Syllabus Fall 2021

Computing in Molecular Biology

Course Rubric: MCB 432 (3 credit hour)

Course Instructor: Mengfei Ho, Ph.D., email: mho1@illinois.edu.
Office Hours: Friday (2-3 PM) via Zoom
Skype for Business through University of Illinois at Urbana-Champaign: +1 217 300 8914
Course TA: Gilberto Padron, email: gpadron2@illinois.edu
Office Hours: Zoom, Wed. (11-12 PM).

Class Location: Nevada Computing Lab, 1203 1/2 W Nevada
Class Time: Tuesday & Thursday 11:00AM-12:20PM

Course Objectives: In meeting the challenge of the coming age of Precision Medicine, knowledge in bioinformatics is becoming more important than ever before in biomedical professions. This course is primarily aimed at helping students build essential entry level computational skills for long-term self-learning and growth in working with bioinformatics. This course includes lectures and hands-on in-class computer workshops.

Personal Computer Requirement: You will need to have a personal computer (either desktop or laptop) that runs current MacOSX or Windows system. There are no minimal requirements for your computer, but in general a better processor and ample memory and storage space will make it more manageable for completing your class work. You also have access to the computers in the computer lab when it is not used for course instructions.

Topics covered in this course:

Working with Excel, R, UNIX/BASH, and moving data among these environments.
Communicating with a computer through command lines.
Simple shell scripts for data trimming and reorganization.
Manipulation of DNA and Protein sequences.
Sequence alignment and phylogenetic tree construction.
Pattern recognition from sequencing data and text string.
Sequence analysis using NCBI-BLAST, MEGA, and some additional packages installed into anaconda/miniconda.
Ordination analysis, including PCoA, PCA, RDA, etc.
Familiarity with R package vegan and others.
Installation of R and python packages.
Using gene browser tools for genome comparison, including Artemis Comparison Tool, BLAST Ring Image tool, etc
Pipeline tools for microbiome analysis: filtering host DNA sequences (bowtie2), sequence alignment and mapping (samtool), genome assembly (megahit), and gene annotation (prokka). More miniconda packages.
Composing shell script program to perform repetitive tasks.

Building script program for data base searching, including MLST assignments, Antibiotic Gene identification, Virulent gene identification.
Four concepts corresponding to the four major homework assignments (cumulatively) will be introduced throughout the class: Using web resources, microbiome/16S rRNA, viral/bacterial genome comparison, and writing shell script programs.

Grading Policy Fall 2021

There will be in-class hands-on problem-solving exercises that contribute to class assignments. Students should upload their script used during class and for each of the assignments as a PDF file. The final product/result of the exercise should be submitted as graphs or data tables in PDF format. If multiple files are submitted, the files should be organized into a zipped folder before submitting. There will also be assignments in video format addressing questions related to the course subjects. The frequency of the video assignments is about once per week. The length of the videos should be limited to no more than 2 minutes and they should be uploaded directly as video clips in a zipped file. Because goal of this course is to gain familiarity and capability in conducting analysis using computers, it is important that you complete each class assignment on time. There will be four major assignments on working with individualized datasets in lieu of in-class written exams.

Each assignment, including class assignments and video essays, will be graded on a scale of 1-5 following this general rubric:

5 Extra effort and great job
4.5 Correct and nice job
4 Good effort with minor errors
3 Good effort but incorrect
2 Insufficient effort
1 Incomplete work
0 No work
-1 Daily penalty for missing or late class assignment.

The class assignment grades will count toward 50% of your final grade. For example, if we have a total of 50 assignments and you have earned a total of 200 points for a maximum of 50x5=250 points, you will have 50x(200/250)=40 points toward your final grade. There will be four major assignments, correspondingly weighted as 5%, 10%, 15% and 20% of your final grade. Major assignment will require some team work. Letter grades will be assigned for your final grade as A ≥ 90, A- ≥ 87, B+ ≥ 83, B ≥ 80, B- ≥ 77, C+ ≥ 73, C ≥ 70, C- ≥ 67, D+ ≥ 63, D ≥ 60, D- ≥ 50, and F < 50.

Any absence or postponement on assignments requires arrangement with the instructor beforehand. Allowed late assignments must be completed within one week of the original due date. Major assignments will be announced well in advance, therefore late submittal of major assignments will result in a daily penalty of 10%.

Class participation and bonus grade:

At the end of the semester, you will have a chance to earn up to two bonus class assignments based on your participation in the class and helping others in the discussion forum.

Course Schedule (Fall 2021)

This tentative schedule may be modified if the need arises.

Aug 24 – Introduction to Computing and Bioinformatics – first day of class, intro and course logistics, using Excell

Aug 26 – Python Graphs with plotly and data extraction from a large file

Aug 31 – Curve Fitting – NLS curve fitting with Excel and in R. Moving data between Excel and R environments. Capability of R in graphing Data.

Sep 2 – Data Format and Multivariate Data – Table and data frame in Excel and R

Sep 7 –Sequence Alignment – Text string and sequence data, fasta and fastq. Sequence Alignment. Online Sequence Alignment Tools and NCBI website, other on-line Alignment tools

Sep 9 – Multiple sequence Alignment – DNA sequence Alignment and Distance between sequences

Sep 14 – Protein multiple sequence alignment – BLASTP, SmartBLASTP, QuickBlastP, and database construction muscle

(Major assignment #1 due — 5%)

Sep 16 – Tree Construction MEGA program – Bootstrapping

Sep 21 – BASH environment and Shell Scripting – trimming and extracting data from a text file

Sep 23 – Phylogenetic evolution models – Tree construction and interpretation

Sep 28 – Shell scripting bootcamp – Extracting information from blast output

Sep 30 – Community microbial contents

(Major assignment #2 due — 10%)

Oct 5 – Dimensionality Reduction – cmdscale, PCoA, PCA, SVD

Oct 7 – Clustering – K-means

Oct 12 – Ordination Methods – R Package vegan

Oct 14 – Using shell script for Processing large size data

Oct 19 – K-mer clustering

Oct 21 – Ordination methods for microbiome analysis

Oct 26 – Script for alignment-based clustering

Oct 28 – K-mer clustering vs alignment-based clustering

(Major assignment #3 — 15%)

Nov 2 – Merge paired reads

Nov 4 – Script for repetitive tasks

Nov 9 – On-line Resource for Bacterial Typing – In silico Serotyping and Phenotyping

Nov 11 – Shell script program – resistant gene search

Nov 16 – Genome assembly – miniconda packages

Nov 18 – Genome assembly – host DNA filtering read assembly.

Nov 23 – Fall Break – No class

Nov 25 – Fall Break – No class

Nov 30 – Metagenome analysis pipeline – assembly and annotation
Dec 2 – Genome Comparison – Artemis and Brig
Dec 7 – Course review
(Major assignment #4 due — 20%)