in the protein's bond angles due to crystal packing and it is reliant on average data which could miss certain structures for protein families. Lower resolution measuring equipment also impacts the frequency are what are known as steric clashes which occurs when two or more atoms which aren't bonded appear too close to one another.

2. Method

I used Python for the entirety of the task. To begin I took two sets of PDP codes for a range of proteins which each had roughly 1600 proteins to process. I then used the tqdm and wget modules to download the .cif file for each protein from the Research Collaboratory for Structural Bioinformatics Protein Data Bank. This was the longest piece of code to run as it took between 30 and 40 minutes per set to download the data. Overall, the data was easy to access and the .cif files appeared to be complete.

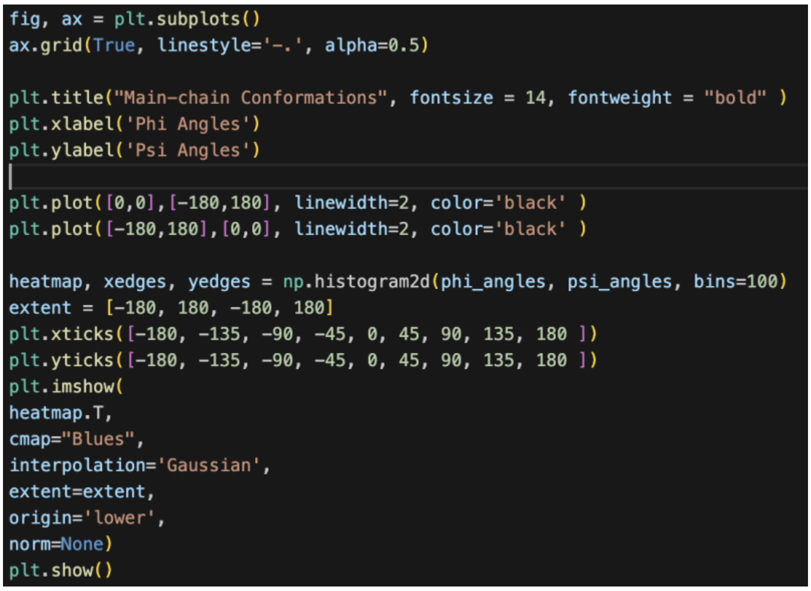
The next step was to use gemmi to create a model from the structure files for each set. I then plugged these models into a function to calculate the omega angles for each protein. I did this to verify that the data was compiled correctly. Most peptide bonds for both sets were trans, and had values close to 180°. I took this as confirmation that the data had been correctly processed.

I then adapted the code used to calculate the omega angles to find both the phi and psi angles for each set as follows:



```
for chain in model2:
    for residue in chain:
        next_res = chain.next_residue(residue)
        prev_res = chain.previous_residue(residue)
        if next_res:
            phi, psi = gemmi.calculate_phi_psi(prev_res, residue, next_res)
            if not isnan(phi) and not isnan(psi):
                phi_angles.append(degrees(phi))
                psi_angles.append(degrees(psi))
```

then used the attained values to create the Ramachandran plots using the following code:



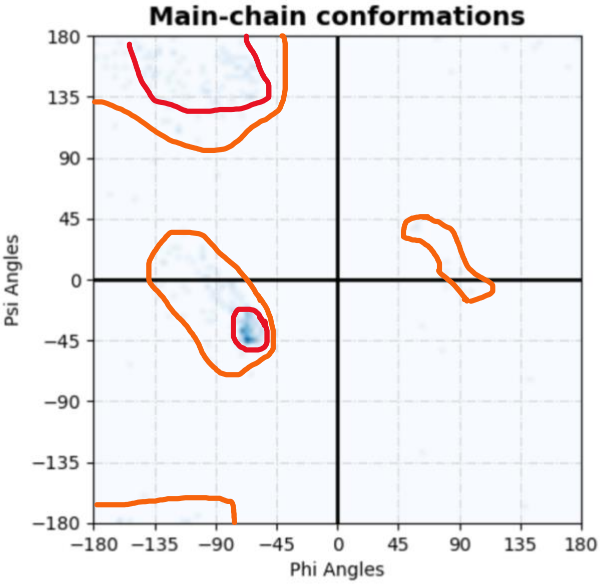
```
fig, ax = plt.subplots()
ax.grid(True, linestyle='-.', alpha=0.5)

plt.title("Main-chain Conformations", fontsize = 14, fontweight = "bold" )
plt.xlabel('Phi Angles')
plt.ylabel('Psi Angles')

plt.plot([0,0],[-180,180], linewidth=2, color='black' )
plt.plot([-180,180],[0,0], linewidth=2, color='black' )

heatmap, xedges, yedges = np.histogram2d(phi_angles, psi_angles, bins=100)
extent = [-180, 180, -180, 180]
plt.xticks([-180, -135, -90, -45, 0, 45, 90, 135, 180 ])
plt.yticks([-180, -135, -90, -45, 0, 45, 90, 135, 180 ])
plt.imshow(
heatmap.T,
cmap="Blues",
interpolation='Gaussian',
extent=extent,
origin='lower',
norm=None)
plt.show()
```

This creates a heatmap-style histogram split into quadrants. The code at the bottom using the imshow command contains variables which I found to create the most accurate, easy to interpret and visually appealing plots.

Overall, the code was reasonably straightforward, and I was able to adapt it using relative file paths to ensure that it should be able to work on other machines immediately and without fault. Instructions for use are included in the repository linked at the end of the document.

3. Results

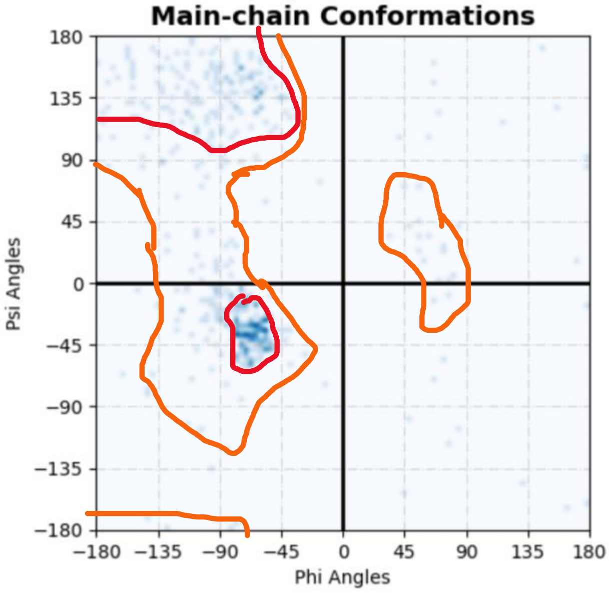
3.1 Set 1



From this Ramachandran plot, we can see that there is a high concentration of points at (ϕ -70, ψ -40), this shows that many of the structures are right-handed α-helices. There is a smaller but still significant area at (ϕ -105, ψ 150). This shows that β-sheets are present in a reasonable number too. The final area of significance is around (ϕ 90, ψ 0). This is an indicator of the presence of left-handed α-helices. These occur in much lower frequencies than the right-handed helices.

I have attempted to show the favoured region in red and the allowed region in orange.

3.2 Set 2



From this Ramachandran plot, we can see that there is a high concentration of points at (ϕ -70, ψ -40), this shows that many of the structures are right-handed α-helices. There is a smaller but still significant area at (ϕ -70, ψ 135). This shows that there are also many β-sheets. The final area of significance is around (ϕ 80, ψ 10). This is an indicator of the presence of left-handed α-helices. These occur in much lower frequencies than the right-handed helices.

I have attempted to show the favoured region in red and the allowed region in orange.

3.3 Comparison Between Sets

Both sets appear similar in the distribution of points. They both point towards a significant presence of α-helices and β-pleated sheets. Right-handed α-helices appear more frequently than left-handed α-helices as they are less prone to influence from side chains i.e. fewer steric clashes, and are therefore more stable.

In the first set, the heatmap showed a weaker set of points compared to the second, however these points were more constrained to the areas we'd typically expect from structures such as α-helices and β-sheets.

In the second set, areas of α-helices and β-sheets were much darker than the first set, showing a much stronger indication for the presence of these structures. There were many angles which we would not typically expect to see however, including many in the lower-right quadrant which shows that there were more unique structures in the second dataset.

The allowed region was much larger in the second dataset when compared to the, however I feel that the favoured region in the first set was more pronounced. This shows that the bond angles and therefore structures in the first set tended to be mainly the ones seen most commonly.

4. Conclusion

Overall, the Ramachandran plot is a very straightforward way to visualise the structures present in proteins. From our data it is clear to see that in terms of named structures, the right-handed α-helix appears most frequently, followed by β-sheets and finally left-handed α-helices. There are also trace amounts of points which correspond to unique structures shown however the occurrence of these happens in a frequency that is related to stability i.e. the most stable structures appear most frequently. The results shown here are extremely typical of what we would expect from as large a dataset as is used here.

Perhaps creating several datasets based on residue type would have provided a clearer distinction between the two datasets or could have possibly explained many of the interesting features of this graph for example would there still be as significant a presence in the lower right quadrant if we removed glycine.

Appendix

Link to Repository: https://github.com/CMR01/Chem_Assessment_1