RSCB PDB Task – Álvaro Maza

Parts 1 & 2 – Read the mmcif file 1DG3 and collect the Cα 3D coordinates.

Even though this part was designed to familiarize me with the library, I encountered some things worth to mention.

When I first iterated through the residues of the 1DG3 protein structure, I noticed that I got a bigger number of residues that the one that appeared as the "Modelled Residue Count". This was due to the water molecules, labeled as HOH; so after filtering the residues I got the desired number of 540 residues.

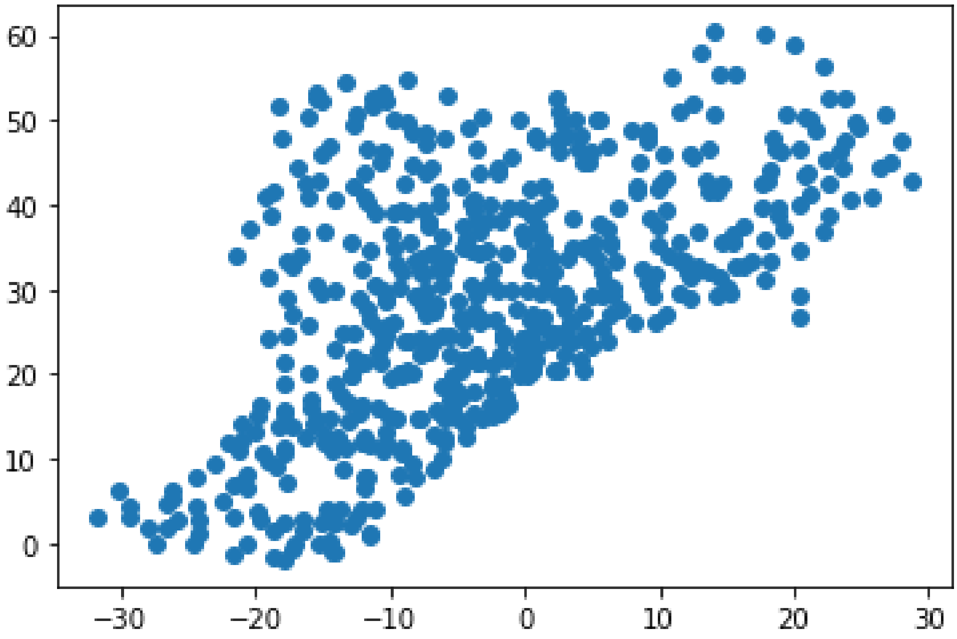
I stored the id of each residue, and the coordinates of its atoms in the list all_atoms.

To store the Cα 3D coordinates I first tried to search them using the full name '.CA.' , but it was not necessary since there was no collision with the label of Calcium Atoms ('CA..'), so the Cα atoms were labeled simply as 'CA'.

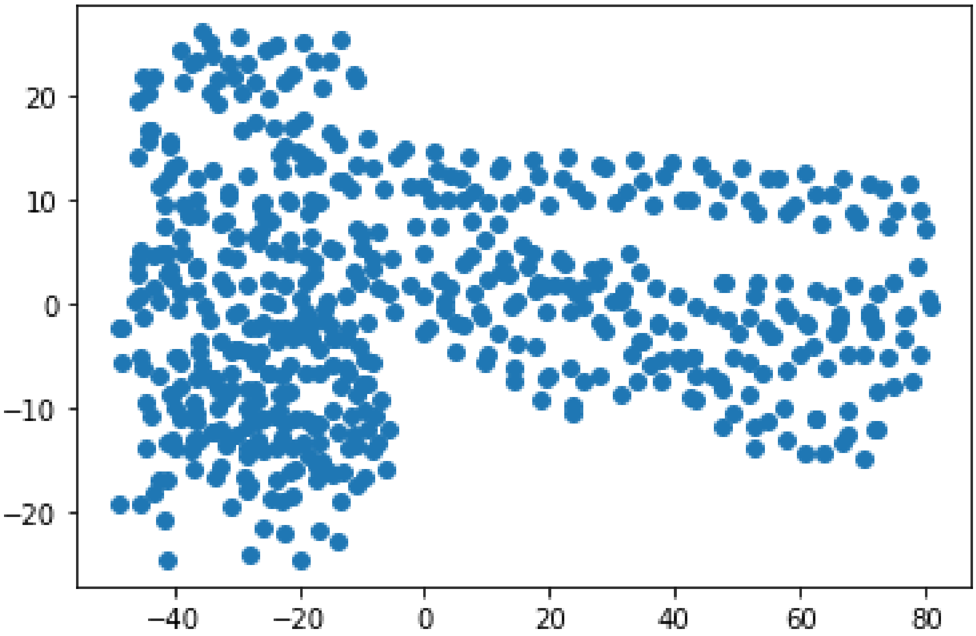
I used get_coord to parse the coordinates of the atoms, but get_vector can also be used and gives more decimals and a vectorized version of the coordinates that can be useful to work with it, but for this particular problem it was not needed.

Part 3 & 4

- Project and plot the Cα coordinates on the plane z=0



- Compute PCA on the Cα coordinates



Part 5 – Documentation of the results

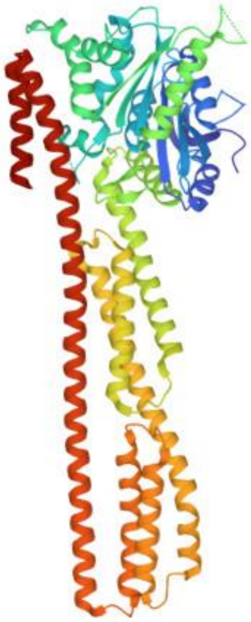
- Discuss why plots are different.

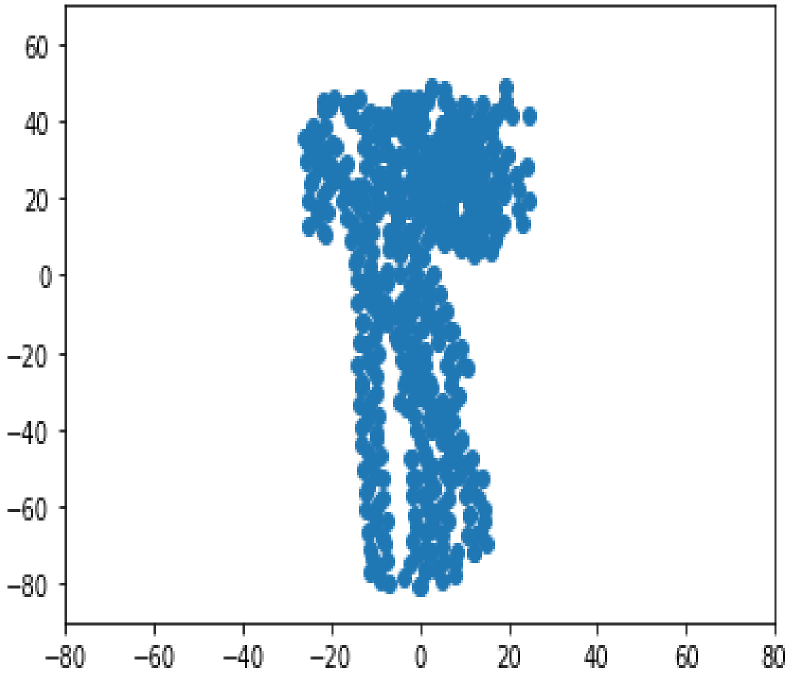
  In the first plot, only X and Y coordinates are taken into account, so the Z coordinate is not considered.

  On the other hand, using PCA, you get new variables that are linear combinations of the original variables (X,Y,Z coords) in such a way that the new variables (PC1,PC2) are uncorrelated. This means that the new variables do take into account the Z coordinate to compute these linear transformations, and maximize the variability of the dataset that can be expressed in two dimensions.

- Open the 3D view for the 1DG3 protein (link), do you find any relationship between the images and the 3D model?

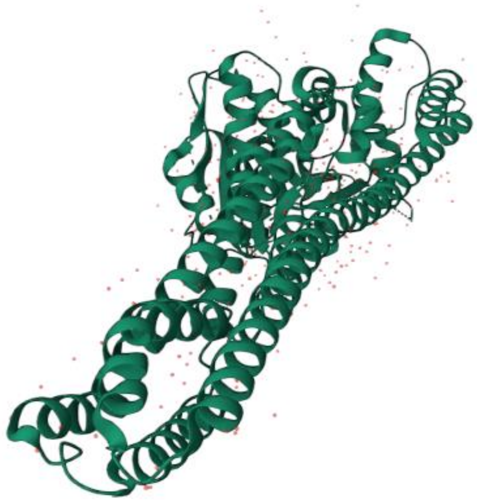
  If you rotate and mirror the PCA plot you can easily see the resemblance between the 1DG3 protein view and the PCA plot:

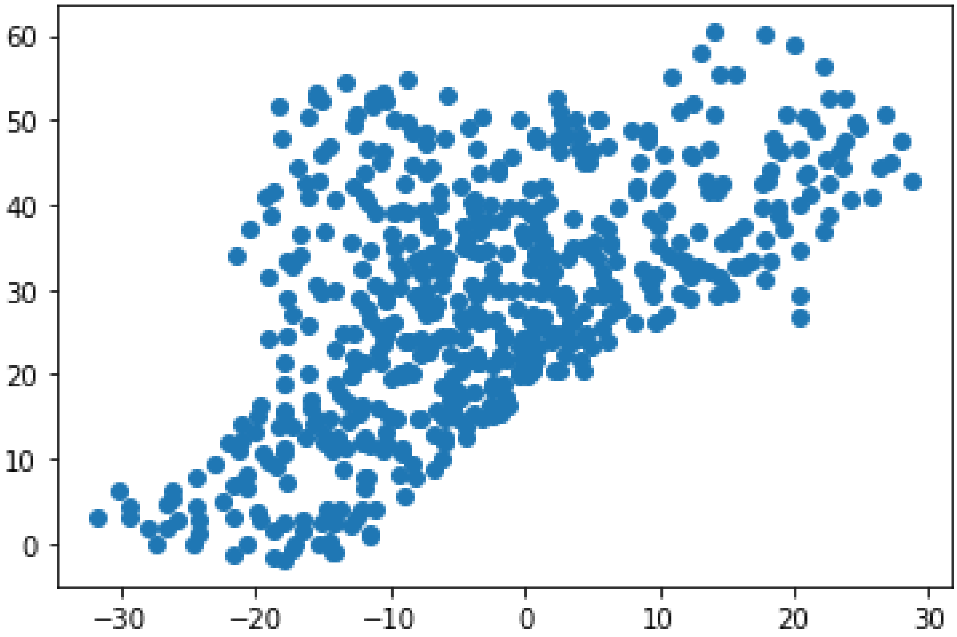




This shows how using PCA can be useful to decide what orientation of the protein will explain better its form in two dimensions.

The original plot resembles the protein structure seen from above, which makes sense since we did not take into account the Z axis:




60
50
40
30
20
10
0
−30
−20
−10
0
10
20
30