Project topics

CS/BioE/CME/Biophys/BMI 279
Nov. 12, 2024
Ron Dror

Project Guidelines

- You are most welcome to pick something that is not mentioned in these slides
- **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. Coding is not required.
  - Alternatively, you can carefully compare the results/ performance of several 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 report and other information
  - Including overlap with projects for other classes
  - Please include analysis/discussion of your results in the report

What if I don’t have an idea for a project?

- Subsequent slides list a number of possibilities
- We’ve posted examples of projects from previous years on the course website
- You can investigate any biological system (e.g., biomolecule, molecular system, cellular process) you happen to be interested in
- You could pick something based on the optional readings on the course website—some of which are review papers—may give you ideas

What if my project turns out to be too hard?

- You don’t need to “solve” the problem you set out to solve in order to get a good grade
- If things don’t work out, please explain in your report what you tried and what went wrong. Hypothesize about what the problem might have been and about how one might be able to overcome it in the future.

What if my project turns out to be too easy?

- If your initial plan requires very little effort, you can always find some next steps
  - For example, do some more sophisticated analysis of the results, or look at other biological systems

Computing resources

- Options for running software include:
  - Run on your own computer
  - Many tools are available as web servers
    - Some links are included in subsequent slides
  - Use FarmShare, particularly the oat and wheat systems
    - See https://uit.stanford.edu/service/sharedcomputing/environments
    - You can install software. Linux operating system.
  - If you wish to use other resources to which you have access (e.g., through a lab), that's also fine.

Biomolecular structure prediction

- Potential topics include:
  - Predict and analyze structures of biomolecules (or complexes) that you find interesting
  - Examine effects of protein modification (e.g., mutation)
  - Structure prediction methodology
    - Compare accuracy of different existing structure prediction packages (or variants of those methods)
    - Improve the scoring function developed in Assignment 2
    - ProteinNet provides some datasets for protein structure prediction benchmarking: https://github.com/aqlaboratory/proteinnet
- Not limited to proteins!

Biomolecular structure prediction web servers

- ColabFold provides a fast Python-based interface for several packages that predict structures of proteins (and, in many cases, protein complexes), including AlphaFold 2, RoseTTAFold 2, and ESMFold:
  - https://github.com/sokrypton/ColabFold
  - https://www.nature.com/articles/s41592-022-01488-1
- The COSMIC2 server also allows one to run deep-learning-based structure prediction algorithms
  - https://cosmic2.sdsc.edu/
- AlphaFold3 server: https://alphafoldserver.com/
  - also predicts structures for RNA, DNA, and more
- D-I-TASSER: https://zhanggroup.org/D-I-TASSER/
- Older servers (still useful, and include recent updates)
  - Phyre2: https://www.sbg.bio.ic.ac.uk/~phyre2/html/page.cgi?id=index
  - SWISS-MODEL: https://swissmodel.expasy.org/
  - Robetta: https://robetta.bakerlab.org/
  - ModWeb: https://modbase.compbio.ucsf.edu/modweb/
  - RNAComposer (for RNA): http://rnacomposer.cs.put.poznan.pl/
  - FlexPepDock (for docking peptides to proteins): http://flexpepdock.furmanlab.cs.huji.ac.il/

Biomolecular structure prediction codes

- Source code is available for nearly all the software packages mentioned on the previous slide.
  - Including Alpha Fold 3 (released yesterday!): https://github.com/google-deepmind/alphafold3
- One can also use Rosetta via the PyRosetta interface. See assignment 2 challenge question.
- These codes generally allow you to do more than the web servers. You can also modify them to experiment with algorithms or extensions.

Molecular dynamics simulation

- You could simulate particular biomolecules of interest
  - For example, beginning with a structure or structures from the PDB (e.g., different proteins, or the same protein bound to two different small molecules)
- Or you can explore methods
  - For example, test effects of changing simulation parameters, or compare molecular dynamics to Monte Carlo sampling

Molecular dynamics software

- Existing software
  - Free, high performance software
    - GROMACS: https://www.gromacs.org/
    - OpenMM: https://openmm.org/
    - NAMD: https://www.ks.uiuc.edu/Research/namd/
  - Desmond can be run through the Schrodinger Maestro graphical user interface.
    - See download instructions at: https://guides.library.stanford.edu/c.php?g=1175377&p=9895759. You'll probably need to wait a few days.
  - 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 many MD simulations take a long time to run
  - For short simulations, you could use the WebGRO server: https://simlab.uams.edu/index.php
  - For longer ones, consider FarmShare

Protein Design

- ColabDesign:
  - https://github.com/sokrypton/ColabDesign
  - Supports several protein design tools (including RFDiffusion and ProteinMPNN) in an interface similar to ColabFold
- Rosetta Design server (older): http://rosettadesign.med.unc.edu/
- See also:
  - Damietta – protein design toolkit (https://damietta.de/)
  - Chroma – generative model for protein design (https://generatebiomedicines.com/chroma)
  - Database of protein design tools: https://github.com/pansapiens/awesome-protein-design-software?tab=readme-ov-file#design

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 spatial organization of molecules within a cell or 3D structures of biomolecules**

Image analysis software

- Some widely used software:
  - Matlab (general-purpose; available on VPTL machines)
  - Fiji or ImageJ (free, widely used for biological image processing)
    - Links: https://imagej.net/software/fiji/, https://imagej.net/ij/
    - Fiji is built on ImageJ, but includes more libraries, and most people find it easier to use.
  - CellProfiler (free, includes support for various machine learning applications)
    - https://cellprofiler.org/
- You can also write your own software
  - For machine learning projects, consider using libraries such as PyTorch or TensorFlow

Cellular image datasets

- 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 specific cellular process, or consider methodological issues (e.g., effect of algorithm choice or parameters on speed or results)
- Write your own code, or use existing software packages:
  - MCell: https://mcell.org/
    - Consider using CellOrganizer or CellBlender to make models or renderings
  - Smoldyn: https://www.smoldyn.org/
  - Simmune: https://www.niaid.nih.gov/research/simmune-project
  - ReaDDy: https://readdy.github.io/
  - Simmune is concentration-based; the others listed above are particle-based but use a variety of algorithms
  - Many of these websites list examples of papers that have used the software to study specific cellular processes

Ligand docking and virtual screening

- Established, free codes: Autodock, Autodock Vina
- Web server: SwissDock
- Machine learning methods
  - AlphaFold 3 can be used for ligand pose prediction
  - GNINA: https://github.com/gnina/gnina
- GLIDE: Powerful commercial software, for which Stanford now has a university-wide license
  - See download instructions at: https://guides.library.stanford.edu/c.php?g=1175377&p=9895759. You'll probably need to wait a few days.
  - You can also access other structural modeling software from the same company (Schrodinger); see the link above

Ligand docking and virtual screening

- DUDE benchmark for virtual screening: https://dude.docking.org/
- ZINC library of purchasable small molecules: http://zinc.docking.org/
- ChEMBL (a large dataset of experimentally measured ligand-binding properties): https://www.ebi.ac.uk/chembl/

Crystallography

- Cambridge structural database: https://www.ccdc.cam.ac.uk/free-products/
  - Offers many free online tools and software to explore and manipulate crystallography data
  - Stanford also has the paid software licensed that students are able to download by following these instructions: https://guides.library.stanford.edu/scilib-software/csd-access
- Phenix software: https://phenix-online.org/
  - Used by many biomolecular crystallography labs
- 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

Cryo-electron microscopy

- Software packages:
  - CryoSPARC (incorporates recently developed machine learning methods, and has a graphical user interface): https://cryosparc.com/
  - Relion (based on Bayesian methods): https://relion.readthedocs.io/en/release-5.0/
  - CryoDRGN (machine learning method for predicting all likely conformations rather than just individual structures): https://ez-lab.gitbook.io/cryodrgn
- COSMIC2 web server: https://cosmic2.sdsc.edu/
- Alternative: implement something yourself
  - Work in two dimensions for simplicity
  - Or tackle early stages in single-particle EM pipeline, such as particle picking
    - CryoPPP dataset:https://www.nature.com/articles/s41597-023-02280-2, http://calla.rnet.missouri.edu/cryoppp, https://github.com/BioinfoMachineLearning/cryoppp
- Raw cryoEM data
  - https://www.ebi.ac.uk/empiar/

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
- ATOM3D
  - A collection of ML tasks and datasets involving 3D molecular structure: https://www.atom3d.ai/
- Neural-network-based 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://evcouplings.org/
  - "Protein and RNA structure, function and fitness predicted from evolutionary sequence covariation"

Some additional information (added in response to student questions)

- Consider the following alternatives to installing AlphaFold 3 yourself, to save some time:
  - Chai-Lab: https://github.com/chaidiscovery/chai-lab
    - Recommended for ease of installation, but requires a relatively recent GPU model (as does AlphaFold 3)
  - HelixFold3: https://github.com/PaddlePaddle/PaddleHelix/tree/dev/apps/protein_folding/helixfold3

Some additional information (added in response to student questions)

- Tools to estimate affinities (binding strengths) of two proteins
  - Prodigy: https://rascar.science.uu.nl/prodigy/
  - FoldX: https://foldxsuite.crg.eu/
  - Rosetta FlexDDG: https://pmc.ncbi.nlm.nih.gov/articles/PMC5980710/