BIMM 143: **Introduction to Bioinformatics** (Spring 2018)

Course Instructor: Dr. Barry J. Grant (bjgrant@ucsd.edu)

Course Website: https://bioboot.github.io/bimm143_S18/

Version: 2018-03-21 (13:43:09 PDT on Wed, Mar 21)

Overview: Bioinformatics - the application of computational and analytical methods to biological problems - is a rapidly maturing field that is driving the collection, analysis, and interpretation of the avalanche of data in modern life sciences and medical research.

This upper division 4-unit course is designed for biology majors and provides an introduction to the principles and practical approaches of bioinformatics as applied to genes and proteins. An integrated lecture/lab structure with hands-on exercises and small-scale projects emphasizes modern developments in genomics and proteomics. 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 comprehensive website containing all reading materials, screencasts and course notes will be maintained throughout the term.

Students completing this course will be able to apply leading existing bioinformatics tools to address biological questions. Our broader goal is to point towards perspectives that bioinformatics can expose for the integration and analysis of complex biological information.

Audience: Biology majors with upper division standing. A familiarity with basic biomedical concepts is essential (students should have successfully completed BILD1 and BILD4 or BIMM 101). 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 laptop computers to access bioinformatics software and data resources that are freely available online. A limited number of classroom computers are also available should the need arise.

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.

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.

Course scope and syllabus outline:

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
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 7
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 8
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 9

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.

Week 10

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. Discussion of the social impacts and ethical implications of how genomic sequence information is used in society. Example topics include: The importance of genomic data de-identification. How to best balance privacy concerns with the desire to share and re-use data generated by taxpayer funded research? Should doctors know and preemptively act based on our genetic disease predispositions? Do YOU want to know your genetic disease predispositions? Should patients be given access to their own data and be able to share it? How are racial DNA differences impacting racial prejudices?

Potential bonus topics

Hands-on with git

Hands-on introduction to git, currently the most popular 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.

Essential statistics for bioinformatics

Data summary statistics; Inferential statistics; Significance testing; Two sample T-test in R; 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.

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

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

8 View and interpret the structural models in the PDB.

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

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

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

12 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.

13 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.

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

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

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

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

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, functional annotation and exploratory data analysis

with R 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.

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.

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)