Overview

Most subdisciplines in modern biology involve the analysis of large amounts of data, inference based on probabilistic models, or both. This course exists to help students gain skills in data analysis and combines elements of applied computational science (data wrangling, computer systems, etc.) with probability and statistics. These are challenging aspects of modern research and can take years to master, which might be one reason why some people neglect to develop their skills in this area. However, you will be able to learn and do new useful things very quickly.

Computational biology can open a world of possibilities for research and employment. The computational and analysis skills we will practice are increasingly imperative for modern biologists (e.g., see the mention of computation, statistics and quantitative analysis in many job advertisements). Without them the scope and ambition for biological research are unnecessarily constrained. In addition, the problems and computational approaches we will study are good application domains for students from outside biology who are interested in pursuing computational science.

This course will be motivated by practical applications of probability, simple mathematics and computational tools to biology. In each section of the course we will begin with biological questions and then investigate computational methods for graphical and statistical analysis of real data sets.

I have a few main goals in this class: 1) to discover the importance of probability, mathematics and computational methods in biology, 2) to understand philosophies and conceptual frameworks underlying commonly-used statistics and analytical methods, and 3) to become proficient with analytical and computational tools that can be applied to biological problems. I hope that by the end of the course you will have an appreciation of the diversity of applications of these analytical tools and the role they play in modern biology, and have begun to develop proficiency in these areas.
Course Materials


2. Additional materials will be distributed in class or linked from the course website.

3. Optional supplemental and potentially useful materials:
   
   
   
c. B. M. Bolker. 2008. *Ecological Models and Data in R*. Princeton University Press. (we will read and use several chapters from this book; I will provide pdf copies of chapters).
   

Assessment and Grading

Assessment of your work in this course will be done regularly by you. You will do independent work (e.g., reading, exercises, problem solving) and determine through interactions with the class what you did correctly, where you need further study or have questions, and so forth. Several small assignments will be graded for completion (not accuracy) to offer an incentive to keep up with classwork. Given that this is an upper-level undergraduate or graduate course, it will be your responsibility to gauge your competency and understanding of these exercises. I will help motivate you by offering challenging and engaging course material, but your success will depend heavily on your personal motivation.

Your performance will be evaluated formally based on three write-ups of lab projects, two exams, and several small assignments (graded for completion). In addition, graduate students will write and present short papers on specific topics in computational biology. Grading will be on a standard scale (i.e., 90–100=A, 80–89=B, etc.).

<table>
<thead>
<tr>
<th></th>
<th>Undergraduate</th>
<th>Graduate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project papers</td>
<td>3 × 60 =</td>
<td>180</td>
</tr>
<tr>
<td>Exams</td>
<td>2 × 30 =</td>
<td>60</td>
</tr>
<tr>
<td>Small assignments</td>
<td>10 × 4 =</td>
<td>40</td>
</tr>
<tr>
<td>Graduate student project</td>
<td>—</td>
<td>30</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>280</strong></td>
<td><strong>310</strong></td>
</tr>
</tbody>
</table>
Project papers

Three of the four sections of the course will culminate in a project paper. The task is to report findings of a computational analysis and to interpret them in the context of the biological questions that motivated the study. The format and length of these will vary slightly from topic to topic. Project reports will be written individually, but they will be based on collaboration among students and the instructor. Due dates will be announced in class and posted on the course’s WyoCourse website. Papers will be submitted electronically, via the WyoCourse website.

Short exams

Two short exams (scheduled for 28 March and 4 May) will be used to assess your understanding of major concepts. In practice people do statistics, mathematics and computation with the assistance of books, computers and colleagues. Thus, for the exams, I will ask you to address concepts and “big-picture” items, things that you would expect to be able to do without books, etc.

Questions on the exam will call for short answers of a few sentences or short essays. We will discuss sample questions before the first exam to illustrate the types of questions I will ask.

Small graded assignments

Exercises and problem sets will be associated with several subjects and will be assigned regularly. Ten assignments will be graded for completion (will be listed on the course’s website) and will be submitted electronically through the website. You will have opportunities to ask questions about assignments during class, but you will require time outside of class to complete the exercises. A subset of these graded assignments will be done as a group and we will compare these critically as one would for a bake- or cook-off at a fair.

Note-taking and questions

You will want to take notes on reading assignments, make outlines of important concepts, and make notes on computer code that is of use to you. In class meetings I will present lectures, answer questions, offer further explanations of material and facilitate discussions. The course’s website is an additional resource for asking questions and discussing topics outside of class, and we will use it in various ways to interact outside of class.

Graduate students

Beyond the above requirements, each graduate student will write a short paper that introduces a specific topic in computational biology (the equivalent of two substantial $8\frac{1}{2} \times 11$ pages, plus any images, references and URLs to key resources, on an individual web page.
associated with the course). Each student will write on a different topic. The purpose of this assignment is for the individual student to dig deeper into a particular subject, and to share the findings with the class and enrich our literacy in computational science.

In addition to the short paper, graduate students will give short presentations on their findings to the class. The presentation and paper will account for 30 points and the course grades of graduate students will be calculated out of 310 possible points. The due date for the written assignment will be 6 April and students will choose a time for their presentations in the last 3 weeks of classes. Below is a list of suggested topics. Students are welcome to suggest and discuss additional topics with me:

1. Data archiving at Dryad and other public repositories
2. The National Science Foundation’s requirement for data management plans
3. Network file systems for scientific computing
4. Types of parallel computing for science
5. High performance computing in a cluster environment (e.g., Mt. Moran computer at UW)
6. Tools and methods for “reproducible research”
7. Version control systems for documents, software development, and other collaborative work (cvs, svn, git, Google docs, etc.)
8. Hardware and software tools for secure data sharing and storage
9. Modern database technologies
10. Programming libraries for scientific computing (HDF5, GSL, etc.)
11. What, if anything, is a supercomputer?
12. What is Cloud computing and what is its relevance for academic computing?
13. The Cyverse (formerly, the iPlant Collaborative)
14. A comparison of machine learning methods to parametric modeling

**Computer use and etiquette**

We are using a room full of computers for this class. Occasionally, a student has difficulty focusing on the class material and not being distracted by the many distractions that a networked computer offers. These distractions are typically detrimental to the attention of the individual user, but also to the students who sit nearby. Therefore I ask that you find a solution to manage and minimize these distractions for yourself. For example, I ask that you do not consult the web while I or anyone else is presenting material to the class, because it will interfere with your ability to follow the presenter. Likewise, there is no reason to use personal email or other messaging software during class. Of course, legitimate uses of these resources during class exist, particularly during those times when I ask you to use the computers. These include transferring a copy of a file to yourself or another student via email, consulting the web for computational science resources related to class work, taking notes, and trying out code as I present it on screen. Please use your best judgment and minimize distractions to yourself and others.
**Additional Items**

- The schedule of topics, assignments, and all other details in this syllabus are subject to change with fair warning, including announcements in class or via university email addresses and the course website.
- Any student who has a disability and is in need of classroom accommodations please contact the instructor and the University Disability Support Services.
- Students whose religious activities conflict with the class schedule should contact the instructor at the beginning of the semester to make alternative arrangements.
- Cheating and other forms of academic dishonesty are listed in University Regulation 802, Revision 2. If you are found to be engaged in academic misconduct, at a minimum you will receive no credit for that exam or assignment. Repeat or serious offenders can expect more serious consequences.

**Overview of Topics**

1. **Project I:** Analysis of body weight and lifespan of laboratory mice  
   - Class meetings 2–8, 3.5 weeks  
   - Introduction to computational science  
   - Exploratory data analysis and simple tests of hypotheses  
   - Probability Theory I  
   - Introduction to R

2. **Project II:** Poisson processes and the distribution of palindromes in cytomegalovirus DNA  
   - Class meetings 9–14, 3 weeks  
   - Probability Theory II – Poisson processes and associated distributions  
   - Statistical analysis of categorical data  
   - Paper on Project II, due 9 March

3. **Project III:** Improbable Wyoming: parameter estimation using likelihood and Bayesian methods  
   - Class meetings 15–20, 3 weeks  
   - Likelihood and Bayesian parameter estimation  
   - Numerical analysis  
   - Paper on Project III, due 11 April

4. **Project IV:** Discovery and quantification of features of a genome  
   - Class meetings 21–28, 4 weeks  
   - Data wrangling and extraction with large data sets  
   - Hierarchical Bayesian modeling  
   - Paper on Project IV, due by 9 May (Tuesday of final exam week)