BIMM-143: INTRODUCTION TO BIOINFORMATICS
http://thegrantlab.org/bimm143

Preparing for the **Final Exam**

Overview: The final exam for BIMM-143 will be an open-book, open-notes 150-minute test consisting of 33 questions.

Questions will be predominantly short answer (typically worth 2 points) with a number of more involved longer answer questions (typically worth 5 points).

The number of points for each question is indicated at the beginning of each question. There are 80 total points on offer.

There will be no questions covering the material from lecture 10 (the git version control system). However, major points from all other lecture material are examinable

General exam guidance and test rules are provided at the end of this document.

Major points from lecture 1.
Understand the increasing necessity for computation in modern life sciences research.

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

Be able to describe how nucleotide and protein sequence and structure data are represented (FASTA, FASTQ, GenBank, SAM/BAM, PDB).

Example question: What database should I visit to help determine the protein domains that my novel protein contains?

Major points from lecture 2.
Be able to describe how dynamic programming works for pairwise sequence alignment

Appreciate the differences between global and local alignment along with their major application areas.

Understand how aligning novel sequences with previously characterized genes or proteins provides important insights into their common attributes and evolutionary origins.

Example question: Fill out the dynamic programming table for determining the optimum global alignment between sequences GATCG and GCTCA. Assume that a match is scored +3 and that mismatches and gaps are scored -1 each?

Major points from lecture 3.

Be able to calculate the alignment score between two nucleotide or protein sequences using a provided scoring matrix.

Understand the limits of homology detection with tools such as BLAST
Be able to perform PSI-BLAST, HMMER and protein structure based database searches and interpret the results in terms of the biological significance of an e-value.

Example question: What is the major heuristic shortcut that BLAST uses to speed up database searches?

***Major points from lecture 4*.**

Understand why we use R for bioinformatics.

Familiarity with R's basic syntax.

Be able to use R to read and parse tab and comma-separated (.csv) formatted files ready for subsequent analysis.

Familiarity with major R data structures (vectors, matrices and data.frames)

Example question: What is the base R function for importing a tab-separated-value (tsv) format file for further analysis in R?

***Major points from lecture 5*.**

Appreciate the major elements of exploratory data analysis and why it is important to visualize data.

Be conversant with data visualization best practices and understand how good visualizations optimize for the human visual system.

Be able to generate informative graphical displays including scatterplots, histograms, bar graphs, boxplots, dendrograms and heatmaps and thereby gain exposure to the extensive graphical capabilities of R

Example question: What is the R code to generate the following scatterplot including axis labels, point colors and lines?

***Major points from lecture 6*.**
Understand the structure and syntax of R functions and how to view the code of any R function.

Understand when you should be writing functions.

Be able to follow a step by step process of going from a working code snippet to a more robust function.

Example question: What are the major component parts of an R function?

***Major points from lecture 7*.**
Be able to find and install R packages from CRAN and bioconductor,

Understand how to find and use package vignettes, demos, documentation, tutorials and source code repository where available.

Be able to write and (re)use basic R scripts to aid with reproducibility.

Example question: What is the R code to install the DESeq2 package and load the help page for the deseq() function?

***Major points from lecture 8*.**
Understand the major differences between unsupervised and supervised learning.

Be able to create k-means and hierarchical cluster models in R

Be able to describe how the k-means and bottom-up hierarchical cluster algorithms work.

Example question: What is the base R functions for hierarchal clustering and k-means clustering and what type of major inputs should each function be given as input?

***Major points from lecture 9*.**
Know how to visualize and integrate clustering results and select good cluster models.

Be able to describe in general terms how PCA works and its major objectives.

Know how to interpret the results of PCA in terms of Eigenvectors and Eigenvalues.

Example question: What do the Eigenvalues from principal component analysis (PCA) represent?

***Major points from lecture 10*.**
The content from this session will not feature on the exam.

How to perform common operations with the Git version control system and use GitHub mechanisms for sharing, updating and collaborating (like a social network)

How Git can help keep your work and software organized and available.

***Major points from lecture 11*.**
Understand the classic Sequence>Structure>Function via energetics and dynamics paradigm,

Appreciate the role of bioinformatics in mapping the ENERGY LANDSCAPE of biomolecules,

Be able to use the Bio3D package for exploratory analysis of protein sequence-structure-function-dynamics relationships.

Example question: Name two major computational methods for investigating protein internal dynamics?

***Major points from lecture 12*.**
Be able to use Bio3D and R for the analysis and prediction of protein flexibility,

Be able to perform *In silico* docking and virtual screening strategies for drug discovery,

Appreciate how bioinformatics can aid drug discovery.

Example question: When might you consider using ligand based based approaches vs receptor based approaches for drug discovery and design?

***Major points from lecture 13*.**
Understand the process by which genomes are currently sequenced and the bioinformatics processing and analysis required for their interpretation.

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

Be able to use the Galaxy platform for basic RNA-Seq analysis from raw reads to expression value determination.

Understand the FASTQ file format and the information it holds.

Understand the SAM/BAM file format and the information it holds.

Example question: Name the four major steps of an RNA-Seq differential expression analysis (from raw reads to significantly expressed gene list) and give examples of the software tools you could use for each step?

***Major points from lecture 14*.**
Given an RNA-Seq dataset, find the set of significantly differentially expressed genes and their annotations.

Gain competency with data import, processing and analysis with DESeq2 and other bioconductor packages.

Understand the structure of count data and metadata required for running analysis.

Example question: Why do researchers often normalize read counts in gene expression studies and what features/effects are they normalizing for?

***Major points from lecture 15*.**
Be able to perform a GO and KEGG based gene set enrichment analysis to identify the pathways relevant to a set of genes (e.g. identified by transcriptomic study or a proteomic experiment).

Use both Bioconductor packages and online tools to interpret gene lists and annotate potential gene functions.

Example question: What are the major difference between GO and KEGG gene-sets and what advantages would each have for a typical geneset enrichment analysis?

***Major points from lecture 16*.**
Essential UNIX for bioinformatics. Understand why we use UNIX for bioinformatics and have a familiarity with 21 key UNIX commands that we will use ~90% of the time.

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

Be able to connect to remote servers from the command line and use existing programs at the UNIX command line to analyze bioinformatics data.

Example question: The alignment file “myaln.fa” is not in your current working directory but it is in your “Downloads” directory. At the UNIX command line what command could you use to view the first 10 lines of this file?

***Major points from lecture 17*.**
Be able to describe the major goals of biological network analysis and the concepts underlying network visualization and analysis.

Be able to use Cytoscape for network visualization and manipulation.

Appreciate that the igraph R package has extensive network analysis functionality beyond that in Cytoscape and that the R bioconductor package RCy3 package allows us to bring networks and associated data from R to Cytoscape so we can have the best of both worlds.

Example question: In the context of protein-protein interaction network why would nodes with a high betweenness centrality be potentially more interesting than those with low centrality values?

***Major points from lecture 18*.**
Be able to describe major cancer genomics resources and bioinformatics tools for investigating the molecular basis of cancer.

Appreciate the emerging field of personalized medicine and cancer immunotherapy.

Be able to describe how genomics and bioinformatics can be used to help harness a patient's own immune system to fight cancer.

Example question: Identify the tumor specific mutations given a patients healthy and tumor sequence in FASTA format. What protein do these sequences correspond to?

Be able to determine how frequent a given mutation is in a particular TCGA project using GDC resources.

General exam guidance and test rules:
Please remember to:

- Make sure you bring some spare pens with you.
- Read all questions carefully before starting.
- Begin by writing your name, UCSD email and PID number on your test.
- Write all your answers on the space provided in the exam paper.
- Please write neatly. Illegible answers will be assumed to be incorrect.
- Remember that concise answers are preferable to wordy ones.
- Clearly state any simplifying assumptions you make in solving a problem.
- No copies of this exam are to be removed from the class-room.
- No talking or communication (electronic or otherwise) with your fellow students once the exam has begun.
- Make sure your cell phone and any messaging app on your laptop is switched off.
- When you finish please raise your hand so your completed test may be collected.
- If you have a question or need more papers, raise your hand and we will come to you.
- Once your test has been collected you may leave the classroom quietly and without without talking.
- **Good luck!**