Administrative Details:

This class is cross listed with three schools: MIT, Harvard School of Public Health and Boston University. The professor who will teach the course is Manolis Kellis, the TA for the course is Matt Edwards.

Online:

The course has a main website: https://stellar.mit.edu/S/course/6/fa10/6.047/materials.html, containing course material and information; and a wiki website, http://6047.wikispaces.com, where students are free to leave questions. The course calendar can be found on Google Calendar: public add "6.047 Students"

Handouts:

Students receive 3 handouts, which are also available on the website:
- Course Information and schedule
- Problem Set 1
- Scribe Policy

There is also a student questionnaire and an Email list sign up. Make sure you're on the email list, an email announcing the recitation room and time was sent out later that day.

Grading: see handout for details.

Final project:

There will be a term-long final project, with many milestones due throughout the semester to ensure sufficient planning. The project mirrors real world scientific process:

Come up with Idea->frame it->propose it->revise it->carry out->present

Projects should use real data and aim to be published. Students are encouraged to talk with TA’s while formulating a final project idea. You should look forward through the lecture notes on future modules to get an idea of what areas will interest you most.

The timeline for final project is as follows:

I. Set-up: a brief overview of your experience and interest. Due 9/27
II. Brainstorming: a list of initial project ideas and partners. Due 10/13
III.a. Proposal: submit a project proposal in the form of an NIH proposal. Due 10/25
III.b. Review: review and critique 3 peer proposals. Due 11/1
III.c. Response: a bullet-list response to peer critiques. Due 11/8
IV. Midterm progress report: write outline of final report. Due 11/22
V. Final project report: write report in conference paper format. Due 12/6
VI. Final class presentation: 10min conference talk. Due 12/9

There will be friday mentoring sessions before each portion of the final project is due, and you are encouraged to find a mentor at the first few sessions who is actively interested in your project and could help you more frequently. The mentoring sessions can be helpful in identifying if unexpected results are the result of a bug or are instead a discovery.

Make sure you start working on the project even while waiting for peer reviews, so that you will have 4-5 weeks to complete the research itself.

The final project will include the following two deliverables:

1) A written presentation, due Mon at 8pm, last week of classes. The written presentation can contain the following elements:
   - Who did what (to reflect trend in publications)
   - The overall project experience
   - Your discoveries
   - What you learned from the experience (introspection)

2) An oral presentation, due Thursday after the written presentation. This allows students three days to prepare the oral presentation.

The grading equation for the final project is:

\[ \text{min}(O,C,R) \times A + P \]

- **Originality** - unoriginal computational experiments don't get published
- **Challenge** - the project needs to be sufficiently difficult
- **Relevance** - it needs to be from biology, can't just reuse something from another field
- **Achievement** - if you don't accomplish anything you won't get a good grade
- **Presentation** - even if you've achieved a good project you have to be able to present it so everyone knows that, and make it look easy. the presentation should show how the project is O, C, and R

O, C, R out of 5 points, A and P out of 10

**Scribe assignments:**

Each student will be required to scribe for one lecture, putting together a summary that will be shared with the class. Scribes are urged to read the lecture material in advance of the lecture. A first draft of scribed notes is due two days after the lecture, if there are no notes up from the previous year, and the final draft of the scribed notes is due six days after the lecture. If either of those dates falls on the day another assignment is due, then you get a two day extension. For example, if the scribe notes are due on a Monday, then you could then submit them on Wednesday. Further details on scribing can be found in the **Scribe Policy** handout.

**Course Information:**

The following three (optional) reference textbooks recommended for the class.
Each book has a different advantage. The first book is a classic one. It is heavy in math and covers much of what is in class. The book is focused on sequence alignment. As part of sequence alignment theory, the book approaches Hidden Markov Models (HMM), pairwise and multiple alignment methods, phylogenetic trees as well as a short background in probability theory.

The second book intends to balance between mathematical rigorous and biological relevance. According to the author, it is a good book for undergrad students. The book includes a table that associates algorithm to biological problems.

The third book is about machine learning. It uses a more engineering approach. It includes machine learning theory, neural network and, as the name suggests, pattern recognition.

Homework:

In general, all homeworks are due Monday at 8pm. Students are encouraged to get work in on time, but are asked to please contact TA’s in advance if homework is to be late. The first homework can be considered a good gauge for the level of difficulty of this course.

Course Outline:

The course is divided into four modules, where in the past the first half of the course would cover foundational material, and the second half of the course would address open research questions. The four modules are: Comparative Genomics, Coding and non-coding genes, Networks and Gene Regulation, and Evolution and Phylogenomics. Each module will be divided in half, with the first half covering the foundations of that module and the second half covering the frontiers.

Some of the major topics that the course will address are:

1) Gene Finding
2) Sequence Alignment
3) Database Lookup: How can we effectively store and retrieve biological information?
4) Genome Assembly: How can we put together the small snippets of sequence produced by sequencing technologies into a complete genome.
5) Regulatory motif discovery: How can we find the sequence motifs that regulate gene expression?
6) Comparative genomics: How can we use the information contained in the similarities and differences between species’ genomes to learn about biological function?
7) Evolutionary Theory: How can we infer the relationships among species using the information contained in their genomes?
8) Gene expression analysis
9) Cluster Discovery: How can we find emergent features in the dataset?
10) Gibbs Sampling: How can we link clusters to the regulators responsible for the coregulation?
11) Protein Network Analysis: How can we construct and analyze networks that represent the relationships among proteins?
12) Metabolic Modeling
13) Emerging Network Properties
Class Goals

The class aims to provide students with an introduction to the field of computational biology and an introduction to the fundamental problems of this field; to teach not only computational skills but also how research in this field is conducted.

During the course, students will have to solve four problem sets, which will require working out computational biology methods and actually applying them to real problems in biology. Students will also conduct a research project for this course, using real data and presented in conference format, as well as evaluating peer research proposals. This is a new field with plenty of new problems, and research projects from this course in the past have been used as the starting point for a published paper.