CSE3810/CSE6800: Computational Genomics – Spring 2023

Lecture:

Mo/Wed/Fri 2:30-3:20PM, ITE 127

Instructor:

Ion Măndoiu
ion@engr.uconn.edu
Office Hours:
Mo/Tu/Wed/Th 11:30am-12:30pm or by appt.
ITE 261

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 genomic 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.

Prerequisites: Undergraduate-level courses in biology, programming, and statistics.

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 course material is covered in the following optional books:

- R. Durbin, S. Eddy, A. Krogh, G. Mitchison, Biological sequence analysis: probabilistic models of protein and nucleic acids, Cambridge University Press, 1998 (library permalink).
- R.C. Deonier, S. Tavare, M.S. Waterman, Computational genome analysis: an introduction, Springer Verlag, 2005 (library permalink).
- K. Lange, Mathematical and Statistical Methods for Genetic Analysis, 2nd ed., Springer Verlag, 2002 (library permalink).
- R. Schwartz, Biological Modeling and Simulation, MIT Press, 2008 (library permalink).

Course website

We will use a course website hosted using Moodle at https://edx.engr.uconn.edu/. Please check this site regularly to access assignments, grades, and course handouts. The Moodle site also includes a discussion forum to ask class-related questions and communicate with the instructor and your peers. 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. You must not post extensive code fragments in public messages – for specific questions about your work you should contact the instructor directly.

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

Quizzes	10%
Theoretical homeworks	20%
Programming assignments	30%
Final project	40%

Assignment submission: Solutions to both theoretical homeworks and programming assignments must be submitted in electronic format via Moodle. The recommended language for solving programming assignments is Python. Solutions to the programming assignments will be automatically checked for correctness on a set of standard test cases, allowing you to receive immediate feedback and fix potential problems before the due date.

Late policy: In-class quizzes are due at the end of the class meeting. All other assignments are due by midnight on the specified due date. For theoretical homeworks and programming assignments 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, theoretical homework, and programming assignment scores will be dropped from the overall grade calculation.

Final Project: The final project aims to give you the opportunity to study a computational genomics application in more depth. Although working individually is acceptable, completing the final project in teams of 2-3 students is strongly encouraged. Suitable project topics include empirical comparison of existing computational genomics tools, novel analyses of existing genomics datasets, implementing and evaluating new statistical models and algorithms, etc. Final project requirements will include several written and oral progress reports as well as a written final report of 15-20 pages and a final presentation during the time allocated for the final exam. Full final project details including a list of potential topics will be provided towards the middle of the semester.

Academic integrity: You are expected to adhere to the highest standards of academic integrity. For homework assignments and programming projects you may discuss ideas and concepts with others, but must not share written solutions or code. All submitted solutions must be your own work. *Submitting solutions from various web sources as your own is considered academic misconduct and will be sanctioned according to the University's Academic Integrity Policy.*

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.