3/3/2020 Week 8 Module 1
Analysis of molecular dynamics simulations 1

- This module will consist of
  - an explanation of key principles in the analysis of biological MD simulations
  - a tour of an analysis of a ubiquitin simulation
    - based on the python package MDAnalysis
    - that you can follow for your own system
- At the end of this module, you should be able to answer the following questions:
  - What is MD used to calculate?
  - What is equilibration? How is the equilibration time determined?
  - Why is structural alignment helpful? How is it done?
  - What is principal components analysis? What is one way to do it?
  - What is clustering and why is it useful? What is one way to do it?
- You should also be able to visualize a MD trajectory in VMD. Hopefully you can modify the scripts that I used to analyze your own systems.

What is MD used to calculate?

- MD simulations may be used to
  - predict events or sequence of events that are physically possible
  - estimate statistical averages of
    - configurational properties, e.g.
      - average distance between two residues
      - histogram of an angle between three domains
      - populations of certain conformations
    - rates
- Statistical estimation is based on the assumption of ergodicity - that the time average is equal to the ensemble average

Installing MD analysis packages w/ conda

- conda create --name mdanalysis
- conda activate mdanalysis
- conda config --add channels conda-forge
- conda install jupyter pandas mdanalysis pymbar
  - juypter - for interactive coding notebooks
  - pandas - for data analysis
  - mdanalysis - for loading and analyzing MD trajectories
  - pymbar - for calculating free energies. also contains equilibration detection.

What is equilibration?

- The early part of a simulation is biased by the initial configuration
  - macromolecule structures start in a local minimum
  - water and ions are placed somewhat arbitrarily
  - box size is probably too large
- Equilibration is the time a system takes to reach a representative configuration
- Generally, simulation results from the equilibration period are ignored
- The samples actually used to calculate averages are known as the production

- See Equilibration.ipynb, which illustrates these points for a simulation of ubiquitin.

How is equilibration time determined?

- Arbitrarily
- Once a key property is stabilized
- By maximizing the effective sample size [Chodera, 2016]
  - A short equilibration leads to a long estimate of the time for the sample to be independent
  - A long equilibration reduces the number of samples
- Equilibration time
  - may look different for different properties
  - if properties are independent, slow equilibration of one may not affect estimation of another

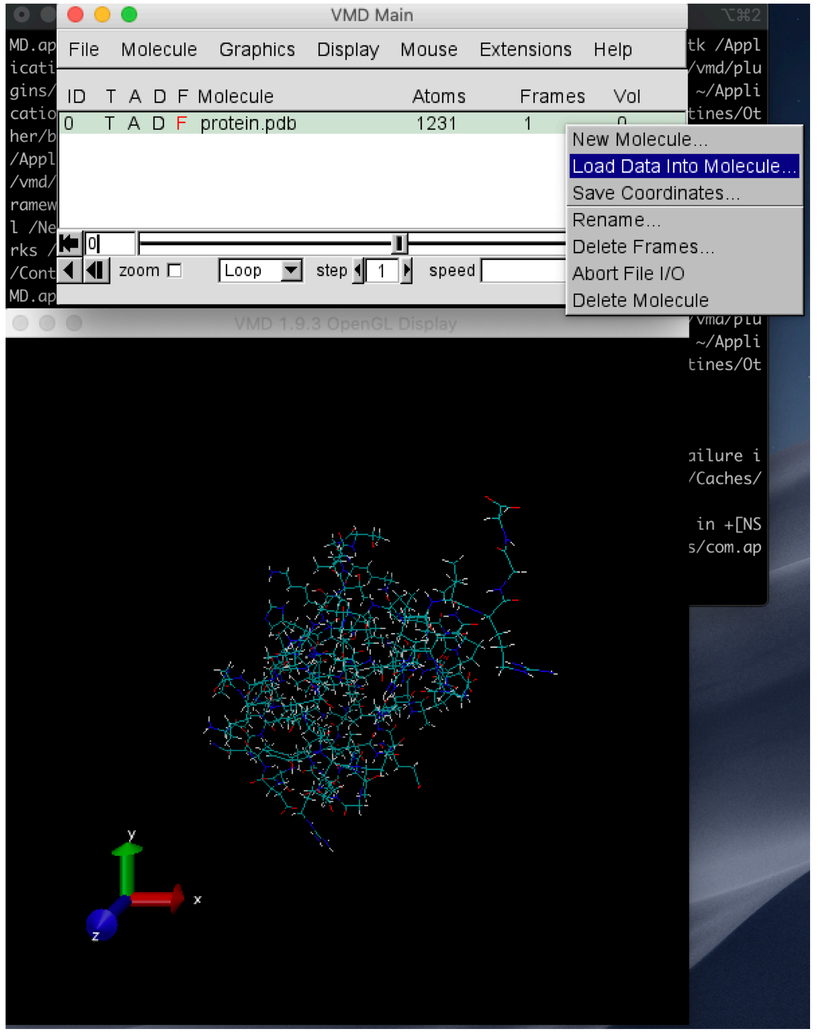
Structural alignment

- Other than identifying equilibration time, structural alignment is usually one of the first tasks of MD analysis
- Why?
  - In most MD simulations, molecules freely diffuse around the box
  - We are usually
    - uninterested in the overall translation and rotation,
    - interested in fluctuations relative to the macromolecule
- Alignment is often based on a rigid-body translation and rotation to minimize the root mean square deviation (RMSD)

- See Alignment.py, which performs a structural alignment for a series of simulations of ubiquitin and outputs a trajectory of the protein by itself.

Visualizing Trajectories

- can be done by
  - loading a model into VMD
  - loading the trajectory into the model
- I will show you unaligned and aligned trajectories of ubiquitin without water
- For an unaligned simulation in explicit solvent
  - molecules, especially water, can be split across a periodic box
  - you probably don't need to see all water



Root mean square analysis

$$\text{RMSD}(v, w) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} ||\, v_i - w_i \,||^2}$$

- Root mean square deviation (RMSD)
  - describes the difference between two structures
  - usually based on a subset of atoms
  - *i* is an index over atoms
- Root mean square fluctuation (RMSF)
  - describes the fluctuations of a specific atom, e.g. alpha carbon, over the course of a simulation
  - usually described per residue, identifying relatively flexible regions of a protein
  - *i* is an index over configurations
- Both require structural alignment

- See RMS.ipynb, which shows different types of RMS analysis for a simulation of ubiquitin.

References

- Chodera, J. D. A Simple Method for Automated Equilibration Detection in Molecular Simulations. Journal of Chemical Theory and Computation 2016, 12 (4), 1799–1805. https://doi.org/10.1021/acs.jctc.5b00784.

Some other software

- MDTraj: http://mdtraj.org/1.9.3/index.html
- ProDy: http://prody.csb.pitt.edu/tutorials/trajectory_analysis/trajectory.html