proteusPy: A Python Package for Disulfide Bond Analysis

4 July, 2023

Summary

proteusPy is a Python package specializing in the modeling and analysis of proteins of known structure with an emphasis on Disulfide Bonds. This package reprises my molecular modeling program `@Pabo_1986`, and relies on the Turtle3D class. The turtle implements the functions `Move`, `Roll`, `Yaw`, `Pitch` and `Turn` for movement in a three-dimensional space. The Disulfide class implements methods to analyze the protein structure stabilizing element known as a *Disulfide Bond*. This class and its underlying methods are being used to perform a structural analysis of over 35,800 disulfide-bond containing proteins in the RCSB protein data bank.

Virtual Environment Installation/Creation

1. *Install Anaconda (http://anaconda.org)*
2. *Build the environment.* At this point it's probably best to clone the repo via github since it contains all of the notebooks test programs and raw Disulfide databases. The source code distribution can be used from pyPi as a normal package, within your own environment.
   - Using pyPi:
     - python3 -m pip install proteusPy
   - From the gitHub repository:
     - Install git-lfs
       - https://help.github.com/en/github/managing-large-files/installing-git-large-file-storage
       - From a shell prompt:

```
$ git-lfs track "*.csv" "*.pkl" "*.mp4"
$ git clone https://github.com/suchanek/proteusPy/proteusPy.git
$ cd proteusPy
$ conda env create --name proteusPy --file=proteusPy.yml
$ conda activate proteusPy
```

```
$ pip install .
$ jupyter nbextension enable --py --sys-prefix widgetsnbextension
```

3. *Profit!* OK, just kidding. I hope you enjoy using proteusPy and would love to hear any success/insights gleaned from it. The Disulfide database is unique as far as I know, and is ripe for mining.

General Usage

Once the package is installed one can use the existing notebooks for analysis of the RCSB Disulfide database. The `notebooks` directory contains all of my Jupyter notebooks and is a good place to start. The `DisulfideAnalysis.ipynb` notebook contains the first analysis paper. The `programs` subdirectory contains the primary programs for downloading the RCSB disulfide-containing structure files, (`DisulfideDownloader.py`), extracting the disulfides and creating the database loaders (`DisulfideExtractor.py`) and cluster analysis, (`DisulfideClass_Analysis.py`).

The first time one loads the database via `Load_PDB_SS()` the system will attempt to download the full and subset database from my Google Drive. If this fails the system will attempt to rebuild the database from the repo's `data` subdirectory (not the package's). If you've downloaded from github this will work correctly. If you've installed from pyPi via `pip` it will fail.

The Future

I am continuing to explore the initial disulfide structural classes gleaned from Hogg *et al.* further using the sextant class approach. This offers much higher class resolution and reveals subgroups within the broad class. I'd also like to explore the catalytic and allosteric classes in more detail to look for common structural elements.

Publications

- https://doi.org/10.1021/bi00368a023
- https://doi.org/10.1021/bi00368a024
- https://doi.org/10.1016/0092-8674(92)90140-8
- http://dx.doi.org/10.2174/092986708783330566

NB: This distribution is being developed slowly. proteusPy relies on my fork of the `Bio` Python package to download and build the database. As a result, one can't download and create the database locally unless the BioPython patch is applied. The changed python file is in the repo's data directory - `parse_pdb_header.py`. Database analysis is unaffected without the patch. Also, if you're running on an M-series Mac then it's important to install Biopython first, since the generic release won't build on the M1. 7/4/23 -egs-

Eric G. Suchanek, PhD., mailto:suchanek@mac.com