Assignment 3: Clustering

Due September 18, 2023 at 13:00

1 Instructions

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 ϕ (phi) and ψ (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 favorable. To understand the problem domain better, please look at:

- (Interactive Tutorial) The Ramachandran Principle: Phi (ϕ) and Psi (ψ) Angles in Proteins
- (YouTube Video) Ramachandran Principle: Protein Atomic Clashes vs. Phi & Psi

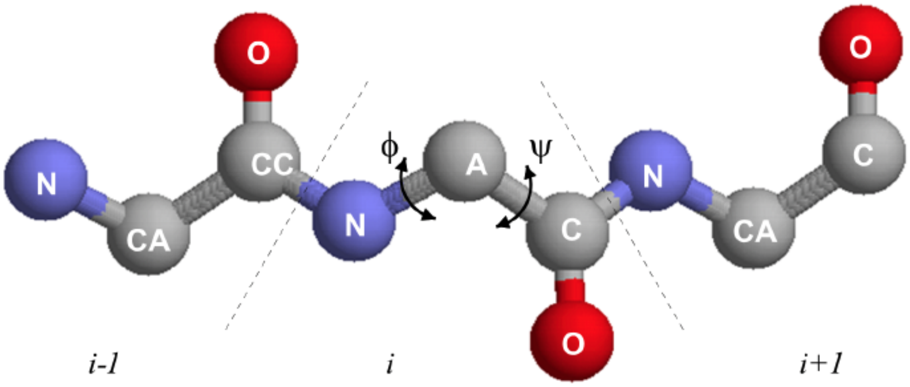

Figure 1: A protein's main chain. The heavy (i.e., non-hydrogen) main chain atoms of three consecutive amino acid residues (i-1, i, and i+1) are represented by spheres, and the covalent bonds between these atoms are represented by rods. Nitrogen and oxygen atoms (N and O) are shown in blue and red, respectively;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

1.1 Protein Angle Dataset

The data file "protein-angle-dataset.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.

Answer the following questions using this dataset:

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 visualization? *Hint: consider what would happen if you shift the range of the x- or y-axis on your plots.*

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 the 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 neighborhood for a point to be considered as a core point, and

      ii. the maximum distance between two samples belonging to the same neighborhood ("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.*

1.2 What To Submit

- A PDF report that includes the figures produced and the descriptions/discussions that are requested in the questions.
- All Python code written. This may take the form of (i) individual Python files (submit the Python source code and append the code as an appendix in your PDF report), or (ii) a Jupyter notebook (submit both the Jupyter notebook and a separate PDF file exported from Jupyter).
- **Do not submit "zip" files**.

In each file that you submit, give the names of the people submitting the work. On the first page of the report state how many hours each person spent working on the assignment. Remember that we check for plagiarism, and we are obliged to report suspected cases.

Ensure that all group members have joined the Assignment Group in Canvas before submitting your solution.

2 Self-Check: *Please read this before submitting your report*

In a data science project, it is usually not sufficient to write a program, run it, and then present a graph or table with results. We should also think carefully about the data that has been used, have a close look at the results, and consider whether these seem reasonable.

For this assignment in particular, it is important to check for yourself the following points:

- □ Did you present your arguments clearly, and describe your results in a pedagogical way? For instance, did you choose to present your results in a specific type of plot? If you chose a plot, what type of plot did you choose, and could there have been any better choices?
- □ Did you make sure that all your code is included in your report, and that it is well-documented (e.g., makes clear use of comments)?

Additionally, for all assignments, **always perform the following self-checks** before submission:

- □ Have you answered all questions to the best of your ability?
- □ Is all the required information on the front page (e.g., you and your partner's names, the hours each partner worked, the correct file name, etc)
- □ Anything else you can easily check? (e.g., clearly labeled axes in figures, clearly labeled units, clear terminology and arguments, clearly stated answers, etc)

Do not submit an incomplete assignment! We teachers are available to help you, and you can receive a short extension if you contact us.