PLS 599 — Practical Programming for the Biosciences
Spring 2015
Draft "Project Description" for Independent Study

Objectives:
Management and manipulation of large-scale data has become a critical skill set 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.

In this course, graduate students will be guided through learning three main topics. Examples will be drawn from standard bioinformatics problems, such as DNA sequence analysis, but the skill set will apply to any realm deploying "big data." No previous knowledge of scripting or programming is required.

The three central, and connected, topics will be:

I. Text manipulation
- Regular expressions (concepts common to text editors, bash, perl, python)

II. The Shell
- Essential commands for files, directories, globs
- Scripting in the shell and your .bash_profile

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)

Approach:
For Spring 2015, the course will be formatted as a 2-unit Independent Study. The instructor(s) will meet once at the beginning of the semester with each student independently to allow an initial assessment of the student's familiarity with any of the topics above. Three weeks into the semester, the instructor(s) will meet with the students independently again, to plan a semester project. Throughout the semester, the students will meet weekly as a group with the instructor(s) (time and location TBA) to go over any problems or questions. Instructor(s) 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:

- Bash Guide for Beginners (at The Linux Documentation Project)

  This site (and/or similar sites) provide exercises in bash scripting and regular expressions. Additional exercises will be provided by the instructor(s) 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.

Additional resources will include access to the book Practical Computing for Biologists, by Haddock and Dunn (2011, Sinauer Associates, Inc.). Although a few years old, this is a good text and will provide additional background for many topics, all from a biologist's point of view. Purchasing the book is recommended but not necessary; relevant material will be made available as needed.

After developing a good understanding of python, students will be encouraged to explore Biopython (biopython.org), an extensive set of libraries for use in sequence manipulation and other relevant tasks.

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 see instructor(s) prior to start of class for help setting up a Windows machine.