Syllabus
Structural Bioinformatics
(BINF-M632)

## 1. Course Information

| | |
|---|---|
| Course# & Title: | BINF-M632: Structural Bioinformatics |
| Semester: | Spring 2023 |
| Meetings: | Wed & Fri, March 29 – May 5, 2023<br>10:30 am – 12:30 pm (Ethiopian Time)<br>https://udenver.zoom.us/j/89490432638 |

## 2. Faculty Information

| | |
|---|---|
| Instructor: | Shimelis Assefa, Ph.D. (Associate Professor) |
| Contact Information: | P. +251-911-340814 Email: Shimelis.Assefa@du.edu |
| Office Hours: | By appointment.<br>https://calendly.com/sassefa/schedule-meeting |

## 3. Course Description

The module aims to provide an in-depth understanding of ***proteins*** with an emphasis on ***structure*** and its connection to ***biological functions***. The module treats the principles that determine these properties and the methods that are used to study them within modem molecular protein science. Fundamentals of ***protein, DNA, RNA sequence***, ***structure and function databases***, experimental ***structure determination*** and structure analysis (***X-ray crystallography, NMR spectroscopy, Electron microscopy, angle X-ray scattering, neutron scattering*** etc.), Structure formats, ***molecular visualization***, calculation of ***RMSD*** and structure superposition, Protein structure evolution and fold classification (***CATH domain structure database***, ***SCOP database***, new methods), inference of function from structure, principle and application of machine learning and deep neural-network learning, ID prediction (***secondary structure***, solvent exposure, neural networks, HMMSTR) & sampling of protein conformations (TorusDBN, directional statistics), methods of ***protein folding*** and ***protein structure prediction*** (homologous modeling & the loop closure problem, threading and ab initio folding), basics of molecular dynamics and Monte Carlo simulations, knowledge based energy functions, de novo prediction methods (ROSETTA, LINUS), ***identifying structural domains in protein***, prediction of protein-protein interaction from phylogenetic analysis, advanced ***sequence and structure comparison and alignment methods,*** structural bioinformatics in ***drug discovery, CASP and CAFASP experiments***, fold recognition methods, AB INITIO

methods. Computer exercises on peptide structure determination based on NMR, X-ray crystallography or EM data, folding of RNA using different time and length scales, binding and folding of ligands and simulation of peptide folding. RNA 2D and 3D structure prediction. ***Python based practicals*** using ***Biopython's Bio.PDB*** module.

## 4. Course Materials

**Recommended Book:**

There is no required textbook for this course. I will compile relevant materials – primarily from widely adopted web-based resources, tools, databases, and software..

A few relevant web-based texts and materials that we rely on include:

- Biopython Tutorial and Cookbook - https://biopython.org/DIST/docs/tutorial/Tutorial.html
- PDB - https://www.rcsb.org
- Uniprot - https://www.uniprot.org/
- Proteopedia - https://proteopedia.org
- PyMOL - a molecular visualization tool - https://pymol.org/2/
- PyMOLWiki - https://pymolwiki.org/
- ChEMBL - https://www.ebi.ac.uk/chembl/
- AlphaFold protein structure database - https://alphafold.ebi.ac.uk/
- HADDOCK - https://wenmr.science.uu.nl/haddock/
- SWISS-MODEL - https://swissmodel.expasy.org/
- The National Center for Biotechnology Information - https://www.ncbi.nlm.nih.gov/
- Fundamentals of Molecular Structural Biology – Multiple chapters from the book
- Algorithms and Methods in Structural Bioinformatics – Complete Book
- The Biology Notes - https://thebiologynotes.com/

## 5. Learning Outcomes

On successful completion of this module, the students will be able to:

- Explain the relation between sequence, structure and function.
- Describe the structural basis for macromolecular dynamics, binding specificity and catalysis.
- Have overview of biological databases, servers and information centers.
- Understand the basic macromolecular structure: three-dimensional structure, PDB co-ordinates, classification of proteins in structure families, programs for analysis and comparison of structures.
- Perform sequence analysis for prediction of secondary and tertiary structures, and homology models of three-dimensional structures based on

sequence data.

- Introduce the theory of classification and comparison of sequences & structures and extraction of common distinctive features (e.g., motifs).
- Know basic aspects of protein physics (folding, physical forces, thermodynamics, statistical physics)
- Know goals and methods of ID structure prediction and parameterization (secondary structure, backbone structure, solvent exposure)
- Know the basics of probabilistic structure prediction.
- Describe how structure translates into function within different biological fields such as catalysis, transport and regulation.
- Estimate the validity of information in macromolecular structure databases and use computer programs to model and analyze macromolecular structures from a functional perspective.
- Implement the RMSD algorithm in mathematical detail.
- Obtain insight in and background for bioinformatic methods for RNA secondary structure prediction.
- Make computer simulations for small biomolecules such as RNA and peptides.
- Know about structural (2D) RNA alignments.
- Understand RNA 3D structure and 3D aspects of RNA structure prediction.
- Interpret the results of RNA structure prediction methods and relate them to that of prediction on "background" sequence.
- Know basic aspects of ligand binding and computer-aided structure-based design of drugs.
- Understand about protein dynamics and flexibility.
- Describe protein structures and 3D-protein models: three-dimensional protein structure alignments and usage of structural databases, like CATH, SCOP, FSSP and MMDB.
- Implement structural bioinformatics algorithms in Python and Biopython's Bio.PDB module.
- Visualize and analyze biomolecular structures using PyMOL.
- Describe how protein folding happens from both an energetic and a structural perspective.
- Describe how protein structure can be determined using x-ray scattering or nuclear magnetic resonance (NMR) experiments.
- Analyze a protein structure with relevant methods and algorithms.
- Analyze a problem in structural bioinformatics and outline a matching algorithm.
- Understand and critically assess scientific literature that treats protein structure and function.

## 6. Course Contents

- 3/29: Introduction to Structural Bioinformatics
  - Broad overview of structural bioinformatics and its applications
  - Overview of molecular biology
- 3/31: Protein Structure and Functions
  - Fundamentals of protein, DNA, RNA sequence
  - Protein and nucleic acid structure and functions
  - Structure and function databases
  - Inference of function from structure
  - FirstGlance in Jmol
- 4/5: Sequence Analysis and Structure Determination
  - Structure databases and search algorithms
  - Protein structure prediction
  - Experimental structure determination and structure analysis
  - CASP and CAFASP experiments
- 4/7: Structure Analysis and Visualization
  - Structure formats
  - Protein structure analysis
  - Molecular visualization, PyMOL
  - Calculation of RMSD and structure superposition
- 4/12: Protein Structure Evolution and Fold Classification.
  - SCOP database - The Structural Classification of Proteins
  - Identifying structural domains in protein
  - CATH structural hierarchy – CATH DB
- 4/14: Methods of Protein Folding.
  - AB Initio, de novo
  - Hierarchical folding
  - AlphaFold protein structure database
- 4/19: Molecular Dynamics Simulation
  - Basics of molecular dynamics and Monte Carlo simulations
  - Force field in molecular dynamics
  - Simulation setup and system preparation
- 4/21: Molecular Modeling
  - Homology modeling
  - Molecular docking
  - SWISS-MODEL
- 4/26: Protein-Protein and Protein-Ligand Interactions
  - Protein-protein interaction
  - Protein-ligand interaction
  - HADDOCK (High Ambiguity Driven protein-protein DOCKing)
- 4/28: Molecular Modeling in Drug Discovery
  - Structural bioinformatics in drug discovery
  - Tools and resources for drug discovery: ChEMBL

- 5/3: Principle and Application of Machine Learning and Deep Neural-Network Learning
    - AI-predicted protein structures Database AlphaFold DB
- 5/5: Course Review and Wrap up.

## 7. Methods of Assessment

Class meets twice a week on Wednesday and Friday morning, from 10:30am – 12:30pm. We can alternate in-person meeting with online meeting (as needed) via zoom (https://udenver.zoom.us/j/89490432638). Course materials, and assignments are published on the course website. Please review descriptions of individual assignments that are available here on the syllabus and under the Assignments page on the website.

**Points Possible:**

| Assignments | Weight (percentage) | Points |
|---|---|---|
| Class Participation | 15% | 100 |
| Assignment 1: Sequence Alignment | 20% | 100 |
| Assignment 2: Homology Modeling | 30% | 100 |
| Assignment 3: Molecular Dynamic Simulation | 35% | 100 |
| Total | 100% | 400 |

**Evaluation:** Grades will be based on points accumulated and converted to 100 percentiles. Letter grades will be awarded according to the Addis Ababa University scale.

## 8. Descriptions of Assignments

**Class participation, 15%**

I Sincerely encourage you to actively participate during class sessions. Active participation can take many different forms – going through the course content before the class, asking questions before, during, and after the each class sessions, and helping answer questions, as well as engaging with me and fellow classmates.

---

**Assignment 1. Sequence Alignment: 15% due date**

1. Go to NCBI Protein and search for the following protein: NP_188876
   - What is the protein name?
   - From which organism is it from?
   - What is its sequence (in FASTA format)

1.1 Go to the FASTA algorithm homepage and perform a protein:protein alignment with this sequence against the swissprot database.

1.2 Go to the BLAST homepage and perform a protein BLAST with this sequence against the swissprot database.

1.3 Can you observe differences between the FASTA and BLAST results? If so, how can you explain them?

1.4 Select all the proteins resulting of the BLAST query and download the file with all their FASTA sequences.
1.5 Go to the ClustalW2 program (e.g., https://www.genome.jp/tools-bin/clustalw) and perform a multiple sequence alignment with the default options using this file as input.
1.6 What can you observe? What does it mean from a biological point of view? Can you make a link between your observations and what we know about the protein we used at the beginning?

**Assignment 2. Homology Modeling: 25%, due date ??**
Proteins that have diverged from a common ancestral gene are known as homologous. Proteins with similar sequences are assumed to be homologous and usually (within certain limits) have similar structures and functions.

If you have the sequence of your target protein, and some hits from your template search, your next step is to use an **homology modelling** software to generate a model of your target protein.

SWISS-MODEL (http://swissmodel.expasy.org) is a web server for homology modelling. You input an alignment of the target with a template of choice, and SWISS-MODEL generates the model according to the alignment. ("Alignment mode"). It provides the following two modes of usage:

- **Automated mode**: Input sequence of target protein. SWISS-MODEL searches for template, performs alignment and then homology modelling for your target protein.
- **Alignment mode**: Input sequence alignment of target protein against a template of your choice (searched through e.g. BLAST). SWISS-MODEL models your target protein according to the alignment you provide.

To continue the task in the 'alignment mode,' you first need to retrieve the template you have chosen in the BLAST search.

Then perform a sequence alignment of your target protein against the template. This would be the instruction for SWISS-MODEL to generate a model for your target.

Use **T-COFFEE** for sequence alignment. It is available at:

- http://tcoffee.crg.cat/
- https://www.ebi.ac.uk/Tools/msa/tcoffee

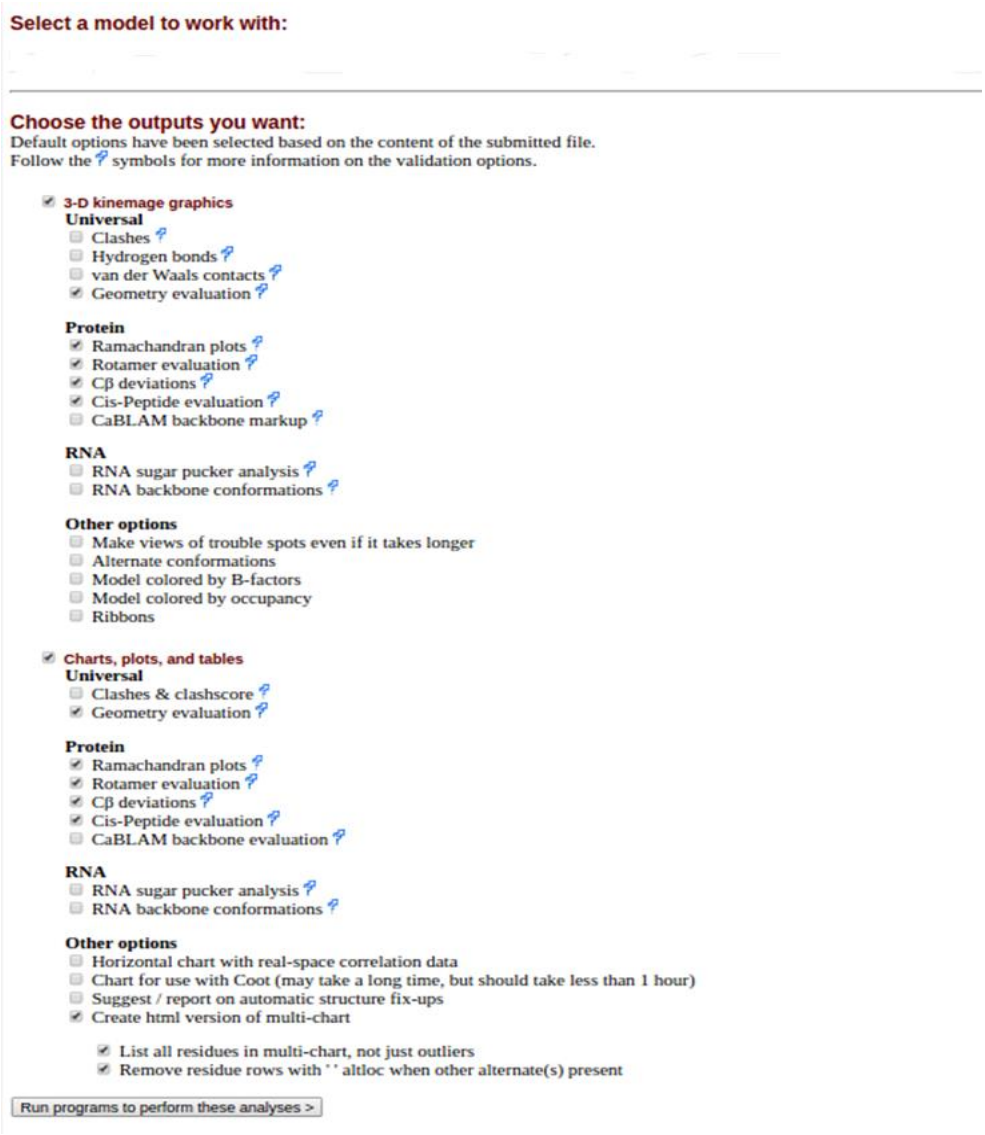
once the sequence and the template alignment instruction is submitted in T-Coffee, download the fasta_aln file from the result page.
Submit the alignment file from T-Coffee (just downloaded) to **SWISS-MODEL (https://swissmodel.expasy.org)** and choose **Target-Template Alignment** on the right.

Assess the model quality of the built model. Download the model and save it on your computer.

Now, submit the model structure to the MolProbity server for quality assessment - http://molprobity.biochem.duke.edu

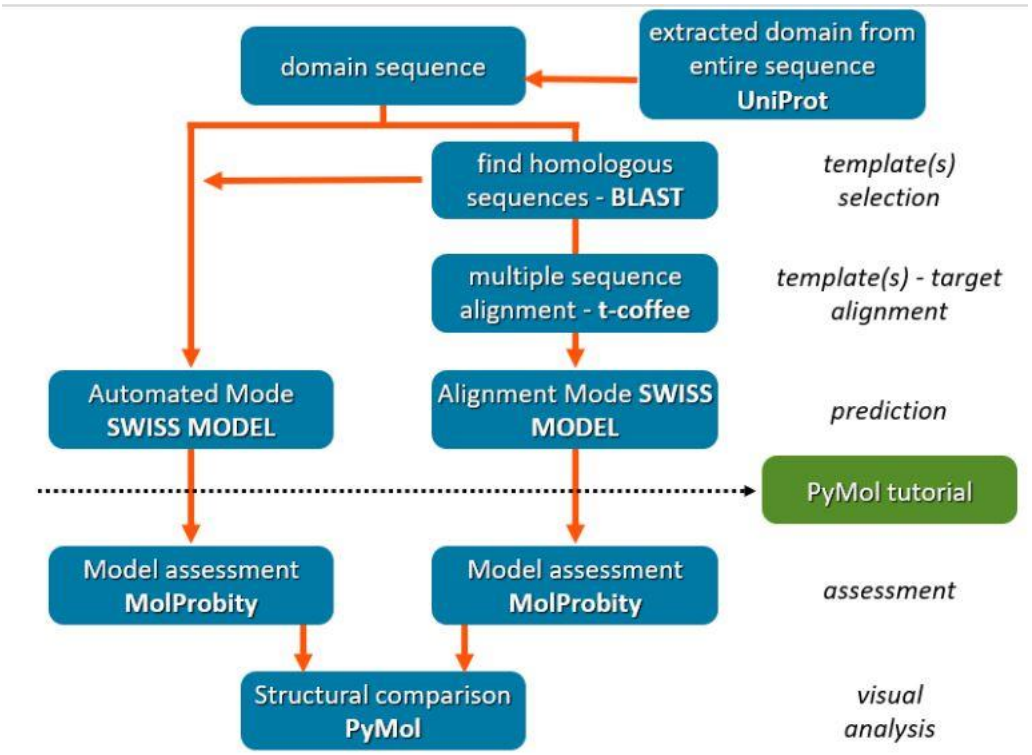
Once you choose the model file and upload to the MolProbity server, select 'Analyze geometry without all-atom contacts,' from the suggested tools.

In the next screen, check all the boxes to perform the analyses, as shown in the screen image below:




From the summary statistics page, Download Ramachandran plot.

The following flow diagram summarizes the steps you need to take either in the automated mode or alignment mode. In this assignment, we did the 'Alignment mode'.




You don't need to compare the results from the 'automated mode' to the 'alignment mode' using PyMol as we only completed the 'alignment mode'.

If you don't have the sequence of your target protein, and some hits from your template search, select a sequence of a human enzyme from one of the protein sequence databases that has a homolog in PDB with at least 30%, but not more than 60% of sequence identity. (Recommended length of the selected sequence is 150-250 residues). Check the Critical Assessment of Techniques for Protein Structure Prediction (e.g. CASP14, https://predictioncenter.org/casp14/index.cgi) for a list of curated and evaluated protein targets.

Submit all downloaded results and summary statistics, including 1) the structure model file, 2) Ramachandran plot, and 3) the multi-criterion chart in the html version.

**Assignment 3. Molecular Dynamic Simulation: 30%, due date??**

Two important concepts are in order to set the stage for this assignment.

1. Molecular recognition is the ability of biomolecules to recognize other biomolecules and selectively interact with them in order to promote fundamental biological events such as transcription, translation, signal transduction, transport, regulation, enzymatic catalysis, viral and bacterial infection and immune response.

2. Molecular docking is the process that involves placing molecules in appropriate configurations to interact with a receptor. Molecular docking is a natural process which occurs within seconds in a cell.

Molecular docking is heavily used in drug discovery and/or drug design because from bioinformatics tools standpoint, we can do virtual screening of drug candidates, molecular fragments, small molecules, and other types of compounds easily and efficiently.

For a detailed instruction / tutorial on docking read this guide on Nature protocols paper available via PubMed central - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4868550/

Use a docking program such as AutoDock Vina - https://vina.scripps.edu/ - for protein-ligand systems or to perform docking simulations.

To perform a molecular dynamics simulation, we need to find a protein-ligand complex. Try one of these complexes for your work - 3HTB , 2BRC.

The overall step to complete this assignment, it is important that you follow the steps below:

1. Get the complex (CPLX) coordinates (i.e. from the PDB).
2. Clean the complex (delete all the water and the solvent molecules and all noninteracting ions).
3. Add the missing hydrogens/side chain atoms and minimized the complex (AMBER Program).
4. Clean the minimized complex (delete all the water and the solvent molecules and all non-interacting ions).
5. Separate the minimized CPLX in macromolecule (LOCK) and ligand (KEY).
6. Prepare the docking suitable files for LOCK and KEY (pdbqt files).
7. Prepare all the needed files for docking (grid parameter file, map files, docking parameter files).
8. Run the docking.
9. Analyze the docking results.

In using AutoDock4 and ADT Tools, I suggest you pay attention to the following steps. For detailed instructions the documentation is an excellent resource - https://autodock.scripps.edu/wp-content/uploads/sites/56/2022/04/AutoDock4.2.6_UserGuide.pdf

- Prepare the file using ADT and run the docking – more instructions on the website – how to get started with docking - https://ccsb.scripps.edu/projects/docking/
- Prepare the macromolecule file
- Prepare the GRID parameter file and run Autogrid4
- Prepare the docking parameter file and running Autodock4

- Analyze the docking result – load the results; visualize the results; cluster the results

Submit all the files generated along the process (steps above – for each session) for protein-ligand docking simulation.