Project topics

CS/BioE/CME/Biophys/BMI 279
Nov. 9 and 14, 2023
Ron Dror

Project Guidelines

- You are most welcome to pick something that is not mentioned in this presentation
- **The key requirement for a project topic is that it should involve 3D structures of biomolecules and/or spatial organization of molecules within a cell**
  - Of course, it should also involve (or at least relate to) computation
  - Machine learning projects are great, *as long as they explicitly involve 3D structures of biomolecules or spatial organization of molecules within a cell*
  - Image analysis is great, *as long as the analysis you do provides information 3D structures of biomolecules and/or spatial organization of molecules within a cell*

Project Guidelines

- Projects can be methods-focused and/or application-focused
  - You can code/modify software, or apply existing software to a biological problem
  - You could also carefully compare the results/accuracy of several algorithms or software packages
- You can work individually, or in groups of 2 or 3
  - In any case, the amount of work *per person* should be similar to assignments 2 or 3.
- See Project Instructions document on website for details on project writeup and other information
  - Group projects, overlaps with projects for other classes, etc.
  - Expected amount of work and deliverables

Protein structure prediction

- Potential topics include:
  - Structure prediction methodology
    - ProteinNet provides some datasets for protein structure prediction benchmarking: https://www.tensorflow.org/datasets/catalog/protein_net
  - Structures of proteins (or protein complexes) of interest
  - Effects of protein modification (e.g., mutation)

Protein structure prediction

- Sample codes and servers:
  - Rosetta/PyRosetta (or Robetta webserver)
  - Phyre2 (web server)
  - Modeller (web server called ModWeb, or download code)
  - SWISS-MODEL (web server)

- Related: RNA structure prediction
  - RNAComposer web server: http://rnacomposer.cs.put.poznan.pl/

Protein structure prediction

- Recent deep-learning methods for protein structure prediction
  - ColabFold package provides a fast Python-based interface for several of these, including AlphaFold 2 and ESMFold (which is based on large language models):
    - https://github.com/sokrypton/ColabFold
    - https://www.nature.com/articles/s41592-022-01488-1
  - Robetta webserver allows use of RoseTTAFold, but multiple sequence alignments must be provided (computed separately)
  - The COSMIC2 server allows one to run AlphaFold 2, ESMFold, and other software for protein structure prediction and cryo-EM data analysis
    - https://cosmic2.sdsc.edu/

Molecular dynamics simulation

- Focus either on simulations of particular molecules, or on methods (e.g., molecular dynamics vs. Monte Carlo)
- Existing software
  - GROMACS, Desmond, NAMD, AMBER (PMEMD module): designed for performance.
    - GROMACS, Desmond, and NAMD are free (for academic use); AMBER is not
    - Desmond can be run through the Schrodinger Maestro graphical user interface
  - 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
- You can write your own code
  - Don't resubmit code you wrote for CS 274 (BIOE/BMI/GENE 214), but you can build on it. For example, increase speed (fast electrostatics methods), improve integrators, add restraints/constraints or other features. Or you could use Tinker for this.
- Note that most MD simulations take a long time to run. For short simulations, you could try the WebGRO server (not yet tested; feedback welcome!):
  - https://simlab.uams.edu/index.php

Protein Design

- Rosetta software is free for academic use
- Rosetta Design server: http://rosettadesign.med.unc.edu/
- ColabDesign:
  - https://github.com/sokrypton/ColabDesign
  - Supports several protein design tools (including RFDiffusion) in an interface similar to ColabFold

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)
  - **The analysis you do should provide information on 3D structures of biomolecules and/or spatial organization of molecules within a cell**
- Useful software:
  - Matlab (general-purpose; available on VPTL 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 PyTorch or TensorFlow

Image analysis

- Sample image sets:
  - https://data.broadinstitute.org/bbbc/
  - https://www.kaggle.com/paultimothymooney/blood-cells
  - https://idr.openmicroscopy.org/cell/
    - Specifically, Cell-IDR, not Tissue-IDR
  - https://www.ebi.ac.uk/empiar/EMPIAR-10592/
    - See https://elifesciences.org/articles/65894
  - https://thecellvision.org/cyclops/
    - See https://academic.oup.com/g3journal/article/5/6/1223/6025272
  - https://ssbd.riken.jp/database/
  - Please let me know of other good ones you find!

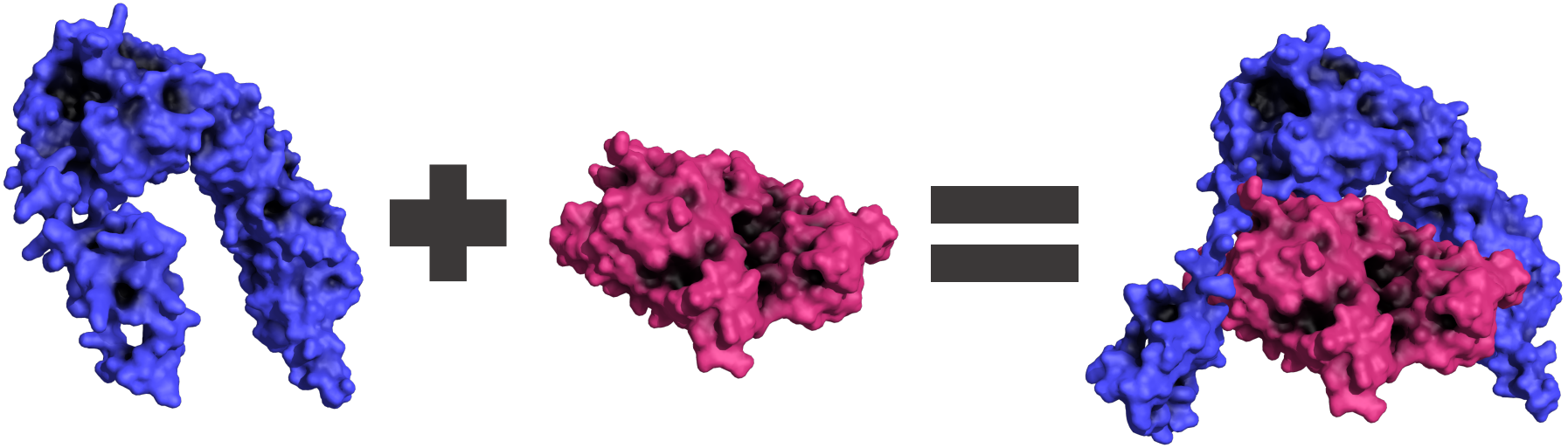
Reaction-diffusion simulation

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

Ligand docking and virtual screening

- Established, free codes and web servers:
  - Autodock Vina
  - SwissDock
- GLIDE: Powerful commercial software, for which Stanford now has a university-wide license
  - See https://guides.library.stanford.edu/c.php?g=1175377&p=9895759
  - You can also access other structural modeling software from the same company (Schrodinger); see the link above
- ZINC ligand library: http://zinc.docking.org/
- ChEMBL (a large dataset of experimentally measured ligand-binding properties): https://www.ebi.ac.uk/chembl/

Protein-protein docking



- Starting with existing structures:
  - ZDock, Haddock (use physics-based scoring functions)
- Starting without existing structures:
  - RoseTTAFold or AlphaFold 2
- Related: docking peptides to a target protein
  - FlexPepDock server
  - You could also try RoseTTAFold or AlphaFold 2

Crystallography

- Structure factors (i.e., primary crystallographic data) are often available in PDB.
  - See https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/structure-factors-and-electron-density
- Phenix software (https://phenix-online.org/)

Single-particle electron microscopy

- Software packages:
  - CryoSPARC (incorporates recently developed machine learning methods, and has a graphical user interface)
  - XMIPP (has a graphical user interface)
  - Relion (based on Bayesian methods)
  - Installation of these packages can be challenging
- Alternative: implement something yourself
  - Work in two dimensions for simplicity
  - Or tackle early stages in single-particle EM pipeline, such as particle picking

Some other machine learning projects

- Protein secondary structure prediction from sequence
  - Some data sets: https://www.compbio.dundee.ac.uk/jpred/about_RETR_JNetv231_details.shtml
- Machine learning on cellular/molecular images (see earlier slide)
- ATOM3D
  - A collection of ML tasks and datasets involving 3D molecular structure: https://www.atom3d.ai/
- Learning force fields
  - ANI-1ccx and ANI-1x data sets: https://www.nature.com/articles/s41597-020-0473-z

Other topics

- CellPack (https://www.autopack.org/): packing proteins into a cell
- Coarse-grained simulation (e.g., assembly of a viral capsid; consider HOOMD-blue software)
- EVCouplings (https://v2.evcouplings.org/): prediction of contacts in protein structures based on covariation across sequences
  - Also see “Distance-based protein folding powered by deep learning” (https://www.pnas.org/content/116/34/16856.short)