1 Format

- approximately 20 T/F and short answer: justify your solution in a sentence or two
- approximately 2 practical problems: work through the steps of an algorithm
- approximately 1 design problem: an open-ended question about approaching a research situation
- covers all lectures and recitations through Friday, November 18th - more emphasis on main lecture/ recitation material, no coverage of esoteric topics
- no calculator, closed-notes, two double-sided handwritten note sheets are allowed

2 Topics

Alignment global alignment (Needleman–Wunsch), local alignment (Smith–Waterman), scoring matrices (PAM, BLOSUM), bounded dynamic programming, linear space alignment, multiple alignment, affine gap penalty

Database search BLAST, hashing (combs), Karp–Rabin

Hidden Markov Models definition (states, transitions, emissions), most likely path (Viterbi), posterior decoding (forward, backward), supervised learning (maximum likelihood), unsupervised learning (Baum–Welch)

Clustering and classification $k$-means, fuzzy $k$-means, hierarchical clustering, naive Bayes, random forest

Gene expression protocols (protein binding microarrays, RNA-seq), read mapping (Burrows–Wheeler transform, suffix trees and arrays), transcript abundance, alignment vs. assembly, assigning reads to transcripts, differential expression, short read sequence analysis
RNA secondary structure (Nussinov, context-free grammars), post-transcriptional regulation (splicing, degradation, codon usage bias), ribosome profiling, RNA world hypothesis

Epigenomics ChIP-seq, chromatin conformation capture (3C, 5C, HiC, ChIA-PET), lamina-associated domains (LADs), read mapping, peak calling, IDR, multivariate HMM, ENCODE, Roadmap Epigenomics

Regulatory motifs transcriptional regulation, motif discovery (expectation maximization, Gibbs sampling, gapped motifs), variation at transcription factor binding sites

Networks types of networks (physical, Bayesian), representing networks (adjacency matrix, Laplacian), structural properties (centrality), SVD, dimensionality reduction (principal components analysis, factor analysis), network modules, learning Bayesian networks, causal networks, D-Separation, label propagation, neural networks (convolutional, recurrent), autoencoder

Evolution phylogenetic trees (genes, species), models of nucleotide evolution (Jukes–Cantor, Kimura), distance metrics (ultrametric, additive, Robinson–Foulds), tree-building algorithms (UPGMA, neighbor joining), alignment scoring (parsimony, maximum likelihood, peeling algorithm), reconciliation (orthologs, paralogs), estimating constraint genome-wide, evolutionary signatures, modeling historical populations

Genome assembly read mapping, error correction, graph representation

Population genetics, association mapping Wright–Fisher process, coalescent process, linkage disequilibrium, haplotype phasing, GWAS, eQTLs, missing heritability, complex trait analysis

3 Practice Problems

1. In the four-nucleotide DNA code, the 20 amino acids are codons of length 3. Suppose Martians have 40 different amino acids and a five-nucleotide code (A,C,G,T,Z). What is the minimum codon length required to encode Martian proteins? Justify your answer.

2. Within our framework for sequence alignment, what alignment (global or local) and costs (gap, match, and mismatch) would you use to find Hamming distance (i.e. number of positions at which two contiguous and equal-length strings are different)?

3. What is the minimum asymptotic amount of space needed to compute the score of a global alignment between sequences of lengths N and M? Describe briefly how this is done.

4. True/False When performing protein BLAST with window size w, a matching target
and query sequence will not necessarily have at least one contiguous substring of length $w$ in common.

5. **True/False** In the Blosum62 scoring matrix, the match score of two rarely occurring amino acids is higher than that of two frequently-occurring amino acids.

6. What are two evolutionary signatures of protein-coding regions that distinguish them from non-coding regions?

7. What evolutionary signatures are associated with a whole-genome duplication?

8. Give an advantage and a disadvantage of posterior decoding over Viterbi for determining the hidden state at each position.

9. Briefly describe why intron invasion poses a problem unique to mapping the short reads obtained from RNA-seq to the genome.

10. When run to completion, what quantity does the forward algorithm compute? When run to completion, what quantity does the backward algorithm compute?

11. We wish to design an HMM, similar to our CpG island detector, that instead uses biases in the usage frequency of codons to distinguish between protein-coding regions and non-coding regions. For example, in protein-coding sequence, the codon AGG is very common, while the stop codon TAG is very rare. How many states would you include in this HMM, and why?

12. What is one limitation of Nussinov’s algorithm for finding RNA secondary structures?

13. How is the out-of-bag estimate in random forest classification calculated?

14. We have seen several algorithms based on the principle of expectation-maximization, a general probabilistic framework for updating model parameters when there is some unknown hidden data. What are the parameters and hidden data in each application?

<table>
<thead>
<tr>
<th></th>
<th>parameters</th>
<th>hidden data</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMMs (Baum-Welch)</td>
<td></td>
<td>hidden state assignments ($\pi_i$)</td>
</tr>
<tr>
<td>motif finding (Gibbs)</td>
<td>motif position weight matrix</td>
<td></td>
</tr>
<tr>
<td>clustering (fuzzy k-means)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

15. Consider one iteration of the K-means and fuzzy K-means algorithms on 3 points with 2 cluster centers. Below you are provided with the probability that each point belongs to each cluster.
(a) Assuming we are performing regular K-means (and that probability is monotonically decreasing with distance), compute updated cluster centers.

(b) Assuming we are performing fuzzy K-means, compute updated cluster centers.

16. Describe the naive Bayes assumption as applied to classification, and how it can lead to double-counting evidence.

17. What is cross-correlation, and how can we use it to assess the quality of ChIP-seq data?

18. What is the difference between Jukes-Cantor and the Kimura 2-parameter model for nucleotide evolution?

19. Is the matrix below additive, ultrametric, both, or neither?

\[
\begin{array}{cccc}
   x & y & P(c_1) & P(c_2) \\
   3 & 5 & 1.0 & 0.0 \\
   8 & 4 & 0.25 & 0.75 \\
   4 & -4 & 0.75 & 0.25 \\
\end{array}
\]