City University of New York (CUNY)

CUNY Academic Works

2019

Bioinformatics II, BIO 3352, Course Outline

Eugenia G. Giannopoulou
CUNY New York City College of Technology

How does access to this work benefit you? Let us know!

BIO3352 Bioinformatics II, FA2019

Lecture and lab material for the first part of BIO3352

Course Description

This course is a continuation of Bioinformatics I. Topics include gene expression, microarrays, next- generation sequencing methods, RNA-seq, large genomic projects, protein structure and stability, protein folding, and computational structure prediction of proteins; proteomics; and protein-nucleic acid interactions. The lab component includes R-based statistical data analysis on large datasets, introduction to big data analysis tools, protein visualization software, internet-based tools and high-level programming languages.

Prerequisites: BIO 3350 and (MAT 1372 or MAT 2572)

Instructor

Prof. Eugenia Giannopoulou

Office: A502B (Academic building)

Phone: 718-260-4971

Email: egiannopoulou@citytech.cuny.edu

Webpage: http://ctp.citytech.cuny.edu/~egiannopoulou/

Site License

Course Objectives

Upon completion of the course, the students will be able to:

1. Use well-established and widely used bioinformatics tools and platforms (e.g., ClustalW, R, GSEA, Gene Ontology).

2. Use basic programming languages (e.g., Unix bash commands, AWK) to perform straightforward computational tasks.
3. Understand the theory and statistical background of commonly available bioinformatics tools, so that they are able to judge the validity of the results provided by these tools.
4. Navigate through internet-based biological databases and genomic browsers.
5. Use online resources to search for scientific literature in the field of bioinformatics.
6. Comprehend specific methodologies and results described in current bioinformatics literature.

Course Materials

Weekly material

Weeks 1-2 – DNA and RNA

- *Explore* DNAs of different organisms in the [UCSC Genome Browser.](#)
- You can *download* the files containing the lengths of the chromosomes of multiple genomes available at the [UCSC Browser](#). These are small text files ending with "chrom.sizes" that can be found under the "Full data set" link at each genome.
- *Obtain* 50K base pairs of DNA in the beta globin locus on human chromosome 11 [(exercise from the Pevsner textbook).](#)
- *Read* about the [ENCODE](#) (Encyclopedia of DNA elements) project.
- *Explore* the most widely used [data file formats](#) in bioinformatics and genomics.
- *Explore* [miRBase](#), the microRNA database.
- [GEO](#) is a public functional genomics data repository for array-based and sequence-based datasets. *Follow* this [tutorial](#) to learn how to download fastq format data from published papers.

Weeks 3-5 – Microarrays

- *Watch* ["DNA microarrays"](#), ["DNA microarray"](#) and ["Microarrays – DNA Chips"](#) for a quick introduction to the microarrays technology.
- *Learn* about the [MIAME](#) (Minimum Information About a Microarray Experiment) project.
- *Read* about [design](#), [normalization methods](#), and [exploratory analysis](#) in a microarray experiment.

Weeks 6-7 – RNA-seq

- *Watch* ["Illumina sequencing by Synthesis"](#) and ["Illumina sequencing technology"](#) for a quick introduction to Illumina's NGS technology.
- *Read* all about RNA-seq on [RNA-seqlopedia](#) an easy to read online resource by the University of Oregon.
- *Read* the introduction to [RNA-sequencing](#) by Illumina
- *Watch* the video ["A gentle introduction to RNA-seq"](#)
- *Read* the publication Wang, Zhong et al. ["RNA-Seq: a revolutionary tool for transcriptomics."](#) *Nature reviews. Genetics* vol. 10,1 (2009): 57-63. doi:10.1038/nrg2484
- *Follow* the Galaxy tutorial on RNA-seq https://galaxyproject.org/tutorials/rb_rnaseq/

Lab

R

The following Bioconductor packages will be used in class. We recommend that you visit the packages through the links below and explore their use.

1. [Bioconductor](#) is an open source and open development software projects for the analysis of high-throughput genomic and proteomic data. It uses the R statistical programming language.
2. [BioMart](#) – genome annotations
3. [Biostrings](#) – manipulation of biological strings
4. [BSgenome](#) – full genomes and SNPs
5. [GOexpress](#) – gene ontology annotations
6. [Limma](#) – microarray data analysis
7. [msa](#) – multiple sequence alignment
8. [rtracklayer](#) – access to the UCSC Genome Browser
9. [stringDB](#) – interacting proteins database
10. [systemPipeR](#) – NGS workflows

UCSC GENOME BROWSER

1. [Genome Browser](#)
2. [Table Browser](#)
3. [YouTube channel](#) with training videos
4. [FAQs](#)
5. [User guide](#)

Syllabus

The syllabus will be available in Aug 2019.

Schedule

Week 1 – DNA: The Eukaryotic Chromosome	Tue 08/27, 2.30-5.50pm
Week 2 – RNA and genome-wide measurement of gene expression	Tue 09/03, 2.30-5.50pm
Week 3 – Microarrays I: Experimental design	Tue 09/10, 2.30-5.50pm
Week 4 – Microarrays II: Gene expression data analysis	Tue 09/17, 2.30-5.50pm
Week 5 – Microarrays III: Descriptive statistics and functional annotation	Tue 09/24, 2.30-5.50pm
Week 6 – Next Generation Sequencing I: Introduction and applications	Tue 10/15, 2.30-5.50pm
Week 7 – Next Generation Sequencing II: RNA-seq	Tue 10/22, 2.30-5.50pm