Syllabus of MCB/ABE/GENE/ECOL/BIOC 416A/516A

Statistical Bioinformatics and Genomic Analysis

Spring 2018

Time and location:

MW 11 am -12:15 pm (Shantz 338)

Instructor information

Dr. Lingling An
Office: Shantz 501
Phone: 621-1248
E-mail: anling@email.arizona.edu
Office Hours: 11am-12pm (Friday) or make an appointment

Course description, objectives, and outcomes:

The course introduces statistical methods and algorithms for analysis of high-throughput experiments in molecular biology. The objectives of the course are:

- Introduce relevant biological concepts, and describes the existing high-throughput technologies and biological questions that these technologies can help answer.
- Discuss statistical methods and models used to analyze high throughput data (e.g., next-generation sequencing data and microarray data) as well as open research problems in this field.
- Discuss data structures and implementation of the methods in the R-based open source project Bioconductor.

The course provides hands-on experience with data analysis and communication of the results. At the end of the course the students will be able to perform independent analysis of biological data in an interdisciplinary environment such as a pharmaceutical company or a research lab.

Primary audience:

Graduate students and senior undergraduate students in life sciences/engineering who have basic quantitative training (e.g. elementary statistics course) and have interests in understanding the intuition and logic underlying the statistical methods; graduate students in statistics, biostatistics, computer science, and mathematics.

Prerequisites: Basic statistical knowledge, e.g. two-sample T test and analysis of variance (ANOVA). Consult the instructor for appropriate prerequisites before enrolling. Prior exposure to R (free software) and/or knowledge of basic biological concepts is desirable, but not required.

Enrollment max: 50 (25 graduate students + 25 undergraduate students)

Textbook: There is no textbook for the course. A series of useful texts are given on the course website through D2L.

Topics:

Module 1: Introduction to statistical methods in molecular biology

- review statistical methods
  - common probability distributions
  - hypothesis testing and confidence interval
  - comparison of group means, ANOVA analysis
  - multiple comparisons and multiple tests
- concepts in molecular biology and scientific questions

Module 2: Introduction to R and Bioconductor

- comparison of R and other statistical languages/software
- frequently used commands and tools
- data structures and visualization

Module 3: Statistical analysis of gene expression microarrays

- introduction to microarray technologies
- signal processing & finding differentially expressed genes
- application of multiple comparisons/multiple tests

Module 4: Statistical machine learning concepts and tools

- supervised classification (i.e., prior information known/given)
- unsupervised cluster analysis (i.e., no prior information known/given)

Module 5: Analysis of next generation sequencing data

- introduction to next generation sequencing technologies
- mapping of reads and quantification of expression
- statistical models in RNAseq data analysis:
  - data normalization
  - differential expression analysis and biomarker detection
  - classification and cluster analyses

Module 6: Biological annotations and analysis of biomolecular networks

- gene sets and ontologies: structures and visualization
- pathway analysis and network analysis

Module 7: Introduction to metagenomic data analysis

- introduction to metagenomics
- methods in metagenomic analysis: differential abundant analysis, classification analysis, network analysis

Software:
Class projects will be carried out using R. Access to R is required. Please install R from http://lib.stat.cmu.edu/R/CRAN/. Instructions for accessing Bioconductor will be provided during the course.

Homework:

There will be no homework assignments.

Projects:

There will be 2 projects for undergraduate students and 3 projects for graduate students.

- **Microarray data analysis:** students will select a paper of interest and duplicate the analysis of microarray data involved. The project can be done in groups of **3 students** along with a report.
- **RNA sequencing data analysis:** students will select a paper of interest and duplicate the analysis of RNAseq data involved. If you already work on a research project in this area you are welcome to use a dataset from your research, provided that you make an extra effort for the class. The project can be done in groups of **3 students** along with a presentation and a report.
- **Metagenomic data analysis (required for graduate students):**
  During the final week every graduate student will present the analysis of metagenomic data and a scientific report due the last day of class (detailed instructions will be announced later).

Exams:
There will be no midterm and no final exam.

Course site: D2L
Syllabus, course information, lecture notes, projects, and announcements etc.

Course mailing list:
The course mailing list will allow us to communicate outside of the lecture hours. You are encouraged to ask and answer questions on the list.

Final grade:
The final grade will be computed as follows:

Quizzes: 3 in-class quizzes (each 20 points) – 60 points
Projects:
Microarray data analysis project – 60 points
RNAseq data analysis project – 60 points
metagenomic data analysis project – 60 points (graduate students only)

The total points for undergraduate and graduate students will be 180 and 240 pts, respectively.

The final letter grades will follow a straight scale:
90%~100 %=A, 80%~89%=B, 70%~79%=C, 60%~69%=D, 0~59%=E.

Subject to change statement

Information contained in the course syllabus, other than the grade and absence policy, may be subject to change with advance notice, as deemed appropriate by the instructor.