Computing in Molecular Biology

Course Rubric: MCB 432 (3 credit hour)

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

Office Hours: By appointment (via Zoom)

Course TA:
Xin Chen, xinc6@illinois.edu
Jingqian Liu, jjl768@illinois.edu

TA Office Hours:
Xin Chen, Wednesday 10:30AM-11:30 AM in the MCB Learning Center (101 Burrill Hall)
Jingqian Liu, Thursday 3:30pM-4:30 pM in the MCB Learning Center (101 Burrill Hall)

Link:TBA

Class Location: 3217 Everitt Laboratory

Class Time: Tuesday & Thursday 11:00AM-12:20PM

Course Objectives: In meeting the challenge of the coming age of Precision Medicine 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 and 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 (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 2023

For each class session, there will be in-class hands-on problem-solving exercises that contribute to class assignments. Attendance is expected. Students should upload their script used and partial work completed during class by the end of the class period. 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 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 and they should be uploaded directly as video clips in a zipped file. 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. 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 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 track of your own remaining late allowance days and clearly state on your late submission how you want this to be counted (use 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 declared.

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 submission 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 2023)

This tentative schedule may be modified if the need arises.

Aug 22 – Introduction to Computing and Bioinformatics – first day of class, intro and course logistics, using Excel

Aug 24 – Excel to R – Model and NLS Curve fit Using Excel and R


Sep 5 – 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 8 – Pairwise sequence Alignment and Sequence assembly – Sequence assembly by hand and automation

Sep 7 – Multiple sequence Alignment – Muscle, Clustal, Kalign, MSA, EBI tools, and Analysis pipeline

(Major assignment #1 due — 6%)

Sep 12 – Hierarchical clustering and Phylogenetic tree – Distance matrix, Hierarchical clustering Phylogenetic clustering. MEGA, HH-suite, BLASTP, SmartBLASTP, QuickBlastP, and database construction and metadata.

Sep 14 – BASH environment and Shell Scripting – trimming and extracting data from a text file

Sep 19 – Phylogenetic evolution models – UPOMA, NJ

Sep 21 – Tree construction and interpretation – Substitution Models, Gap options, Bootstrap

Sep 26 – Shell scripting bootcamp – Extracting information from blast output

Sep 28 – Stand alone shell script program – Parsing blast output editing data format

(Major assignment #2 due — 10%)

Oct 3 – Dendrogram vs Scatterplot – cmdscale, PCoA

Oct 5 – Ordination Methods and Dimensionality Reduction – For-loop in R and Eigen-decomposition

Oct 10 – Multivariate data and Ordination methods – PCA, SVD, Clustering, R Package vegan

Oct 12 – Fasta and Q-scores – Illumina sequencing data

Oct 17 – Fastp and merging pair-ended reads – 16S rRNA marker gene

Oct 19 – CD-hit and sequence clustering – Pipeline for analysis

Oct 24 – Loop in shell script – Multiple-step processing of multiple files

(Major assignment #3 due — 15%)

Oct 26 – Hierarchy of folder and files – Moving among folders

Oct 31 – Clustering vs classification – Shell script, blast database

Nov 2 – Election Day – No class

Nov 7 – Biom and UniFrac – Community Microbiota and OTUs

Nov 9 – NCBI Dataset – NCBI database and Database retrieval

Nov 14 – Genome comparison – Tree, ordination and alignment

Nov 16 – Genome comparison – BRIG and Atenris-ACT

Nov 21 – Fall Break – No class

Nov 23 – Fall Break – No class

Nov 28 – Genome Comparison – BlastDistMat shell script program

Nov 30 – Pheatmap and Course review

Dec 5 – Course review – Last minute help for final project

Dec 12 (Major assignment #4 due — 20%) No Late submission Allowed.