BGGN 213: **Foundations of Bioinformatics** (Fall 2017)

Course Instructor: Dr. Barry J. Grant (bjgrant@ucsd.edu)
Course Website: https://bioboot.github.io/bggn213_f17/

DRAFT: 2017-08-10 (15:02:30 PDT on Thu, Aug 10)

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: Participants must bring a laptop with specific software installed.

Schedule: Lectures are on Tuesday and Thursday at 9:00 - 12:00 pm in Warren Lecture Hall 2015 (WLH 2015, UCSD Map Bldg #625). These lectures will include hands-on sessions requiring both individual and small group based computational work. A detailed lecture schedule with topic outlines is provided below.

Class announcements: All announcements regarding the course will be by email to your UCSD address.

Office hours & location: TBD – For now email me for a time and we will make it happen.

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.

Lecture Schedule:

Fall 2017 BGGN 213: **Foundations of Bioinformatics** Lectures (TuTh) 9 - 12 pm		
Th, 09/28	**Welcome to '*Foundations of Bioinformatics*'** (Course introduction, Leaning goals & expectations, Biology is an information science, History of Bioinformatics, Types of data, Application areas and introduction to upcoming course segments, Student computer setup)	1
Tu, 10/03	**Bioinformatics databases and key online resources** (NCBI & EBI resources for the molecular domain of bioinformatics, Focus on GenBank, UniProt, Entrez and Gene Ontology. Hands on with BLAST, GenBank, OMIM, GENE, UniProt, Muscle, PFAM and PDB bioinformatics tools and databases)	2
Th, 10/05	**Sequence alignment fundamentals, algorithms and applications** (Homology, Sequence similarity, Local and global alignment, classic Needleman-Wunsch, Smith-Waterman and BLAST heuristic approaches)	3
Tu, 10/10	**Advanced database searching** (Database searching beyond BLAST, PSI-BLAST, Profiles and HMMs, Protein structure comparisons)	4
Th, 10/12	**Introduction to UNIX for bioinformatics** (Why do we use UNIX for bioinformatics? UNIX philosophy, 21 Key commands, Understanding processes, File system structure, Connecting to remote servers)	5
Tu, 10/17	**Working with Unix** (Bioinformatics on the command line, Redirection, streams and pipes, Workflows for batch processing, Shell scripting, Organizing computational projects)	6
Th, 10/19	**Bioinformatics data analysis with R** (R language basics and the RStudio IDE, Major R data structures and functions, Using R scripts from the command line)	7
Tu, 10/24	**Data exploration and visualization in R** (Import 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), Building more complex charts with ggplot)	8
Th, 10/26	**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 genome-scale data with bioconductor)	9
Tu, 10/31	**Structural Bioinformatics** (Protein structure function relationships, Protein structure and visualization resources, Modeling energy as a function of structure, Homology modeling, Predicting functional dynamics, Inferring protein function from structure)	10
Th, 11/02	**Bioinformatics in drug discovery and design** (Target identification, Lead identification, Small molecule docking methods, Protein motion and conformational variants, Molecular simulation and drug optimization)	11

Tu, 11/07	**Mid Term: Find a gene project assignment** (Principles of database searching and sequence analysis)	12
Th, 11/09	**Genome informatics and high throughput sequencing** (Searching genes and gene functions, Genome databases, Variation in the genome, Sequencing technologies past, present and future (Sanger, Shotgun, PacBio, Illumina, toward the $500 human genome), Biological applications of sequencing, Bioinformatics analysis methods)	13
Tu, 11/14	**Major bioinformatics resources for genomics.** (Databases, tools and visualization resources from NCBI, EBI & UCSC, The Galaxy platform for quality control and analysis; FASTQ, SAM and BAM file formats; Sample workflows with FASTQC and bowtie2)	14
Th, 11/16	**Transcriptomics and the analysis of RNA-Seq data** (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)	15
Tu, 11/21	**Genome annotation and the interpretation of gene lists** (Gene finding and functional annotation, Functional databases KEGG, InterPro, GO ontologies and functional enrichment)	16
Th, 11/23	**Happy Thanksgiving!** (No class) **N.B.** Find a gene assignment due on Monday 11/27!	
Tu, 11/28	**Systems and network modeling** (Analysis of protein-protein interactions, Pathways and networks, Computational methods of network modeling, Hands on with Cytoscape)	17
Th, 11/30	**Continuing genomic advances and bioinformatics challenges** (From genome to phenotypes, Integration of heterogenous high throughput genome-wide data sets into their functional context, Data mining and hypothesis generation in the era of "big data", deep learning and artificial intelligence)	18
Tu, 12/05	**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**)	19
Th, 12/07	**Course summary** (Summary of learning goals, Student course evaluation time and exam preparation)	20
Th, 12/12 Date-TBD	**Final exam!**	

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:

		Lecture(s):
1	Appreciate and describe in general terms the role of computation in hypothesis-driven discovery processes within the life sciences.	1, 2, 20
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).	2, 12, 13
3	Describe how nucleotide and protein sequence and structure data are represented (FASTA, FASTQ, GenBank, UniProt, PDB).	3, 10
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.	4, 5

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.	5, 10
6	Use UNIX command-line tools for file system navigation and text file manipulation.	6, 7, 10, 11, 24, 15
7	Use existing programs at the UNIX command line to analyze bioinformatics data.	7, 10, 11, 13, 14, 15, 16
8	Use R to read and parse comma-separted (.csv) formatted files ready for subsequent analysis.	8, 9, 10, 11, 13, 15, 16
9	Perform elementary statistical analysis on boimolecular and “omics” datasets with R and produce informative graphical displays and data summaries.	9, 10, 11, 13, 15, 16
10	View and interpret the structural models in the PDB.	10, 11
11	Explain the outputs from structure prediction algorithms and small molecule docking approaches.	11
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, 14, 15
13	Understand the process by which genomes are currently sequenced and the bioinformatics processing and analysis required for their interpretation.	13
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.	14
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.	15, 16
16	Perform a GO analysis to identify the pathways relevant to a set of genes (e.g. identified by transcriptomic study or a proteomic experiment).	16
17	Use the KEGG pathway database to look up interaction pathways.	17
18	Use graph theory to represent biological data networks.	17, 18

19	Understand the challenges in integrating and interpreting large heterogenous high throughput data sets into their functional context.	19
20	Have an appreciation for the social impacts and ethical implications of how genomic sequence information is used in our society	20

Homework assignments, mid-term project and final exam:
Weekly homework will consist of online knowledge assessment quizzes and application assignments 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 35% of the course grade.

A total of 20% of the course grade will be assigned based on the mid-term "*find-a-gene project assignment*". The purpose of this mid-term assignment is for you to grasp the principles of database searching, sequence analysis and functional annotation that we cover in the course (see additional details online). Further details will be given in class.

There will be one final exam that accounts for 45% of the final grade for the course.

Ethics Code.
You are encouraged to collaborate with your fellow students. However, all material submitted to the instructor must be your own work.

> *"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, 2017)