Practical Computation for Biosciences
Spring 2016

PLS 599 Independent Study
Project Description

Instructors:

Eric Lyons

Marley 441F (and Keating 3rd floor)

elyons.uoa@gmail.com

Shelley McMahon

Marley 441G (and Herring 104)

mcmahonm@email.arizona.edu

Course type: Project-based
Credits: 2
Schedule: Tue 2-3pm; location TBA

Objectives:
Management and manipulation of large-scale data have become critical skills for professional biologists. Creating custom pipelines, formatting data for input to various published analytical tools, and developing new tools for data validation are becoming more and more important as the rate of data creation, such as from high-throughput genomic technologies, skyrockets. This **Independent Study** course will provide a framework for students to learn how to manage and process their data from the command line.

This course will begin with lessons on using the unix shell, manipulate text from the command line, and program in Python. Students will develop a data processing project relevant to their research interests. Our goal is that each student will learn successful project management (e.g., using github and wikis), develop confidence in using high performance computing, and advance their thesis projects. No previous knowledge of scripting or programming is required.

Four fundamental topics to be covered:
- I. The Unix Shell
  - Essential commands for files, directories, globs
  - Shell environment and variables
  - Scripting in the shell and your .bash_profile
- II. Text manipulation
  - Regular expressions (concepts common to text editors, bash, perl, python, sed, awk)
- III. Programming with Python
  - General concepts in programming (variables, control structures)
  - Python particulars (dictionaries and complex data structures, input/output, classes, Perl-type regex)
  - Application of Python to problems in biology (e.g., through use of Biopython)
- IV. External resources
  - UA HPC
  - Virtual Machines and cloud computing
  - Code repositories and versioning (git, github, sourceforge, svn)
  - Documentation for using programs and reproducible science

Approach:
Computation skills are best learned by doing. Many computationally competent post-docs and faculty in the biosciences were self-taught (in many cases, out of desperation), but this takes time. This course aims to make that process more efficient and productive by providing guidance and structure for concentrated yet independent learning, as well as an opportunity to learn from other students facing similar challenges. Therefore the course is formatted as a **2-unit project-based Independent Study**. The instructors will provide short lectures at the beginning of the course, and online tutorials will be required. Three weeks into the semester, students will begin to plan their semester projects. Throughout the semester, the class will meet weekly as a group for students to report on their project, discuss any problems or questions, and have ad hoc lectures on specific topics. Instructors will also be available for consultation via email.

Online tutorials will be the primary sources for learning the skills. Students will be expected to complete selected tutorials at the following sites:
- Ryans Tutorials: Linux
  - This site (and/or similar sites) provide exercises in bash scripting and regular expressions. Additional exercises will be provided as needed.
- Codecademy (codecademy.com)
  - This is a good general introduction to programming in Python. Completing the online course will be required.
- Rosalind (rosalind.info)
  - This site will allow the use of the skills developed so far to solve problems in bioinformatics. A brief Python tutorial/refresher is available, and several problems from the "Bioinformatics Stronghold" will be assigned. Students will be encouraged also to explore the "Bioinformatics Armory" in preparation for the semester project.
- Software Carpentry (software-carpentry.org)
  - Students will be guided to particularly relevant portions of the curricular material on this site and will be encouraged to attend a workshop as feasible.

Registration:
Enrollment in this section of PLS 599 will be capped at 10 this year. Grades will be S/P/F. Email Dr. Lyons or Dr. McMahon to indicate interest, and consult with Georgina Lambert in SPLS, glambert@email.arizona.edu, for registration details.

Required:
Access to a computer running Unix or Linux (including Mac OSX) will be needed; best would be a laptop that can be brought to group sessions. Please contact instructors prior to start of class for help setting up a Windows machine *(we recommend setting up Virtual Box to run linux virtual machines)*.