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.

   False. Bounded dynamic programming is not guaranteed to find the optimal solution.

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 $S(i) = \sum_k f_k(i)b_k(i)$.

   False. This is $P(\pi_k = i \mid x_1, \ldots, x_n)$.

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.

   False. The classifier on all five features will perform worse than the classifier on only the first three features because the three length features are simply linear transformations of each other and violate the Naive Bayes assumption.

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

$$S \to aSt$$
$$S \to tSa$$
$$S \to cSg$$
$$S \to gSc$$
$$S \to \epsilon$$

False. We can generate it as

$$S \to aSt \to aaStt \to aacSgtt \to aactSagtt \to aact\epsilon agtt$$

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.

False. We can use soft margins to not penalize points which violate the margin of the separating line. We can additionally use the kernel trick to project the data into a higher space (potentially infinite-dimensional) where it is separable by some hyperplane.

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.

False. If the distance is ultrametric, the distance from the root to every leaf is equal. This means every subtree of an ultrametric tree is also ultrametric. Therefore, when we merge two clusters the pairwise distances between points are all equal, which means using the minimum, the average, or the maximum distance between points will give the same result.

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

False. Paralogous genes arise from duplications. Orthologous genes arise from speciations.

Short answer (4 points each)

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

DNA polymerase can slip during replication meaning that once a gap has been opened, extending it should not be penalized as much. We can compute alignments with affine gap penalties in $O(n^2)$ time where we would require $O(n^3)$ time for general gap penalties.

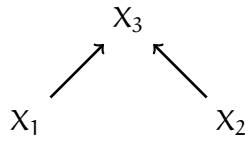
9. When aligning proteins, explain why the score of a match between two rarely occurring amino acids has larger absolute value than the score of a match between two commonly occurring amino acids.

The score of a match is $-\log \frac{p_{aa}}{q_a q_a}$ where p is the probability of a particular substitution and q is the probability of observing a particular amino acid. q_a is smaller for a rare amino acid, making the score larger.

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

The axes of the low rank approximation produced by sparse PCA are more easily interpretable in terms of the original dimensions of the space. For example, in the setting of reducing the dimensionality of a gene expression dataset, PCA could produce principal axes which involve all genes where sparse PCA would produce axes which involve only some small number of genes.

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



$$P(X_1, X_2, X_3) = P(X_1)P(X_2)P(X_3 \mid X_1, X_2)$$

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

miRNA binding motifs are less degenerate because the miRNA seed region binds to mRNA using sequence complementarity. Transcription factors bind to DNA by interacting with atoms on the sugar backbone which allows for degeneracy of individual positions.

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.

The genetic variant could affect the shape of the DNA, specifically the roll (deformation of the helix) or helix twist (non-constant rotation between subsequent base pairs). The variant could fall within the binding site of a different transcription factor which cooperatively binds with the given transcription factor and directly disrupt its binding. The variant could create a new binding site for a different transcription factor which competes with the given transcription factor for occupancy.

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.

Uncertainty in gap lengths; uncertainty in repeat lengths; polymorphisms (SNPs, indels); long haplotypes with recent mutations; structural variants (large deletions, rearrangements)

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

The cytosine in CpGs is more likely to become methylated, which in turn makes it more likely to mutate to a TpG or CpA. Therefore, those CpGs which remain in the genome are more likely to be either unmethylated (not silenced) or under selection.

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

UPGMA assumes the mutation rate is constant over the whole tree where neighbor joining assumes the mutation rate is only constant over each branch.

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

We can use the fact that substitutions in coding regions are more likely to occur in every third position of the coding sequence.

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?

We say M mediates the link between X and Y if X is associated with Y, M is associated with Y, but X is not associated with Y conditioned on M. In this case, genotype causes a change in methylation, which in turn causes a change in disease status.

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

An eQTL could affect the binding of a transcription factor at an enhancer element which is brought near the gene in three-dimensional space, directly disrupt the gene encoding a transcription factor whose binding is required for transcription of the target gene, or indirectly disrupt the pathway regulating the expression of such a transcription factor.

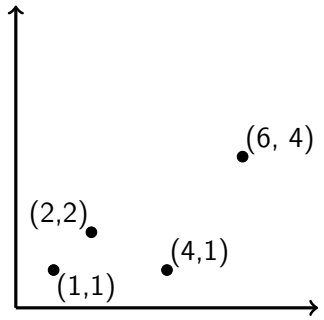
20. Explain why Mendelian diseases are more likely to be caused by rare variants in protein-coding exons.

Mendelian diseases have large effects on reproductive fitness so the variants which cause them are likely to have large effects and be under strong selection. Variants with large effects are more likely to alter proteins. Variants under strong selection are less likely to rise to high frequency in the population.

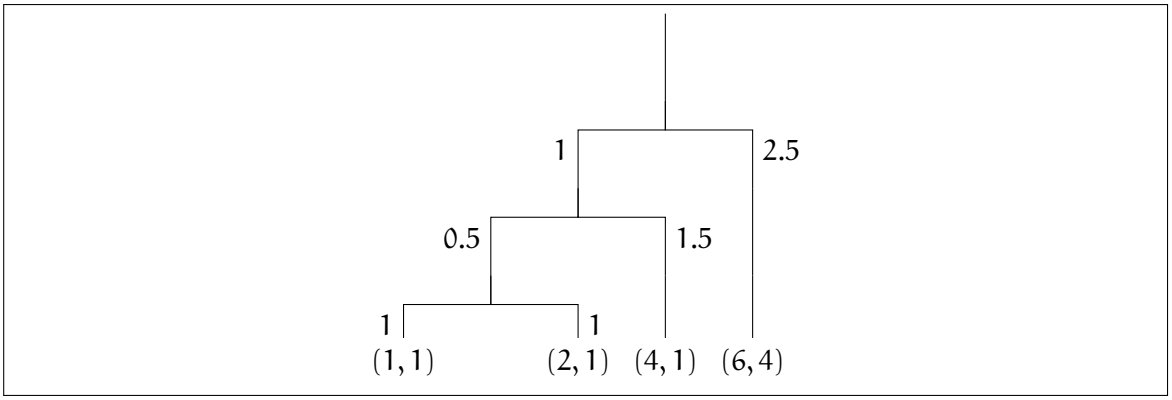
Practical Problems (12 points)

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:

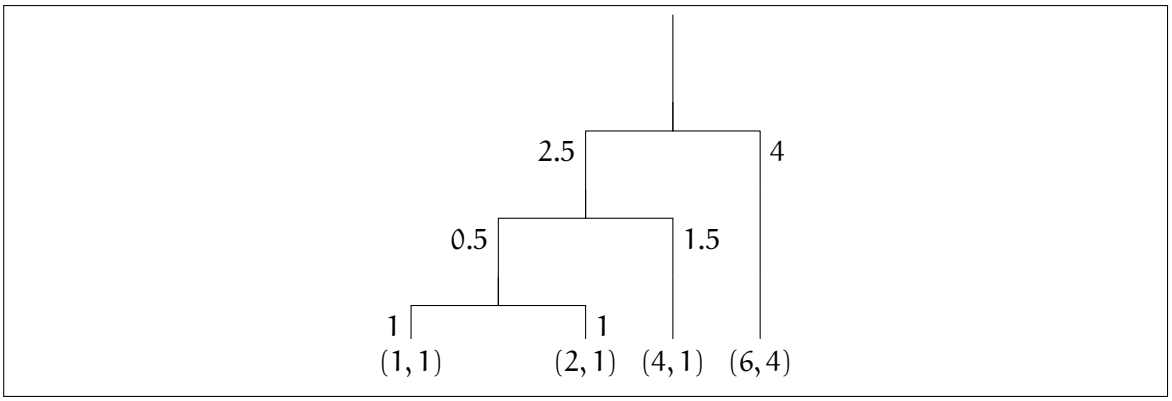
$$|(x_2 - x_1)| + |(y_2 - y_1)|$$



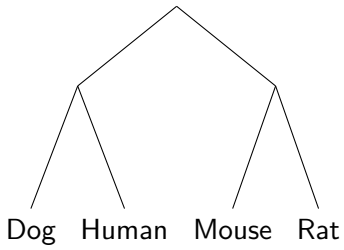
(a) Draw the tree built using *single linkage* where the distance between two clusters is defined as the minimum distance between pairs of points in the two clusters. Include the branch lengths.



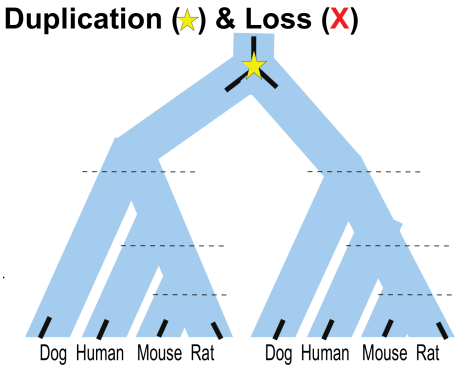
(b) Draw the tree built using *complete linkage* where the distance between two clusters is defined as the maximum distance between pairs of points in the two clusters. Include the branch lengths.



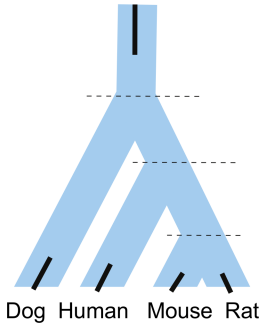
22. Consider the gene tree tree given below:



(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.

    (1) If we knew the causal cell types, we could use the enhancer annotations in those cell types to estimate the probability a variant is causal. If we knew the causal variants, we could use the enhancer annotations to find the causal cell types.

    (2) The causal cell types are hidden variables, so we use EM to compute the posterior probability of a SNP being causal given the hidden variables, integrating over all possible configurations weighted by their joint probability. The observed variables are binary indicator variables for each SNP–cell type pair which take value 1 if that SNP falls into an enhancer region in that cell type. The parameters of the model are the probability that each SNP is causal. The GWAS p-values are prior probabilities of a SNP being causal.

    (3) In the E step, we compute the expected value of the hidden variables using the current estimate of the parameters. In this case, we compute the probability of each cell type being causal based on the probability that the variants falling in the enhancers of that cell type are causal.

    In the M step, we re-estimate the parameters from the expected value of the hidden variables. In this case, we compute the probability of each SNP being causal based on whether it falls in an enhancer in a causal cell type, integrating (summing) over all possible combinations of causal cell types weighted by their probability.