# DAT405/DIT407 Introduction to Data Science and AI, LP3 2023

## Assignment 3: Clustering

These questions concern the main conformation of proteins. Part of a protein's main chain is shown in Figure 1. A protein chain is able to fold into its native conformation by rotation around two of the bonds in the main chain, designated $\phi$ (phi) and $\psi$ (psi). Some combinations of phi and psi values are impossible (e.g. some atoms clash into each other if we try to force the main chain to have a particular combination of phi and psi values). Some other combinations of phi and psi values are very common since they are energetically favourable. To understand the problem domain better, please look at:

- http://bioinformatics.org/molvis/phipsi/
- http://tinyurl.com/RamachandranPrincipleYouTube

The data file “data_assignment3.csv” contains a list of phi and psi combinations that have been observed in a large set of proteins. The angles are measured here in degrees.

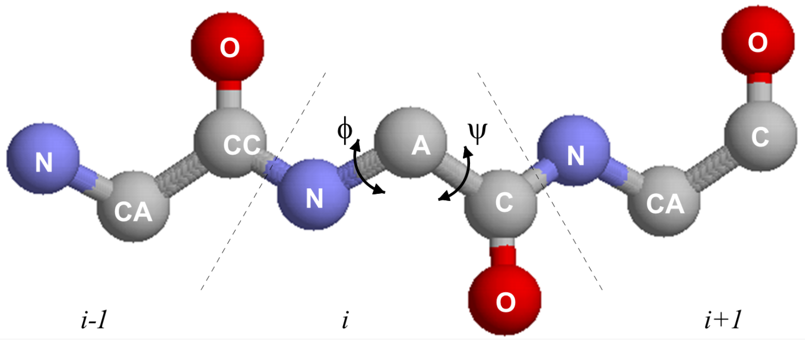

*Figure 1. A protein's main chain. The heavy (i.e. non-hydrogen) main chain atoms of three consecutive amino acid residues (i-1, I and i+1) are represented by spheres, and the covalent bonds between these atoms are represented by rods. Nitrogen and oxygen atoms (N and O) are shown in blue and red respectively; carbon atoms are shown in grey. The central carbon atom (the alpha carbon, or C$\alpha$, labelled CA) is the main chain atom to which a side chain (not shown) is attached. Rotation can occur around the bonds labelled $\phi$ (phi) and $\psi$ (psi).*

1. Show the distribution of phi and psi combinations using:
    a. A scatter plot
    b. A 2D histogram

   Make sure the plots are nice and clean. Can you modify them for better visualisation?

2. Use the K-means clustering method to cluster the phi and psi angle combinations in the data file.
    a. Experiment with different values of K. Suggest an appropriate value of K for this task and motivate this choice.
    b. Do the clusters found in part (a) seem reasonable?
    c. **(Optional question, if you are interested and have time)** The top edge of a Ramachandran plot wraps round to the bottom edge, and the right edge wraps around to the left edge (we can think of the 2D Ramachandran plot being

mapped onto the surface of a torus). Ideally, this should be considered when clustering the data points on a Ramachandran plot. Repeat questions a and b taking this into account.

3. Use the DBSCAN method to cluster the phi and psi angle combinations in the data file.
    a. Motivate the choice of:
        i. the minimum number of samples in the neighbourhood for a point to be considered as a core point, and
        ii. the maximum distance between two samples belonging to the same neighbourhood ("eps" or "epsilon").

       Compare the clusters found by DBSCAN with those found using K-means.
    b. Highlight the clusters found using DBSCAN and any outliers in a scatter plot.
    c. How many outliers are found? Plot a bar chart to show how often each of the amino acid residue types are outliers.

4. The data file can be stratified by amino acid residue type. Use DBSCAN to cluster the data that have residue type PRO. Investigate how the clusters found for amino acid residues of type PRO differ from the general clusters (i.e., the clusters that you get from DBSCAN with mixed residue types in question 3).
   Note: the parameters might have to be adjusted from those used in question 3.

## Self-check

Ensure that you any plots are complete and readable (labels on the axes, etc.).

## Submitting work

The **most convenient format for submitting your work is by extracting a pdf from your Jupyter notebook** (File -> Download As.. -> pdf). This way, you can include both code, figures and text in one file, and we can easily view it directly in Canvas, meaning marking will be quicker and you get your feedback sooner.
The submission should contain:

- All Python code written.
- Figures produced and the descriptions/discussions that are requested in the questions, as text. Remember to motivate all steps and decisions taken.

In each file that you submit (but we strongly prefer it to be one), give the names of the people submitting the work. At the beginning, also state how many hours each person spent working on the assignment.

Deadline: Tuesday 7 February at 23:59.