Syllabus Spring 2020

Computing in Molecular Biology

Course Rubric: MCB 432 (3 credit hour)
Course Instructor: Mengfei Ho, Ph.D., email mho1@illinois.edu
Office Hours: Appointment arranged through email
Office Hours: (TBA)
Class Location: Wohlers Hall Computer Lab 70B Wohlers Hall (map)
Class Time: Tuesday & Thursday 9:30 – 10:50 AM

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. There are several Computer Labs on campus open to students when they are not used for instructions. However, only those computers in the assigned classroom for MCB 432 will have the program packages installed for this class.

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, to a limited scope, some packages installed into miniconda.

Ordination analysis, including PCoA, PCA, RDA, etc.

Installation of R packages and familiarity with R package vegan and, to a limited scope, phyloseq, ape, etc.

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.

Using 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, bacterial genome comparison, and writing shell script programs.

Course Schedule (Spring 2020)

This tentative schedule may be modified if the need arises.

Jan 21 – Precision Medicine and Bioinformatics-first day of class, intro and course logistics

Jan 23 – Curve Fitting-NLS curve fitting with Excel and in R. Moving data between Excel and R environments. Capability of R in Graphing Data.

Jan 28 – Data Format and Multivariate Data-Table and data frame in Excel and R

Jan 30 – Sequence Alignment-Text string and sequence data, fasta and fastq. Sequence Alignment. Online Sequence Alignment Tools and NCBI website, other on-line Alignment tools

Feb 4 – Multiple sequence Alignment-DNA sequence Alignment and Distance between sequences

Feb 6 – Protein multiple sequence alignment – BLASTP, SmartBLASTP, QuickBlastP, and database construction

(Major assignment #1-5%)

Feb 11 – Welcome to Conda and beyond-EMBL-EBI resource, HMM, Bash and miniconda3 and muscle

Feb 13 – Tree Construction MEGA program-Bootstrapping

Feb 18 – Phylogenetic evolution models-Tree construction and interpretation

Feb 20 – BASH environment and Shell Scripting-trimming and extracting data from a text file

Feb 25 – Shell scripting bootcamp-Extracting information from blast output

Feb 27 – Community microbial contents

(Major assignment #2-10%)

Mar 3 – Dimensionality Reduction-cmdscale, PCoA, PCA

Mar 5 – Clustering-K-means

Mar 10 – Ordination Methods-R Package Vegan

Mar 12 – Using shell script for Processing large size data-

Mar 17 – Spring Break – No class

Mar 19 – Spring Break – No class

Mar 24 – K-mer clustering vs alignment-based clustering-

Mar 26 – Ordination methods for microbiome analysis-

Mar 31 – Script for alignment-based clustering-

(Major assignment #3-15%)

Apr 2 – K-mer clustering vs alignment-based clustering-

Apr 7 – SerotypeFinder-Shell script for in-silico serotype gene finder

Apr 14 – ResFinder-Shell script for resistance gene finder

Apr 16 – Artemis – Genome comparison

Apr 21 – Brig – Blast Ring Plot

Apr 23 – Metagenome analysis pipeline – miniconda environment, host DNA filtering read assembly.

Apr 28 – Metagenome analysis pipeline – assembly and annotation

May 5 – Recap – final day of the class

(Major assignment #4 – 20%)

Major Assignment #1

MCB432 Major Project Assignment #1

Grading: This assignment will count toward 5% of your final grade.

Due: by Midnight on Monday Feb 10

Rules:

You need to produce and submit your own report by the deadline.
You are encouraged to work with partners in the class.
You are free to consult any publications and on-line resources.
Posting your assignment on-line to solicit help or answers is not allowed.
Submit your report as a pdf document – include the alignment output as your raw data.
Submit your processed data file in text format.
Include a statement that you have read the assignment and have followed the rules of the assignment.

This assignment has two parts:

Part I: Select a data set from the data folder. The data set contains sequencing reads corresponding to the 16S RNA genes of bacteria isolated from soil samples. Align and joint the four sequence reads for each sample into a continuous sequence. Submit your sequence at the NCBI blast website and assign a most likely bacterial species for each sample, with unique genus and species names.

Part II: Find 3 to 5 other students to form a group. Share and compare your findings with that of your group members and present the combined data from the group in a tabulated form including a simple interpretation of the data. Identify the "metadata" associated with each sequence and include them in your table. Generate a text only data file of FASTA sequences including your joined reads and the 16S hits. Align the sequences in your combined sequence file. Construct a neighbor-joining tree.

Submit your own report including both parts and the names of your group members.

Grading rubric:

Each of the following items will be graded according to the 5 point scale detailed in the grading policy for MCB 432.

Raw data for your analysis.
Results of your analysis.
A text file of your joined sequences and 16S hits in FASTA format.
A distance tree of your FASTA sequences.
Tabulation of data (sample, assignment, meta data, etc).

Competence in using NCBI blast tool.
Competence in using Jalview tool.
Competence in using alignment to join sequence reads.
Collaboration with your groups members.
Presentation quality of your report.

Major Assignment #2

The recent outbreak of Coronavirus disease (COVID-19) in China and other areas of the world has become a major health threat and has made a noticeable impact on the global population. The complete genome sequence of several SARS-CoV-2 (or 2019-nCoV) virus isolates are now available in Genbank. We will use our knowledge of Bioinformatics to analyze Coronavirus genomes and identify potentially useful information for outbreak management and therapeutic strategies.

Part 1:

1. Visit the CDC website to read about the current outbreak of COVID-19.
2. Visit the news articles at major science journal websites, nature, science, NEJM and others, for example,
   https://www.nejm.org/coronavirus
   https://www.nature.com/articles/d41586-020-00154-w
   https://science.sciencemag.org/content/367/6479/727
3. Write a single page statement to describe the current state of COVID-19 and the Bioinformatics Analysis you will be presenting in your report.

****This Major assignment #2 will be updated with additional activities, see the most recent version.