CSE3810/CSE6800: Computational Genomics
Spring 2019

Lecture: M/W/F 2:30-3:20PM, ITE 125

Instructor:

Ion Măndoiu
ion@engr.uconn.edu
Office:
ITE 261
Office Hours:
M/W 3:30pm-5pm or by appt.

Course Description: Started in 1995 by the completion of the first genome sequence of a free-living organism, *H. influenzae*, the genomic era has led to thousands of complete genome sequences deposited in public databases and many more genome projects at various stages of completion. The large-scale availability of genome data is revolutionizing biological and medical research, with data-driven computational approaches taking a central role. This course covers fundamental computational methods for genomic data analysis, with a main emphasis on statistical methods and current applications in genomics and genetic epidemiology.

Tentative list of topics to be covered: Basic probability theory and statistics; statistical modeling of biological sequences; EM and Gibbs sampling algorithms for DNA motif discovery; Markov chains; profile HMMs for representing sequence families; models of DNA and protein evolution; likelihood methods in phylogenetics; bootstrapping; basic principles of population genetics; genotype phasing and haplotype frequency estimation; computation of Mendelian likelihoods; Elston-Stewart and Lander-Green algorithms; admixture mapping; association studies; next-generation sequencing data analysis. The list of topics may change according to progress and student interest.

Textbooks: There is no required textbook for this course. Most of the covered material appears in the following optional books, which are placed in reserve at the Babbidge library:

- R. Durbin, S. Eddy, A. Krogh, G. Mitchison, Biological sequence analysis: probabilistic models of protein and nucleic acids, Cambridge University Press, 1998 (call number **QP620 .B576 1998**).
- R.C. Deonier, S. Tavare, M.S. Waterman, Computational genome analysis: an introduction, Springer Verlag, 2005 (call number **QH438.4.M33 W378 2005**).
- K. Lange, Mathematical and Statistical Methods for Genetic Analysis, 2nd ed., Springer Verlag, 2002 (call number **QH438.4.M33 L36 2002**).
- R. Schwartz, Biological Modeling and Simulation, MIT Press, 2008 (call number **QH323.5 .S364 2008**).

Prerequisites: Undergraduate-level courses in biology, programming, and statistics. However, required background in these areas will be provided as needed via lectures and supplementary readings.

Grading: Grading will be based on in-class quizzes given throughout the semester, theoretical homework assignments and programming assignments reinforcing the material covered in lectures, a final project, and class participation, according to the following breakdown:

In-class quizzes	10%
Theoretical homework assignments	20%
Programming assignments	20%
Final project	40%
Class participation	10%

Assignment submission: Solutions for theoretical homework assignments and deliverables for the final project must be submitted in electronic format via Moodle (see below). Programming assignments must be submitted electronically via the Rosalind site at rosalind.info. Rosalind is a repository of intellectually stimulating problems of varying difficulty that are extracted from real challenges of molecular biology. Solutions can be prepared using any high-level programming language. You will be asked to process a dataset generated by Rosalind on your own computer and then upload or copy-paste the solution to Rosalind along with your source code. Each submitted solution is automatically checked for correctness, allowing you to fix potential problems before the due date.

Late policy: All assignments are due by midnight on the specified due date. Late submissions are allowed for up to three days with a 10% penalty for each late day. Assignments that are more than three days late and make-up quizzes will not be allowed, however, to accommodate unforeseen circumstances that may prevent timely submission, the lowest quiz, homework assignment, and programming assignment scores will be dropped from the overall grade calculation.

Final project: The final project will give you the opportunity to study a computational genomics problem in more depth. You are encouraged to devise your own final project topic; suitable topics include surveys of computational genomics topics not covered in the lectures, design and implementation of novel algorithms, theoretical analyses, and empirical evaluation of existing methods. Project requirements will include submitting 2-3 intermediate progress reports and a written final report of 15-20 pages. You will also be required to give a short presentation on your project at the end of the semester. Although working individually is acceptable, completing the final project in teams of 2-3 students is encouraged.

Class participation: Each student will be expected to scribe notes for 1-2 lectures during the semester and participate in discussions of progress reports and final project presentations of other teams.

Moodle site: Course announcements and class related materials including scribe lecture notes, handouts, assignments, grades, etc. will be distributed using Moodle. To get access to the Moodle site you must first *create a new account* at https://dna.engr.uconn.edu/moodle/login/ then self-enroll for the Computational Genomics course using enrolment key "DNA". *Note that this is a local Moodle installation and you will not be able to login using your netid info.* You can also use Moodle to ask class-related questions and communicate with your peers and the instructor. Please observe basic etiquette by keeping your postings polite, concise, and on-topic. Appropriate questions are general questions about the covered material and clarifications on the assignments. For questions that are specific to your own work you should contact the instructor directly.

Academic integrity: You are expected to adhere to the highest standards of academic integrity. All submitted solutions must be your own work. For homework assignments and programming projects you may discuss ideas and concepts with others, but must not share written solutions or code. Use of published materials (including web resources) is allowed, but all sources should be explicitly acknowledged in your solutions. Violations will be reviewed and sanctioned according to university policies.

Students with disabilities: If you have a documented disability for which you are or may be requesting an accommodation, you are encouraged to contact the instructor and the Center for Students with Disabilities or the University Program for College Students with Learning Disabilities as soon as possible to better ensure that such accommodations are implemented in a timely fashion.