PROGRAMMING FOR BIOINFORMATICS – BIOL 7200

Instructor: ***King Jordan;*** *Teaching Assistant:* ***Lu Wang***
School of Biology, Engineered Biosystems Building, Office: 2109
king.jordan@biology.gatech.edu ; lu.wang@gatech.edu

Course Summary: The fields of Bioinformatics and Computational Biology occupy the intersection of the life sciences and information technology. Over the last decade, there has been an explosion of data in the life sciences and the proliferation of raw information promises to continue at an even more rapid pace. Computers are needed to handle and assimilate this massive amount of information. More importantly, the role of bioinformatics is to convert information, in the form of raw data, into biological knowledge. In order to do this, bioinformaticists and/or computational biologists must be adept at the use of computers. In other words, you must know how to code.

This active-learning, project-based course will provide a rigorous and intensive introduction to programming for bioinformatics. We will begin by introducing you to the command line environment in the Unix / Linux operating system – this is where real scientific computing gets done. This will include a fairly broad coverage of Unix / Linux utilities as well as shell scripting. The course will then go on to use the Perl programming language to illustrate the fundamentals of bioinformatics programming.

This class meets for lecture sessions on **Mondays from 4:05-5:55 pm in ES&T L1105** and for exercise/problem sessions on **Thursdays from 4:35-6:25pm in Cherry Emerson 206**. All required and recommended readings, lectures and exercises will be made available on the course T-square site. This is an exclusively practical and active learning class. Students will complete exercises in order to learn how to code and how to do bioinformatics. The only way to learn the course material is by doing. Accordingly, attendance and participation are mandatory and critical. Students are required to attend all lecture sessions on Mondays; exercise/problem sessions on Thursdays are driven by student participation and questions and are intended to provide additional support for students' code development. Students who show up late or miss lecture sessions without an Institute approved excuse will lose 10% of the class participation for each incident. Participation in lecture sessions will be judged by the degree to which each student participates in class discussions and exercise sessions. Students will also have the opportunity to demonstrate and explain their code to the class. Students will also be required to post their code to the course T-square site for evaluation.

Learning Outcomes: Students will be able to:

1. Fluidly navigate the Unix/Linux command line computing environment
2. Accurately use and apply a large set of Unix/Linux utilities and commands
3. Create Unix/Linux shell scripts for data analysis and pipelining

4. Understand the fundamental concepts of software installation including environment variables and dependencies
5. Download and install a variety of bioinformatics software applications based on these principles
6. Navigate and retrieve large-scale data sets from major bioinformatics databases, such as NCBI-Genbank and the UCSC Genome Browser
7. Understand the data models that underlie, and readily convert among, widely used bioinformatics file formats
8. Understand the fundamental concepts and principles that underlie programming in Perl
9. Write Unix/Linux shell scripts and/or Perl programs to parse large-scale bioinformatics data sets and extract useful subsets of information
10. Write Unix/Linux shell scripts and/or Perl programs to integrate and automate bioinformatics analysis pipelines
11. Write Unix/Linux shell scripts and/or Perl programs to solve bioinformatics analysis challenges
12. Understand the concepts that underlie object oriented programming and apply them in order to utilize pre-existing Perl code modules (*e.g.,* from CPAN or BioPerl)
13. Understand and apply the concepts that underlie code parallelization and threading

Honesty and Integrity Policy: We support and encourage collaboration in this course. Students may work together to formulate general problem solving strategies and/or algorithmic approaches. Students may also consult outside resources to help formulate conceptual approaches to the course coding assignments. However, all coding must be done individually. To repeat, the code that you turn in must be your own. Anything taken from the web is not your own. All students' code will be automatically reviewed for plagiarism using state-of-the-art software for code comparison. In addition to these guidelines, you should review Georgia Tech's Academic Honor Code http://www.honor.gatech.edu/content/2/the-honor-code, which you are required to uphold at all times.

Learning Accommodations: We will make classroom accommodations for students with documented disabilities. These accommodations must be arranged in advance and in accordance with the Office of Disability Services (http://disabilityservices.gatech.edu/).

Course Evaluation and Grading:

Class participation (attendance)	20 %
Class demonstrations	20 %
Exercises, quizzes & code	60 %

Schedule of lecture / exercise sessions

Date	Topic	Room
Mon Aug 18	Introduction to *nix environment	ES&T L1105
Thu Aug 21	Open exercise session	CE 206
Mon Aug 25	Basic system administration in *nix	ES&T L1105
Thu Aug 28	Open exercise session	CE 206
Thu Sep 4	Open exercise session	CE 206
Mon Sep 8	Introduction to bioinformatics databases	ES&T L1105
Thu Sep 11	Open exercise session	CE 206
Mon Sep 15	Utility compilation and installation with *nix	ES&T L1105
Thu Sep 18	Open exercise session	CE 206
Mon Sep 22	Advanced regex and file handling	ES&T L1105
Thu Sep 25	Open exercise session	CE 206
Mon Sep 29	Shell scripting and pipelining basic	ES&T L1105
Thu Oct 2	Open exercise session	CE 206
Mon Oct 6	SNP calling pipeline	ES&T L1105
Thu Oct 9	Open exercise session	CE 206
Thu Oct 16	Open exercise session	CE 206
Mon Oct 20	Introduction to Programming Concepts and Perl	ES&T L1105
Thu Oct 23	Open exercise session	CE 206
Mon Oct 27	Perl for Bioinformatics I	ES&T L1105
Thu Oct 30	Open exercise session	CE 206
Mon Nov 3	Perl for Bioinformatics II	ES&T L1105
Thu Nov 6	Open exercise session	CE 206
Mon Nov 10	Regular Expressions in Perl	ES&T L1105
Thu Nov 13	Open exercise session	CE 206
Mon Nov 17	High-throughput analysis with Perl	ES&T L1105
Thu Nov 20	Open exercise session	CE 206
Mon Nov 24	Code optimization with Perl	ES&T L1105
Mon Dec 1	Object oriented Perl & BioPerl	ES&T L1105
Thu Dec 4	Open exercise session	CE 206

Note that the syllabus is subject to change depending on the speed at which the class progresses and the performance of the students.