Structural Bioinformatics level III - Report session 1

Tanguy Lallemand

M2 Bioinformatics, NANTES

February 4, 2019

We are looking to construct a data set of protein structures gathering structures filtered at a cut-off value of **20% of sequence identity, a resolution cut-off of** 1.6 Å **and a R-factor cutoff of 0.25**. This cannot be achieved with a simple request on PDB database, and requires the use of a specialized server (PISCES). This server provide a data set grouping a list of 3429 PDB IDs of proteins and in addition some additional informations like resolution or R-factor. First part of script was written in order to retrieve the list of IDs and download the associated files from the PDB database. All downloaded pdb files are then compressed into a file called **all_pdb_files.zip**.

It remains therefore to compute all the files of this archive using DSSP, a tool allowing to assign secondary structures and get a lot of informations about structure of each proteins. To begin with, we ran DSSP on a single file. The output provides a lot of informations for each residue. The residues are inline and the different informations are placed in the columns of the file. We are looking for particular information, namely:

- The angles **phi** and **psi** gathering the torsion angles of the peptide skeleton. Those informations can be found in **columns 105 to 116**.
- The **amino acid sequence**. This information is given in **column 14**.
- The **accessibility of the solvent** which is concretely the number of water molecule in contact with this residue. This information can be found in **columns 36, 37 and 38**.

The script follows with a loop allowing to launch the DSSP program on all pdb files. Next part of script enabling to parse needed informations from every DSSP results and output them in different directories and files.

In concrete terms, amino sequences are outputted in **.seq** files saved in **./dssp_output/sequence/** directory. seq files are constituted by one line gathering all residues of structure. Accessibility of each residues is stored in **.acc** in **./dssp_output/solvent_accessibility/**. To finish angles are saved in **./dssp_output/angles** in **.ang** files. A sample of each files is provided in archive given with this report.

Please see next figure (figure 1) to have a global view of script.

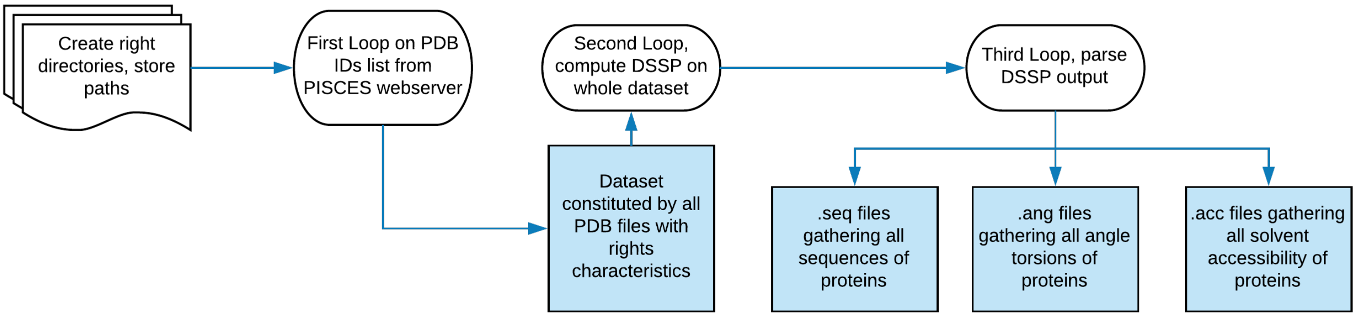
Create right directories, store paths
First Loop on PDB IDs list from PISCES webserver
Second Loop, compute DSSP on whole dataset
Third Loop, parse DSSP output
Dataset constituted by all PDB files with rights characteristics
.seq files gathering all sequences of proteins
.ang files gathering all angle torsions of proteins
.acc files gathering all solvent accessibility of proteins

Figure 1: Global schema of script's sctructure