(a) 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.

Codons of length 3 will give $5^3 = 125$ possible combinations, enough to encode the 40 Martian amino acids. (Codons of length 2 give only 25 combinations, which is insufficient).

(b) 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)?

Global alignment with match cost = 0, mismatch cost = 1, gap cost = $\infty$. Note that applying the sequence alignment algorithm is not the most efficient way to compute the Hamming distance between two strings.

(c) 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.

$O(N+M)$ using the linear-space alignment algorithm that we saw in lecture 2.

(d) 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.

True. When performing protein BLAST, after we split the query into overlapping words of size $w$, we find neighboring words for each word up to a threshold $T$. An alignment seed corresponds to an exact match between any of these neighboring words and the database sequence, so a matching query and target string do not necessarily have a word of length $w$ in common.

(e) 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.

True. Two rarely occurring amino acids are less likely to be aligned by chance than two frequently occurring amino acids; therefore, their match score will be higher.

(f) What are two evolutionary signatures of protein-coding regions that distinguish them from non-coding regions?

Possible answers include reading frame conservation (gaps are multiples of three), codon substitution frequencies, and mutations falling primarily in the third codon position.
(g) What evolutionary signatures are associated with a whole-genome duplication?

Gene interleaving of sister regions, duplicate mapping of chromosome regions (two regions in the species with the whole-genome duplication map to one region in the species without the WGD), duplicate mapping of centromeres.

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

Posterior decoding gives us the likelihood of each hidden state at each position, but the sequence of states may be an invalid path.

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

If a read we are trying to map contains parts of two exons that have been spliced together, it is difficult for a matching algorithm to find the sequence in the genome that could have generated it, as it doesn’t know if it has to account for the intron gap.

(j) When run to completion, what quantity does the forward algorithm compute? When run to completion, what quantity does the backward algorithm compute?

When run to completion, both the forward and the backward algorithms compute the total probability of the emission sequence over all possible paths. The forward algorithm computes the probability of observing the previous emission sequence and ending in a particular state; the backward algorithm computes the probability of starting in a particular state and observing the subsequent emission sequence.

(k) 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?

We could include $2 \times 4^3$ states, one for each possible codon in CpG islands and in negative regions.

(l) What is one limitation of Nussinov’s algorithm for finding RNA secondary structures?

Possibilities include that Nussinov’s algorithm cannot identify certain types of RNA secondary structures, such as those that include pseudoknots, and that the algorithm does not account for stacking interactions between neighboring base pairs.

(m) How is the out-of-bag estimate in random forest classification calculated?
Since each tree in RF is built using n samples drawn from our pool of n samples with replacement, about one-third of the samples won’t be used to build the tree. After we have finished building each tree, have it vote on the samples not used to build it. Once we have built all the trees, compare the majority class for each sample tabulated from all the trees it was left out of to its actual class.

(n) 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>transition and emission probabilities</td>
<td>hidden state assignments ($\pi_i$)</td>
</tr>
<tr>
<td>motif finding (MEME)</td>
<td>motif position weight matrix</td>
<td>motif positions</td>
</tr>
<tr>
<td>clustering (fuzzy k-means)</td>
<td>cluster centers</td>
<td>cluster assignments</td>
</tr>
</tbody>
</table>

(o) 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.

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>P(c₁)</th>
<th>P(c₂)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>5</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>0.25</td>
<td>0.75</td>
</tr>
<tr>
<td>4</td>
<td>-4</td>
<td>0.75</td>
<td>0.25</td>
</tr>
</tbody>
</table>

(a) Assuming we are performing regular K-means (and that probability is monotonically decreasing with distance), compute updated cluster centers.

$c₁ : (\frac{7}{2}, \frac{1}{2}); c₂ : (8, 4)$

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

$c₁ : (4, \frac{3}{2}); c₂ : (7, 2)$

(p) Describe the Naive Bayes assumption as applied to classification, and how it can lead to double-counting evidence.

The Naive Bayes algorithm assumes that input features are independent. If the features are correlated then we will overcount evidence.

(q) What is cross-correlation, and how can we use it to assess the quality of ChIP-seq data?
We will observe forward-strand reads only at one end of a fragment, reverse-strand reads only at the other end. This leads to a correlation peak between forward and reverse signals at the fragment length offset. If we don’t perform ChIP, or if the dataset is low-quality, we will see a phantom peak at the length of a read since if position $x$ is mappable on the positive strand, position $x + r - 1$ is mappable on the negative strand. In high quality data, the fragment-length peak should dominate the read-length peak.

**(r)** What is the difference between Jukes-Cantor and the Kimura 2-parameter model for nucleotide evolution?

The Kimura 2-parameter model allows different rates for transitions and transversions, while the Jukes-Cantor model has only a single rate for all types of nucleotide changes.

**(s)** Is the matrix below additive, ultrametric, both, or neither?

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0</td>
<td>2</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>b</td>
<td>2</td>
<td>0</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>c</td>
<td>8</td>
<td>8</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>d</td>
<td>8</td>
<td>8</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

Both. Recall that if a matrix is ultrametric, then for all points $i$, $j$, and $k$, two distances are equal and the third is smaller. An additive distance matrix is one that exactly corresponds to distances drawn from a tree. If a distance matrix is ultrametric, it will be additive as well.