6.047/6.878/HST.507
Fall 2014 Quiz

November 25, 2014

Name: □

No books, notes, or electronic aids are permitted. Please turn off your phone. There is a clock in the back of the room above the door. Exam begins at 1:05pm and ends at 2:25pm. You have 80 minutes to earn 90 points.

Section	**Questions**	**Points**	**Score**
Write your name	1	2	
True/False	7	14	
Short answer	13	52	
Practical problems	2	12	
Design problems	1	8	
Total	24	88	

True/False with justification (2 points each)

Read each statement carefully, circle the correct answer, and write a brief justification of your answer.

1. **True / False** We can find an optimal global alignment of two sequences in $O(kn)$ time using bounded dynamic programming where k is the number of off-diagonal entries of each row of the dynamic programming table we compute.

2. **True / False** Consider the tables $f_k(i)$ containing the forward probabilities for state k at position i and $b_k(i)$ containing the backward probabilities for state k at step i. The quantity $f_k(i)b_k(i)/S(i)$ is equal for all positions i and all states k, where $S(i) = \sum_k f_k(i)b_k(i)$.

3. **True / False** Suppose we are classifying fish based on their color, weight, length, the latter measured in inches, centimeters, and smoots. The performance of a Naive Bayes classifier trained on all five features will be equal to that of a classifier trained with only the first three features.

4. **True / False** The string $aactagtt$ cannot be generated by the context-free grammar below:

$$\begin{aligned} S &\to aSt \\ S &\to tSa \\ S &\to cSg \\ S &\to gSc \\ S &\to \epsilon \end{aligned}$$

5. **True / False** When classifying points in the two-dimensional plane, a support vector machine will only perform well on data which is perfectly separated by a straight line.

6. **True / False** Given an ultrametric distance matrix, hierarchical clustering will produce different results depending on whether we use the minimum distance, the average distance, or the maximum distance between points of two clusters as the distance between clusters.

7. **True / False** In a gene tree–species tree reconciliation, the most recent common ancestor of two paralogous genes can be a speciation event.

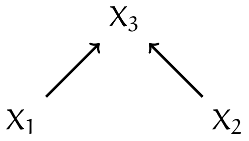
Short answer (4 points each)

8. Give a biological motivation and a computational motivation for computing alignments with affine gap penalties.

9. When aligning proteins, explain why the score of a match between two rarely occurring amino acids has larger absolute value than the score of a match between two commonly occurring amino acids.

10. What is one advantage of sparse principal components analysis over ordinary principal components analysis?

11. Give the factorization of the joint distribution $P(X_1, X_2, X_3)$ implied by the Bayesian network below:



12. Why are motifs for miRNA binding sites less degenerate than motifs for transcription factor binding sites?

13. Explain how a genetic variant near a specific motif instance but not within it could still affect the binding of a transcription factor to that instance.

14. A graph representation of a genome assembly has several advantages over a linear representation. Give two types of uncertainty which cannot be represented by a linear assembly but can be represented by a graph.

15. Explain why CpG islands in the human genome are more likely to contain functional elements.

16. What assumptions about mutation rates are made by UPGMA versus neighbor joining?

17. Explain how we can use comparative genomics to identify the open reading frame of translation of an mRNA transcript.

18. Suppose we perform a combined genome/epigenome-wide association study where we measure the genotype and methylation of a cohort of disease cases and healthy controls. Consider the genotype at a particular SNP X, the methylation at a particular CpG site M, and the disease status Y. Under what conditions can we say that methylation causes disease rather than being a consequence of disease?

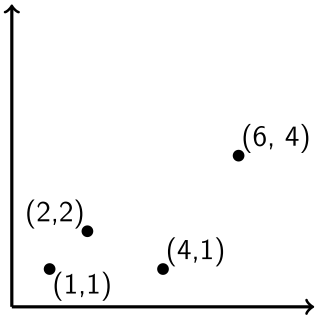
19. Explain how an eQTL could affect the expression level of a gene on a different chromosome.

20. Explain why Mendelian diseases are more likely to be caused by rare variants in protein-coding exons.

Practical Problems (12 points)

21. Consider hierarchically clustering the two-dimensional points below using *Manhattan distance*. The Manhattan distance between points (x_1, y_1) and (x_2, y_2) is defined as:

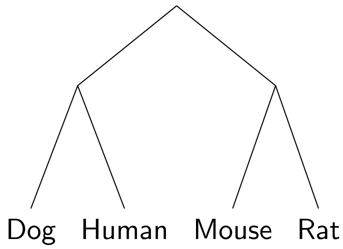
$$|(x_2 - x_1)| + |(y_2 - y_1)|$$



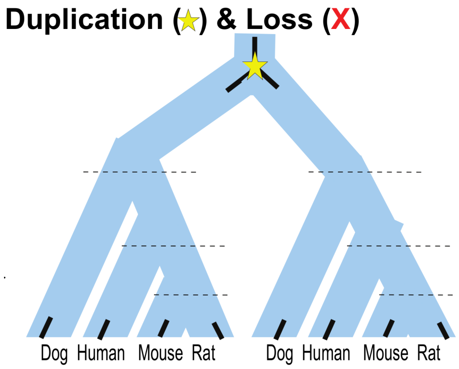
(a) Draw the tree built using *single linkage* where the distance between two clusters is defined as the minimum distance between pairs of points in the two clusters. Include the branch lengths.

(b) Draw the tree built using *complete linkage* where the distance between two clusters is defined as the maximum distance between pairs of points in the two clusters. Include the branch lengths.

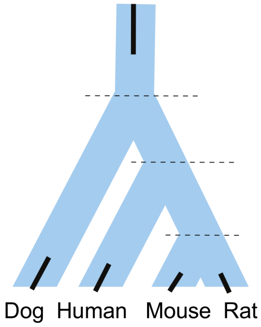
22. Consider the gene tree tree given below:



(a) Reconcile the gene tree with the species tree. Fill in the nesting of the gene tree inside the species tree (blue), labeling duplications and losses as shown below.



(b) Draw a deep coalescence scenario which can explain the gene tree.



Design problem (8 points)

23. One important question in understanding the genetic basis of human disease and developing therapeutics is identification of the causal nucleotides in regions of linkage disequilibrium. Your colleagues from the ENCODE and Roadmap Epigenomics consortia have produced genome-wide maps of regulatory region annotations, including promoters and enhancers across a large number of human cell types, and they reach out for your help. You know that enhancers are enriched for cell type–specific variants associated with disease and that specific variants falling within enhancers are more likely to have causal roles in disease, but you don't know the relevant cell types or the causal variants, so you decide to use an expectation-maximization (EM) approach.

    Your goal is to compute the posterior probability of each SNP being causal, given (a) the p-value of disease association for each common variant in the human genome from a given genome-wide association study, and (b) a binary matrix storing whether or not each common variant falls in an enhancer in each cell type. Describe an EM approach for solving this problem, including (1) an overview of your rationale, (2) what the hidden variables represent and what the observed variables represent, and (3) what is computed in the E and M steps.