Comp790-166: Computational Biology

Lecture 1

December 15, 2020

Outline for Today

- Introductions
- Course Logistics
- Bioinformatics vs computational biology
- Exciting problems for computational biology
- Computational challenges for modern biological datasets

Introductions

Let's introduce ourselves with the following info.

- Name
- Department, Graduate or Undergrad?
- One thing you are hoping to learn about here
- An interesting or funny fact about you.

Class Meetings, Course Webpage

- We meet here on zoom on Tuesdays and Thursdays from 9:30am-10:45am
- Office Hours will be at 11-12 on Fridays and by appointment (use same zoom link as for this class).
- Course website: `https://github.com/stanleyn/Comp790-166-Comp-Bio`
- Notes will be available at least an hour before the lecture in the git repo

Prerequisites

- This is a graduate CS course, so I will not assume any knowledge of biology
- Mathematical Foundations: comfort with linear algebra and probability
- Strong Programming (one of Python, Julia, R)
- Comfortable reading and ideally implementing ideas from research papers
- Feel free to chat with me if you feel weaker in one area, but still think that the course is beneficial for you.

Course Structure

- Grading will be based on two homework assignments, a course project, and weekly reading summaries
  - **HW:** 20% each, a mixture of programming and light math to practice implementing and interpreting output of methods we will discuss.
  - **Project Proposal:** 10%. This is a writeup of your proposed project and a presentation to the class.
  - **Project Writeup and Final Presentation:** 30%. You will turn in a writeup, link to your code, and present to the class.
  - **Reading Questions:** 20%. I will have a fixed set of questions to answer for one of the assigned papers for the day.

Helpful References

We will mostly use research papers as references. I have listed a few books on the course webpage. In general, I find the following the most helpful for most things.

- Pattern Recognition and Machine Learning by Chris Bishop
- Spectral Learning on Matrices and Tensors by Janzamin *et al.*
- The Matrix Cookbook `http://matrixcookbook.com/`

Course Project Overview

- The goal is to ask a question, implement an idea, and apply it to a biological dataset
- I have compiled a list of publicly available datasets here, `https://github.com/stanleyn/Comp790-166-Comp-Bio/blob/main/Datasets.md`
- You can either propose a new method, or apply existing methods on data and interpret the results.
- Feel free to work alone or in teams of 2-3.
- Come talk to me if you need some ideas.

Course Project Writeup

This will be good practice in writing up research and putting your results in context in comparison to what has already been done.

- **Proposal:** You will write up a brief document of your question, how it relates to what has been done, and perhaps from preliminary result
- **Project Writeup:** This will be structured like a regular research paper and will include background, results, discussion, and a link to your code.
- **Presenting Proposals and Write-ups:** You will present your project proposals (mid-semester) as well as your final project (end of semester).
- **Code Belonging to Your Project:** Create a git repository for your project and create a README with an example to run your method on a subset of data or to reproduce at least some of your results.

Reading Summaries

Before each class you will turn in reading summary questions by email. There may be several papers assigned for a particular day, so you only need to write the summary about one paper.

- Please explain in 2 sentences or less what the problem being solved is.
- What were the main contributions of the authors in this work? (You can answer in a few bullet points).
- Please describe 1-2 computational experiments that the authors implemented to test their method.
- Were the authors the first to attempt this particular problem? If not, did they compare their results to other baselines? Do you think that their evaluation was objective?
- Do you think that the authors provided enough evidence for why their developed method is an important contribution? If yes, please describe their reasoning here. If you do not think they adequately justified why they worked on this particular problem, please describe your thoughts on that here.
- What is one follow-up idea or extension from this work?

Homework

Two homework assignments, which will mostly be coding, visualization, and some light math.

- If I provide code to use with the homework assignment, it will be in Python. However you are free to submit your code in either Python, Julia, or R.
- Please submit your ultimate homework writeup as a PDF. I will provide a LaTeX template, but you can choose not to use it as necessary.
- You can send a link to your code (dropbox, google drive, git) labeled by hw problem, submit PDF writeup and code as zip, or if it is only a few lines you can paste it or add a screenshot to your writeup.
- It is ok to use thing implemented by other people (for example building a kNN graph using K-d trees, or PCA, etc). Just acknowledge where you got this code.
- Don't forget to label your axis on your plots. :)

What is this course?

We will explore the mathematical foundations and theory of algorithms that commonly used to analyze modern biomedical datasets. This means that we will center lectures around a particular task, but dive into the mathematical details as well.

What is this course (and what is it not)?

- This is **not** a course about bioinformatics pipelines or how to use packages
- This course does **not** serve as a comprehensive overview of computational biology and its evolution as a field
- The content here will be biased towards single-cell data, systems immunology, combining datasets from multiple modalities, visualization, and how to benchmark and compare particular algorithms.
- Because we are combining theory with modern research, we will actively question what we are learning. (e.g. What could the authors have done better, are we convinced their approach is the best approach?)

Topic and Application Overview

- **Important Background:** Linear algebra and graph fundamentals
- **Analysis of Single Cell Data:** Automated cell population discovery, imputation and batch effect correction, linking to external information, trajectory inference
- **Benchmarking:** How do we objectively compare bioinformatics algorithms
- **Combining Multiple Modalities:** How do we jointly represent biological samples or features?

Mathematical Foundations of Most Algorithms that we will Study

- **Graph-based analysis and manifold learning.** We will get used to thinking about distances and how distances can be preserved. We will be frequently thinking about efficient ways to build graphs, diffusion on graphs, etc.
- **Numerical linear algebra.** Dimensionality reduction, matrix decomposition, spectral clustering, etc
- **Deep Learning.** We are at the beginning of a deep learning revolution in computational biology. It is just starting to make its entrance here (behind many other applications)

Let’s get started with content now.

- A few examples of exciting advancements in technology, and how we can study biology.
- Bioinformatics vs computational biology
- Computational challenges and considerations

Living in the high-throughput era of biomedicine

We have a variety of technologies that are becoming increasingly lower in cost, such that we can tractably measure diverse aspects of biology.

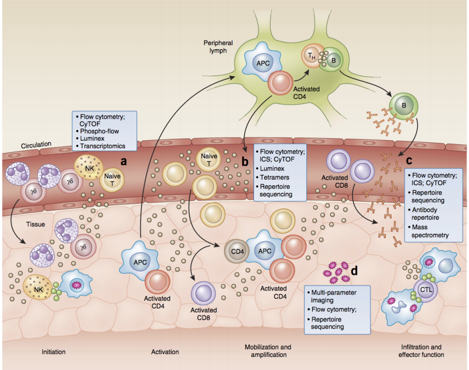


Figure: from Davis *et al.* 2018 'Systems Immunology, Just Getting Started'

The Immune System: A Diverse Set of Cell Populations

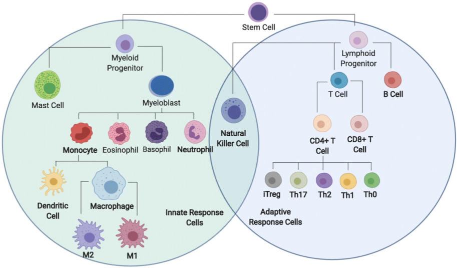


Figure: Torang *et al.* BMC Bioinformatics. 2019. Modern technologies help us to characterize the phenotype and function of these diverse immune cell types.

Deep Learning Meets Structural Biology

- DeepMind used a deep learning approach to predict protein structure from amino acid sequence.
- Strong accuracy on this task can have important implications in applications, such as, drug discovery.

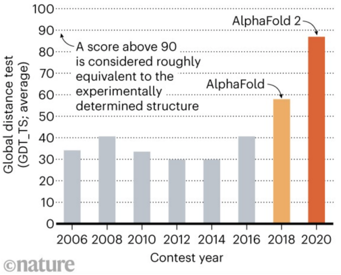


Figure: Nature (News) 2020. AlphaFold2 achieves state-of-the-art performance on the protein structure prediction problem.

Single-Cell Analysis to Understand COVID Severity

Single-cell gene and protein expression assays have been used in the past year to identify biomarkers of COVID severity.

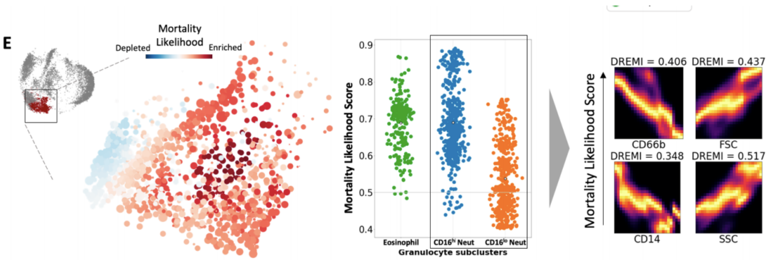


Figure: Kuchroo *et al.*. BioArXiv 2020. Understanding the Connection of Particular Cell-Populations with COVID severity.

Combing Multiple Biological Modalities

How do you determine a joint representation for a set of patients/samples/examples, according to multiple sources of information?

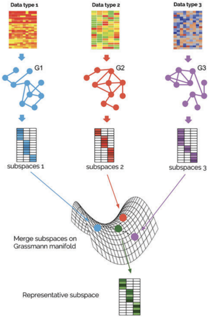


Figure: Ding *et al.* 2018. Bioinformatics.

Single Cell Imaging Modalities

Now we can take pictures of tissues and simultaneously measure the expression of proteins in individual cells

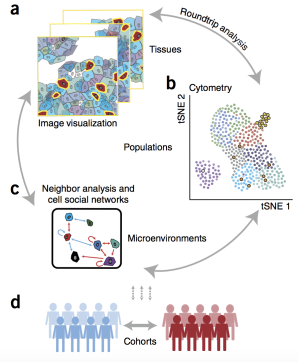
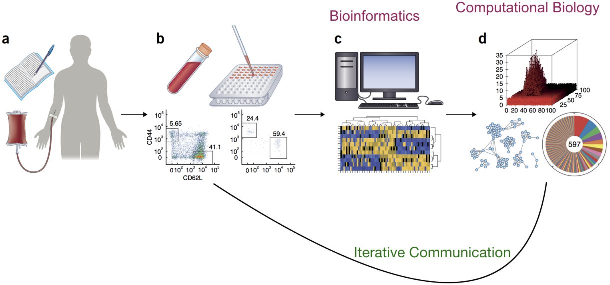


Figure: Schapiro *et al.* 2017. Nature Methods.

The Difference Between Bioinformatics and Computational Biology

You may hear these terms used interchangeably, but there are subtle differences.



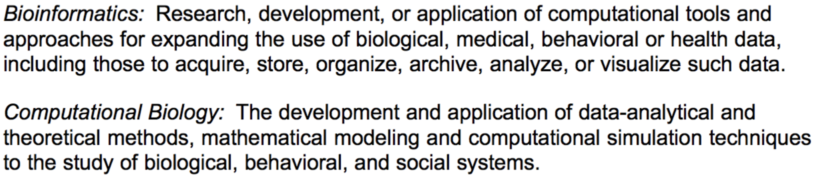
Bioinformatics

Computational Biology

Iterative Communication

Figure: adapted from Davis *et al.* Nature Immunology 2018. Bioinformatics → software engineering and efficiently storing or representing information. Computational biology → modeling and prediction.

Official Definitions of Bioinformatics and Computational Biology



Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

Computational Biology: The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

Figure: NIH definitions of Bioinformatics and Computational Biology

A Couple of Computational Problems

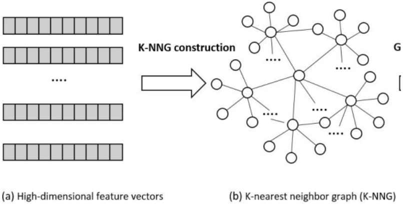
Lots of Dimensions, Few Samples

Very high dimensional and typically low sample size data.

- Few samples/patients/examples relative to the number of feature measured (P>>N)
- Danger of overfitting with a model.
- A low number of samples/patients/example limits the application of modern ML techniques, like deep learning.
- Variability in the model depending on which subset of examples you train with.
- Typically need to use cross validation

Single Cell Data: Combining Potentially Millions of Cells

If you collect hundreds of thousands of cells over hundreds of patients, you quickly end up with millions of cells that you need to collectively consider.



(a) High-dimensional feature vectors

(b) K-nearest neighbor graph (K-NNG)

Figure: Tang *et al.* 2018. ArXiv. How do you calculate distances between millions of cells without computing all pairwise distances? How do you ensure that the representation preserves both local and global between-cell similarities?

Recap

Today

- Course Logistics
- Motivating Examples

Next time:

- Building graphs from data
- Graph Laplacian
- Some numerical linear algebra...