Syllabus for GEN220: High Throughput Biological Data Processing

Course Description

This course focuses on computational skills for processing data using programming language Python and UNIX environment. No prior programming experience is required, but some basic computer skills will be useful.

With the advancement of high throughput data generation methods, a major challenge that graduate students in life sciences have to face today is to analyze large amount of biological data. The objective of this course is to provide an opportunity for graduate students with no computer science background to learn the basic skills of handling high throughput biological data. It covers the Linux/Unix environment and the importance of the command line interface; the Python programming language; program design, implementation, and testing; BioPython; Strategies for analyzing genome resequencing, RNASeq, sequencing data. Students build hands-on skills by analyzing real high throughput biological data through homework assignments and team projects.

Units: 3

Instructor: Jason Stajich (jason.stajich@ucr.edu)

Time and location: W 4:10-5:00PM, F 3:10-5:00PM, ULB104

Office Hours: By Appointment, 1207K Genomics

Prerequisites

- Coursework in genetics or molecular biology or permission of instructor

Resources

None of these texts are required for completion of the course but they will provide a great deal of helpful background and examples that will improve your ability to master UNIX or Programming in Python.

1. *Bioinformatics Data Skills: Reproducible and Robust Research with Open Source Tools*. Vince Buffalo. 2015 O'Reilly & Associates. Available from O'Reilly and Associates, Amazon
2. *Unix and Perl to the Rescue: A Primer*. Keith Bradnam and Ian Korf. Unix and Perl Primer for Biologists
3. *Unix and Perl to the rescue!* Bradnam and Korf. Amazon
4. Rosalind - An online platform to learn bioinformatics and programming in Python.
5. Software Carpentry - https://software-carpentry.org/ and Data Carpentry - http://www.datacarpentry.org/.
6. Berk Ekmekci, Charles E. McAnany, Cameron Mura. An Introduction to Programming for Bioscientists: A Python-Based Primer. PLoS Comp Bio. DOI: 10.1371/journal.pcbi.1004867

Grading

- Programming Homework assignments (5 in total): 50% of grade
- Team project: 50% of grade

Homework

- Homework is due before class on the date listed in the syllabus or if amended when assignment is given.
- There will be a programming assignment every two weeks during the first half of the course. Programming assignments must be prepared along with any necessary input files or documentation to demonstrate program usage.
- Code should be runnable as turned in. You will deposit your code in your github repository or if not possible, by iLearn. You can make one private personal repository to deposit and should organize a folder for each homework assignment (e.g. hw1, hw2, hw3).

Projects

- Project Topics will be discussed in October and teams will select a project idea to focus on.
- Project will be 1 or 2 individuals working together.
- A presentation will be made by each team - last day(s) of class.

- A final report with the details will be turned in by the group.
- The report needs to detail what each person's contribution is to the project.

Schedule

Date	Day	Lecture Topic	Notes
Sep-28	F	Course Outline / UNIX I: Cmdline, GitHub	
Oct-3	W	UNIX II: Biocluster HPCC, Running programs	
Oct-5	F	UNIX III: Tools for data processing	HW1 Due
Oct-10	W	UNIX IV: Advanced UNIX and data processing	
Oct-12	F	Python 1 - Variables, running, cmdline, strings, math	HW2 Due
Oct-17	W	Python 2 - Logic, loops, lists, iterator	
Oct-19	*F*	*No Class*	HW3 Due
Oct-24	W	Python 3 - I/O reading/writing files, directories	
Oct-26	F	Python 4 - Dictionaries, Arrays, functions	HW4 Due
Oct-31	W	Python 5 - Libraries, packages, BioPython	
Nov-2	F	Python 6 - Structured data (CSV, XML, GFF, BED)	HW5 Due
Nov-7	W	Bioinformatics 1 - BLAST, cmdline & automation	
Nov-9	F	Bioinformatics 2 - Aligning short reads, coverage, identifying variants	
Nov-14	W	Bioinformatics 3 - Genome Assembly	
Nov-16	F	Data Plotting and R graphics	
Nov-21	W	Bioinformatics 4 - RNASeq	
Nov-23	*F*	*Thanksgiving Holiday - No class*	
Nov-28	W	Bioinformatics 5 - Protein Sequence analyses	
Nov-30	F	Bioinformatics 6 - Microbiome analyses	
Dec-5	W	Bioinformatics 7 - Phylogenetic Trees	

Date	Day	Lecture Topic	Notes
Dec-7	F	Presentations	