Course Rubric: MCB 432 (3 credit hour)

Course Instructor: Mengfei Ho, Ph.D., email: mho1@illinois.edu.

Office Hours: By appointment (via Zoom or in person at MCB Learning Center)

Course TA:
Oraya Zinder, email: ozinder2@illinois.edu
Rajendra K C, email: rkc5@illinois.edu

TA Office Hours: In person at MCB Learning Center
Oraya Zinder, Fridays 11:00 am, in the MCB Learning Center (101 Burrill Hall)
Rajendra K C, Wednesdays 9:00 am, in the MCB Learning Center (101 Burrill Hall)

Class Location: 106B1 Engineering Hall

Class Time: Tuesday & Thursday 11:00AM-12:20PM (08/27/24-12/10/24)

Course Objectives: In meeting the challenge of the coming age of Precision Medicine, One Health and unprecedented global pandemics, 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 laptop computer that runs current MacOSX or Windows system. You need to use your own laptop during and after the class for all assignments. There are no minimal requirements for your computer, but in general a better processor, ample memory and storage space will make it more manageable for completing your class work. You will need to install Ubuntu WSL2 on Windows.

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.

HMM model and homology modeling

Sequence analysis using NCBI-BLAST, MEGA, HMMER 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, sequence alignment and mapping, genome assembly, and gene annotation. More miniconda packages.

K-mer based sequence comparison and clustering.

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 2024

For each class session, there will be in-class hands-on problem-solving exercises that contribute to each class assignments. Attendance is expected. Students should upload their script used and partial work completed during class, and answered to any quick questions by the end of the class period. No late submission for in-class work will be accepted. The script and completed assignments should be uploaded as a PDF file. The final product/result of the exercise should be submitted as graphs or data tables in PDF format. If the assignment is a shell script, it should be in plain text. 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 3 minutes or other designated length and they should be submitted as a video link. If you submit your video through YouTube or other video sharing platform, your video link must be valid through the end of the semester. Because the 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.

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 Late more than 3 days
-1 Penalty for missing in-class work or one-day late class assignment.
-3 Penalty for two-day late class assignment.

*** NOTE: You have a credit of 10 late days to use at your discretion without penalty. You can use these 10 late days for a single or for multiple assignments. This late credit is not applicable for in-class work. You should keep tract of your own remaining late allowance days and clearly state on your late submission how you want this to be counted (used as the penalty or the late credit allowance). In the absence of a declaration, a late penalty will be applied automatically. Two absences due to illness or other conflicts with documented excuse will not be charged for missing in-class work, however on-time submission of full assignment is required unless a charge for late submission is declaired. 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 from a maximum possible 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 assignments 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.

Major assignments will be announced well in advance, therefore late submittal of major assignments will result in a daily penalty of 10%. NOTE: No late submissions are allowed for the final (4th) assignment, which counts as your final project.

Class participation and bonus grade:
At the end of the semester, you will have an opportunity to earn up to two bonus class assignments based on your participation in the class during the semester and helping others in the discussion forum.

Course Schedule (Fall 2024)

This tentative schedule may be modified if the need arises.

Aug 27 – Introduction to Computing and Bioinformatics – first day of class, intro and course logistics, using Excell

Aug 29 – Excel to R – Model and NLS Curve fit Using Excel and R

Sep 3 – Data Format – Multivariate Data with Excel and in R. Moving data between Excel and R environments. Capability of R in graphing Data. Data extraction from a large file using script. Merging data frames in Excel and in R.

Sep 5 – R graph for large data set – Table and multivariate data. Graphing in Excel, R and python.

Sep 10 – 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 12 – Pairwise sequence Alignment and Sequence assembly– Sequence assembly by hand and automation

(Major assignment #1 due — 5%)

Sep 17 – Multiple Sequence Alignment – Muscle, Clustal, Kalign, MSA, EBI tools, and Blastp and Blastn

Sep 19 – Sequence analysis tools — BlastX, MEGA, HH-suite, HMMER, Alpha-fold, blast database construction and stand-alone blast.

Sep 24 – Hierarchical clustering and Phylogenetic tree – Distance matrix, Hierarchical clustering Phylogenetic clustering.

Sep 26 – Phylogenetic evolution models – UPGMA, NJ, ML tree method

Oct 1 – Phylogenetic trees and Maximum likelihood - Phangorn package DNAbin, phyDat, pml, evolution models

Oct 3 – Tree construction and interpretation – Substitution Models, Gap options, Bootstrap

(Major assignment #2 due — 10%)

Oct 8 – Kmeans and kmer – clustering

Oct 10 – Bootsrapping and Maximum likelihood

Oct 15 – Dendrogram vs Scatter plot – alignment based vs alignment-independent, plotly and 3D plot

Oct 17 – Clustering of sequences by cd-HIT and kmer – kmer cd-hit and script writing

Oct 22 – Multivariate Data and Dimensionality Reduction – Eigen-decomposition and PCA R Package vegan **Ordination methods –** Eigen-decomposition PCA, SVD, PCoA, cmdscale and NMDS

Oct 24 – Fastq and Q-scores - Illumina sequencing data Fastp and merging pair-ended reads -16S rRNA marker gene

Oct 29 – CD-hit and sequence clustering - Pipeline for analysis

Oct 31 – Loop in shell script - Multiple-step processing of multiple files - Hierarchy of folder and files

(Major assignment #3 due — 15%)

Nov 5 – Blast hits table to OTU table - Parse blast output - extract data for csv file - Moving among folders

Nov 7 – Biom Phyloseq and UniFrac - Community Microbiota comparison

Nov 12 - Clustering vs classification – Community Microbiota and OTUs

Nov 14 – NCBI Dataset – NCBI database and Database retrieval (Viral genome project)

Nov 19 – Genome comparison – Viral genome

Nov 21 – Genome comparison – Tree, ordination and alignment

Nov 26 – Fall Break – No class

Nov 28– Fall Break – No class

Dec 3 – Genome comparison – BRIG and Atemis-ACT

Dec 5 – Genome Comparison – BlastDistMat shell script program

Dec 10 – Pheatmap and Course review

Dec 15 (Major assignment #4 due — 20%) No Late submission Allowed.