This README contains the instruction to run the script protein\_info\_extract.py

1. Description:

Rosetta is a protein modeling/designing software developed by the Baker Lab of the University of Washington. PyRosetta is a Rosetta version that is compatible with Python which allows users to access Rosetta functions easier. Under normal circumstances, PyRosetta intakes cleaned protein PDB files (without non-protein molecules). If a protein small-molecule ligand is not parameterized in the PyRosetta database, it will automatically discard it. protein\_info\_extract.py is a Python script that uses the functions in PyRosetta to:

* Download PDB files from <https://www.rcsb.org/> based on the input 4-letter PDBIDs. The downloaded PDB files will be cleaned and will be input for extracting the parameters.
* Assign secondary structures to the input crystal structures based on Dictionary of secondary structure of proteins (DSSP) algorithm by recognizing hydrogen bonding patterns extrapolated from the atomic coordinates of the input PDB files [1].
* Calculate the solvent accessible surface area (SASA) of each amino acid residue using the Lee-Richards molecular surface [2-3].
* Calculate the total residue energy (Rosetta energy unit, REU) of each residue based on Rosetta score function (ref2015 energy function is used as the default score function) [4-5].
* Extract the b-factor of each residue stored the input crystal structures

All parameters will be recorded into a .csv file. This script will automatically separate the downloaded PDB files, the cleaned PDB files and the result data files into three separate directories.

1. Installations and set up:
   1. Installations of BASH, PyRosetta and Python:

PyRosetta can only be run via a BASH system. If a Linux computer is not available, BASH can also be installed on Windows 10. The instructions to install BASH, PyRosetta and Python can be found on <https://www.pyrosetta.org/downloads/windows-10>

* 1. Setting up the Path to PyRosetta program:

This setup is to let BASH and Python know where PyRosetta is, so it does not have to run the program in the same directory as PyRosetta.

To do this:

* + 1. Start a BASH terminal.
    2. Type cd and hit enter.
    3. Type vi .bashrc and hit enter.
    4. Hit shift+G to reach the bottom of the file.
    5. Hit i to start editting.
    6. To the bottom of the file type export PYTHONPATH=$PYTHONPATH:/path/to/the/directory/of/PyRosetta/

Example:

export PYTHONPATH=$PYTHONPATH:/mnt/c/Users/me/Desktop/Pyrosetta4

* + 1. Hit Esc and type :wq to save and exit out of the file.
    2. Type souce .bashrc
  1. Install Pandas:

Pandas is a Python package for manipulating Excel files. To do this, execute ipython and type pip install Pandas

1. instruction to run the script:
   1. Start a BASH terminal.
   2. Change directory to where the protein\_info\_extract.py script is located by typing

cd /path/to/the/script/

* 1. Execute ipython
  2. Run the script with input arguments
     1. If only one structure is required to be calculated, then just pass the 4-letter PDBID as the argument.

Example: in the terminal type run protein\_info\_extract.py 4M7T

* + 1. By default, this script will calculate the parameters of all protein chains available in the input PDB file. However, to save time a "homo" argument can be passed to only calculate one protein chain (chain A) of the input PDB file. This assumes all protein chains are homo-multimer (all subunits are the same).

Example: run protein\_info\_extract.py 4M7T homo

* + 1. Multiple PDBIDs can be stored in one .csv file. If this .csv file is passed as one of the arguments, this script will process the parameters of each PDBID indicated in the file automatically (run time varies based on the amount of the PDBIDs in the file). The .csv file should be in the following format:

|  |  |
| --- | --- |
|  | A |
| 1 | PDBID |
| 2 | 4R33 |
| 3 | 1TV8 |
| 4 | 2A5H |

Example: In the terminal type run protein\_info\_extract.py Input\_PDB\_ID.csv

run protein\_info\_extract.py Input\_PDB\_ID.csv homo

Reference:

1. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 22 (12): 2577–637. December 1983. [doi:10.1002/bip.360221211](https://onlinelibrary.wiley.com/doi/10.1002/bip.360221211)
2. Lee, B; Richards, FM. The interpretation of protein structures: estimation of static accessibility. 1971. J Mol Biol. 55 (3): 379–400. February 1971. [doi:10.1016/0022-2836(71)90324-X](https://www.sciencedirect.com/science/article/abs/pii/002228367190324X?via%3Dihub)
3. Shrake, A; Rupley, JA. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J Mol Biol. 79 (2): 351–71. September 1973. [doi:10.1016/0022-2836(73)90011-9](https://www.sciencedirect.com/science/article/abs/pii/0022283673900119?via%3Dihub)
4. Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, Shapovalov MV, Renfrew PD, Mulligan VK, Kappel K, Labonte JW, Pacella MS, Bonneau R, Bradley P, Dunbrack RL Jr, Das R, Baker D, Kuhlman B, Kortemme T, Gray JJ. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J Chem Theory Comput. 2017 Jun 13; 13(6):3031-3048. [doi: 10.1021/acs.jctc.7b00125.](https://pubmed.ncbi.nlm.nih.gov/28430426/) Epub 2017 May 12.
5. Park H, Bradley P, Greisen P Jr, Liu Y, Mulligan VK, Kim DE, Baker D, DiMaio F. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. J Chem Theory Comput. 2016 Dec 13;12(12):6201-6212. [doi: 10.1021/acs.jctc.6b00819.](https://pubmed.ncbi.nlm.nih.gov/28430426/) Epub 2016 Nov 7.