Bioinformatics Syllabus

NYU School of Medicine, Fall 2013

Course title: Bioinformatics
Instructor: Dr. Stuart Brown,
Associate Professor, NYUMC Center for Health Informatics & Bioinformatics
Contact Information: stuart.brown@nyumc.org
212-263-7689

Schedule
Lectures/Lab: TBA (1-4:30 pm)
Location - NYUMC Center for Bioinformatics: Verizon Bldg, 227 E 30th St., 7th fl conf room.

Textbooks:
Understanding Bioinformatics
by Marketa Zvelebil and Jeremy Baum
Beginning Perl for Bioinformatics
by James Tisdall
Also Recommended:
Bioinformatics, A Practical Guide to the Analysis of Genes and Proteins
by A.D. Baxevanis and B.F. Ouellette
Bioinformatics: Sequence and Genome Analysis
by David W. Mount
Bioinformatics for Biologists
by Pavel Pevzner and Ron Shamir
Blast
by Ian Korf, Mark Yandell, and Joseph Bedell
The Linux Command Line: A Complete Introduction
by William Shotts

Course overview
This is a practical course in Bioinformatics which will emphasize how to analyze genetic and genomic data. Prerequisites include a thorough understanding of theoretical and practical aspects of molecular biology, and some University level mathematics and statistics, but no prior knowledge of computer programming or computer hardware is necessary. The course will provide a broad introduction to basic bioinformatics concepts including data structures, functional motifs and pattern searching, alignment, clustering, evolution and phylogenetics, gene expression, and sequence variation.

Weekly readings will include book chapters, key journal articles, and software tutorials. Weekly exercises will include basic computing skills in Unix, Perl, R, and the use of software including Blast, Clustal, Mega, and Galaxy.

Course Format, Requirements, Assignments
The course will consist of one weekly lecture/lab session. Students will discuss readings and begin work on practical computing exercises (problem sets) with the opportunity to
ask questions and help each other. Assignments will be completed on their own time and handed in by the end of the week. Weekly readings should be completed before lecture to allow for clarification and in depth discussion. One problem set (Perl programming) will be graded as a midterm and one objective exam will be held that will focus on concepts and readings. The final research project may be conducted in small groups and results will be presented to the class.

Lecture 1. Introduction to the Course: Bioinformatics Algorithms & Data Structures
- Overview of the challenges of Bioinformatics in the era of Genomics
- What is an algorithm
- What is a data structure, how is genomic data structured

Reading:
- Why Biologists Want to Program Computers, by James Tisdal

Lecture 2. Bioinformatics Databases, UNIX Basics (part 1)
- Using ENTREZ to find and retrieve sequences and MEDLINE references
- Genome databases
- Other online genetic databases: mutations, SNPs, etc
- Intro to the Linux shell, basic UNIX commands
- Intro to SGE shell on the HPC cluster.

* The NCBI Handbook: Genbank,
Entrez Tutorial

Lecture 3. Alignment, UNIX Basics (part 2: text editors)
- Pairwise alignment algorithms: Needleman-Wunch, Smith-Waterman
- Dynamic programming
- Scoring systems, optimization
- Gap penalties
- Unix text editors: Nano, emacs

Reading:
- How Perl Saved the Human Genome Project, by Lincoln Stein
- Emacs Guide, tutorial

Lecture 4. Similarity Searching (BLAST), Perl (part 1)
- Using alignment for database searching
  Word/seed/hash based methods:
  - FASTA alignment algorithm
  - BLAST alignment algorithm
- BLAST2 alignment algorithm (gapped BLAST)
- Using BLAST to search databases
- Perl basics

Reading:
- Altshul et al, (1990) Basic local alignment search tool
- Altshul et al, (1997) Gapped BLAST and Psi-BLAST
- Searching Databases by Similarity
- Korf, BLAST (chap 4)
- Similarity Searching on the Web

Lecture 5. Advanced Alignment methods, Perl (part 2)
- Translated BLAST
· BLAT
· psiBLAST
· large scale alignment: synteny (MAUVE)
· Perl loops, decisions, file i/o

Reading:
· Altshul & Koonin (1998) Iterated profile searches with PSI-BLAST
· Advanced Similarity Searching on the Web
· Searching Sequence Databases A novel Src kinase in the C.elegans genome

Lecture 6. Multiple Alignment

· Challenges of applying pairwise alignment to multiple sequences
· Progressive pairwise algorithms
· CLUSTAL
· Other multiple alignment tools
· Challenges to optimize multiple alignment for biological function

Reading:
· Thompson et al. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment
· Multiple Alignment on the Web
· Protein structure tutorial with Cn3D

Lecture 7. Patterns and Motifs in biological sequences

· Patterns in DNA: promoters, transcription factor binding sites
· Patterns in proteins: structural motifs
· Hidden Markov models
· Databases of biological motifs
· Perl regular expressions

Reading
· Sonnhammer, Eddy, Durbin (1997), Pfam: a comprehensive database of protein domain families based on seed alignments.

Lecture 8. Computing Evolution: Phylogenetic Analysis

- From Multiple Alignment to Phylogeny
- Sequence-based taxonomy: Overview and Assumptions
- Basic concepts of clustering algorithms
- Distance methods
- Parsimony methods
- Maximum likelihood methods
- Detecting selection in sequences
- Introduction to MEGA

Reading
- MEGA tutorial

Lecture 9. Genomics II: Analysis of Microarray Data
• Introduction to the "R" programming language for bio-statistical computing
  • Introduction to the microarray platform for gene expression
  • Data normalization
  • Differential expression: the t-test, non-parametric approaches
  • Multiple testing correction
  • Introduction to R

Reading:
  • R tutorial
  • Golub et al (1999), Molecular Classification of Cancer

Lecture 10. Genomics I: SNPs and Haplotypes
  • What is a SNP
  • Genomics assay platforms for sequence variants
  • Large scale SNP detection
  • Haplotypes (HapMap)
  • GWAS

Reading
  • SNP/HapMap PowerPoint slides
  • 1.42 Million SNPs
  • Users Guide to the HapMap
  • Hapmap Tutorial
  • Comparative Genomics of Regulatory Regions
  • GWAS

Lecture 11. Next-Generation Sequencing and Alignment
  • Next Generations sequencing molecular biology and concepts
  • Alignment with the Burrows-Wheeler transform

Open Helix tutorial on UCSC Genome Browser
Genomics Tutorial paper using UCSC and Galaxy
UCSC Tutorial paper

Lecture 12. Finding Sequence Variants with NGS & Galaxy Genomics Toolkit
  Finding sequence variants by alignment of NGS reads
  Allele frequency in populations
  Cancer specific variants
  Intro to the Galaxy platform for genomics analysis

Reading
Understanding Cancer Genomes w/ NGS
Galaxy ChIP-seq exercise
Galaxy Bushman/Bantu project
Galaxy video find unique Bantu SNPs

Lecture 12. NGS applications: ChIP-seq and RNA-seq

Reading
  • Zhang et al Model-based Analysis of ChIP-Seq
  • Computational Analysis of ChIP-seq Data
Lecture 13. Metagenomics
   sequencing microbial populations
   Introduction to data analysis with Qiime.

Lecture 14. Final course meeting: Poster/project presentations