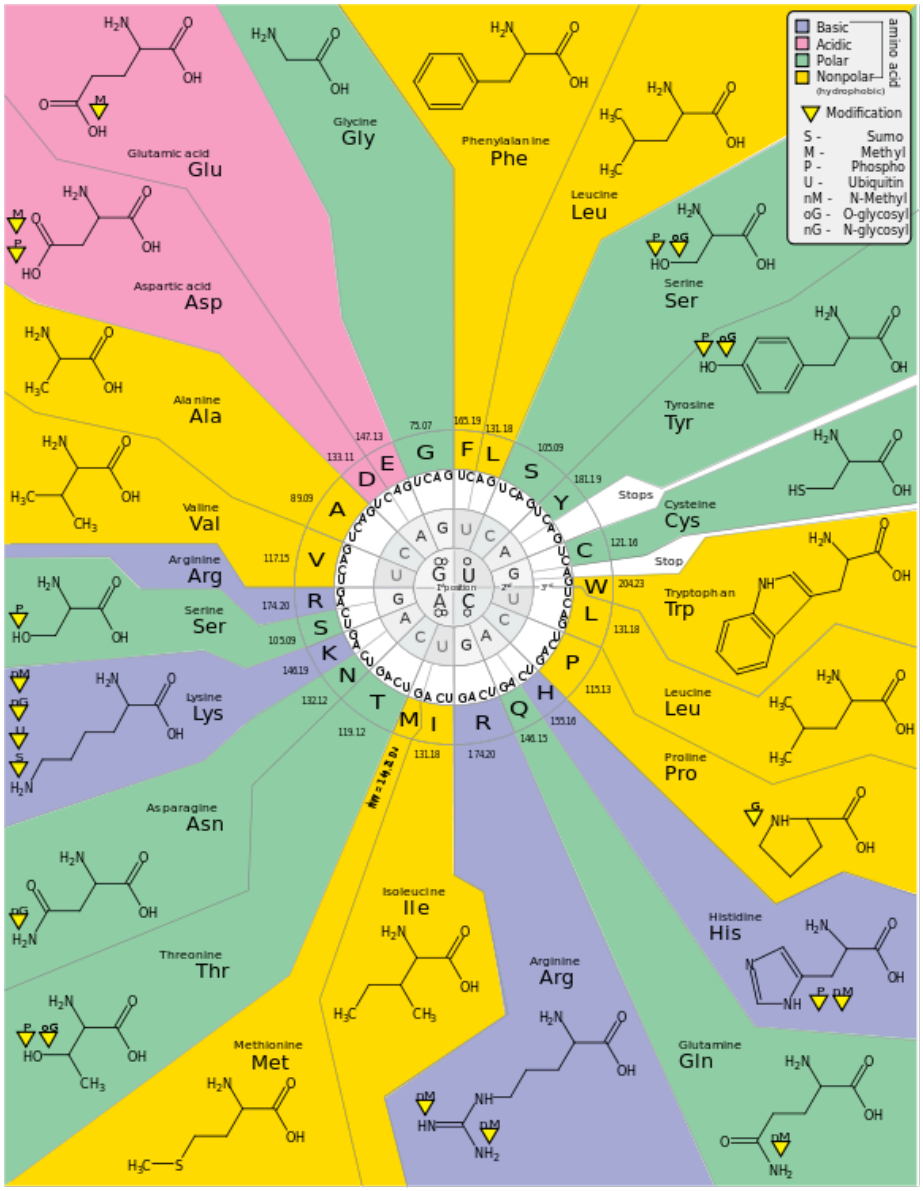
Lecture 16. Protein Structure Prediction

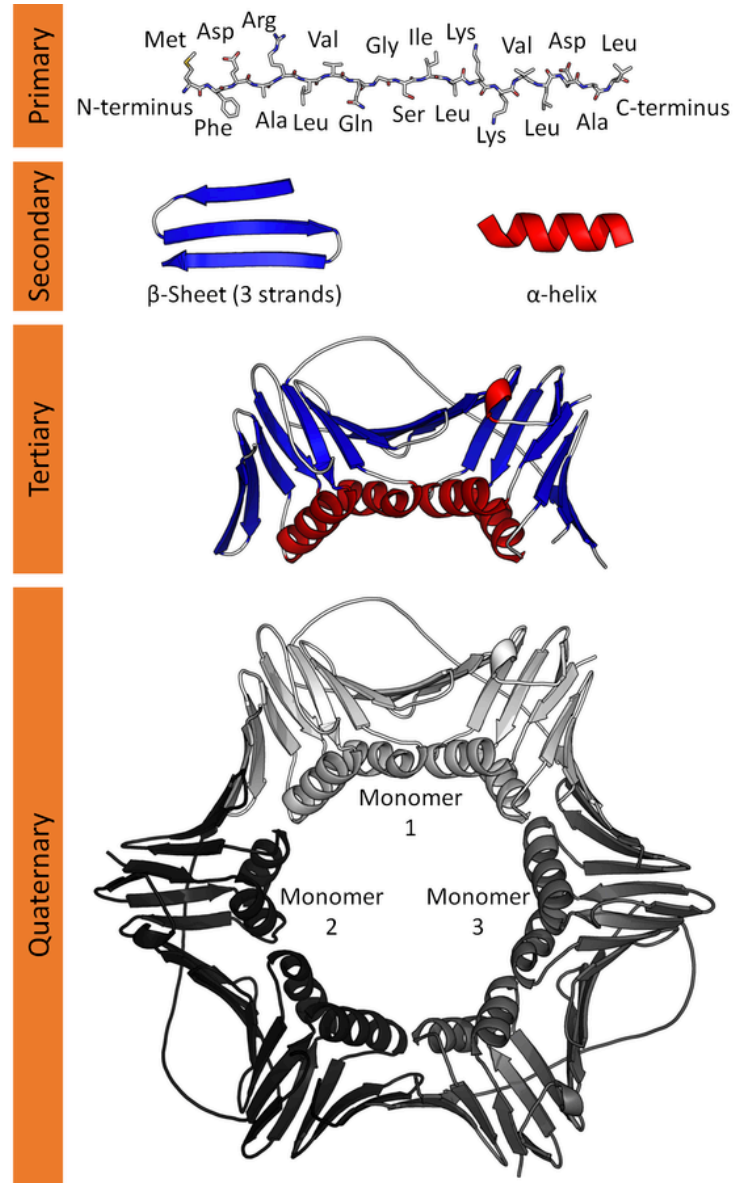
Michael Schatz

April 4, 2017

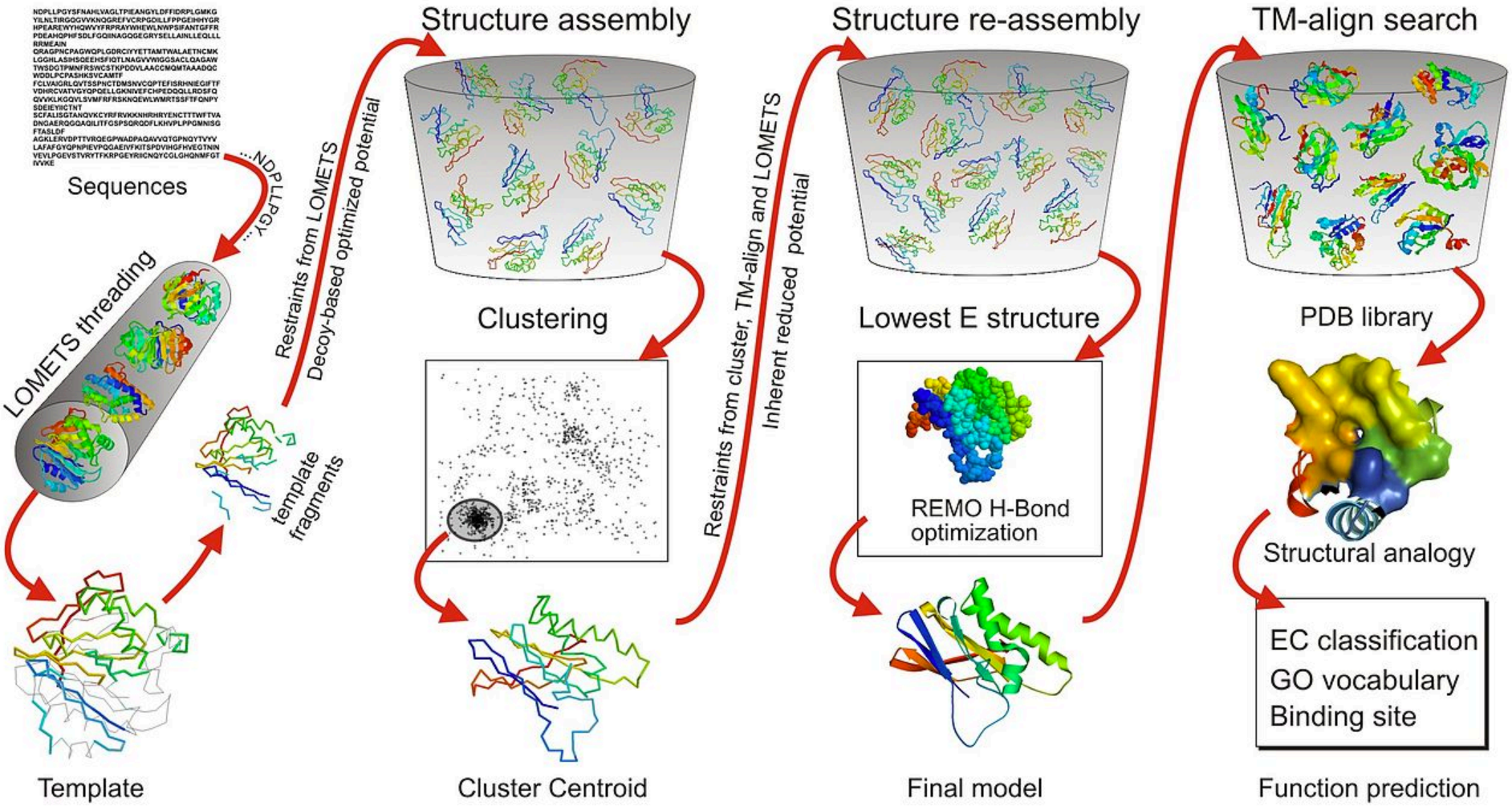
JHU 600.649: Applied Comparative Genomics

Protein Structure



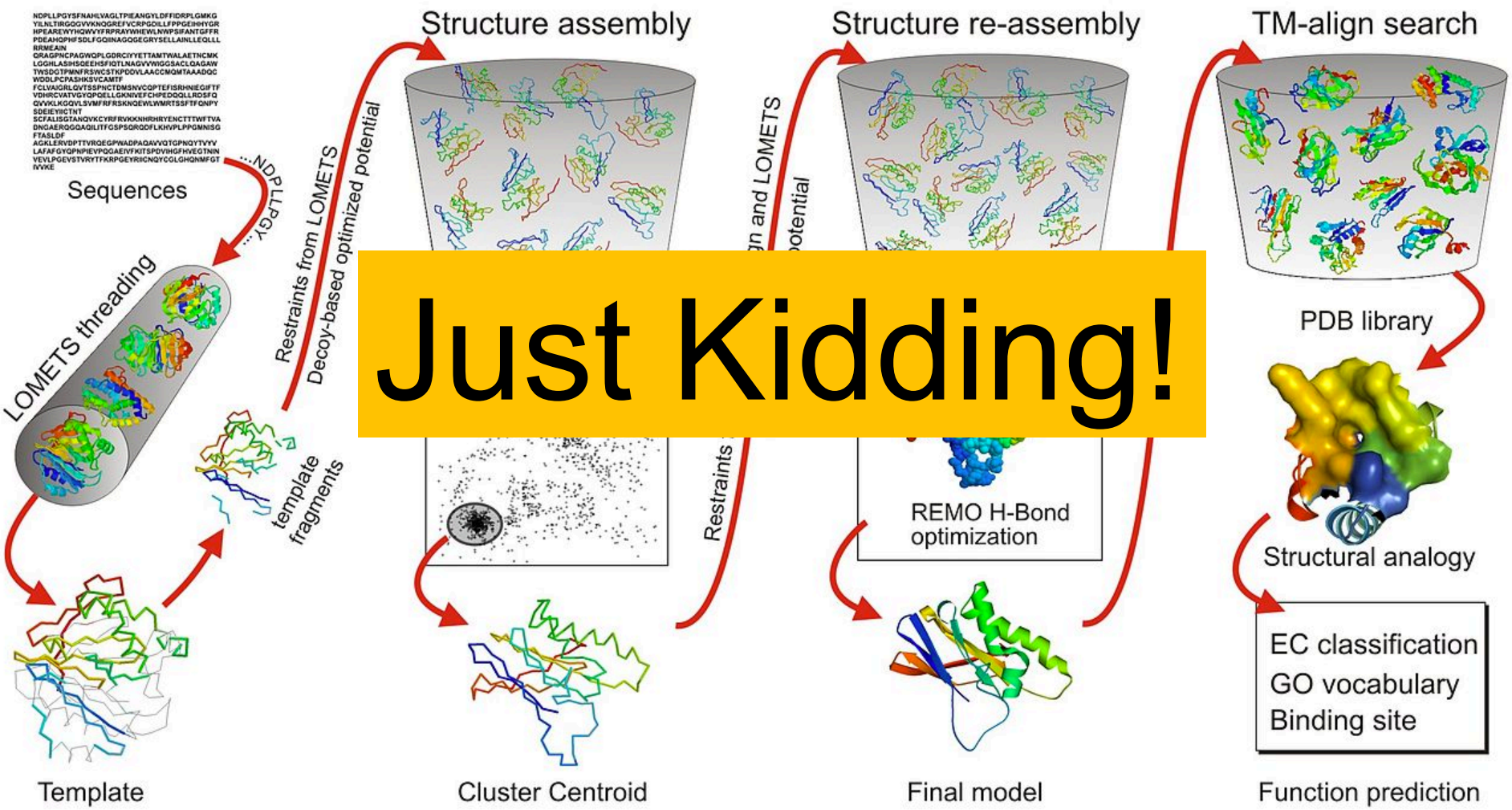


Protein Structure Prediction



I-TASSER: a unified platform for automated protein structure and function prediction
Roy et al (2010) *Nature Protocols* 5, 725–738. doi:10.1038/nprot.2010.5

Protein Structure Prediction



Just Kidding!

I-TASSER: a unified platform for automated protein structure and function prediction
Roy et al (2010) *Nature Protocols* 5, 725–738. doi:10.1038/nprot.2010.5

Review!

Assignment 1: Due Thursday @ 11:59pm

Email PDF to: jhuappliedgenomics@gmail.com

appliedgenomics/README.m

GitHub, Inc. [US] https://github.com/schatzlab/appliedgenomics/blob/master/assignments/assignment1/README.md

This repository Search Pull requests Issues Gist

schatzlab / appliedgenomics

Unwatch 5 Unstar 8 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Wiki Pulse Graphs Settings

Branch: master appliedgenomics / assignments / assignment1 / README.md Find file Copy path

mschatz Update README.md 31eccf2 10 days ago

1 contributor

138 lines (96 sloc) 8.07 KB Raw Blame History

Assignment 1: Genome Assembly

Assignment Date: Thursday, Feb. 9, 2017
Due Date: Thursday, Feb. 23, 2017 @ 11:59pm

Assignment Overview

In this assignment, you are given a set of unassembled reads from a mysterious pathogen that contains a secret message encoded someplace in the genome. The secret message will be recognizable as a novel insertion of sequence not found in the reference. Your task is to assess the quality of the reads, assemble the genome, identify, and decode the secret message. If all goes well the secret message should decode into a recognizable english text, otherwise doublecheck your coordinates and try again. As a reminder, any questions about the assignment should be posted to Piazza

Some of the tools you will need to use only run in a linux environment. Allpaths, for example, will *not* work under Mac, even though it will compile. If you do not have access to a linux machine, download and install a virtual machine following the directions here: https://github.com/schatzlab/appliedgenomics/blob/master/assignments/virtualbox.md

Question 1. Coverage Analysis [10 pts]

Download the reads and reference genome from:
https://github.com/schatzlab/appliedgenomics/raw/master/assignments/assignment1/asm.tgz

Assignment 2

Due: Thursday March 16 @ 11:59pm

appliedgenomics/assignments

GitHub, Inc. [US] https://github.com/schatzlab/appliedgenomics/tree/master/assignments/assignment2

schatzlab / appliedgenomics

Code | Issues 0 | Pull requests 0 | Projects 0 | Pulse | Graphs | Settings

Branch: master appliedgenomics / assignments / assignment2 /

mschatz Add assignment 2 — Latest commit d96b865 13 seconds ago

README.md — Add assignment 2 — 13 seconds ago

README.md

Assignment 2: Variant Analysis

Assignment Date: Tuesday, March 7, 2017
Due Date: Tuesday, March 14, 2017 @ 11:59pm

Assignment Overview

In this assignment, you are will identify variants in a human genome and then analyze the properties for them. **Make sure to show your work in your writeup!** As before, any questions about the assignment should be posted to Piazza

Some of the tools you will need to use only run in a linux environment. If you do not have access to a linux machine, download and install a virtual machine following the directions here:
https://github.com/schatzlab/appliedgenomics/blob/master/assignments/virtualbox.md

Question 1. Gene Annotation Preliminaries [10 pts]

Download the annotation of build 38 of the human genome from here: ftp://ftp.ensembl.org/pub/release-87/gtf/homo_sapiens/Homo_sapiens.GRCh38.87.gtf.gz

- Question 1a. How many many GTF data lines are in this file? [Hint: The first few lines in the file beginning with "#" are so-called "header" lines describing thing like the creation date, the genome version (more on that later in the course), etc. Header lines should not be counted as data lines.]
- Question 1b. How many annotated protein coding genes are on each chromosome of the human genome? [Hint: Protein coding genes will contain the following text: transcript_biotype "nonsense_mediated_decay"]
- Question 1c. What is the maximum, minimum, mean, and standard deviation of the span of protein coding genes?
- Question 1d. What is the maximum, minimum, mean, and standard deviation in the number of exons for protein coding genes? [Hint: you should separately consider each isoform]

Question 2. Genome Sequence Analysis [10 pts]

Download chromosome 22 from build 38 of the human genome from here: http://hgdownload.cse.ucsc.edu/goldenPath/hg38/chromosomes/chr22.fa.gz

- Question 2a. What is the length of chromosome 22? [Hint: You should include Ns in the length]
- Question 2b. How many Ns are in chromosome 22? What is the GC content? [Hint: You should exclude Ns when computing GC content)
- Question 2c. Restriction enzymes cleave DNA molecules at or near a specific sequence of bases. For example, the HindIII enzyme cuts at the "/" in either this motif: 5'-A/AGCTT-3' or its reverse complement, 3'-TTCGA/A-5'. How many perfectly matching HindIII restriction enzyme cut sites are there on chr22?
- Question 2d. How many HindIII cut sites are there on chr22, assuming that a mutant form of HindIII will tolerate a mismatch in the second position? Think about ways in which you could best test for all the possible DNA combinations. [Hint: There are many valid approaches]

Question 3. Small Variant Analysis [10 pts]

Download the read set from here: https://github.com/schatzlab/appliedgenomics/blob/master/assignments/assignment2/sample.tgz

For this question, you may find this tutorial helpful: http://clavius.bc.edu/~erik/CSHL-advanced-sequencing/freebayes-tutorial.html

- Question 3a. How many single nucleotide and indel variants does the sample have? [Hint: Align reads using `bwa mem`, identify variants using `freebayes`, filter using `vcffilter -f "QUAL > 20"`, and summarize using `vcfstats`]
- Question 3b. How many of the variants fall into genes? How many fall into exons? [Hint: `bedtools`]
- Question 3c. What is the transition/transversion ratio of the variants in protein coding genes? What is the ratio of variants in the exons? [Hint: try `bedtools` and `vcfstats`]
- Question 3d. Does the sample have any 'nonsense' or 'missense' mutations? [Hint: try the Variant Effect Predictor using the `Gencode basic transcripts`]

Question 4. Structural Variation Analysis [10 pts]

For this question, you should use the same reads and `bwa mem` alignments as question 3.

- Question 4a. Plot the copy number status of the sample across the chromosome divided into 10kbp bins [Hint: your plot should show how many reads align to bases 1-10k, 10k-20k, 20k-30, etc]

Topics

Genomics

- Genomics Technologies
- Genome Assembly
- Whole Genome Alignment
- Read mapping
- Variant Identification
- Gene Finding
- RNA-seq
- Methyl-seq
- Chip-seq
- Genome Annotation
- ENCODE

Quantitative Techniques

- Normal Distribution
- Poisson Distribution
- Binomial Distribution
- P-value
- Z-statistic
- Quality Values (Phred Scale)
- Differential Expression
- Expectation Maximization
- Full text indexing
- Plane Sweep Alg
- Hidden Markov Models

What is the goal? What is the approach? What are the key challenges?