

1. Course description (50 words or less)

This short course introduces basic computational environments and tools to graduate students with limited prior experience. It will provide an introduction to UNIX systems, software compilation / installation, cluster job management as well as data formats, management, and visualization. A brief introduction to scripting programming languages will also be presented.

2. Course goals

Students enrolled in the class will develop skills to accelerate their research in computational research environments. Topics will include an intensive introduction to (a) UNIX systems and software management, (b) data processing and simple programming, (c) data formats and visualization, and (d) software version and cluster control. This training will provide a computational foundation that will allow students to focus on the theoretical and biological aspects of their research.

3. Course competencies

After completing this class, students are expected to be able to attain the following competencies:

Core Competencies:

- Navigate and organize UNIX files and folders
- Compile and install software in UNIX environments
- Understand basic programming data structures and processes
- Create simple scripts to manage and analyze data
- Utilize and apply popular file formats to modern large-scale data sets
- Apply proper visualization tools and strategies to view data
- Utilize software versioning technologies for documenting and organizing software
- Utilize high-throughput computing clusters for parallel data processing

4. Course syllabus

Day 1 (Monday): Introduction to UNIX

Session I - Unified Setup and Installation

Session II - UNIX Command Line and File Structure/Permission

Session III - UNIX System Structure and File Manipulation/Editing

Session IV - Applied workflow of UNIX-based commands

Day 2 (Wednesday): Introduction to Data Processing

Session I - Programming Concepts

Session II - Fundamentals of Data Structures

Session III - Control Structures and Functions

Session IV - Automated workflow for pipeline construction

Day 3 (Tuesday): Data Formats and Visualization

Session I - Data management and manipulation best practices

BIOSTAT 606/HG 606/Bioinfo 606 Syllabus

Session II - Data formats and conversions

Session III - Overview of visualization strategies and practicality

Session IV - Applied workflow using to visualize data formats

Day 4 (Thursday): Versioning and Cluster Computing

Session I - Data and software version control

Session II - Concepts in Cluster Computing

Session III - Cluster job management and submission

Session IV - Parallelization and consolidation of previous workflows

Day 5 (Friday): Unified Analytical Walkthrough

Session I - Divide into groups and describe problem(s)/data sets

Session II - Group Work

Session III - Group Work

Session IV - Presentation of Results

5. Course reading list

None.