**Introduction to Computational Biology and Biological Modeling**

**Course number:** BIOL 437/CIS 436/GCB 536/CIS 536

**Schedule:** Fall 2018, Mon-Wed: 3:30-5:00, Fagin Hall 218

**Instructor:** Junhyong Kim ([Junhyong@sas.upenn.edu](mailto:Junhyong@sas.upenn.edu); Lynch 304G)

**Office Hours:** By appointment

**Teaching Assistants:** Ammon Perkes ([perkes.ammon@gmail.com)](mailto:perkes.ammon@gmail.com)) and Yongjun Li (yjli@sas.upenn.edu)

**Office Hours:** TBA

**Intended Audience:** Upper level undergraduates. MS students, and first year PHD students.

**Pre-requisites:** Calculus, Part I (Math 104), Statistics (BIOL 446 or equivalent), Intermediate level Biology (BIOL 121 + 200 level course);

**Recommended:** Calculus, Part II and III, knowledge of programming

**Synopsys:**

Biology is flooded with data that cannot be understood without computational analysis and modeling. For example, new instruments are sequencing the DNA and RNA at such a rate that less than 0.001% of the data will be ever seen by human eyes. Computational Biology is a growing field that develops algorithms, statistical analysis methods, and ultimately biological models for these kinds of data. That is, Computational Biology is a ***subfield of natural science*** where quantitative approaches are used to discover and understand biological and medical phenomena.

The goal of this course is to develop a deeper understanding of techniques and concepts used in Computational Biology. The course will strive to focus on a small set of approaches to gain both theoretical and practical understanding of the methods. We will aim to cover practical issues such as programming and the use of programs, as well as theoretical issues such as algorithm design, statistical data analysis, theory of algorithms and statistics. This course WILL NOT provide a broad survey of the field nor teach specific tools but focus on a deep understanding of a small set of topics.

**Topics to be discussed (not in order):**

Algorithms, script programming, probability theory, multivariate statistics, geometry, string algorithms, molecular evolution, Hidden Markov Models, and Machine Learning. Other topics as time permits.

## Course Web page

This course will utilize Penn’s “CANVAS” web software at: <https://canvas.upenn.edu/>. All updates, materials, discussion, and grades will be posted here. Visit the site often. This is a secure site requiring your PennKey.

## Lectures and Exercises

This is a lecture-based course with two lectures a week. The lectures will run from 2:00-3:30 with a short break in between. Reading materials and lecture notes will be uploaded prior to each lecture. The lectures will assume that all reading material will have been read before attending the lecture. Class participation will be expected based on the reading materials. There will be tracking of class participation for bonus points. Occasional Exercises will be assigned for bonus points. See below for bonus points and grades.

***Office Hours***

Office hours are meant to clarify reading and lecture material, not as a substitute for these. Office hours are more productive when you have clearly formulated questions about the material and are not a substitute for independent studying. If you are struggling or have doubts about the material, come to the TA office hours or email me to set up an appointment. If you email me, you will get an appointment for the same week.

**Debugging support will be not provided for programming.**

In lieu of appointments the days before the midterms, office hours will be held (TBA) to clarify any lingering questions in preparation for the mid-term exams. The location of these office hours will be announced at least a week before the mid-term. Emails less than 24 hours prior to the exam may not be answered.

***Grading***

Grades will be compiled from: (1) Mid-term exam I (20 pts); (2) Mid-term exam II (20 pts); (3) Final exam (60 pts); (4) Participation points and Homework (see rules below).

**GCB536/CIS535: PHD students taking 500-level credits will be assigned special final projects, which will count for another 25 pts (total points will be normalized to 100). Details of the projects will be announced during the semester.**

**All grades for GCB536/CIS536 will be considered separately from BIOL437/CIS436 registrants.**

*Mid-term exam I and II:* These exams will be multiple choices/short answers and consist of problems derived from lecture notes and lecture slides.

*Final exam:* This will be a take home exam with problems derived from the entire semester. See example final exam. You will be asked to pledge to an honor system.

*Participation points:* Throughout the semester you will be periodically given a challenge problems or other kinds of participation invitations. You will receive points for completion or vigorous participation in debates/discussions. Participation points will be used with homework points (see below) to modify your final letter grade.

*Homework points:* Home works will be given throughout the semester. See example homework. They will not be individually graded but checked for quality and completion. If you have collected 80% of the maximum number of homework/participation points given out during the semester, I will raise your letter grade if your numerical grade is within 2 points of a letter grade cutoff. For example, if your score is 79 and 81 is the cutoff for a B, you will receive a B, if you have received 80% of the points.

**Letter Grades**

Numerical grades will be converted to letter grades based loosely on absolute standards. If your numerical grade is sufficiently close to a letter grade cutoff, I will use your homework and participation points to raise your letter grade (see above). Otherwise, all grade conversions are absolute.

**Grade disputes**

I welcome all grading disputes. However, all grading disputes must be submitted in writing. You must write me a letter stating in scientific manner why your answer is correct. I will not accept arguments of the form “but it’s all there” or “I should have received more points for this particular answer.” No grade dispute letter will be accepted after the Reading Period.

## Recommended Reading

There are no required textbooks for this course. You must read all posted material that is marked for reading. The following is a list of books you may use to complement the lecture notes. Lecture notes will be posted weekly.

An introduction to bioinformatics algorithms (Computational Molecular Biology)

N.C. Jones and P.A. Pevzner

A Primer of Genome Science   
Greg Gibson and Spencer Muse, North Carolina State University

**Other books of interest:**

Machine learning approaches to bioinformatics, Zheng-rong Yang, World Scientific (a fairly comprehensive survey of machine learning approaches)

Algorithms on strings, trees, and sequences: computer science and computational biology, Dan Gusfield, Cambridge Univ Press. (A very comprehensive algorithms book with easy to read style)

Computational Molecular Biology: An Algorithmic Approach

Pavel A Pevzner (Somewhat hard core computational biology)

Molecular Biology of the Cell, Alberts et al. (Again standard textbook—long)

Recombinant DNA, Watson et al., W.H. Freeman. (Quite outdated but still a great introduction to molecular biology)

Statistical methods in bioinformatics: An introduction, Grant and Ewens, Springer-Verlag. (This is one of the few statistically focused books on bioinformatics. And, you can directly talk to the author!)

**Draft schedule for BIOL 437 (may change dependent on progress and other factors)**

|  |  |  |
| --- | --- | --- |
| Date | Topic | Lecturer |
| 8/29 | Introduction to Computational Biology | JK |
| 9/5 | Python programming primer | AP |
| 9/10 | Warm up: string finding problem | JK |
| 9/12 | Suffix/Prefix trees and Finite State Automata | JK |
| 9/17 | Sequence alignment | JK |
| 9/19 | Sequence alignment 2 | JK |
| 9/24 | Graphs in Computational Biology Problems | JK |
| 9/26 | Next Generation Sequencing Part 1 | JK |
| 10/1 | Next Generation Sequencing Part 2 | JK |
| 10/3 | **Mid-Term Exam I (Multiple Choice)** | JK |
| 10/8 | **End of Drop Period** |  |
| 10/8 | Probabilities, stochastic models | JK |
| 10/10 | Hidden Markov Models I | JK |
| 10/15 | Hidden Markov Models II | JK |
| 10/17 | RNA sequencing and multi-variate data | JK |
| 10/22 | Differential gene expression and hypothesis testing | JK |
| 10/24 | Distances and geometry | JK |
| 10/29 | Visualizing high-dimensional data | JK |
| 10/31 | **Mid-Term Exam II (short questions)** | JK |
| 11/5 | Learning approaches to genome-scale data 1 | JK |
| 11/7 | **Last day to withdraw** | JK |
|  | Learning approaches to genome-scale data 2 |  |
| 11/12 | Learning approaches to genome-scale data 3 | JK |
| 11/14 | Trees, clusters, and graph models | JK |
| 11/19 | Guest Lecture | TBA |
| 11/19 | **500-level course special assignment** | Due 12/16 11:59 pm |
| 11/26 | Discovery and hypothesis generation and statistical hypothesis testing | JK |
| 11/28 | Molecular evolution and the comparative method | JK |
| 12/3 | TBA | JK |
| 12/5 | TBA | JK |
| 12/10 | TBA | JK |
| 12/13 | **Final Exam Assigned (Take Home)** | Due 12/15 11:59 pm |