Syllabus of CS6222 (2025 Spring)

Course Name:

CS6222: Advanced Topics in Computational Biology

Course Description:

This course introduces fundamental concepts and methodologies of bioinformatics and computational biology. Topics covered include sequence, structure and function databases of nucleic acid and protein molecules; advanced sequence and structure alignment methods; molecular dynamics and Monte Carlo simulations; AI and deep learning methods; and protein and RNA folding and structure prediction (homologous modeling, threading, *ab initio* folding, and deep learning).

Special emphasis is placed on the latest breakthroughs in deep learning biomolecular structure predictions brought about by Google DeepMind and other leading teams of the field, with a focus on introducing the core technologies behind the breakthroughs, including but not limited to convolutional neural networks, graph neural networks, multimodal networks, transformers, language models, and diffusion models. The classes emphasize understanding computational biology concepts and their practical application, aiming to equip students with the skills to utilize cutting-edge bioinformatics tools/methods to solve problems in their own research projects.

Instructor:

Prof Yang Zhang (Email: zhang@nus.edu.sg; Phone: 6601-1241)

Orgainzation:

Department of Computer Science
School of Computing
National University of Singapore

Schedule and location:

Monday 4-6pm from 2025/1/13 to 2025/4/8, at COM1-VCRM

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. Monte Carlo Simulation and Local Energy Minimization

6.1. Introduction: Why Monte Carlo simulation?
6.2. Monte Carlo sampling of probabilities
6.2.1. Random number generator
6.2.1.1.How to test a random number generator?
6.2.2. Sampling of rectangular distributions
6.2.3. Sampling of probability distribution
6.2.3.1.Reverse transform method
6.2.3.2.Rejection sampling method
6.3. Boltzmann distribution
6.4. Metropolis protocol
6.5. Advanced Metropolis methods
6.5.1. Replica exchange simulation
6.5.2. Simulated annealing
6.6. Local energy minimization
6.6.1. Gradient descent
6.6.2. Quasi-Newton
6.6.3. L-BFGS optimization

7. Artificial Intelligence and Deep Learning (Zhiyuan Liu & Yang Li)

7.1. Introduction to Key Machine Learning Concepts
7.1.1. Supervised Learning
7.1.1.1.Linear Models: Linear Regression, Logistic Regression
7.1.1.2.Support Vector Machines
7.1.1.3.Trees: Decision Trees, Random Forests
7.1.1.4.Summary of Loss Functions
7.1.2. Unsupervised Learning
7.1.2.1.Clustering: K-Means, Hierarchical clustering
7.1.2.2.Dimensionality Reduction: PCA, t-SNE
7.1.3. Training a Machine Learning Model
7.1.3.1.Learning curve, overfitting, regularization, validation set,
7.1.3.2.Gradient descent: mini-batch, epoch, learning rate
7.1.3.3.Data augmentation
7.2. Introduction to Graph Neural Networks and Multi-Modal Language Models
7.2.1.1.Graph Neural Networks
7.2.1.2.Language models
7.2.1.3.Multi-modal LMs
7.3. Introduction to deep learning
7.3.1. Why deep?
7.3.2. Basic Building Blocks
7.3.2.1.Fully connected layer
7.3.2.2.Convolutional layer
7.3.2.3.Recurrent neural networks
7.3.3. Training deep learning models
7.3.3.1.Regularization techniques
7.3.3.2.Adaptive learning rate

    - 7.3.4. Advanced Architectures
        - 7.3.4.1.Attention mechanisms
        - 7.3.4.2.Transformers and the "Attention is All You Need"
        - 7.3.4.3.BERT
        - 7.3.4.4.Potts Model
        - 7.3.4.5.From Potts Model to BERT
    - 7.3.5. Diffusion model

- **8. Protein Folding and Protein/RNA 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. AlphaFold
      - 8.6.4.1.AlphaFold2: End-to-end protein structure prediction

        - 8.6.4.2.AlphaFold3: Diffusion based complex structure prediction
    - 8.6.5. ESM
        - 8.6.5.1.ESM1: protein language model
        - 8.6.5.2.ESM2 and ESM-Fold
        - 8.6.5.3.ESM3: a multimodal generative language model
- 8.7. RNA structure prediction
- 8.8. CASP: A community-wide blind experiment on protein structure predictions
    - 8.8.1. History of CASP
    - 8.8.2. Current start of the art of protein structure prediction
    - 8.8.3. Progress and challenge of proteins and RNA structure prediction