Computational Genetics - Bio 399

Instructor: Andy Kern (adkern@uoregon.edu)

Office: Pacific Hall 301

GE: Danny Brown (dabrown@uoregon.edu)

Introduction

Biology is now a computational field, where large, multidimensional datasets are the norm. In this class we introduce quantitative approaches to problems that arise naturally in genetics and genomics through a computational lens. The course is project based, with each week centering around a new problem in biology. Students will write code in Python to visualize and statistically analyze datasets that result from modern experiments such as genome sequencing and RNA transcription profiling. Students are **not required** to have any previous experience with programming– if you do have previous experience consider this an excellent opportunity to practice "Beginner's mind". The course will be taught in an interactive, workshop-like environment where we will go over code together. This environment will benefit the active participant, so please do you best during class to concentrate on the material at hand rather than blowing it off until later (even if Andy is terribly boring). Below is the tentative schedule for the course. Depending on the speed we go through topics there might be shifts in the schedule.

Learning objectives for this class

- Gaining some level of fluency with the Python programming language
- Develop skills at dealing with biological data, from importing large datasets, to visualization, to fitting statistical models.
- Gain appreciation for a broad survey of quantitative methods in genetics / genomics
- Do some actual research on SARS-CoV-2!

Slack space

In addition to Canvas and normal email I'll be putting up a slack chat space for the class. If you have not yet used Slack it is an excellent way to message people that takes a middle ground approach between emails and text messages, and it's particularly useful for code based efforts.

Requirements and Evaluation

The class is organized around weekly problem sets which our in class work and discussion will prepare you for. Problem sets will be a combination of coding and analytical work, and will be turned in via preparing a jupyter notebook and

sending either a path on our server or the notebook itself via email to Andy. My goal is to have the problem sets be challenging and interesting for you but not evaluative– think pass/fail. There will be no quizes or tests. There will be a final group project where you will partner with 1 or 2 other students (depending on enrollment) to do a bit of research on your own and then present to the class.

Discussion Section

Discussion will be a chance for further review of the programming material with me or the GE. Discussion section is meant to fill in details of what we cover in lectures, not go beyond what we have learned in class. Having said that the goal of discussion section is to give you all a chance to ask questions beyond what we might have time for in class.

Final Project

The last week of class will be small groups presenting their final projects to the class. **We will organize into groups for this in Week 5.** The idea will be for the group to come up with it's *own research idea* or *piece of science communication*, centered around the themes of the class and then take that idea to Andy for approval. With approval in hand the group will go off, do their bit of research or project, and then present it to the class in a 15min presentation during the last week of class. Some example ideas for the final project might be like: creating a phylogenetic tree of a SARS-CoV-2 and interpreting it, doing simulations of the SIR model, developing a classifier to differentiate genetic variation, creating an animation describing the spread of COVID-19 or some other pathogen, or perhaps even creating an audio piece that debunks bad science. This should be fun and creative, not a slog.

Class Courtesy

Our classroom is a learning environment, and as such should be a safe, inclusive and respectful place. Being respectful. Disrespecting fellow students as well as combative approaches, tones and/or actions are not acceptable. Please make me aware if there are classroom dynamics that impede your (or someone else's) full engagement.

Students with Disabilities

I strongly encourage students with disabilities, including "invisible" disabilities like chronic diseases, learning disabilities, and psychiatric disabilities to discuss with me as soon as possible what appropriate accommodations might be helpful to them. You are also encouraged to contact the Accessible Education Center in 164 Oregon Hall at 541-346-1155 or uoaec@uoregon.edu.

Week 1. Computational Genetics Intro

Basics of nomenclature, our server, jupyter notebooks, python baby steps, Printing, data structures, indexing, control flow, conditionals, looping, functions. Case study: plotting COVID cases

Week 2. Tracking the spread of SARS-CoV-2

Using numpy and scipy to boost our productivity. ndarrays, indexing, etc. Case study: fitting exponential models to growth of SARS-CoV-2

Week 3. Dealing with sequence data

Using Biopython, managing sequences, alignment Case study: aligning SARS-CoV-2 genomes, extracting genomic features

Week 4. Computational Phylogenetics

Trees and tree thinking, clustering, Maximum likelihood principle, ML phylogenetics Case study: building trees from SARS-CoV-2

Week 5. Unsupervised clustering

Hierarchical clustering, PCA, UMAP Case study: clustering SARS-CoV-2 and visualizing these clusters

Week 6. Modeling evolution

Using our skills to simulate some the population genetic model of genetic drift and natural selection. Case study: calculating time to fixation of neutral mutations vs. selected mutations

Week 7. Modeling infectious disease

SI, SIR, and SEIR models, numerical integration of ordinary differential equations, SciPy Case study: simulating SIR model, estimating COVID params

Week 8. Genome-wide association studies

Genotype to phenotype mapping, linear models, GWAS Case study: host genetic variation and its impact on COVID

Week 9. Machine learning

IMPORTANT – no class on 5/29.

Intro to machine learning, Support Vector Machines, Random Forests, Neural Nets Case study: country of origin prediction of genomes using ML

Week 10. Final Project Presentations

Yay let's see everyone's cool work.