**Training plan: scanning for proline-compatible positions in the protein structure**

Hongjun Bai @ VGS, MHRP

June 28, 2020

**[Prerequisite knowledge/skills]**

**Basics of amino acid, protein and protein structure**

Key concepts: amino acid, side chain, classification of amino acids, peptide bond, **alpha-helix, parallel/antiparallel beta sheet, primary/secondary/tertiary/quaternary structure**, non-covalent interactions (hydrophobic interactions, hydrogen bonds, salt-bridge, van der Waals interaction), disulfide bond

* <https://www.khanacademy.org/science/biology/macromolecules#proteins-and-amino-acids>
* <https://www.khanacademy.org/test-prep/mcat/biomolecules#amino-acids-and-proteins1>
* dihedral angles in the backbone: Ψ and Φ
* Accessible surface area (ASA) of residues

**Necessary computational skills**

* Operation system: Mac OS or Linux
  + Ubuntu/Debian
  + The file tree
* Command line environment – useful commands (not a complete list)
  + echo, ls, cp, mv, mkdir, find, history, pwd, ln, du, df, top, kill, sudo, su, ps, chmod, less, man, which
  + wc, head, tail, cat, grep, diff, sort, uniq, split, sed, cut
  + git, pacman/apt-get, nano/vim, awk
* Proficiency in python programming language
  + python 3.8: <https://docs.python.org/3.8/tutorial/>
* PyCharm, Git, GitHub
  + https://clt.champlain.edu/knowledgebase/configuring-git-with-pycharm/
  + Install git program
    - Necessary commands - diff, clone, revert, commit, push, pull
  + Create a GitHub account, set up a repository, integrate the Git, GitHub, PyCharm and practice the usage
    - Same time while learning python unittest, implement warm up task 1 and 2.
* Test driven development
  + Python unittest
    - <https://www.digitalocean.com/community/tutorials/how-to-use-unittest-to-write-a-test-case-for-a-function-in-python>
  + Warm up task 1 and 2 as test cases
* Necessary python models
  + Biopython PDB
    - <http://biopython.org/DIST/docs/tutorial/Tutorial.html#sec182>
    - <https://biopython.org/wiki/The_Biopython_Structural_Bioinformatics_FAQ>
  + numpy
  + Matplotlib
    - <https://matplotlib.org/>
  + Pandas

**[Additional bioinformatics tools and online resources]**

* pymol
  + windows install: <https://pymolwiki.org/index.php/Windows_Install>
  + Tutorial part 1: <https://www.youtube.com/watch?v=ai7p9Neguks>
  + Tutorial part 2: <https://www.youtube.com/watch?v=uxa-9UYnIAw>
* RCSB PDB
  + URL: <https://www.rcsb.org/>
  + PDB format: <https://www.cgl.ucsf.edu/chimera/docs/UsersGuide/tutorials/pdbintro.html>
* DSSP – secondary structure assignment
  + Maybe too much work to get it installed on windows, use the webserver instead.
  + <https://swift.cmbi.umcn.nl/gv/dssp/DSSP_3.html>
  + <https://github.com/PDB-REDO/dssp>
  + Webserver: <https://www3.cmbi.umcn.nl/xssp/>
* Aliview, the alignment viewer
  + URL: <https://ormbunkar.se/aliview/>

**[The project]**

**scanning for proline-compatible positions in the protein structure**

**Timeline** (to be considered)

2-3 weeks: learn the basics, lay out the problem

3-4 weeks: working on the problem

1-2 weeks: wrap up / poster presentation

**TASKS**

Warm up tasks (to be updated):

1. Function to calculate distance between two points/atoms in 3D space
2. Function to calculate dihedral angle for four given points/atoms in 3D space
3. Clean up non-amino acid residues in a PDB structure
4. Read and parse PDB, get the xyz of a given atom in each residue
5. Calculate Ψ and Φ angles for a given site in the structure
6. Read and parse DSSP output
   * Extract accessible surface area (ACC/ASA), phi, psi and secondary structural assignment a specific site

Tasks

* Superimpose two residues by their backbone atoms (CA, C, N and O)
  + Bio.PDB.Superimposer
  + Test function: compare with superimposition result from pymol (or a man-made example)
  + RMSD for the evaluation of superimposition
    - <https://en.wikipedia.org/wiki/Root-mean-square_deviation>
* Typical proline conformations in high resolution structures. Extract proline residues from a given structure (7DWY), check
  + Phi, psi angles (two different combinations)
  + Side chain atom positions (two different positions)
  + Save the information
    - Typical structures
    - A plot of proline conformations: 3D plot, x - Phi, y - psi, and z - sidechain RMS; expecting four clusters.
* Get proline-compatible positions
  + Trans proline phi-psi contour (Ramachandran plot):
    - [rama8000-transpro.data](https://github.com/rlabduke/reference_data/blob/master/Top8000/Top8000_ramachandran_pct_contour_grids/rama8000-transpro.data)
    - data source
      * Publication: <https://pubmed.ncbi.nlm.nih.gov/27018641/>
      * Data: <https://github.com/rlabduke/reference_data>
* Mutate a given site to proline
* Calculate the cost of the proline replacement
  + Count number of neighbor atoms of a given residue
* Compare automatically identified sites to the 6 proline locations of SARS-CoV-2 splike-6P

**Routine meeting:**

Daily (Mon-Thu, 10 am)