BINF200 H2025 learning goals

Tom Michoel

November 7, 2025

1 Molecular biology

Know essential concepts of molecular biology necessary for understanding biological sequences and structures.

- Be able to explain the central dogma
- Be able to exaplain what base pairing, genes, introns, exons, codons, etc. are.
- Be able to explain what the genetic code is.
- Be able to construct the complimentary sequence of a given sequence.
- Be able to to translate an RNA sequence into a protein sequence using a genetic code table.

2 Global pairwise sequence alignment

Understand the Needleman-Wunsch algorithm.

- Be able to explain the principle of dynamic programming and why it can be used to maximize a pairwise alignment score.
- Be able to fill a dynamic programming table given a scoring scheme.
- Be able to backtrack in a dynamic programming table to find one or more optimal global alignments.

3 BLAST

Understand the basic steps of the BLAST algorithm.

- Be able to compile a list of n-grams for a given input sequence.
- Be able to create a look-up table of n-grams for a given input sequence.

- Be able to create a score table for matching n-grams for a given score matrix.
- Be able to create a look-up table of matching n-grams.
- Be able to search a database for sequences with matching n-grams using a score threshold.
- Be able to explain the concept of statistical significance and the E-value.

4 Multiple sequence alignment

Understand the principle and main approach of multiple sequence alignment (MSA).

- Be able to explain different MSA scoring methods (sum of pairs, entropy-based).
- Be able to explain why a dynamic programming solution exists in theory but is not practical.
- Be able to explain and apply the progressive multiple sequence alignment algorithm.

5 Phylogenetics

Understand what is phylogenetics and how phylogenetic trees are constructed.

- Be able to explain what evolutionary relationships are.
- Be able to explain the difference between physiological trait-based and molecular sequence-based phylogenetics.
- Be able to explain the major assumptions of molecular phylogenetics.
- Be able to define the root of a tree using an outgroup or midpoint rooting.
- Be able to explain the different types of trees.
- Be able to explain the different steps in tree construction.
- Be able to explain and apply the different steps of the UPGMA algorithm for constructing a tree from a matrix of pairwise distances.
- Be able to explain and apply the principle of maximum parsimony.

6 Sequence motifs

Understand what sequence motifs are, how known motifs can be detected in a sequence, and how new motifs can be discovered.

- Be able to give examples of types of regulatory sites in a genome and their biological function.
- Be able to explain what position-specific count, probability, and log-odds score matrices are.

- Be able to detect known motifs in a sequence by scoring all sequence positions using a position-specific probability or log-odds score matrix.
- Be able to explain the concept of statistical significance for deciding a score cutoff.
- Be able to explain the motif discovery problem and why it is hard.
- Be able to explain the Expectation-Maximization algorithm, incl. the general idea of the algorithm, the main mathematical equations or principles defining the steps of the algorithm, and high-level pseudo-code.
- Be able to explain the Gibbs sampler algorithm, incl. the general idea of the algorithm, the main mathematical equations or principles defining the steps of the algorithm, and high-level pseudo-code.

7 Hidden Markov models

Understand what hidden Markov models (HMMs) are, how the hidden state path can be inferred from a sequence of observations, and what HMMs are used for in biological sequence analysis.

- Be able to explain what a Markov chain is, what a HMM is, and what the difference between them is.
- Be able to give a formal definition of a HMM.
- Be able to compute the joint probability of observing a sequence of states and emitted symbols given the state transition and emission probabilities of a HMM.
- Be able to explain the Viterbi algorithm for finding the most probable hidden state path from a sequence of observations, incl. the general idea of the algorithm, the main mathematical equations or principles defining the steps of the algorithm, and high-level pseudo-code.
- Be able to explain the general idea of posterior decoding and how it differs from finding the most probably state path.
- Be able to give examples of biological sequence analysis tasks that can be solved using HMMs.

8 RNA secondary structure

Understand the physical origin of RNA folding, how RNA secondary structure can be represented, and how folded structures can be predicted from an RNA sequences.

- Be able to explain the physical origin of RNA folding.
- Be able to explain the assumption of no tertiary interactions and what it means for defining secondary structures.
- Be able to convert the secondary structure for a given RNA sequence into its dot-bracket notation, and vice versa.

- Be able to explain the Nussinov folding algorithm, incl. the general idea of the algorithm, the main mathematical equations or principles defining the steps of the algorithm, the main difference with standard dynamic programming algorithms for sequence alignment, and high-level pseudo-code.
- Be able to explain why maximizing the number of base pairs does not necessarily result in the thermodynamically most stable structure.

9 Protein structure analysis

Know fundamental structures of biological molecules, how structures are determined, and where structural information can be found.

- Be able to explain the fundamental structure of DNA, RNA, and proteins.
- Be able to explain multiple mechanisms by which structures are tuned.
- Be able to list important knowledgebases to find information about genes, proteins, and protein structures and interactions.
- Be able to explain the basic building blocks of protein structures.
- Be able to list multiple fundamental experimental methods for protein structure determination.
- Be able to explain why we want and can predict protein tertiary structure from sequence.
- Be able to explain what is CASP.
- Be able to explain classical methods for protein tertiary structure prediction.
- Be able to explain what is alphafold and why it has revolutionized the field of protein tertiary structure prediction.