M21-550 Course Syllabus

Summer 2018 (July 31-August 13)

Course Title:	**Introduction to Bioinformatics**
Course Master:	C. Charles Gu, Ph.D.
Guest Lecturer(s):	Gary Stormo, Ph.D., Ting Wang, Ph.D.
Teaching Assistant:	Will Yang, Ph.D.
Schedule:	*R-Primer July 9~16 required of ALL students.* Bioinformatics Class - **Every weekday, 9:00am-12:00pm** and **1:30-4:30pm**.
Place:	MSIBS classroom & Computer Lab (Room 502 & 501, 5th floor, Becker's Library)
Format:	Small group lecture and extensive computer labs using real-world data
Grade Criteria:	Numerical score based on:

Quizzes	35%
Lab projects	45%
Final exam	20%

Reference Text:

1. Andreas Baxevanis and Francis Ouellett (eds.)**,** Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins**,** 3rd Edition, Wiley & Sons, 2005.
2. Robert Gentleman, Vincent Carey, Wolfgang Huber, Rafael Irizarry, Sandrine Dudoit (eds.), Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Springer, 2005.
3. Malcolm Campbell and Laurie Heyer, Discovering Genomics, Proteomics and Bioinformatics (2nd Edition), Benjamin Cummings, 2007.
4. Richard Durbin, Sean Eddy, Anders Krogh, and Graeme Mitchison, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press, 1998.

Other Readings:

Additional reading materials and papers will be assigned by lecturers.

Description: This course provides a general introduction to basic concepts, methodologies and applications of bioinformatics. Students will learn online bioinformatics databases & data mining tools, and acquire understanding of mathematical/computational algorithms in classic sequence analysis (alignment, gene finding, and hidden Markov models), next-generation sequence analysis (short-read data format and processing, variant calling algorithms), and in selected topics of high-dimensional analysis in the context of mining high-throughput biomedical experiments data such as gene expression microarrays (data QC & normalization, univariate & multivariate differential expression analysis). Students will become familiar with popular bioinformatics software, online tools, and specialized R/BioConductor packages. We will discuss methods for high-dimensional data analysis including classification and clustering analysis, principal component analysis (PCA), statistical/machine learning, and Bayesian inference. There also will be seasonal additional lectures on topics such as proteomics and applications of bioinformatics to real studies of complex diseases.

As an important component of this course, students will conduct hands-on computer labs to learn basics of online bioinformatics databases and tools, and to practice computer programming. The labs require using the statistical computing environment R (i.e., the R primer is required) though introduction to BioConductor basics will be provided. Students will use specialized software and R packages to accomplish tasks including designing experiments, low-level analysis of expression levels, univariate differential expression analysis, and various multivariate analysis techniques taught in class. A variety of software will be used for NGS data analysis covering alignment, variants calls, differential analysis, and visualization of results. Through the lab exercises, students will learn how state-of-the-art computational tools are applied to solving bioinformatics problems in real studies of human diseases.

Schedule of Lectures:

Day	Morning (9:00am -12:00)		Afternoon (1:30pm - 4:30pm)	
	Topics	**Instructor**	**Computer lab**	**Instructor**
D#1 Tu 7/31 *Qz-1**	**Course overview:** summary of topics, grading policy, expectations **Sequence analysis I:** homology detection; pair-wise sequence alignment, alignment score statistics, BLAST search	Gu	**Lab #1:** online databases & tools, BLAST	Gu
D#2 W 8/1 *Qz-2*	**Sequence Analysis II:** multiple alignment, Hidden Markov Models (HMM) & applications to motif and gene finding	Gu	**Lab #2:** seq alignment, database/R tools; HMM motif/gene prediction	Gu
D#3 Th 8/2 *Qz-3*	**DNA variants:** genetic variations & diseases; DNA variants, copy number variations (CNVs), CNVs detection & analysis; impact of SVs; ***Guest (Stormo):*** *DNA cryptography*	Gu, Stormo	**Lab #3:** R refresher; R tools for CNV (*DNAcopy, ReadDepth,* etc.)	Gu
D#4 F 8/3 *Qz-4*	**NGS data analysis I:** intro to next generation sequencing (NGS), sequencing technology & platforms, short reads alignment algorithms & software;	Gu	**Lab #4:** BWA, Bowtie; R tools for exploring sequencing data	Gu
D#5 M 8/6 *Qz-5*	**NGS data analysis II:** NGS file formats, SAM/BAM; SAMtools; variants calling, VCF format; QC of NGS reads; NGS study design and statistical issues	Gu	**Lab #5:** SAMtools, more R tools for exploring sequencing data	Gu
D#6 Tu 8/7 *Qz-6*	**NGS data analysis III:** RNA-seq based expression analysis, count-based normalization, transcript-level differential expression	Gu	**Lab #6:** R tools for RNA-seq analysis (*edgeR, DESeq2*); TopHat, Cufflinks;	Gu
D#7 W 8/8 *Qz-7*	**High-dimensional analysis I:** microarray & transcriptomics; high-dimensional analysis, bias & normalization; robust statistics; differential expression, false discovery rate, resampling statistics	Gu	**Lab #7:** microarray data preprocessing (*ReadAffy*); differential expression analysis (*siggenes*);	Gu
D#8 Th 8/9 *Qz-8*	**High-dimensional analysis II:** multivariate analysis & dimension reduction, PCA (principal component analysis); dissimilarity measures, clustering analysis (hierarchical, k-means)	Gu	**Lab #8:** R tools for PCA (*prcomp*), clustering analysis (*heatmap, hclust*)	Gu
D#9 F 8/10 *Qz-9*	**High-dimensional analysis III:** statistical learning (supervised/unsupervised), model/feature selection; classification analysis, LDA (linear discriminant analysis), Naïve Bayes classifier	Gu	**Lab #9:** R tools for LDA (*lda*), naive Bayes classification (*naiveBayes*)	Gu
D#10 M 8/13	**Bioinformatics in medicine:** genetic circuits, environments & interactions; EHR & bigdata; precision medicine & public health initiatives; ***Guest (Wang):*** *epigenetics & diseases*	Gu, Wang	**FINAL EXAM (2:00 - 4:00pm)**	Gu

* *Qz-1* is an assessment of your proficiency in R; it will not be counted toward your final grade.