Applied Bioinformatics
PATH9577B and BIOL9919B

Regularly Scheduled Class Meeting Time
January 9, 2024 – April 4, 2024
Tuesdays and Thursdays 2:30pm – 4:30pm
Room: Kresge K103

Course Description

This course will train students to process biological data sets by developing basic 'command-line' and script-based computing skills. Topics covered include the UNIX file system, basic UNIX commands, running programs and building pipelines, and data exploration and visualization in R. Students will also develop the skills needed to understand the purpose, the parameters, and interpret the output of bioinformatic tools for their field of study.

Course Instructors and Coordinators

- **Dr. Christina Castellani**
  Department of Pathology and Laboratory Medicine, Schulich School of Medicine & Dentistry
  christina.castellani@schulich.uwo.ca

- **Dr. Art Poon**
  Department of Pathology and Laboratory Medicine, Schulich School of Medicine & Dentistry
  apoon@uwo.ca

- **Dr. Vera Tai**
  Department of Biology, Faculty of Science
  vtai4@uwo.ca

Course Summary

Bioinformatics has become an essential skill set in modern biology and biomedical research. This is largely driven by new genetic sequencing technologies that generate gigabytes of data overnight, but other technologies such as remote sensing and image/signal processing are driving similar challenges in broader fields including ecology, pathology and neuroscience. Although commercial software feature graphical user interfaces and 'one click' analysis workflows, these are often expensive, proprietary (closed source) programs that are constrained to a narrowly defined selection of the most popular analyses. However, biology is diverse (different organisms don't play by the same rules as model species), and research is driven by innovation, customization and asking new questions. Consequently, there is a widespread demand for the ability to customize bioinformatic workflows.

The objectives of this course are to provide students with diverse backgrounds and no prior experience with programming with a basic foundation in bioinformatics. However, greater emphasis will be placed on programming and developing skills to customize analyses, than on using existing programs. The course is structured to resemble a standard bioinformatic workflow: data files are obtained and managed within a UNIX-like system; open source tools and a scripting language such as bash or Python is used to manipulate and clean the data; and

the processed data are analyzed and visualized using a language like R. Students will also investigate existing bioinformatics tools or programs and develop skills to understand their parameters, how they are operated, and how to interpret their output.

Course Learning Objectives

1. understand the role of bioinformatics in modern biological and biomedical research
2. develop knowledge and skills to operate bioinformatics tools and interpret their output
3. manage files with a user account in a remote UNIX-like computing environment
4. run programs on the command line with arguments, and control data streams
5. analyze genomic sequence data using various bioinformatic tools
6. understand how to reduce a complex problem down to the simplest steps
7. write concise and informative comments and documentation within and outside a script
8. import and manipulate data frames in R
9. explore data by drawing and customizing plots using the base R graphics package
10. generate visualizations of three or more variables using projections and dimensionality reduction methods in R

Course Materials

There is no textbook for this course. For course readings, we refer students to a small number of online resources that can be accessed at no cost, including materials that are developed and maintained by the course instructors:

- Course readings on data exploration and visualization with R by Art Poon: http://filogeneti.ca/ApplBioinf/
- basic UNIX commands: https://github.com/PoonLab/courses/blob/master/PATH9577Q/Readings/basicunixcommands.md
- The Linux Command Line by William Shotts, https://linuxcommand.org/index.php
- Fundamentals of Data Visualization by Claus O. Wilke, https://serialmentor.com/dataviz/

Course Schedule

The course schedule comprises two 2-hour sessions per week for 12 weeks (one term).

Day	Topic	Instructor
January 9	Introduction, review of syllabus. What is bioinformatics? The history and philosophy of UNIX. Connecting to a remote system.	VT
January 11	UNIX file system structure and navigation. Absolute and relative paths. Wildcards. Managing files (cp, mv, rm) and directories (mkdir, rmdir). Editing files with nano.	VT
January 16	User permissions. Running programs on the command line. PATH. Redirecting data streams.	VT
January 18	Working with plain text files and understanding data formats, with wc, grep, awk, sed, cut.	VT
January 23	Running standalone BLAST. Bash scripting.	VT

January 25	Next generation sequence (NGS) analysis. Reference mapping.	VT
January 30	Open source: the missing lecture. Licenses. Checksums. File compression (zip, gzip, xz). Compiling with make, cmake. Package managers and distributions (apt/yum, homebrew, pip, CRAN, Bioconductor). Virtual environments.	AP
February 1	An introduction to R. Vector types, variable naming and assignment. Arithmetic.	AP
February 6	Lists. Data frames. Importing tabular data. Summary statistics. Logical operators. Subsets and sorting.	AP
February 8	Control flow. Loops and conditional statements. Functions and functional programming (apply).	AP
February 13	Dates are awful: as.Date, strptime and lubridate in R.	AP
February 15	Theory of data visualization in R. Base plot functions (points, lines, rect). Layers. Space, shape and colour.	AP
February 20	*Spring reading week*	
February 22	*Spring reading week*	
February 27	Data visualization in R with ggplot2	CC
February 29	Data exploration and quality control	CC
March 5	*Student presentations*	
March 7	*Student presentations*	
March 12	*Student presentations*	
March 14	*Student presentations*	
March 19	Hierarchical clustering	CC
March 21	Linear models and linear mixed effects models	CC
March 26	Introduction to differential expression analysis	CC
March 38	Extracting biological information from gene lists	CC
April 2	*Oral examinations*	
April 4	*Oral examinations*	

Methods of Evaluation

- **Paper review (10%).** Find an exemplary paper from your research field of interest and write a short (1-2 page) description of the role that bioinformatics played in that study.
    - **Due February 1st 2024**

- **Assignments (50%).** Students will be expected to complete homework assignments using the methods presented in class to carry out bioinformatic data processing and analysis.

- **Oral presentations (15%).** Based on their paper review assignments, students will choose or will be assigned a bioinformatic program or tool to present to the class. Presentations should be 15 to 20 minutes and include a description of the tool's purpose, what features distinguish it from similar tools, an explanation of its parameter settings, and how its output can be interpreted. Students will sign up to present over a number of weeks during the term.

- **Oral examination (25%).** Students will be evaluated on an individual (one-on-one) basis by one of the course instructors on methods covered in the class assignments.

Statement on Academic Offences
Scholastic offences are taken seriously and students are directed to read the appropriate policy, specifically, the definition of what constitutes a Scholastic Offence, at the following Web site: http://www.uwo.ca/univsec/pdf/academic_policies/appeals/scholastic_discipline_grad.pdf

All required papers may be subject to submission for textual similarity review to the commercial plagiarism-detection software under license to the University for the detection of plagiarism. All papers submitted for such checking will be included as source documents in the reference database for the purpose of detecting plagiarism of papers subsequently submitted to the system. Use of the service is subject to the licensing agreement, currently between The University of Western Ontario and Turnitin.com (http://www.turnitin.com).