No books, notes, or electronic aids (such as calculators) are permitted. Please turn off your phone network/wifi connections. There's a clock in the back of the room above the door. Exam begins at 9:35am and ends at 10:55am. You have 80 minutes.

<table>
<thead>
<tr>
<th>Problem/Section</th>
<th>Parts</th>
<th>Points</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Write your name</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>T/F - Justify</td>
<td>10</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Short Answer</td>
<td>15</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>Practical Problems</td>
<td>3</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
True/False with justification (2 points each)

Read each statement or question carefully and circle the correct answer. Write a brief (one sentence or even a few words) justification of your answer so that we know you did not just guess right.

1. **True / False** We can find the optimal global alignment of three sequences A, B, and C, each of length n, in $O(n^2)$ time.

2. **True / False** Global sequence alignment with appropriate gap, mismatch, and match costs is a reasonable algorithm for finding the minimum edit distance between two sequences (i.e., the minimum number of insertions, deletions, and substitutions needed to transform one string into the other).

3. **True / False** Gibbs sampling can return the different discovered motifs when run multiple times until convergence on the same dataset.
4. **True / False** Increasing the word length $W$ in the BLAST algorithm will increase both the sensitivity and the run time.

5. **True / False** For a genome containing long identical repeats, increasing the number of reads sufficiently will allow us to construct a complete assembly, even if each individual read is very short.

6. **True / False** One shortcoming of the affinity propagation algorithm is that it requires the number of clusters $k$ to be pre-selected.

7. **True / False** For $k > 2$, the expected time to the first coalescence (going from $k$ to $k-1$ lineages) is less than the expected time to the last coalescence (going from 2 lineages to 1 lineage).

8. **True / False** Neighbor joining always outperforms UPGMA for reconstructing a phylogenetic tree from an ultrametric distance matrix.
For questions 9 and 10, consider the following diagram of a sequence of emissions and states generated by a Hidden Markov Model.

9. **True / False** If we observe the value of the hidden state $\pi_{i-1}$, additionally observing emission $x_{i-1}$ changes the posterior probability of the value of the hidden state $\pi_i$.

10. **True / False** If we observe the value of the hidden state $\pi_{i-1}$, additionally observing emission $x_{i+1}$ changes the posterior probability of the value of the hidden state $\pi_i$. 
Short Answer (4 points each)

11. Recall that a bounded dynamic programming solution to the global alignment problem only considers paths within a distance k of the central diagonal of the dynamic programming matrix. Name one advantage and one disadvantage of using bounded DP for solving the global alignment problem.

12. Suppose you select the top N instances from a ranked list of genes that you have predicted are regulated by a given transcription factor. What is the likely effect on sensitivity and specificity of increasing the number of selected regions from N to 2N?

13. Describe two evolutionary signatures that can be used to detect functional regulatory motifs in the genome.
14. Recall that an entry in a protein-substitution matrix reflects the log-odds ratio \( s(a, b) = \log \left( \frac{p_{a,b}}{q_a q_b} \right) \).
What do \( p_{a,b}, q_a, \) and \( q_b \) represent?

15. Suppose you are hashing all substrings of a reference sequence using the Karp-Rabin update rule for hashing adjacent strings. Will halving the length of the substring that you are hashing affect the runtime, and if so, how?
16. (a) After sequencing the phoenix genome, you hypothesize that the phoenix has undergone a whole-genome duplication at some time during its evolutionary history. What evidence to support your hypothesis could you observe by examining the phoenix genome only?

(b) You further sequence the griffin, which you believe branched from the lineage that led to phoenixes prior to the date of the whole-genome duplication. Describe one feature that you would expect to observe when comparing the phoenix and griffin genomes if there was a whole-genome duplication in the phoenix lineage.

17. For mapping short reads to a genomic sequence using the Burrows-Wheeler transform, do we perform the BWT on the genomic sequence, the sequences of the short reads, or both? Justify your answer.

18. Name one advantage of the Zucker algorithm for predicting RNA secondary structure over the simpler Nussinov algorithm.
19. The ENCODE project developed a rank-based method (IDR) to combine peaks across replicate pairs of ChIP-seq experiments. Name one advantage of IDR over combining replicates by simply taking the union of the peaks.

20. Explain the biological relevance of identifying densely connected subgraphs in regulatory networks.

21. Given a set of cluster centers, what is the rule in the fuzzy k-means algorithm for assigning points to clusters? How does this differ from the corresponding step of the k-means algorithm?

22. Suppose you are performing classification using a set of highly correlated features. What effect will this have on the confidence of predictions generated by the Naive Bayes classifier?
23. Draw a tree with four nodes in which the distances are additive but not ultrametric. Indicate the length of each branch in your diagram.

24. Even in the absence of duplication, loss, or horizontal transfer, the phylogenetic “gene” tree of a region of the genome may differ from the corresponding species tree. For example, although humans and chimpanzees are more closely related to each other than to gorillas, a substantial fraction of the human genome is closer to the gorilla than to the chimpanzee. Explain how this is possible.

25. Consider a Wright-Fisher model with a population of $N$ haploid individuals. What is the probability that two randomly selected individuals share the same parent in the previous generation?
Practical Problems (6 points each)

26. (a) **Genome-wide association studies** Psyphilia is a disease characterized by an inability to stop dancing gangnam-style. You determine that psyphilia is a heritable genetic trait. Describe a scenario in which linkage-based family studies will outperform genome-wide association studies for identifying the genes that are involved in this disorder.

   (b) Describe the importance of multiple hypothesis testing for ensuring the validity of the results of a genome-wide association study. What does a hypothesis in a genome-wide association study correspond to?

   (c) Based on your research, you develop a diagnostic test for psyphilia. Suppose that one in every thousand people in the population has psyphilia. Everyone who is ill with psyphilia, as well as ten percent of healthy individuals, tests positive using your diagnostic test. Write an expression for the likelihood that you have psyphilia, if you test positive. (You do not need to compute the value of the expression).
27. HMMs

(a) We want to design an HMM, similar to our CpG island detector, that uses biases in k-mer frequency to identify likely promoter regions in the genome. In terms of k, how many states would you include in this HMM, and why?

(b) Suppose we have constructed an HMM that considers k-mers of length $q$. How will the runtime of the Viterbi algorithm scale when we construct an HMM that considers k-mers of length $q + 1$?
28. **Expectation maximization**

(a) Consider an iteration of the EM algorithm for a motif of length 3. Below we have provided you with the M matrix (the position weight matrix representing the motif) and a set of sequences. Fill in the missing entries in the matrix Z, where $Z_{ij}$ gives the probability that position $j$ in sequence $i$ is the start of the motif. (Note that the Z matrix should be normalized such that each row sums to 1).

\[
\begin{array}{c|c|c|c}
 & 1 & 2 & 3 \\
\hline
A & 0 & 0 & 0 \\
C & 0 & 1 & 0.25 \\
G & 0 & 0 & 0.75 \\
T & 1 & 0 & 0 \\
\end{array}
\]

\[
\begin{array}{c|c|c|c|c}
\text{Position} & 1 & 2 & 3 & 4 \\
\hline
\text{Sequence 1} & G & T & C & G \\
\text{Sequence 2} & T & C & C & A \\
\end{array}
\]

\[
\begin{array}{c|c|c}
\text{Z} & 1 & 2 \\
\hline
\text{Sequence 1} & & \\
\text{Sequence 2} & & \\
\end{array}
\]

(b) Now, update the M matrix based on the Z matrix that you derived in the previous step.

\[
\begin{array}{c|c|c|c}
M & 1 & 2 & 3 \\
\hline
A & & & \\
C & & & \\
G & & & \\
T & & & \\
\end{array}
\]