BIMM 143: **Introduction to Bioinformatics**

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.

Hands-on exercises and small-scale projects emphasize 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.

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

Why do we use R for bioinformatics? R language basics and the RStudio IDE, Major R data structures and functions, Using R interactively from the RStudio console. Import biomolecular data in various formats (both local and from online sources).

Week 4

Data exploration and visualization in R

The exploratory data analysis mindset, Data visualization best practices, Simple base graphics (including scatterplots, histograms, bar graphs, dot chats, boxplots and heatmaps), Building more complex charts with ggplot.

Week 5

Writing your own R functions and working with R packages for bioinformatics Using R scripts and Quarto/Rmarkdown files, Import data in various formats both local and from online sources, The basics of writing your own functions that promote code robustness, reduce duplication and facilitate code re-use, Obtaining R packages from CRAN and Bioconductor, Working with Bio3D for molecular data, Managing genome-scale data with bioconductor..

Week 6

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 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. The Galaxy platform along with resources from the EBI & UCSC.

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

Structural Bioinformatics (AlphaFold)

Comparative structure and sequence analysis in R. The importance of Multiple Sequence Alignments (MSAs). Combining knowledge based and physics based approaches for structure prediction and modeling functional motions, AlphaFold2, LLMs and the new age of structural biology.

Week 10

Course wrap up, project completion

Summary of learning goals, Student course evaluation time; Find a gene assignment due. Open study or a student selected topic from those below:

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. Hands-on with Cytoscape and igraph packages.

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.

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, BitBucket and GitLab.

Essential statistics for bioinformatics

Review of data summary statistics; Inferential statistics; Significance testing; Two sample T-test; Power analysis; Multiple testing correction; and almost everything you wanted to know about p values but were afraid to ask! Extensive R examples and applications.

Unix for bioinformatics

Bioinformatics on the command line, Why do we use UNIX for bioinformatics? UNIX philosophy, 21 Key commands, Understanding processes, File system structure, Connecting to remote

servers, Redirection, streams and pipes, Workflows for batch processing, Organizing computational projects.

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-separated (.csv) formatted files ready for subsequent analysis.
7	Perform elementary statistical analysis on biomolecular and “omics” datasets with R and produce informative graphical displays and data summaries.
8	Appreciate and describe in general terms the rapid advances in sequencing technologies and the new areas of investigation that these advances have made accessible.
9	Understand the process by which genomes are currently sequenced and the bioinformatics processing and analysis required for their interpretation.
10	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.

11	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.
12	Perform a GO analysis to identify the pathways relevant to a set of genes (e.g. identified by transcriptomic study or a proteomic experiment).

13	Use the KEGG pathway database to look up interaction pathways.
14	Understand the challenges in integrating and interpreting large heterogenous high throughput data sets into their functional context.
15	Have an appreciation for the social impacts and ethical implications of how genomic sequence information is used in our society