Syllabus of BIOINF 528 (2021 Fall, Bioinformatics Program)

Course Name:

Structural Bioinformatics

Course Description:

This course introduces fundamental concepts and methods of bioinformatics and structural bioinformatics. Topics covered include sequence, structure and function databases of DNA and protein molecules, advanced sequence and structure alignment methods, principle and application of machine learning and deep learning, basics of molecular dynamics and Monte Carlo simulations, methods of protein folding and protein structure prediction (homologous modeling, threading and *ab initio* folding), and techniques of protein structure determination (X-ray crystallography, NMR and cryo-EM). We will particularly discuss the newest breakthrough recently brought by the Google DeepMind team on deep-learning structure prediction and the core techniques behind the breakthrough. Emphasis of the classes is on the understanding of bioinformatics concepts and the practical utilization, with the objective to help students to use cutting-edge bioinformatics tools/methods to solve problems in their own research. For this term, three world's top experts are invited to give lectures on NMR spectroscopy (Prof. Tomek Cierpicki), Cryo-electron microscopy (Prof. Melanie Ohi), and X-ray crystallography (Prof. Zhaohui Xu).

Instructor:

Yang Zhang, Email: zhng@umich.edu, Phone: 734-647-1549

Invited Lecturers:

Prof. Tomek Cierpicki (on NMR spectroscopy)
Prof. Melanie Ohi (on Cryo-electron microscopy)
Prof. Zhaohui Xu (on X-ray crystallography)

Schedule and location:

9:00 am - 12:00 noon, Friday; #2036 Palmer Commons

Textbook:

No textbook is required for this course. All teaching materials will be posted on the course website.

Presentation, homework, & grades:

There will be homework assignments, including code writing and literature reading and presentation. Final grade consists of literature presentation (20%), homework (30%), and exam (50%).

Table of content

2.4.2. BLAST
2.5. Statistics of sequence alignment scores
2.5.1. E-Value
2.5.2. P-Value

3. Phylogenic Tree & Multiple Sequence Alignments

3.1. Neighbor-joining method and phylogenetic tree
3.2. How to construct multiple sequence alignments?
3.2.1. ClustalW
3.2.2. PSI-BLAST
3.2.2.1. PSI-Blast pipeline
3.2.2.2. Profile pseudocount
3.2.2.3. PSSM-position specific scoring matrix
3.2.2.4. Installing and running PSI-Blast programs
3.2.2.5. Interpret PSI-Blast output
3.2.3. Hidden Markov Models
3.2.3.1.Viterbi algorithm
3.2.3.2.HMM based multiple-sequence alignment
3.2.3.2.1. Creating HMM by iteration
3.2.3.2.2. SAM
3.2.3.2.3. HMMER/Jackhmmer
3.2.3.2.4. HHblits
3.2.4. DeepMSA
3.3. Sequence profile & profile based alignments
3.3.1. What is sequence profile?
3.3.2. Henikoff weighting scheme
3.3.3. Profile-to-sequence alignment
3.3.4. Profile-to-profile alignment

4. Protein Structure Alignements

4.1. Structure superposition versus structural alignment
4.2. Structure superposition methods
4.2.1. RMSD
4.2.2. TM-score
4.3. Structure alignment methods
4.3.1. DALI
4.3.2. TM-align
4.4. How to define the fold of proteins?
4.5. Number of protein folds in the PDB

5. Protein Secondary Structure Predictions

5.1. What is protein secondary structure?
5.2. Hydrogen bond
5.3. How to define a secondary structure element?
5.4. Basics of machine learning and neural network methods
5.5. Methods for predicting secondary structure
5.5.1. Chou and Fasman method
5.5.2. PHD
5.5.3. PSIPRED

5.5.4. PSSpred

6. Machine Learning & Deep Neural-Network Learning

6.1. Introduction of machine learning
- 6.1.1. Basic concept of machine learning
- 6.1.2. Supervised learning
  - 6.1.2.1. Classification
  - 6.1.2.2. Regression
- 6.1.3. Semi-supervised learning
- 6.1.4. Transfer learning
- 6.1.5. Unsupervised learning
- 6.1.6. Reinforcement learning

6.2. Basic structure for deep learning models
- 6.2.1. Why deep?
- 6.2.2. Fully connected layer
- 6.2.3. Convolutional layer
  - 6.2.3.1. Sparse connectivity
  - 6.2.3.2. Parameter sharing
- 6.2.4. Recurrent structure
- 6.2.5. Training of deep learning models
  - 6.2.5.1. Early stopping
  - 6.2.5.2. Regularization
  - 6.2.5.3. Dropout
  - 6.2.5.4. Adaptive learning rate

6.3. Unsupervised deep learning
- 6.3.1. Auto-encoder
- 6.3.2. Variational autoencoder
- 6.3.3. GAN

6.4. Applications of deep learning in structural bioinformatics
- 6.4.1. Protein secondary structure prediction
- 6.4.2. Protein contact-map prediction
- 6.4.3. Protein design

7. Monte Carlo Simulation and Local Energy Minimization

7.1. Introduction: Why Monte Carlo simulation?

7.2. Monte Carlo sampling of probabilities
- 7.2.1. Random number generator
  - 7.2.1.1.How to test a random number generator?
- 7.2.2. Sampling of rectangular distributions
- 7.2.3. Sampling of probability distribution
  - 7.2.3.1.Reverse transform method
  - 7.2.3.2.Rejection sampling method

7.3. Boltzmann distribution

7.4. Metropolis protocol

7.5. Advanced Metropolis methods
- 7.5.1. Replica exchange simulation
- 7.5.2. Simulated annealing

7.6. Local energy minimization

7.6.1. Gradient descent
7.6.2. Quasi-Newton
7.6.3. L-BFGS optimization

8. Protein Folding and Protein Structure Modeling

8.1. Basic concepts
8.2. Ab initio protein structure prediction
8.2.1. Anfinsen thermodynamic hypothesis
8.2.2. Molecular dynamics simulation for protein folding
8.2.2.1. CHARMM
8.2.2.2. AMBER
8.2.3. Knowledge-based free modeling (FM)
8.2.3.1. Bowie-Eisenberg approach
8.2.3.2. ROSETTA
8.2.3.3. QUARK
8.2.3.4. Why is beta-protein so difficult to fold?
8.3. Comparative modeling (homology modeling)
8.3.1. Principle of homology modeling
8.3.2. PSI-BLAST
8.3.3. Modeller
8.4. Threading and fold-recognition
8.4.1. What is threading?
8.4.2. Threading programs
8.4.2.1.Bowie-Luthy-Eisenberg
8.4.2.2.HHpred
8.4.2.3.MUSTER
8.4.3. Meta-server threading
8.4.3.1.3D-jury
8.4.3.2.LOMETS
8.5. Composite structure modeling approach
8.5.1. TASSER/I-TASSER
8.5.1.1. Force field design
8.5.1.2. Search engine: replica-exchange Monte Carlo simulation
8.5.1.3. Major issues and recent development
8.6. Deep-learning based approaches
8.6.1. Contact-map prediction
8.6.1.1.Mutual information
8.6.1.2.Direct coupling
8.6.1.3.Deep-learning coupled with direct coupling
8.6.2. Distance-map prediction
8.6.3. D-I-TASSER: combining deep-learning with MC simulations
8.6.4. trRosetta: combining deep-learning with L-BFGS
8.6.5. AlphaFold2: End-to-end deep-learning based approach
8.6.5.1. Attention in neural networks
8.6.5.2. Local-global protein structure mapping
8.6.5.3. End-to-end learning
8.7. CASP: A community-wide blind experiment on protein structure predictions

8.7.1. History of CASP
8.7.2. Current start of the art of protein structure prediction
8.7.3. Progress and challenge of proteins structure prediction

9. Principle of X-ray Crystallography & Molecular Replacement (Prof. Zhaohui Xu)

9.1. Introduction
9.2. Protein purification and crystallization
9.3. Diffraction
9.3.1. X-ray sources and data collection
9.3.2. Diffraction theory
9.3.3. Phase problem
9.4. MIR and MAD methods of phasing
9.4.1. MIR
9.4.2. MAD
9.5. Molecular replacement
9.5.1. Rotation function
9.5.2. Translation function
9.6. Phase improvement
9.7. Structure refinement and validation

10. Introduction to Nuclear Magnetic Resonance (Prof. Tomek Cierpicki)

10.1. Basics of NMR spectroscopy
10.2. Determination of protein structures from NMR data
10.2.1. Protein NMR spectra (homonuclear and heteronuclear)
10.2.2. Assignment of protein NMR spectra
10.2.3. Structural information from NMR
10.2.4. Structure calculation based on NMR data
10.3. Use of NMR to study large proteins
10.4. Residual dipolar couplings
10.5. Application of NMR to complex biological systems

11. Cryo-Electron Microscopy for Protein Structure Determination (Prof. Melanie Ohi)

11.1. What is cryo-EM and what is it good for?
11.2. Why use electrons?
11.3. Basics in image formation
11.4. Basics in image detection
11.5. What are the limits of using cryo-EM for biological samples?
11.6. Basic steps for determining high resolution structures using cryo-EM
11.7. General image processing steps
11.8. Common challenges encountered during image processing and structure determination
11.9. Future advances of cryo-EM