

Project topics

CS/BioE/CME/Biophys/BMI 279

Oct. 22, 2020

Ron Dror

Project Guidelines

- **You are welcome (and encouraged) to pick something that is not mentioned in this presentation**
- Projects should involve structure or spatial organization at a molecular or cellular level
 - Machine learning projects are great, as long as 3D structure or spatial organization factor in explicitly
 - Image analysis is great, as long as images are at a molecular or cellular level

Project Guidelines

- Projects can be methods-focused and/or application-focused
 - I.e., you can code/modify software, or apply existing software to a biological problem
 - You could also carefully compare the performance of several algorithms or software packages
- See project assignment sheet on website for details on project writeup and other information
 - Joint projects, overlaps with projects for other classes, etc.

Protein structure prediction

- Sample codes:
 - Rosetta/PyRosetta (or Robetta webserver)
 - Phyre2 (web server)
 - I-Tasser (web server or download code)
 - Modeller (web server called ModWeb, or download code)
 - SWISS-MODEL (web server)
- Topics of interest include
 - Structure prediction methodology
 - Structures of proteins of interest
 - Effects of protein modification (e.g., mutation, phosphorylation)
- Related: RNA structure prediction
 - RNAComposer web server:
<http://rnacomposer.cs.put.poznan.pl/>

Molecular dynamics simulation

- You can write your own code
 - Don't resubmit code you wrote for BMI/BIOE/GENE 214/CS 274, but you can build on it. For example, increase speed (fast electrostatics methods), improve integrators, add restraints/constraints or other features
- Or use existing software
 - GROMACS, Desmond, NAMD, AMBER (PMEMD module): designed for performance.
 - GROMACS, Desmond, and NAMD are free (for academic use); AMBER is not
 - Tinker—slow, but designed to be easy to work with the code (also free)
 - Most of these are designed for Linux, but Windows and Mac executables are available for Tinker
- Either focus on simulation of a particular molecule, or on methods (e.g., molecular dynamics vs. Monte Carlo)

Protein Design

- Rosetta software is free for academic use
- Rosetta Design server:
<http://rosettadesign.med.unc.edu/>

Image analysis

- Image classification, segmentation, or denoising; disease diagnosis; cell counting; measurement of protein concentration/localization, and detection of colocalized proteins of different types; or other useful tasks
 - Project should involve cellular or molecular images (as opposed to traditional medical MRI or x-ray images, for example)
- Useful software:
 - Matlab (general-purpose; available on LTS machines)
 - ImageJ (free, widely used for biological image processing)
 - CellProfiler (free, includes machine learning applications)
- Or write your own software
 - For machine learning projects, consider using libraries such as TensorFlow or PyTorch

Image analysis

- Sample image sets:
 - <https://data.broadinstitute.org/bbbc/>
 - <http://www.cellprofiler.org/>
 - <https://www.kaggle.com/iarunava/cell-images-for-detecting-malaria>
 - <https://www.kaggle.com/paultimothymooney/blood-cells>
 - <https://www.proteinatlas.org/about/download>
 - Please let me know of other good ones you find!

Reaction-diffusion simulation

- Use existing codes:
 - MCell, Smoldyn, Simmune
 - For MCell, consider using CellOrganizer or CellBlender to make models or renderings
- Write your own code
- Build a model of a cellular process, or consider methodological issues

Ligand docking and virtual screening

- Established, free codes and web servers:
 - Autodock Vina
 - SwissDock
- Rosetta Dock (newer; can use in PyRosetta framework)
- GLIDE: Powerful commercial software, for which Stanford now has a university-wide license
 - See <https://library.stanford.edu/science/software/schrodinger>
 - You can also access other structural modeling software from the same company
- ZINC ligand database: <http://zinc.docking.org/>
- Related: Protein–peptide docking (e.g., with FlexPepDock or Backrub servers) or protein–protein docking (e.g., with ZDock, Haddock)

Crystallography

- Structure factors (i.e., primary crystallographic data) are often available in PDB.
 - See http://www.rcsb.org/pdb/101/static101.do?p=education_discussion/Looking-at-Structures/structurefactors.html
- CNS software (<http://cns-online.org/v1.3/>)

Single-particle electron microscopy

- Software packages:
 - XMIPP: has a graphical user interface, somewhat easier to use
 - Relion: more mathematically sophisticated (Bayesian methods)
- Most use MPI, which complicates installation.
- Alternative: implement something yourself
 - Work in two dimensions for simplicity
 - Or tackle early stages in single-particle EM pipeline, such as particle picking

Other topics

- CellPack (<http://www.autopack.org>): packing proteins into a cell
- Coarse-grained simulation (e.g., assembly of a viral capsid; consider HOOMD-blue software)
- EVFold (<http://evfold.org/>): protein structure prediction based on covariation across sequences
 - Also see “Distance-based protein folding powered by deep learning (<https://www.pnas.org/content/116/34/16856.short>)”
- Secondary structure prediction from sequence (e.g., using a neural network)
 - Some data sets:
<https://www.princeton.edu/~jzthree/datasets/ICML2014/>