NYU School of Medicine, Spring 2014

Course title – High Performance Computing in Biomedical Informatics, number, 3 credits

Time and Place – TBD

http://www.nyuinformatics.org/services/hpcf/hpccourse

Instructor

• Efstratios Efstathiadis PhD, Research Assistant Professor
• Contact information:
  o Email: efstratios.efstathiadis@nyumc.org
  o Phone: 212-263-9245
  o Office: Translational Research Building (TRB)
    227 East 30th Street, 7th floor, Rm 735

Prerequisites

• Background in Computer Science, Mathematics, or Statistics recommended.

Course overview

High-performance computing (HPC) is becoming increasingly important for efficient research in BioMedical Informatics (BMI). The course will cover a range of topics including a broad coverage of the fundamental concepts in HPC, parallel computing, high-throughput computing, and utility computing. It will provide an understanding of HPC architectures and technologies enabling students to apply this knowledge in understanding algorithms and analyzing data in relevant informatics areas, such as modern genomics and next-generation DNA sequencing.

The course will have a strong practical focus. Students will be provided with access to a variety of computer hardware architectures where they will be able to develop, optimize and execute their applications and analyze relevant biomedical datasets as part of a course project.

Learning objectives

At the conclusion of the course, the student will be able to:

1. Access and efficiently use powerful computing resources.
2. Understand modern computing architectures and algorithms, their possibilities and limitations, and their application in biomedical informatics.
3. Manage and analyze large biomedical dataset using best practice algorithms.
High Performance Computing in Biomedical Informatics

Required readings for the course


Software

The students will not need to download and install software packages on the HPC resources that will be used for the course. All needed packages will be pre-installed by the administrators of the HPC resources. Most of the software packages used in the course are open source packages. Examples are: BWA, QIIME, Matlab, and R.

Course Format, Requirements, Assignments

Readings and participation: Students are required to attend class, to complete reading assignments and to participate in discussions and engage in healthy exchange of ideas.

Discussion lead: Each student is required to lead at least one reading from the assigned weekly readings. This discussion lead will be graded.

Exam
There will be one exam in this class. It will primarily focus on the readings. It will be more like a quiz and less like a final exam.

Group project
Depending on the number of students in the class and mutual interests, there will be several teams formed. Each team/group will have at least 2 members. The group will be responsible for sharing the work on the group project, providing drafts of the works in progress and presenting the final research to the class.

Final project
There will be a final project. It may be related to the group project and will be assigned 4-5 weeks prior to the due date.

Grading/ Student Evaluation
Students’ grades will be based on their class participation, 3 analysis, exam, and final group project. The components will be weighted as follows:
Participation in class and discussion lead  10%
Exam                          20%
Group project               40%
Final project             30%

Missed Exams and Grade Appeals

Make-up examinations (for final only) will be given under special circumstances. Documentation will be required to verify a student’s claim. If a make-up exam is permitted, a different exam will be written for that student and may have a different format than the regular examination.

The projects must be turned in on time and no late assignments will be accepted.

If there is a time that you believe that there is a mistake in grading of an assignment/exam, you will have a chance to appeal your exam grade within a week after you receive your grade. If you think this is the case, you must write a note describing the error, attach it to the original exam, and give it to me within a week of the return of your exam. I will review your argument and my initial grading, and then return your exam with a decision to you in a timely manner.

General Policies

Late/missed work
You must adhere to the due dates for all required submissions. If you miss a deadline, then you will not get credit for that assignment/post. Try to avoid last minute submissions.

Incompletes
No “Incompletes” will be assigned for this course unless we are at the very end of the course and you have an emergency.

Responding to Messages
I will check e-mails daily during the week, and I will respond to course related questions within 48 hours.

Announcements
I will make announcements throughout the semester, so make sure you check updated messages in the ‘Announcement’ section. In most cases, I will also email the announcement to you. Make sure that your email address is updated; otherwise you may miss important emails from me.

Safeguards
Always back up your work on a safe place (electronic file with a backup is recommended) and make a hard copy. Do not wait for the last minute to do your work. Allow time for deadlines.

Plagiarism
Plagiarism, the presentation of someone else's words or ideas as your own, is a serious offense and will not be tolerated in this class. The first time you plagiarize someone else's work, you will receive a zero for that assignment. The second time you plagiarize, you will fail the course with a notation of academic dishonesty on your official record.
Tentative calendar of topics
(Course Schedule)

**Week 1:** Overview of High Performance Computing. HPC challenges in Biomedical Informatics. Chapter 5 in [3] (Required Reading).

**Week 2:** Parallel Architectures and Programming Models

**Week 3:** Linux for cluster computing. Shell scripts and Batch Job submission. Access to local HPC resources. Version Control.
Reference material is available online at: http://webdoc.nyumc.org/nyumc/files/chibi/attachments/linux_for_cluster_computing.pdf
Exercises involve job submission, monitoring and version control.

**Week 4:** Basic concepts of parallel computing. Message Passing and communication Operations using distributed memory computing.

**Week 5:** Applications of Parallel programming in Biomedical research.

**Week 6:** Map Reduce and the Hadoop implementation in analyzing biomedical research data.
Exercise: Analyze a large data set using a map-reduce implementation framework.

**Week 7:** Sharing large biomedical datasets. Data encryption. Sharing of data involving Personal Health Information (PHI) and HIPAA compliance.

**Week 8:** Cloud Computing.
Exercise: Deploy a cluster of virtual machines with biomedical informatics software pre-installed on public clouds.

**Week 9:** An overview of the available US cyber-infrastructure and shared resources for biomedical research. Grid computing.

**Week 10:** Analyze a large Next-Generation DNA Sequencing data set on the local HPC cluster using best practices algorithms I: File formats, file types, protocols and schemes. DNA read alignment using the Burrows Wheeler Transformation.
Chapter 4: DNA Sequence Alignment from [4] in Required Reading List.

**Week 11:** Analyze a large Next-Generation DNA Sequencing data set on the local HPC cluster using best practices algorithms II: Identify SNPs and small insertions-deletions (indels) using the Genome Analysis Tool Kit (GATK).
Chapter 8: Using NGS to detect sequence variants, from [4] in Required Reading List.

**Week 12:** Invited/Guest speaker

**Week 13:** Presentations of projects