BGGN 213: **Foundations of Bioinformatics** (Fall 2025)

Course Instructor: Dr. Barry J. Grant (bjgrant@ucsd.edu)
Course Website: https://bioboot.github.io/bggn213_F25/

Version: 2025-09-19 (16:41:27 PDT on Fri, Sep 19)

Overview: Bioinformatics is driving the collection and analysis of big data in the biosciences. This course is designed for bioscience graduate students and provides a hands-on introduction to the computer-based analysis of genomic and biomolecular data.

Major topics include: Genomic and biomolecular bioinformatic resources, Advances in sequencing technologies; Genome informatics, Structural informatics, and Transcriptomics. Computational tools, techniques and best practices that foster reproducible bioinformatics research will also be introduced. A guest lecture from a genomic scientist at Illumina Inc., Synthetic Genomics Inc., Human Longevity Inc., or the La Jolla Institute for Allergy and Immunology will feature subject to student voting preferences. A comprehensive website containing all reading materials, screencasts and course notes will be maintained throughout the term.

Students completing this course will be able to evaluate new genomic and biomolecular information using existing software and gain experience in combining bioinformatic approaches to answer specific biological questions.

Audience: Bioscience graduate students and others familiar with basic molecular biology concepts. No formal programming training or high level mathematical skills are required.

Requirements: To fully participate in this course students will be expected to use their own computers with specific software installed.

Schedule: Class meets in-person on Wednesday and Friday at 1:00 - 4:00 pm in **TATA 2501** (Map). Please check our class website for schedule updates. Supporting video lectures and screencast hands-on lab sessions will be available on a weekly basis. See the class website for full details.

Class announcements: All announcements regarding the course will be by email to your UCSD address. We will also be using Piazza to facilitate course communication, particularly around questions and answers. If you have a question outside of class or office hours, first check if it has already been asked on Piazza and if not post there. If you have a question or concern you don’t feel comfortable posting on Piazza feel free to reach out via email.

Office hours: We will use walk-in (i.e. in person) and Zoom on a weekly basis at a time to be determined from student polling. At the time of writing this is set to Friday at 4:00pm San Diego time. Again, see the class website for full up-to date details.

Textbook: There is no textbook for the course. **Lecture notes, homework assignments, grading criteria, pre-class screen casts** and required **reading material** will be available from our public facing course website.

Piazza "clubhouse" forum: We will use **Piazza** to help facilitate course communication. If you have course related questions first check if they have already been asked on Piazza and if not post there. I also encourage students to post suggestions and strategies that can help their classmates cope with this most peculiar of quarters. Feel free to post something funny that relates to our course materials, your study skill suggestions, memes about our topics, videos, and, yes, the occasional complaint. My hope is that this "clubhouse" forum will allow us to support one another and give us all a feeling of camaraderie that we would experience in a traditional in-person class.

Course scope and syllabus outline: This is a tentative course schedule. The flow of topics might change slightly depending on how quickly/slowly it feels right to progress through them.

Furthermore, our schedule may change due to campus efforts to contain Covid-19. Any schedule changes will be posted on the course website and to Piazza. Please make sure to frequently check the website to keep updated.

Week 1
Introduction to bioinformatics and key online bioinformatics resources: NCBI & EBI
Biology is an information science, History of Bioinformatics, Types of data, Application areas: Introduction to upcoming segments, NCBI & EBI resources for the molecular domain of bioinformatics, Focus on GenBank, UniProt, Entrez and Gene Ontology.

Week 2
Sequence alignment, DNA and protein database searching
Homology, Sequence similarity, Local and global alignment, Database searching with BLAST, PSI-BLAST, Profiles and HMMs, Protein structure comparisons.

Week 3
Bioinformatics data analysis with R
R language basics and the RStudio IDE, Major R data structures and functions, Data exploration and visualization in R, Import biomolecular data in various formats (both local and from online sources), The exploratory data analysis mindset, Data visualization best practices, Simple base graphics (scatterplots, histograms, bar graphs and boxplots).

Week 4
Writing your own R functions and working with R packages for bioinformatics
Extending functionality and utility with R packages, Obtaining R packages from CRAN and bioconductor, Working with Bio3D for molecular data, Managing and analyzing genome-scale data with bioconductor.

Week 5

Machine learning for bioinformatics

Unsupervised learning, K-means clustering, Hierarchical clustering, Heatmap representations. Dimensionality reduction, Principal Component Analysis (PCA). Longer hands-on session with unsupervised learning analysis of cancer cells further highlighting practical considerations and best practices for the analysis and visualization of high dimensional datasets.

Week 6

UNIX and git for bioinformatics

Why do we use UNIX for bioinformatics? UNIX philosophy, 21 Key commands, Understanding processes, File system structure, Connecting to remote servers. Bioinformatics on the command line, Redirection, streams and pipes, Workflows for batch processing, Shell scripting, Organizing computational projects. Hands-on introduction to the git version control system. We will learn how to perform common operations with git and RStudio. We will also cover the popular social code-hosting platforms GitHub and BitBucket.

Week 7

Structural bioinformatics and bioinformatics in drug discovery and design

Protein structure function relationships, Protein structure and visualization resources, Structural genomics, Homology modeling, Inferring protein function from structure. Small molecule docking methods, Protein motion and conformational variants, Bioinformatics in drug discovery.

Week 8

Genome informatics and high throughput sequencing

Searching genes and gene functions, Genome databases, Variation in the Genome, High-throughput sequencing technologies, biological applications, bioinformatics analysis methods.

Week 9

Transcriptomics, RNA-Seq analysis, and the interpretation of gene lists

RNA-Seq aligners, Differential expression tests, RNA-Seq statistics, Counts and FPKMs and avoiding P-value misuse, Hands-on analysis of RNA-Seq data with R. Gene function annotation, Functional databases KEGG, InterPro, GO ontologies and functional enrichment analysis.

Week 10

Biological network analysis

Network based approaches for integrating and interpreting large heterogeneous high throughput data sets; Discovering relationships in ‘omics’ data; Network construction, manipulation, visualization and analysis; Major graph theory and network topology measures and concepts (Degree, Communities, Shortest Paths, Centralities, Betweenness, Random graphs vs scale free); Hands-on with Cytoscape and igraph packages.

Guest lecture

Student selected guest presentation with possible topics including: Metagenomics / Pharmacogenomics / Epigenomics / Personal genomics / Genome evolution / Genome editing and synthetic genomics / Social impacts and ethical implications of continuing genomic advances.

Potential bonus topics

Essential statistics for bioinformatics
Data summary statistics; Inferential statistics; Significance testing; Two sample t-tests, ANOVA, Regression, and Chi-Square and Power analysis in R; Multiple testing correction; and almost everything you wanted to know about p-values but were afraid to ask! Extensive R examples and applications.

Cancer genomics
Cancer genomics resources and bioinformatics tools for investigating the molecular basis of cancer. Mining the NCI Genomic Data Commons; Immunoinformatics and immunotherapy; Using genomics and bioinformatics to help design a personalized cancer vaccine. Implications for personalized medicine.

The future of bioinformatics
Applications of bioinformatics to translational medicine and the social impacts and ethical implications of how genomic sequence information is used in society.

Systems biology
From genome to phenotypes. Integration of genome-wide data sets into their functional context, Analysis of protein-protein interactions, pathways and networks, Modeling and simulation of systems and networks, Computational methods of network modeling.

Course Objectives:
At the end of this course students will:

- Understand the increasing necessity for computation in modern life sciences research.
- Be able to use and evaluate online bioinformatics resources including major biomolecular and genomic databases, search and analysis tools, genome browsers, structure viewers, and select quality control and analysis tools to solve problems in the biological sciences.
- Be able to use the UNIX command line and the R environment to analyze bioinformatics data at scale.
- Understand the process by which genomes are currently sequenced and the bioinformatics processing and analysis required for their interpretation.
- Be familiar with the research objectives of the bioinformatics related sub-disciplines of Genome informatics, Transcriptomics and Structural informatics.

In short, students will develop a solid foundational knowledge of bioinformatics and be able to evaluate new biomolecular and genomic information using existing bioinformatic tools and resources.

Specific Learning Goals
Teaching toward the specific learning goals below is expected to occupy 60%-70% of class time. The remaining course content is at the discretion of the instructor with student body input. This includes student selected topics for peer presentation as well one student selected guest lecture from an industry based genomic scientist.

All students who receive a passing grade should be able to:

1 Appreciate and describe in general terms the role of computation in hypothesis-driven discovery processes within the life sciences.

2 Be able to query, search, compare and contrast the data contained in major bioinformatics databases and describe how these databases intersect (GenBank, GENE, UniProt, PFAM, OMIM, PDB, UCSC, ENSEMBLE).

3 Describe how nucleotide and protein sequence and structure data are represented (FASTA, FASTQ, GenBank, UniProt, PDB).

4 Be able to describe how dynamic programming works for pairwise sequence alignment and appreciate the differences between global and local alignment along with their major application areas.

5 Calculate the alignment score between two nucleotide or protein sequences using a provided scoring matrix and be able to perform BLAST, PSI-BLAST, HMMER and protein structure based database searches and interpret the results in terms of the biological significance of an e-value.

6 Use UNIX command-line tools for file system navigation and text file manipulation.

7 Use existing programs at the UNIX command line to analyze bioinformatics data.

8 Use R to read and parse comma-separted (.csv) formatted files ready for subsequent analysis.

9 Perform elementary statistical analysis on boimolecular and “omics” datasets with R and produce informative graphical displays and data summaries.

10 View and interpret the structural models in the PDB.

11 Explain the outputs from structure prediction algorithms and small molecule docking approaches.

12 Appreciate and describe in general terms the rapid advances in sequencing technologies and the new areas of investigation that these advances have made accessible.

13 Understand the process by which genomes are currently sequenced and the bioinformatics processing and analysis required for their interpretation.

14 For a genomic region of interest (e.g. the neighborhood of a particular gene), use a genome browser to view nearby genes, transcription factor binding regions, epigenetic information, etc.

15 Given an RNA-Seq data file, find the set of significantly differentially expressed genes and use online tools to interpret gene lists and annotate potential gene functions.

16 Perforn a GO analysis to identify the pathways relevant to a set of genes (e.g. identified by transcriptomic study or a proteomic experiment).

17 Use the KEGG pathway database to look up interaction pathways.

18 Use graph theory to represent biological data networks.

19 Understand the challenges in integrating and interpreting large heterogenous high throughput data sets into their functional context.

20 Have an appreciation for the social impacts and ethical implications of how genomic sequence information is used in our society

Homework assignments and project:
Weekly homework will consist of online knowledge assessment quizzes and application assignments (a.k.a. "hands-on lab sections") together with pre-class reading and video screen-casts.

Specific grading criteria (assessment rubrics) for each homework will be given at the time of assignment. Weekly grades will be posted online. Each student is responsible for checking to ensure that a grade has been entered for their submissions. Documents submitted by email do not always arrive at their intended destination and late submissions will not be accepted after one week past the original due date. Collectively, homework performance will account for 65% of the course grade.

A total of 35% of the course grade will be assigned based on the“*find-a-gene project assignment”*. The purpose of this project assignment is for you to grasp the principles of database searching, sequence analysis, functional annotation and exploratory data analysis with R that we cover in the course (see additional details online). Further details will also be given in class.

There will be no final exam for the Fall 2025 quarter. I understand this is a challenging time and that you may have difficulties with accessing the course material, adapting to online-only learning, and taking online quizzes and exams. I also understand the stress and uncertainty of our current daily reality. My goals are to reach, teach and engage you in the course material. I also aim to fairly test your knowledge of this material, and grade you accordingly while keeping these challenges in mind.

Accommodations
Students requesting accommodations and services due to a disability for this course need to provide a current Authorization for Accommodation (AFA) letter issued by the Office for Students with Disabilities (OSD), prior to eligibility for requests. Receipt of AFAs in advance is necessary for appropriate planning for the provision of reasonable accommodations. Please note that instructors are unable to provide accommodations unless they are first authorized by OSD. For more information, contact the OSD at (858) 534-4382 (voice), osd@ucsd.edu, or visit osd.ucsd.edu."

Diversity and Inclusion.

I would like to create a learning environment for my students that supports a diversity of thoughts, perspectives and experiences, and honors your identities (including race, gender, class, sexuality, religion, ability, etc.) To help accomplish this:

- If you have a name and/or set of pronouns that differ from those that appear in your official UCSD records, please let me know!
- If you feel like your performance in the class is being impacted by your experiences outside of class, please don't hesitate to come and talk with me. I want to be a resource for you. Remember that you can also submit anonymous feedback (which will lead to me making a general announcement to the class, if necessary to address your concerns).
- I (like many people) am still in the process of learning about diverse perspectives and identities. If something was said in class (by anyone) that made you feel uncomfortable, please talk to me about it. Again, anonymous feedback is always an option.
- As a participant in course discussions, you should also strive to honor the diversity of your classmates.
- Please contact me (in person or electronically) or submit anonymous feedback if you have any suggestions to improve the quality of the course materials.

AI Policy.

Generative AI technologies (e.g., ChatGPT, Claude, GitHub Copilot, etc.) are rapidly becoming part of professional practice in bioinformatics. In this course, I will encourage you to engage with these tools, and in some assignments you will be required to use them. Learning to use AI effectively is an emerging skill, and I will provide resources and guidance as required. I am also available to discuss effective strategies during office hours.

Expectations for Use

Encouraged Use: You may use AI to explore ideas, test knowledge, generate code, improve writing, or visualize concepts. AI can help you experiment, avoid simple errors, and deepen your engagement with the course material.

Critical Responsibility: AI outputs may contain errors, bias, or repetition, and often need refinement. You are responsible for validating information, testing code thoroughly, and ensuring your final submission meets the standards of the course.

Documentation and Transparency is key: For your assignments, including the find-a-gene project, include a short paragraph at the end of your assignment explaining how you used AI and provide the key prompts. For code, include the prompt(s) and significant parts of the AI's response in your comments. Transparency is key: we are learning how to use these tools effectively together.

Things to Keep in Mind

Garbage in, garbage out: Minimal prompts will produce low-quality results. Refining prompts takes thought and practice.

AI is not a fact-checker: Assume that factual information, numbers, or citations may be wrong unless you verify them independently.

AI is not always appropriate: Use it when it enhances your work, but recognize when other tools, resources, or your own reasoning are more suitable.

Why We Do This
Our goal as a learning community is not only to master the course content, but also to understand how emerging AI tools can augment and support our research, study, and professional performance. By using AI openly, responsibly, and critically, we are preparing for the environments you will encounter in your academic and professional futures.

Ethics Code.
You are encouraged to collaborate with your fellow students. However, all material submitted to the instructor must be your own work. AI tools can be used to augment and support your work. However any use of AI must be openly acknowledged (see previous section).

> *"Academic Integrity is expected of everyone at UC San Diego. This means that you must be honest, fair, responsible, respectful, and trustworthy in all of your actions. Lying, cheating or any other forms of dishonesty will not be tolerated because they undermine learning and the University's ability to certify students' knowledge and abilities. Thus, any attempt to get, or help another get, a grade by cheating, lying or dishonesty will be reported to the Academic Integrity Office and will result sanctions.*
>
> *Sanctions can include an F in this class and suspension or dismissal from the University. So, think carefully before you act. Before you act, ask yourself the following questions: a) is my action honest, fair, respectful, responsible & trustworthy and, b) is my action authorized by the instructor? If you are unsure, don't ask a friend—ask your instructor, instructional assistant, or the Academic Integrity Office".*

You can learn more about academic integrity at academicintegrity.ucsd.edu
(Source: UCSD Academic Integrity Office, 2022, 2023, 2024, 2025)