6.047 / 6.878 Computational Biology: Genomes, Networks, Evolution
Fall 2008

6.047/6.878 Fall 2008 Midterm Exam

October 21, 2008

Name:

No books, notes or electronic aids (such as calculators) are permitted. Both 6.047 and 6.878 have the same exam and scoring rubric, but they will be considered separately in determining final grades.

True/False and Multiple Choice (2 points each)

Read each statement or question carefully, and circle the correct answer.

1. **True / False** Baum-Welch is an iterative algorithm to learn the maximum likelihood parameters of an HMM from a labeled training set.
2. **True / False** In Bayes Formula, $P(B_i|A) = \frac{P(A|B_i)*P(B_i)}{\sum_j P(B_j)*P(A|B_j)}$, the B_js are mutually exclusive.
3. **True / False** In the Jukes Cantor model, a $C \rightarrow T$ transition is more likely than a $C \rightarrow G$ transversion.
4. **True / False** Kullback-Leibler is a true distance measure.
5. **True / False** One shortcoming of Generalized HMMs is that state durations are geometrically distributed.
6. **True / False** At every iteration, Gibbs sampling for motif finding chooses the most likely position for the start of the motif in each sequence.
7. **True / False** Manipulating the window size parameter W of a BLAST search will dramatically affect its running time.
8. **True / False** A long haplotype with low frequency is evidence of recent positive selection.
9. Which position in the following Position Frequency Matrix has the most information given a low-GC background?

	1	2	3	4
A	0.40	0.25	0.10	0.30
G	0.10	0.25	0.40	0.30
T	0.40	0.25	0.10	0.30
C	0.10	0.25	0.40	0.10

(a) Position 1

(b) Position 2

(c) Position 3

(d) Position 4

10. What is the runtime of sequence alignment with affine-gap cost between sequences of length N and M, where $N > M$?

    (a) $\Theta(NM)$

    (b) $\Theta(N^2M)$

    (c) $\Theta(N^2M^2)$

    (d) None of the above.

Short Answer (4 points each)

11. What is the minimum asymptotic amount of space needed to compute the score of a global alignment between sequences of lengths N and M, ignoring traceback? Describe briefly how this is done.

12. Describe the Naive Bayes Assumption as applied to classification, and how it can lead to double-counting evidence.

13. Is K-Means (in Clustering) more similar to Viterbi Learning or Baum-Welch Learning (in HMMs)? Justify.

14. Assume two populations have recently combined. The combined population has genotype frequencies AA 40%, Aa 40%, and aa 20%. If the assumptions of the Hardy-Weinberg model are true, how many generations are needed until the genotype frequencies reach equilibrium, and what are the equilibrium genotype frequencies?

15. In the BLOSUM62 matrix, a conserved Tryptophan position has score S(W,W)=11, but a conserved Leucine position has score S(L,L)=4. Explain at least one reason these values differ.

16. How many branches (edges) are in a rooted binary phylogenetic tree with n leaves?

17. Give an additive distance matrix (or its equivalent tree) of four genes for which UPGMA would produce the wrong tree topology.

18. If you apply the dN/dS computation on an intergenic region, do you expect it to be > 1, < 1, or $= 1$? Justify your answer.

Practical Problems (8 points each)

19. 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.

x	y	$P(c_1)$	$P(c_2)$
3	5	1.0	0.0
8	4	0.25	0.75
4	-4	0.75	0.25

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

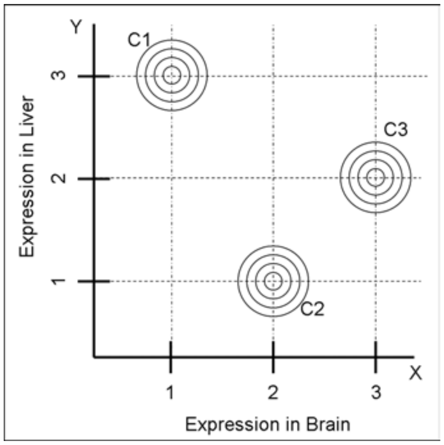
	x	y
μ_1		
μ_2		

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

	x	y
μ_1		
μ_2		

20. Derive an equation for the probability that an HMM was in hidden states l and k at times i and $i+1$ respectively, and emitted sequence $x_1 \dots x_n$, with remaining states unspecified; namely, the probability $P(x_1, \dots, x_n, \pi_i = l, \pi_{i+1} = k)$. You may define it in terms of the Forward, Backward, and Viterbi variables $F_k(i)$, $B_k(i)$, and $V_k(i)$, respectively.

21. Classification problems are frequently plagued by missing data, and measurements of some of the features are sometimes missing for some of the objects to be classified. In this problem, you are trying to classify the function of a gene in one of three classes C1, C2, C3, based on its expression value as measured in two tissues, Brain and Liver. The figure below shows the known distributions for genes in each class: each is normally distributed with unit variance, and the three classes have equal priors and different means, centered at (1,3), (2,1), and (3,2) respectively for C1, C2, and C3.



(a) Given a new gene A with expression measurements $[A_X, A_Y] = (2, 3)$, identify the most likely class for A (i.e. the one with the highest posterior probability $P(C_i|A_X, A_Y)$). Justify your answer.

(b) Assume now that we are given a new gene B, but that the expression level B_X for Brain is missing. We would like to estimate the expected value B_X, based on all other points currently observed. Assuming these are drawn from the distributions in the figure, what is the expected value for B_X, with no prior knowledge of B_Y? Given this value, and $B_Y = 1.9$, then what is the most likely class, i.e. maximizing $P(C_i|X = E[B_X], Y = B_Y)$. Justify your answer.

(c) Assume again that we are given gene B, with known Liver expression $B_Y = 1.9$ and missing Brain expression, as before. Instead of assuming a particular value for B_X, this time we will integrate (i.e. marginalize) over all values of B_X given $B_Y = 1.9$. Given this new approach, what is the most likely class for B? Justify your answer.

(d) Which class prediction is preferred, that of part (b) or part (c)? Justify your answer.

Design Problem (14 points)

22. Using dynamic programing, design a global alignment algorithm with linear gap penalty, which only allows insertions and deletions with lengths that are multiples of three. Give the initialization, recurrence, and termination for your dynamic programing table. Briefly describe the order in which entries in the table must be computed. Lastly, indicate for what input sequences an alignment would be undefined?