RCSB PDB - SDSC Software Development Assistant

The primary purpose of this exercise is to prove that you can learn and understand the basic concepts and tools that are needed for the position: SDSC Software Development Assistant. Many of the concepts and/or tools required for this task might be new to you. However, the motivation of this task is not to achieve a deep understanding of the biology or mathematics behind the models and tools but to get an idea of their meaning and how to use them.

Brief description:

In this task, you will have to collect a particular set of atomic positions (3D coordinates) from a protein file and apply principal component analysis to project them into the 2D plane defined by the first two components (dimensionality reduction).

Task goals:

- Understanding mmcif file format
- Working with mmcif files
- Extract and transform protein features
- Visualization and interpretation of results

Task evaluation:

- Source code
- External libraries and resources used
- Documentation of the software
- Documentation of the results

Introduction:

Proteins are large molecules that comprise one or more chains of amino acids (a.k.a. residues). For this exercise, let's consider only monomeric proteins, thus, proteins that are composed of a single chain of amino acids. There are 20 different types of amino acids. These are known as the building blocks of proteins. A protein can be described as sequential combinations of these amino acids that are bound together in a linear fashion. In brief, a single amino acid structure can be divided into the main chain and the side chain (warning: we use again the term chain but in this case, it identifies a subset of atoms in a single amino acid). The main chain is common for all the amino acids and the side chain is what makes each type different. For this task, we will be interested only in the main chain and in particular the carbon-alpha (Cα) atom, the central atom of the amino acid. A brief introduction to proteins can be found here.

Generally speaking, proteins fold into a stable 3D conformation, determined by their sequence of amino acids. For this study, we can consider proteins as a set of 3D coordinates (x,y,z).

The Protein Data Bank (PDB) is the single repository of protein 3D structures. Its main goal is to keep 3D structural data available for research and educational purposes. We are going to use the data and visualization tools provided by RCSB PDB, one of the partners of the PDB

consortium. PDB proteins are identified by 4 alphanumeric characters, e.g. 1ACB, 2UZI, or 1XXX. In this task, you are going to work with the entry: 1DG3.

In this exercise, you will collect the atomic (3D) positions of the carbon-alpha (Cα) atoms of the 1DG3 protein structure, apply different geometric transformations, and plot the transformation results.

Task summary:

1. Read the mmcif file of 1DG3
2. Collect the Cα 3D coordinates
3. Project and plot the Cα coordinates onto the plane z=0
4. Compute PCA on the Cα coordinates and project them onto the first and second components (dimensionality reduction)
5. Documentation of the results
6. Make source code and documentation available in any open-source repo

Task Details:

You can use any programming language. I recommend using python, the data analysis (scikit-learn) and visualization (matplotlib) libraries will make this task easier. However, feel free to choose any approach. Manual parsing of files is highly discouraged (see 1). Source code and results have to be available on GitHub, GitLab, or any other open-source repo. Results documentation and how to replicate the analysis need to be included.

1. Read the mmcif file 1DG3

The 1DG3 mmcif file can be downloaded here. There are different libraries for working with mmcif files.

- Python: Bioython
- Java: Biojava

As a test, try to use one of these libraries to open the 1DG3 mmcif, collect a list of its residues and print their atom coordinates.

Hint: In general, a mcif file can contain multiple models (identified by integers) and each of them multiple chains (in most cases identified by a single uppercase character). Data organization level: Model > Chain > Residue > Atom. The 1DG3 contains a single model and a single chain.

Parsing code example:

```
parser = MMCIFParser()
structure = parser.get_structure("myFile", "1dg3.cif")
chain = structure[0]['A'] #uniqe model: 0 and uniqe chain: A
chain.get_atoms()
```

2. Collect the Cα 3D coordinates

The next step is to collect the coordinates of Cα atoms. Open the 1DG3 mmcif file with a text editor and scroll down until you see lines starting with “ATOM” (see below).

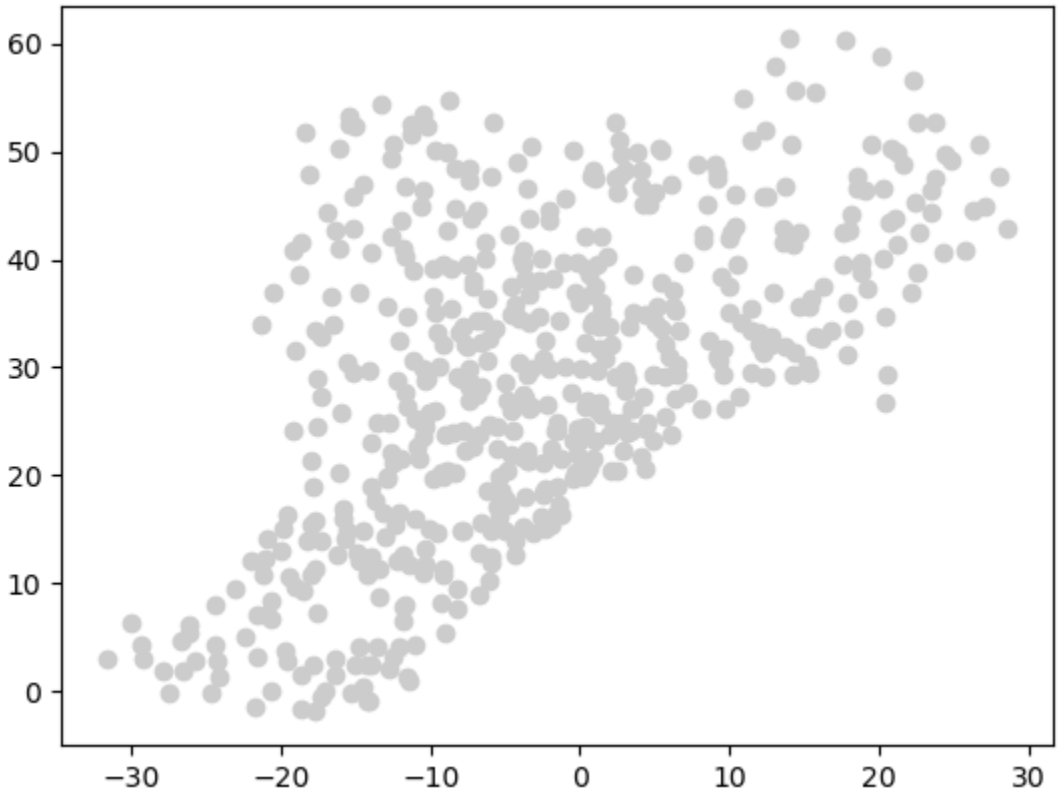
```
_atom_site.pdbx_PDB_model_num
ATOM   1    N  N   . VAL A 1 1   ? 6.204   16.869  4.854   1.00 49.05 ? 1   VAL A N   1
ATOM   2    C  CA  . VAL A 1 1   ? 6.913   17.759  4.607   1.00 43.14 ? 1   VAL A CA  1
ATOM   3    C  C   . VAL A 1 1   ? 8.504   17.378  4.797   1.00 24.80 ? 1   VAL A C   1
```

The field highlighted in green indicates the atom type and in yellow the atomic 3D coordinates (x,y,z). Use the library API to extract those coordinates. Please, manual parsing of the file is highly discouraged.

3. Project and plot the Cα coordinates on the plane z=0

Project the Cα coordinates on the plane z=0. Use your favorite chart library to display a scatter plot of the results (include an image as part of the documentation)

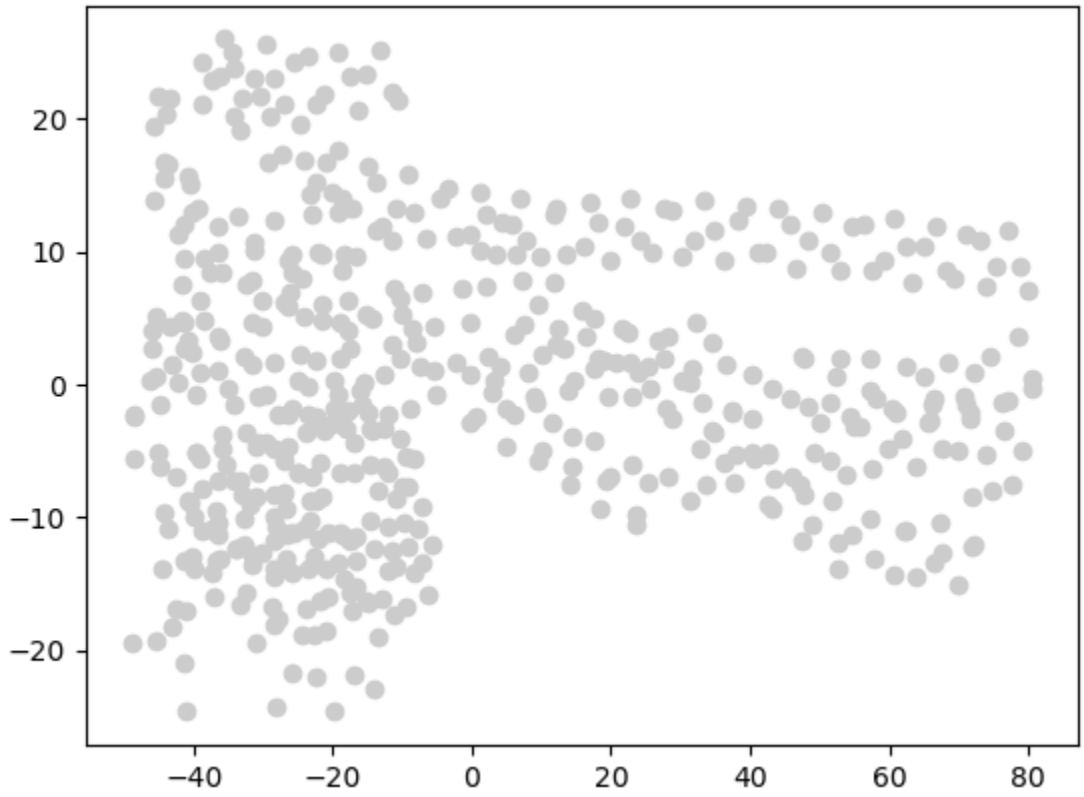
Hint: Results should look similar to



4. Compute PCA on the Cα coordinates

Using the original Cα coordinates compute principal component analysis and apply dimensionality reduction to project the coordinates onto the first two components. Display a scatter plot of the results and include an image as part of the documentation.

Hint: Results should look similar to



5. Documentation of the results

Discuss why plots are different. Open the 3D view for the 1DG3 protein (link), do you find any relationship between the images and the 3D model?

6. Make source code and documentation available in any open-source repo

Make source code and documentation available in any open-source repo (recommended: GithHub).

Please, feel free to contact me if you have any questions or if you need any help (joan.segura@rcsb.org)

Useful links

Guide to protein structure:
https://www.nature.com/scitable/topicpage/protein-structure-14122136/

Guide to mmcif format:
https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/beginner%E2%80%99s-guide-to-pdb-structures-and-the-pdbx-mmcif-format

RCSB PDB entry:
https://www.rcsb.org/structure/1DG3

1DG3 mmcif file:
https://files.rcsb.org/download/1DG3.cif