Syllabus
Bi/BE/CS 183 - Computation Biology
Winter 2020

Grading: 60% homeworks, 15% midterm, 25% final.
Homeworks are due Thursdays at noon via Moodle.

Collaboration Policy: If you get stuck with a homework problem, you are encouraged to discuss it with other students (offline or online on Piazza) and consult/use any resources. But you will have to prepare and submit your solution by yourself. Please write down on the homework who you worked with and explain what, if any, help you received. Also list all resources that were used for the homework and explain how they helped. No collaboration is allowed on the Midterm and Final exams.

Homework submission: All programs must be written in R or Python. Code must be submitted along with examples demonstrating that the code runs correctly on basic tests. Please submit notebooks as .html files. Points may be deducted for homeworks that are not easily accessible.

Extension Policies: Extensions can be used on a maximum of two homeworks. For each extension, we will accept homeworks handed in Sundays by 5pm via Moodle. There will be no exceptions to this rule. Any homeworks submitted late after two extensions are used or after Sunday 5pm will not be accepted. Extensions can be taken automatically, i.e. no permission needs to be requested. Just indicate on homework when extension is used.

Q&A: We will address all questions on Piazza. You can enroll by finding courses at Caltech and looking for "Bi/BE/CS 183." Please ask questions on Piazza that in lieu of emailing, so that we can have a centralized repository of all questions asked.

Resources: This course does not closely follow a textbook, resources for each class will be posted on Moodle. To dive deeper into some sections of the material, you might be interested in the following books:
- **Algebraic Statistics for Computational Biology** by L. Pachter and B. Sturmfels
- **Information Theory, Inference, and Learning Algorithms** by David MacKay
  Available for free at www.inference.org.uk/itila/book.html
- **An introduction to statistical learning** by James, G., Witten, D., Hastie, T., & Tibshirani, R.

Office Hours:
Lior: 320 Annenberg, Mondays 1-2PM
TA office hours:
Monday 10-12 AM - SFL 326 (Jialong and Charles)
Wednesday 7-9 PM - SFL 326 (Lambda and Charles)

Planned Schedule

Lec 1 - Tue Jan 7
Introduction to the class and the gene

Lec 2 - Thu Jan 9
Introduction to hidden Markov models

Lec 3 - Tue Jan 14
Hidden Markov models (continued) and applications to gene finding

Lec 4 - Thu 16- **HW1 due**
Sequence alignment, the Needleman-Wunsch algorithm and Introduction to bulk RNA-seq

Lec 5 - Tue Jan 21
Overview of read alignment and pseudoalignment. Walkthrough of kallisto

Lec 6 - Thu Jan 23 - **HW2 due**
Regression and principal component analysis

Lec 7 - Tue Jan 28
Methods for differential expression analysis

Lec 8 - Thu Jan 30 - **HW3 due**
Introduction to single-cell RNA-seq

Lec 9 - Tue Feb 4
Single-cell RNA-seq primitives

Lec 10 Thu Feb 6 - **Midterm out**
Processing of single-cell RNA-seq data

Lec 11 - Tue Feb 11
Gene expression distributions and noise models

Lec 12 - Thu Feb 13 - **HW4 due**
Comparing gene expression distributions

Lec 13 - Tue Feb 18 - **Midterm due**
Dimensionality reduction, linear methods, sparsity

Lec 14 - Thu Feb 20 - **HW5 due**
Nonlinear dimensionality reduction, NNMF, tSNE

Lec 15 - Tue Feb 25
Graph based methods, cluster identification, Louvain

Lec 16 - Thu Feb 27 - **HW6 due**
Spectral clustering, Isomap, and pathfinding

Lec 17 - Tue Mar 3
Probabilistic models: Mixture models

Lec 18 - Thu Mar 5 - **HW7 due**
Probabilistic models: neural networks

Lec 19 - Tue Mar 9
New measurements and applications of single cell RNA-seq

Lec 20 - Thu Mar 12 - **HW8 due, final out**
Learning predictive regulatory network models from single cell data

Final due Mar 17