PQHS 430: Design and Analysis of High-Dimensional Data

Course description: This is an exciting genomic revolutionary era when scientists can use high-throughput data to extract the genetic basis of complex diseases such as cancers. High-dimensional high-throughput data are often encountered in the fields of genomics, proteomics, system biology and bioinformatics. Through this course students will learn how to analyze the high-dimensional genomic data necessary for personalized medicine, using interdisciplinary approaches that combine statistics, computer science, molecular biology, and genomics. While this particular course will focus mostly on statistical methods for designing and analyzing molecular studies, those who take it will come from a wide variety of disciplines. The instructional design will be one of active experimental learning: the course will include in-class lectures, group discussion and brainstorming, homework, simulations, and collaborative projects on real and realistic problems in human health tied directly to the student's own professional interests. Review of some multivariate methods, including statistical learning and inference methods when the number of measures far exceeds the number of subjects ("high-dimensional data"). Topics include (but not limited to) designing high-throughput studies, sample size and power analysis, low-level preprocessing of microarrays, basic exploratory genomics and proteomics data analyses, classification and supervised learning, cluster analysis and unsupervised learning methods. These statistical methods will be applied to gene and protein expression data, and next generation sequencing data. This course stresses how the core statistical principles, computing tools, and visualization strategies are used to address complex scientific aims powerfully and efficiently, and to communicate those findings effectively to researchers who may have little or no experience in these methods. Basic knowledge in biology will be helpful however required molecular biology will be reviewed.

Learning Objectives:

1. Gain proficiency in designing high-throughput studies and statistical learning methods.
2. Hone skills by applying statistical methods in solving high-dimensional data analysis problems.
3. Acquire competency in standard and cutting edge high-dimensional methods and algorithms.

PQHS 430 (3 Credits Hours): Course Information

Time and Place	TBA
Instructors:	Abdus Sattar, PhD, Assistant Professor, PQHS Yu Liu, PhD, Senior Research Associate, CPB
Office:	BRB: Suite G-19

E-mail/Phone:	sattar@case.edu/ 216-368-1501
Office Hours:	TBA
Course Web Page:	http://blackboard.case.edu/
Text (required):	No required text
Reference	-Data mining for genomics and proteomics *by* Darius M. Dziuda -Statistics and Data Analysis for Microarrays *by* Sorin Draghici -Bioinformatics for High Throughput Sequencing *Edited by* Rodriguez-Ezpeleta, Hackenberg, Aransay

Prerequisites:

- This course is designed for advanced undergraduate students, and graduate students in Biostatistics or other quantitative sciences with background and adequate preparation in statistical methods (at least one statistics, similar to PQHS431, course experience).
- Some programming experience. Knowledge in statistical computing or statistical software package is helpful. We aim to use R and some commonly used NGS software (e.g. BWA, Bowtie, Tophat).

Disability Help: If you have a disability and need help, please contact me and the Office of Educational Support Services at disability@case.edu, 216.368.5230 as early as possible in the term.

Academic Integrity: You are expected to maintain the highest integrity in your work for this class. This includes not passing off anyone else's work as your own, even with their permission. Your homework solutions must be your own work, not from outside sources, consistent with the university rules on academic honesty. I expect you to follow this policy scrupulously. Evidence of academic dishonesty may lead to loss of credit for the assignment, and possibly failure of the course.

PQHS 430: Course Requirements & Grading

Homework: Homework will be assigned biweekly. There will be approximately 6 - 7 homework assignments. No late homework will be accepted unless you have a university-excused absence. There will be high-dimensional data analysis problems drawn from the scientific studies, which will provide the opportunity for you to demonstrate your statistical knowledge and computational skills in hypothesis testing and making statistical inferences.

Project: There will be a final project in this course.

Grading Scale: The course grade will be determined according to the following,

- Homework 70%
- Project 30%

Tentative PQHS 430: Course content and Timeline

Week 1 - 7 will be devoted in the following topics:

- Introduction to the molecular biology, genomics, and proteomics
- Review of statistical (supervised and unsupervised) learning methods
- Multiple comparisons (p>>n problems)
- Design of high-throughput experiment
- Low-level processing (normalization, background correction, etc) of microarray data
- Feature selection methods: random forests, support vector machines

Week 8 - 10 will be devoted in the following topic:

- Introduction to proteomics and Mass Spectrometry
- Preprocessing of Mass Spectrometry Data
- Statistical Analysis of protein expression data

Week 11 - 15 will be devoted in the following topics:

- Introduction to next generation sequencing data
- DNA-seq, RNA-seq, ChIP-seq data analysis