BIO321G - Principles of Computational Biology / Computing in the Biological Sciences

Instructor
Vagheesh Narasimhan
vagheesh@utexas.edu
Lectures: TTh 9:30am – 11:00am
Office hours: Th 10:30am-noon at https://utexas.zoom.us/my/vagheesh1 or by appointment

TA
Carly Scott
Office hours: W 9:30am – 11:00am at https://utexas.zoom.us/j/99526308692 or by appointment

Some of these office hours for particular weeks will be recitation sections to help with getting setup and getting dependencies loaded for projects and will be held at the same time as the office hours stated above. The timing of these will be announced on Canvas and in class.

Pre-requisites
Biology 325 or 325H; Statistics and Data Sciences 328M; and Mathematics 408C, 408S, or 408R. As CS 303E or SDS 348

Class Overview

The will serve as a first in a series of courses planned to build student expertise in computational analysis of biological datasets, which are rapidly increasing in size and complexity. Our primary goal is to teach you how to program using Python and R, particularly working with biological datasets. In later classes (BIO 382K, BCH 394P/BCH364C) you will use this knowledge to solve specific problems in different areas of biology. However, we also have a very important secondary goal: to teach you how a computer program is written and how it functions independent of the programming language. So we will teach you "algorithms" (which is actually a very well defined word in this context) and a bit about how computers actually work. The assignments and lectures will all use examples from many domains of biology and you will learn to work with sequence, matrix and imaging data all of which are becoming ubiquitous in the field. Interspersed throughout the course will be lectures about how different kinds of modern genomic data (next-generation sequencing, gene expression, DNA methylation) are generated and represented, and you will learn tools that have been built to rapidly index, manipulate and store these types of datasets. You will also learn to use the Texas Advanced Computing Cluster (TACC), the 9th fastest supercomputer in the world (https://www.top500.org/lists/top500/list/2020/11/) to analyze large datasets. There are no textbooks for the class, you will learn everything you need from the lectures and completing the programming assignments which will be designed to slowly ramp up in difficulty and build on your previous knowledge.

The class is organized into three different components.

1. Introduction to the basics of Python
2. Introduction to algorithms and their software implementations in Python and R
3. Biological applications

Class Schedule / Topics (roughly in order)

We will cover the following topics over the semester, interspersing algorithms with the analysis of different types of biological datasets as your knowledge of computing progresses. The programming assignments are designed to use the data specified in the topics of biological application below.

Topics in computing	Language
Introduction to computation	Python
Loops and iteration	Python
More loops and strings	Python
Functions and abstraction	Python
Compound data types	Python
Recursion, dictionaries	Python
Error correcting and debugging	Python
Object-oriented Programming	Python
Python classes and inheritance	Python
Timing and complexity	Python
Examples of complexity	Python
Simple algorithms	Python
Vectors, matrices and data frames	R
Statistics	R
Data visualization	Tableau

Topics in biological analysis of large, complex data

Biological sequence analysis
Gene expression data (normalization, imputation, differential analysis)
Hidden Markov Models
Principal Components Analysis
Population genetics and ancestry inference
Genome wide association studies
Time series analysis of ancient DNA data

Grades

Your grade is made up of the following components:

Programming assignments	70%
Mini-project	20%
In-class quizzes	10%

There will not be any exams in this class.

Programming assignments

There will be a number of programming assignments throughout the semester. These are the most important aspect of the class and therefore contribute the majority of the class grade. Each of these assignments will have specific tasks in increasing order of difficulty. You can submit them with some but not all of the tasks completed and will be assigned a grade for them based on the sections that you managed to get to work successfully. The grading will be completely electronic (we will compare your output file with a reference file), so you will receive specific instructions on how you format your output. If you format the files incorrectly, you will not receive partial credit, so remember to check that you follow the instructions of each assignment carefully. While you are welcome to verbally consult with your peers on many of these assignments, it is vitally important that you write the code on your own. The programming skills needed, both in later classes and in your scientific career, will only be developed if you learn to write these programs on your own.

Mini Project

You have to options two options to pick from for your mini project that is designed to be completed from Spring Break till the last day of the class. Option 1 is to pick an area of biological application / hypothesis and analyze publicly available data that you will download and analyze. Option 2 is to write a review paper about a specific area of biology where computational methods have made recent headway. Specific formats for each of the two options will be uploaded on Canvas. Students should think about an area that interests them and discuss their mini project with the instructor during one of the office hour sessions or by appointment and be approved to work on the assignment prior to spring break.

Quizzes

There will be quizzes during the class (which can also be completed in the week of each lesson) offline if need be on the canvas site. The quizzes are intended to test your understand of the lectures at the end of each week. If you follow what is happening in the lectures, you will be able to answer the questions in the quizzes.