6.047/6.878/HST.507
Fall 2017 Quiz

November 21, 2017

Name:

Two double-sided handwritten sheets of notes are permitted. No books or electronic aids are permitted. Please turn off your phone. There is a clock in the back of the room. Exam begins at 1:05pm and ends at 2:25pm. You have 80 minutes to earn 110 points. Good luck!

Section	Questions	Points	Score
True/False	5	15	
Short answer	15	60	
Practical problems	2	20	
Design problems	1	15	
Total	23	110	

True/False with justification (3 points each)

Read each statement carefully, circle the correct answer, and write a brief justification of your answer. No points will be given if you do not provide justification.

1. **True / False** In random forest classification, one way to maximize the diversity of trees is to select at each decision node a different subset of variables during decision tree construction.

2. **True / False** Increasing read length reduces the number of unresolved repetitive regions during genome assembly.

3. **True / False** In supervised learning, we should use all of the data available to train our model in order to generate the best model.

4. **True / False** Polynomial time affine gap alignment requires more than one dynamic programming matrix.

5. **True / False** Consider the problem of identifying transcription factor motifs in a set of DNA sequences. Gibbs sampling will increase the data likelihood monotonically during its execution.

Short answer (4 points each)

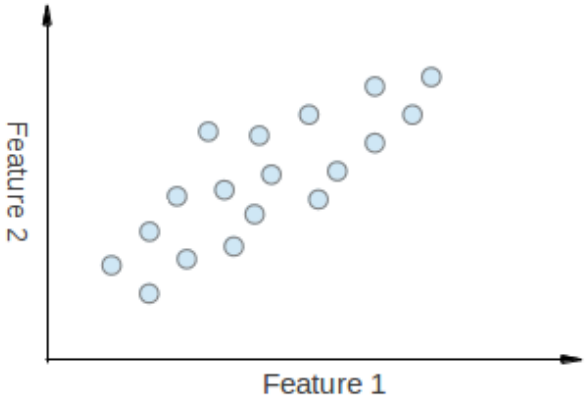
6. In the Needleman-Wunsch algorithm for global alignment, what does the $(i, j)^{th}$ entry in the matrix used for dynamic programming store once it's been filled?

7. Recall that in the BLAST algorithm, we split our query into W-mers and generate a neighborhood if these W-mers within a similarity threshold, T. Describe how the choice of W and T affects the sensitivity and specificity of the BLAST algorithm.

8. In the reverse operation of the Burrows-Wheeler Transform (BWT), when obtaining the original string from the compressed string, how do you compute the second column, when you have the first and the last?

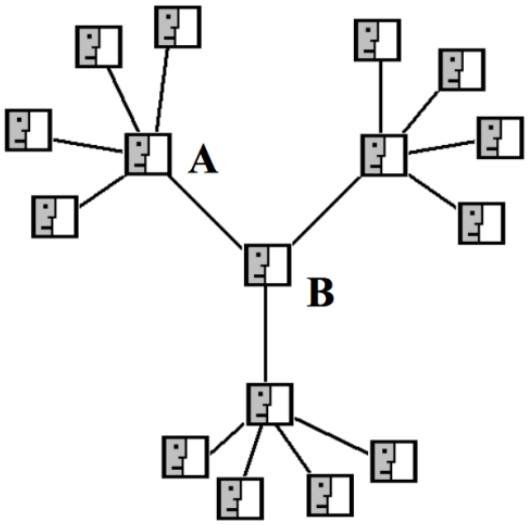
9. Describe the difference between the path returned by the Viterbi algorithm and that returned by posterior decoding. What is each method maximizing?

10. Recall Nussinov's algorithm for predicting RNA structure. Describe two ways in which Nussinov's algorithm does not accurately model biologically realistic folding.



11. Sketch the two eigenvectors that you would expect to obtain from PCA on the data in the plot above. Label the first principal component and the second principal component. Which would you expect to have the larger corresponding eigenvalue?

12. What is the source of the checkerboard/plaid pattern of genome-wide three-dimensional interactions patterns observed in high-throughput chromatin conformation capture (Hi-C) interactions?



13. For each of the following measures of network centrality, indicate whether node A or node B in the above figure is more central: (a) degree, (b) betweenness, and (c) closeness.

14. In eQTL discovery, what fundamental statistical problem do we overcome by only searching SNPs within a one-megabase window surrounding each target gene?

15. In GWAS, why is the identification of a disease-associated region insufficient to form a theurapeutic program in most cases? What are two missing pieces of information required to move forward once a disease-associated region is discovered?

16. Why are long reads (vs. short reads) and paired-end reads (vs. single-end reads) useful in determining alternatively-spliced isoforms from RNA-seq?

17. Describe how haplotype structure (for example, blocks of linkage disequilibrium extending up to the megabase range) can both help and hinder efforts to map disease genes with large scale association studies.

18. What is the 'missing heritability problem'? Describe two hypotheses put forward to explain this problem.

19. Recall that we can use evolutionary signatures to infer functions of the conserved regions. Describe two signatures which would characterize a protein coding region, versus two signatures which would characterize regions with important RNA structure.

20. Explain how recombination rate and time both contribute to the gradual decrease in linkage disequilibrium.

Practical Problems (20 points total)

21. You are given the following sequences and would like to determine the evolutionary relationship between them.

Human	A	A	C	T	C
Chimp	A	A	G	T	C
Gorilla	T	A	G	T	T
Zebrafish	C	C	T	C	C

(a) (2 points) Using a cost of 1 for a mismatch and 0 for a match, and assuming no gaps, create a distance matrix giving the pairwise distances between each sequence.

(b) (3 points) Is the matrix you constructed additive, ultrametric, both, or neither? Explain.

(c) (3 points) Construct a phylogenetic tree of these sequences using the UPGMA algorithm. You need not show every step.

(d) (2 points) What does your answer to (b) tell you about the correctness of the tree that you created in (c)?

22. (5 points) You are using Gibbs sampling to discover a 5-base motif in the following five sequences:

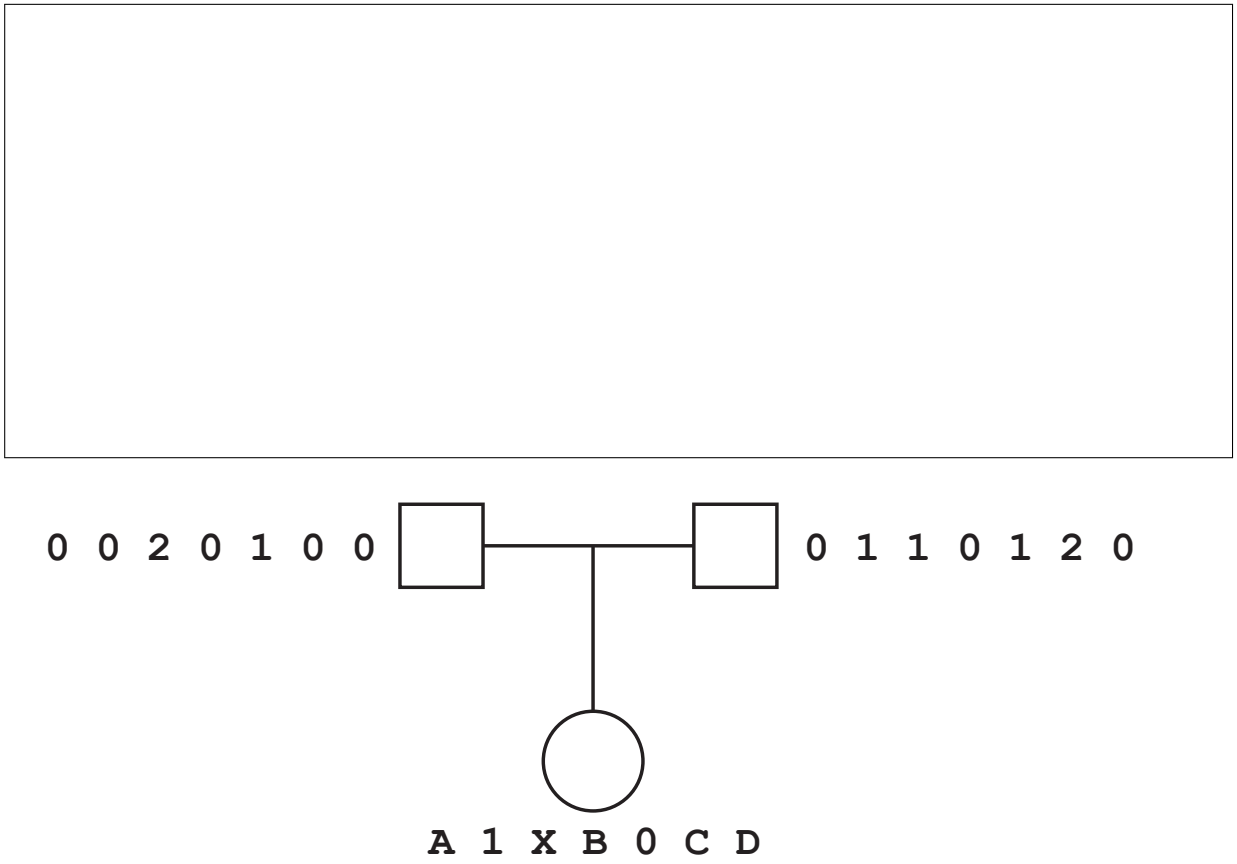
Sequence 1: TTTTGAGTAC
Sequence 2: GCAGAATTCT
Sequence 3: ATTATTCTCG
Sequence 4: CAGATTGTGG
Sequence 5: GTTTTTTCTA

After t iterations, the maximum-score starting positions for the motif are 2,6,4,4,5 respectively for the five sequences (as underlined).

On iteration $t+1$, the algorithm is going to estimate a new motif starting position for **sequence 2**. Calculate the motif model (in probability format) for this step. Include a pseudocount of 0.25 for all bases at all positions.

	1	2	3	4	5
A					
G					
C					
T					

23. (5 points) Given the following maternal and paternal genotypes and partial child genotype, impute the missing genotypes of the child's genotype at positions A, B, C, D. For position X, give the possible genotypes and their frequencies. Lastly, resolve the two haplotypes of the child, except for X.



Design problem (15 points)

24. You seek to build a Hidden Markov Model (HMM) to determine which genomic locations are bound by the CTCF regulator using multiple lines of evidence.

    (a) (4 points) Describe the architecture of your model. How many hidden states does your model have and what are they? What types of emission and transition probabilities? What do they represent?

    (b) (3 points) List at least three data types that you expect will be informative for your problem at hand. (**Hint:** Not all lines of evidence need to be cell type specific.)

You now seek to extend your model to handle the same lines of evidence across multiple cell types:

(c) (2 points) What are the advantages of training a different HMM for each cell type? How would you carry out the training?

(d) (2 points) What are the advantages of training a single HMM across multiple cell types? How would you carry out the training?

(e) (4 points) Describe a learning strategy that combines advantageous features of (c) and (d).