Course Instructor: Goran Bozinovic, Ph.D.
Contact: gbozinovic@ucsd.edu

Instructional Assistants:
Yichi Zhang yiz814@ucsd.edu
Ray Huang crhuang@ucsd.edu

Course Schedule:
Lecture: Tue / Thu 3 - 3:50 TATA 2501
*Labs Tue / Thu 4 - 6:50 TATA 2501

Virtual office hours: by appointment

Required material: lecture slides (posted on Canvas), manuscript readings, online tutorials, computation labs, and supplementary materials

Lecture and Computer Lab Virtual Participation: Each class consists of both lecture and computer lab. Lectures are in-person (attendance is highly recommended); Lab attendance is mandatory, and you will need to complete all computation labs on time. Your IAs will be available to assist you during scheduled lab hours on Thu 4-6:50 PM in TATA 2501.

Course Website: http://canvas.ucsd.edu

Course Description:
Bioinformatics (upper division 4-unit course for biology majors) - the application of computational and analytical methods to biological problems - is based on collection, management, analysis, and interpretation of data in life sciences and medical research. This course introduces the principles and applications of bioinformatics focused on genes and proteins. A lecture/lab format with computational lab exercises emphasizes recent developments in genomics and proteomics. Students completing this course will be able to apply bioinformatics tools to address biological questions and topics, including:

- Advances in sequencing technologies,
- Genome informatics,
- Structural informatics,
- Transcriptomics, and
- Bioinformatics data analysis with R.

The lecture/computer lab format is structured on bioinformatics primary scientific literature and utility of custom data sets emphasizing recent developments and analytical applications in genomics and proteomics. Course topics, supplemented by relevant biostatistical concepts and applications, include genomic and biomolecular bioinformatics resources and databases, advances in sequencing technologies, genome and structural informatics, phylogenetics and transcriptomics. Computational tools and applications promoting best analytical practices will be emphasized. A personal computer is required. Students will learn how to manage and analyze bioinformatics data by utilizing published primary life science literature. We will use "R", a common biostatistics programming language, to explore experimental designs, fundamental bioinformatics (and relevant biostatistics concepts), manage and analyze biological, sequencing, and gene expression data sets.

Course Requirements / Eligibility: There is no textbook for this course. A familiarity with basic biological concepts is essential. No formal programming training or advanced mathematical skills are assumed or required. Students must have their own computers to access class material and utilize free bioinformatics software and data resources. All course materials (lectures, computer labs, videos, science manuscript .pdf files, data sets, and supplementary materials.) will be accessible via Canvas course website.

Learning Objectives:

Students completing this course will be able to apply leading bioinformatics and biostatistics tools to test hypotheses and infer biological relevance using genes and proteins sequencing data. By the end of the course, student will be able to:

- Establish a testable null and alternative hypothesis
- Explain the importance of proper data collection, experimental design, hypothesis testing, and interpreting results
- Critically evaluate scientific writing, experimental design, and analysis
- Comprehend scientific primary literature, interpret data and experiments
- Understand the experiments with adequate biological, technical replication and statistical power
- Critically assess primary scientific literature focused on biostatistics and bioinformatics concepts
- Perform gene finding and sequence alignments, DNA and Protein Database searching
- Become familiar with the bioinformatics databases, including NCBI / EBI, PDB, UCSC Genome browser, DAVID Bioinformatics, and STRING protein network bioinformatics
- Explore and visualize data and perform bioinformatic analysis, including principal component analysis (PCA) using R
- Perform phylogenetic and hierarchical clustering analysis and quantify intra- and inter-species variation and estimate divergence
- Understand the high throughput sequencing methods and applications
- Analyze transcriptomics data using a RNASeq data set
- Collaborate to learn and communicate bioinformatics concepts and relevant fundamental statistical analysis
- Improve written / oral communication applying bioinformatics concepts

Course structure and content access: The course is structured as about 30 min lecture/concept introduction, 10 min manuscript discussion/concept applications, Q/A session, and Computer Lab Exercises. Lectures will cover the bioinformatics (and relevant biostatistics) theory and introduce concepts, data bases, and data analysis applied during the computer labs. All the course materials are available through the course website. Students should access this site regularly. Once you are enrolled in the class you will have access to CANVAS using your ACS username and password. Be sure to check the course website frequently for announcements and updates. Use the "Discussion Board" to ask relevant course content questions. Your instructional assistants will check the Discussion Board frequently, but students are encouraged to answer questions also.

Materials:

1) Online lecture slides, manuscript PDFs and computer labs (available on Canvas)
2) A computer with permissions to download/install applications

Classroom and Computer Lab Etiquette:

Please ask questions! Student discussion during lectures is vital to course effectiveness. Lectures will be "paused" periodically to allow for your questions and/or clarification and data analysis / critical thinking exercise. Outside of class, you can email the instructor and IAs to ask questions or set up office hours. If you have a comment or question, please be considerate of class time. To make sure all the questions are addressed, the last 10' of each lecture will be reserved for review and discussion.

Course Requirements and Grading: Your final grade for the class will be calculated using the following criteria:

Bioinformatics Labs (L2,L4-9)	70
Biostatistics Lab (L3)	10
Professional Development Workshop Lab (L10)	10
Midterm Exam	50
Final Exam	60
Participation	10
Group Presentation	20
* Surveys, Pre-post assessments – **mandatory to successfully complete the course**	10
Total Points	**240**

Bioinformatics lab: Students will complete seven weekly bioinformatics computer labs. The goal of these exercises is to introduce students to several major bioinformatics websites / databases / resources, manage and utilize sequencing data to apply molecular biology concepts, review primary scientific literature, analyze the data, and interpret results. Some of these labs will be completed in the R programming language as an introduction to programming as applied to bioinformatics.

Biostatistics lab: A single 2-part biostatistics lab (L3) using R programming language will introduce students to fundamental and bioinformatics-relevant statistics concepts. Scores will be assigned based on checkpoint questions and R code functionality. Please note that biostatistics concepts will be introduced thought the course, as sequencing and gene expression analysis is integral in the field of bioinformatics.

Professional Development Workshop Lab: following the lecture during week 9, students will complete L10 by creating the 1-page cover letter and 1-page resume highlighting skills and application learned during this course. The cover letter and resume content can be combined / added to an existing resume, but the emphasis will be placed on the resume and the cover letter sections relevant to the content of this course. Detailed instructions will be provided during the lecture.

Midterm Exam and Final Exam: students will have a full lab period (2 hrs 50 mins) to complete an open-notes / resources, in-person midterm and final exam. Students will be allowed to refer to course materials, including their lab reports submissions, during the exams. Note that the final exam includes the lecture material and labs presented after the midterm exam.

Participation: Active contribution to class discussion is highly encouraged.

Group presentation: One group presentation during the last week (week 10) of class covering a real-world application of bioinformatics concepts. Details will be provided at the end of week 5.

Surveys, Pre-post assessments: Anonymized assessments taken during the first and last weeks of class. Graded only on completion: your scores will not affect your course performance – the purpose of surveys is to assess and improve course content and students' experience. **Students who do not complete all surveys and assessments will not pass the class.**

% Point Cutoffs for Grade Assignments: (cutoffs may be lowered at the instructor's discretion)

>92	A	78-79.99	C+
90-91.99	A-	72-77.99	C
88-89.99	B+	70-71.99	C-
82-87.99	B	60-69.99	D
80-81.99	B-	<60	F

Schedule

Day	Lecture Topic (subject to change)	Lab	Manuscript
4.4 Tu, week 1	• Course introduction and Discussion	• Pre-course survey • Pre-assessment • Install R and RStudio; make sure they work (no submission)	N/A
4.6 Th, week 1	• Introduction to Bioinformatics; History • how to read the science paper: review vs. original Manuscript; what to avoid and what to focus on; • Sources of Data for computer lab (3 manuscripts overview)	**L1**: • Intro to "R" • how to read the original science manuscript • Read manuscript 1	Divergent low-density lipoprotein receptor (LDLR) linked to low VSV G-dependent viral infectivity and unique serum lipid profile in zebra finches
4.11 Tu, week 2	• Data Types • Databases (NCBI, EBI, PBD) • Intro to R • Paper 1 introduction	**L2-1** • Bioinformatics databases: NCBI, EBI, UCSC genome browser, PDB, and UniProt via accession number	
4.13 Th, week 2	• Introduction to Biostatistics Hypothesis testing; Independent vs dependent variables • Type I and II errors • Data types and data representation • Data visualization – tables and graphs	**L2-2** Hypothesis testing in R using data from Paper 1; descriptive stats output and graphing in R. Intro to R Markdown	
4.18 Tu, week 3	• Law of large numbers • Central limit theorem T-distribution • Student's t-test	**L3-1** R - stat testing – two-sample and paired t-test, 1-way ANOVA and post hoc test with interpretation	
4.20 Th, week 3	• F-distribution • 1-way ANOVA & post-test correction	**L3-2** Graphs is ggplot 2	Evolution of Melanoma Antigen-A11 (MAGEA11) During Primate Phylogeny
4.25 Tu, week 4	• Sequence alignment / similarity • Local and global alignment • Paper 2 introduction • Dot plots • Needleman-Wunsch Algorithm	**L4-1** Alignment by hand (Needleman-Wunsch algorithm) Journal Club (paper 2)	
4.27 Th, week 4	• Smith-Waterman • BLAST • Homology	**L 4-2** Alignment by hand (Smith-Waterman algorithm and dot plots)	
5.2 Tu, week 5	• synonymous/nonsynonymous mutations, • Ka/Ks ratios Kimura: Neutral Theory of Molecular Evolution	**L5** Using BLAST and online alignment tools (MUSCLE)	
5.4 Th, week 5	• Phylogenetics • Hierarchical clustering	**Midterm Exam**	

	• k-means clustering • Phylogenetic trees		
5.9 Tu, week 6	• High throughput sequencing • Reference genomes • UCSC genome browser tour	**L6-1** Determining Ka/Ks ratios by hand (small-scale) and via bioinformatics tools (large-scale) • Hierarchical clustering and k-means clustering in R • Creating and interpreting phylogenetic trees	
5.11 Th, week 6	• RNA-sequencing • PCA • Paper 3 Introduction	**L6-2** PCA creation and interpretation in R: scree plots, PCA scatter plots, loading plots. Using PCA for clustering.	Staphylococcus aureus induces a muted host response in human blood that blunts the recruitment of neutrophils
5.16 Tu, week 7	• UNIX / command line • RNA-seq data processing and file types (Phred scores, trimming, alignment to reference genome, counting reads)	**L7-1** Command line basics. FASTA/FASTQ and SAM/BAM files. Trimming and alignment with dummy data.	
5.18 Th, week 7	• RNA-seq normalization • DEG analysis • Genome annotation	**L7-2** DESeq2 normalization. DEG analysis and visualization (volcano plots) with DESeq2.	
5.23 Tu, week 8	• GO/KEGG • Enrichment analysis • Pathway analysis • STRING	**L8-1** Tour of GO/KEGG databases. Functional enrichment analysis and STRING	
5.25 Th, week 8	• Protein structure • Structural bioinformatics • Comparative structure analysis Structure prediction (AlphaFold2) • Protein motion and conformational variants	**L8-2** Comparative structure analysis. Alpha Fold	
5.30 Tu, week 9	*Post-UCSD professional workshop development: Bioinformatics-focused career opportunities and transition – part 1*	**L9** *Bioinformatics-focused Resume* Work on group Presentations and Paper 3 Group Questions	
6.1 Th, week 9	*Post-UCSD professional workshop development: Bioinformatics-focused career opportunities and transition – part 2*	***L9** – Bioinformatics-focused Cover Letter* Post-Assessment and Survey	
6.6 Tu, week 10	• Group Presentations (1)	Group Presentations (2)	
6.8 Th, week 10	Final Exam	**Final Exam**	