EE 372: Data Science for High Throughput Sequencing

Course Overview

Course Description

Extraordinary advances in sequencing technology in the past decade have revolutionized biology and medicine. Many high-throughput sequencing based assays have been designed to make various biological measurements of interest. This course explores the various computational and data science problems that arises from processing, managing and performing predictive analytics on high throughput sequencing data. Specific problems we will study include genome assembly, haplotype phasing, RNA-Seq assembly, RNA-Seq quantification, single cell RNA-seq analysis, multi-omics analysis, and genome compression. We attack these problems through a combination of tools from information theory, combinatorial algorithms, machine learning and signal processing. Through this course, the student will also get familiar with various software tools developed for the analysis of real sequencing data. The target audience for the course include

1. students specializing in information theory / algorithms/ signal processing / machine learning who want to learn of applications in biology and get exposure to real data

2. students specializing in computational biology, who want to strengthen their knowledge of basic information theory / signal processing / machine learning

Course news and assignments will be posted at `ee372.stanford.edu`, which redirects to the course's GitHub website.

Prerequisites

- Undergraduate level probability
- Some programming experience. We will be using Python.
- Some undergraduate background in algorithms would be beneficial
- NO PRIOR background in biology will be assumed

Lecture Times

Monday, Wednesday 3:00 PM - 4:20 PM at McCullough 115

Course Staff

Instructor: David Tse (dntse@stanford.edu)
Teaching assistants: Govinda Kamath (gkamath@stanford.edu) , Jesse Zhang (jessez@stanford.edu)
Office hours to be decided.

Course Grading

The grading for the course will be broken down as follows:

- Attendance 10%
- Scribing 10%
- Assignments 40%
- Project 40%

There will be 4 assignments. The assignments will involve a theory component and a programming component. The programming component is aimed at exposing students to the messiness involved in real data, and various popular tools (so that they can gauge the pros and cons of these tools). The programming assignments will include

- experiments demonstrating biases of different types in various types of data,
- implementing simple algorithms for assembly, alignment, and quantification,
- using popular software packages to perform simple experiments.

All programming assignments will require require only laptop-level computing. The main language used to code will be Python. UNIX/LINUX/OS X may be needed for some of the software packages. We recommend that windows users use the Corn cluster. Projects can be theoretic or practical in nature (ideally a mix of the two). A list of possible projects will be put up shortly. Students can also come up with project topics that they are interested in (in consultation with the teaching staff)