# CPSC 565
# Perl and UNIX for Bioinformatics
# Matthew Hudson

**Course Information**

**Instructor:** Matthew Hudson
Office: Room 334, National Soybean Research Center
Email: mhudson@uiuc.edu
Phone: (217) 244 8096

**Classroom lecture-laboratory**

**Second half of spring semester**
Lecture / discussion section
M 2 - 4:50pm N-120 Turner Hall
Lab section
W 3 - 4:50pm N-120 Turner Hall.

**Course name:** CPSC 565 – Perl and UNIX for Bioinformatics
**Course web address:** http://stan.cropsci.uiuc.edu/courses/
**Credit hours**: 2

**Prerequisite:** Graduate status or consent of instructor. In addition, familiarity with DNA and protein sequence data, and basic Windows computing skills are required.

**Objectives:** This intensive course is an introduction to high-throughput bioinformatics and genome data analysis. An introduction to programming with Perl and Bioperl will be given, and students will learn to write scripts relevant to their own research goals. We will also cover the use of UNIX and Perl for automating and customizing bioinformatics tools.

The course is designed to be taken by biologists who wish to develop more advanced practical bioinformatics skills. Students should be familiar with the data produced by DNA sequencing, and understand the biological interpretation of DNA and protein sequence. No prior programming or UNIX experience is assumed, but familiarity with GenBank and BLAST, Windows computer skills and the ability to type with reasonable speed are assumed.

This is an intensive course. Students are expected to attend all the sessions and to work independently during the week of the course. Assessment will be based both on the amount of material learned and on the ability to apply the material to real problems. An

‘A’ grade student will be able to write scripts that allow automated processing of bioinformatics data, and process large genomics files, by the end of the course.

## Syllabus Details

The first six weeks will consist of a lecture with examples and interactive questions (one hour) followed by practical exercises using the material from the lecture on the laboratory computers (two hours). The last two weeks will be devoted to the student’s programming projects. Students are encouraged to bring with them problems that need scripting (for example, multiple BLAST searches, extracting data from big FASTA or GenBank sequence files, parsing microarray data, annotation). Projects will be assigned to students who do not have a suitable problem within the scope of the course.

**Each week is 4 contact hours**

Week 1:

Review of key bioinformatics applications (BLAST, FASTA, HMMER, CLUSTALW). Advanced BLAST. Making BLAST databases. Windows, UNIX, DOS and MacOSX. Using the command line. Using ssh to connect to a server. Paths. Shells. Basic UNIX commands (ls, cd, rm, rmdir, nano, less, head, tail, grep, top, ftp, ssh, wget). Running bioinformatics applications on the command line. Redirecting output.

Exercise: Command line bioinformatics. BLAST of a multiple protein sequence file against databases. Different BLAST programs. Redirecting output and grepping the output files.

Week 2:

Shells and paths. The root user and your home directory. Installing and compiling. UNIX power tools. Parallel processing and supercomputers. Binaries and scripts. Shell scripts. Running a script. Sequence files. Phred and phrap.

Exercise: Compile a bioinformatics application from source code and install it for your own use. Write a shell script to execute long commands. Handling multiple processors.

Week 3:

Interpreted languages. Running a ready-made Perl script. Introduction to programming. Introduction to Perl. A basic Perl script. Variables. Precision. Data output (print). How scripts are used.

Exercise: Using Perl in command-line bioinformatics

Week 4:

Data input (STDIN). Performing operations on variables (numerical and string). Running a script. Crashes and die. Files and filehandles. Reading from and writing to files. Iterating through a file. Bioinformatics data files. Parsing.

Exercise: A script to parse the names and hits from a BLAST output file.

Week 5:

Loading information into memory for processing. Data structures (arrays and hashes). Algorithms. Modules. References. System commands and backticks.

Exercise: A script to extract a particular sequence from a FASTA file, BLAST it, and output the top hit.

Week 6:

Using ready-made modules. Bioperl. Real world bioinformatics examples. Perl for parallel processing and clusters.

Exercise: Reformatting files for compatibility, translating DNA to protein, dealing with real sequence data.

Weeks 7 and 8:

Using bioinformatics in the real world.

Exercise: Projects (student's own, or assigned).

**Total contact hours: 32**

Grading and class activities:

Attendance, participation and completion – 20%. Students will be penalized for non-attendance without a valid excuse.

Quizzes (5 in total) – 50%

Programming projects – 30%

Grades are on the plus/minus system
There are no final exams

***Academic Integrity***

Rule 33 of the Code on Campus Affairs and Handbook of Policies and Regulations Applying to All Students (http://www.uiuc.edu/admin_manual/code/rule_33.html) gives complete details of rules governing integrity for all students. Students are responsible for knowing and abiding by these rules.

***Policies on computer resources and copyrights***

All students must adhere to the rules and policies indicated by the software, websites and computer laboratories used for course related purposes. The policy on course notes and related printed and internet materials (e.g. published articles, website information) copyrights follows The General Rules Concerning University Organization and Procedure (University of Illinois Board of Trustees, 1998) and can be found at http://www.vpaa.uillinois.edu/policies/patents.htm and any other rule mentioned in the materials.

***Reading materials***

Handouts will be posted on the course website and will cover all subjects in the Syllabus. There is no single book that can address appropriately all the required knowledge, and no books are necessary to successfully complete the course. Three books are however recommended for students desiring further reading in order to reinforce programming knowledge:

Taylor, D. (2005). Teach Yourself Unix in 24 Hours (4th Edition). Sams.

Tisdall, J. (2001). Beginning Perl for Bioinformatics. O'Reilly.

Schwartz, R. (2005). Learning Perl ($4^{th}$ Edition). O'Reilly.