CSE8803 - Machine Learning in Computational Biology

General information

Instructor: Xiuwei Zhang, School of CSE, College of Computing
Contact: xiuwei.zhang@gatech.edu
Class timings: Tuesday and Thursday, 9:30-10:45

Description

The large scale biological data, including DNA and RNA sequencing data and other measurements like gene-expression levels, require computational methods to perform data mining and prediction in order to lead to biological discoveries. We consider the following problems: (1) which subsequences in our genome are genes, or are motifs with certain patterns that affect whether a gene has RNA or protein products? (2) can we predict the 3D structure of RNA or protein molecules from their sequences? (3) how do molecules like RNAs and proteins interact to determine mechanisms in a living organism? (4) what cell types do we have and how does each cell function? Machine learning methods have an important role to play in these problems. This course will touch these topics and introduce how different machine learning methods are used in studying each of these problems.

Tentative topics

- Learning from DNA sequences:
    - Gene finding, motif finding (HMM models)
- Structure of molecules
    - Protein structure prediction (deep neural networks)
    - RNA structure prediction (a dynamic programming method, Stochastic context-free grammar, deep learning models)
- Learning from high dimensional data (eg. single cell gene-expression data)
    - Dimension reduction methods (PCA, MDS, auto-encoders, VAE, visualization in low dimensions, diffusion maps)
    - Clustering cells to find new cell types (k nearest neighbor graphs, graph based clustering methods, matrix factorization)
- Interactions between molecules (biological networks)
    - Learning network structure and causality between molecules (Bayesian networks, decision trees, random forests)
    - Comparison between multiple networks (probabilistic graphical models)

Evaluation

Students will be evaluated based on homeworks and paper presentations. There may be programming assignments in homeworks.

Background and Prerequisites

This is a graduate-level course and will be highly multidisciplinary. Students are required to have taken probability and statistics, algorithms, and linear algebra courses. Background knowledge in data analytics will be helpful for this course. Students should be able to program in at least one of the following languages: Python, R, Matlab. PyTorch may be required to implement deep learning models.