



Lecture 1

Introduction



MCB/ABE/GENE 416A/516A
Statistical Bioinformatics and Genomic Analysis

Prof. Lingling An
Univ of Arizona

Basic Info

- Time: MW 11am-12:15pm
- Location: Shantz 338
- Lecturer: Dr. Lingling An
- Office hour: Fri 11-12 or by appointment
- Office: Shantz 501
- Email: anling@email.arizona.edu

What to expect?

This course

- is not a regular course with a regular textbook

- is a course

 - to familiarize you with this developing field

 - to equip you with useful knowledge and methods

 - to inspire your research interests

 - to connect your knowledge to real world
problems

- is a fun class to enjoy! (I hope)

Course description

- We'll introduce:
 - relevant biological concepts
 - high throughput microarray & sequencing technologies
- and introduce:
 - statistical methods (applied to analyze high throughput data)
 - research problems
 - how to use Bioconductor (based on the free software R) and R to write your own code

Cont' d

- hands-on experience for data analysis
- team projects:
 - presentations & reports
- at the end of the semester:
 - you'll be able to perform independent analysis of biological, biomedical data
 - trained how to present your project and results to a very diversely trained audience

-
- Prerequisites:
 - basic statistical knowledge
 - biological knowledge is not required but will help you in this class

We'll convey a survey afterwards ...

Textbook

- There is no textbook for the course. A series of lectures will be posted at the course website.

Topics to cover

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

Topics to cover - 2

Module 2: Introduction to R and Bioconductor

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

Topics to cover - 3

Module 3: Statistical analysis of gene expression microarrays

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

Topics to cover - 4

Module 4: Statistical machine learning concepts and tools

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

Topics to cover - 5

Module 5: Analysis of next generation sequencing data

- intro to next generation sequencing technologies
- mapping of reads and quantification of expression
- statistical models in RNAseq data analysis
 - Data normalization
 - Differential expression analysis biomarker detection
 - Classification and cluster analyses

Topics to cover - 6

Module 6: Biological annotations and analysis of biomolecular networks

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

Topics to cover - 7

Module 7: Introduction to metagenomic data analysis

- introduction to metagenomics
- methods in metagenomic analysis:
 - differential abundant (taxonomic/function) analysis
 - classification analysis
 - network analysis

Homework, quiz, and exams

- No homework
- No midterm; No final exam
- 3 quizzes!

Projects

- Aim: turning statistical techniques you learn in class to become your weapon in fighting real world problems
- 2 team projects for undergraduate and 3 for graduate students
 - **Project 1: 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.

-
- **Project 2: RNAseq 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.

■ **Project 3: Metagenomic data analysis (required for graduate students):**

- 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. During the final week every graduate student will present the analysis of metagenomic data and scientific report due the last day of class (detailed instructions will be announced later).

Grading Policy

Final grade will based on:

Quizzes: 60 points (=3*20)

Projects:

project 1 – 60 points

project 2 – 60 points

project 3 – 60 points (graduate students only)

Total points:

180 pts -- undergraduate

240 pts -- graduate students

The final letter grades will follow a straight scale:

90%~100 %=A, 80%~89%=B, 70%~79%=C, 60%~69%=D,

0~59%=E

Attendance

- Students are expected to attend and participate every class
- Students are responsible for all materials covered during lectures, including quizzes.

Any questions?