02-710/10-810/MSCBIO2070 Computational Genomics: Midterm Exam

March 23, 2011

Name	

Instructions:

- There are 6 questions in this exam (X pages including this cover sheet).
- Show your steps clearly.
- If you need more room to work out your answer to a question, use the back of the page and clearly mark on the front of the page if we are to look at what's on the back.
- This exam is open book and open notes. Calculators are allowed, but no computers, PDAs, or other communication devices.
- You have 1 hour and 30 minutes. Good luck!

No.	Topic	Max Score	Your score
1	Biological background (Short answers)	8	
2	Genome sequencing and assembly	9	
3	Sequence Alignment	18	
4	Molecular Evolution and Phylogenetics	21	
5	Motif Discovery and MEME	10	
6	Hypothesis Testing	12	
7	Normalization	12	
8	Clustering & Classification	10	
	Total	100	

1 Biological Background (8 points)

Short answers, 1-2 sentences

1. (3 + 3 points) So you always hear that the human genome is approximately 3 billion nucleotides long. Based on this, can you estimate the total genomic DNA bases in a single human skin cell, that is not undergoing cell division? Do you think the answer will highly vary (by high, we mean > 1 billion) depending on the cell type? Explain your answer using estimates for normal human liver, sperm, and egg cell types.

2. (2 points) Name four different types of mammalian RNAs and explain their functions with one sentence.

2 Genome sequencing and assembly (9 pts)

1. (5 points) In this question, we'll formalize finding a shortest superstring as a Hamiltonian path problem in an overlap graph. A hamiltonian path is a path that visits each node in a graph exactly once. Construct an overlap graph using the maximum overlap possible between pairs of sequences shown below, without allowing for mismatches. You don't have to consider the reverse complement. You don't have to show arcs with zero weights. Find one superstring defined by the Hamiltonian path on the graph that you constructed.

   a = TGCGAA
   b = CGATAA
   c = AACCTG
   d = CTGTTCGA

2. (1 points) Does a shortest superstring always define a Hamiltonian path in the overlap graph? Argue.

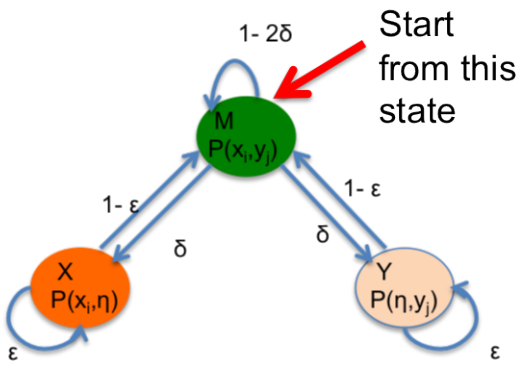
3. (3 points) Suppose your alphabet, L, consists of only two letters, L = {A,T}. Construct a shortest superstring for all strings of size 4, defined on this alphabet, L. Hint: the strings are as follows: AAAT, AATA, ATAA, TAAA, AATT, ATAT...

3 Sequence Alignment (18 pts)

1. **a.** Align the following sequences using Smith-Waterman algorithm. Show your alignment matrix and intermediate steps (3 points).
   Assume the following scoring scheme: match = +1, mismatch= -1, gap penalty: -1
   Sequence 1: GCGGT
   Sequence 2: CGG

   b. If you are going to use Needleman-Wunsch algorithm to align the following two sequences, how can you modify the matrix initialization and traceback in global alignment so as NOT to penalize for terminal gaps at the BEGINNING and END on the Sequence-2 ? Assume the following scoring scheme: match = +1, mismatch= -1, gap penalty: -1
   In this problem, you don't need to show the full matrix, just explain which row/column will be affected and how they will be affected. Also explain if and how the traceback needs to be modified. (3 points)
   Sequence 1: GCGGT
   Sequence 2: CGG

2. (3 points) Figure below illustrates an HMM that can be used as a model for global alignment. This HMM has 3 states: State M, emits two aligned characters from sequences x and y; state X, emits one character from sequence x aligned with gap, and state Y emits one character from sequence y aligned with gap. Transition and emission probabilities are depicted on the figure. Consider the following alignment:
   Sequence 1: HEVPDK- E
   Sequence 2: VE - - DASE
   Starting from state M, write down the probability of obtaining this alignment.



3. (6 points) Now suppose you are given Sequence 1 and Sequence 2 and your goal is to obtain the above alignment, We can solve this using a Viterbi algorithm, but unlike the Viterbi algorithm we implemented for the homework 2, for this question, we'll have to account for the additional sequence. In this question we'll write the modified Viterbi algorithm, by filling the blanks below. To get you started, we have provided the initial steps:

   $\nu_M(0,0) = 1$; $\nu_M(\text{i},0) = 0$; $\nu_M(0,\text{i}) = 0$;
   $\nu_X(0,0) = 0$; $\nu_X(\text{i},0) = 0$; $\nu_X(0,\text{j}) = 0$;
   $\nu_Y(0,0) = 0$; $\nu_Y(\text{i},0) = 0$; $\nu_Y(0,\text{j}) = 0$;

   for i = 1...m, j = 1...n
   $\nu_M(i,j) =$
   $\nu_X(i,j) =$
   $\nu_Y(i,j) =$
   return max[ν_M(m,n),ν_X(m,n),ν_Y(m,n)];

4. (3 points) In the above model can you explain why the transition between X and Y are not needed?

4 Molecular Evolution and Phylogenetics (21 points)

1. **Synonymous vs Non-synonymous mutations (3 points)**
   What is a synonymous mutation and non synonymous mutation? Here are two aligned protein coding DNA sequences (codons are separated by hyphens). Suppose the second sequence is a result of cumulative mutations from the first one. Based on the sequence below, explain whether mutation/substitution is neutral, advantageous or deleterious. (3 points)
   DNA1 = CAT – – ACA – – GAG – – AAG – – GGG – – GTC – – TAT
   DNA2 = CAC – – ACT – – GAC – – ACA – – GGG – – ATC – – TAC

   Here is a codon table that you may find useful:

1st letter	Second Letter: U	Second Letter: C	Second Letter: A	Second Letter: G	3rd letter
U	UUU Phe, UUC Phe, UUA Leu, UUG Leu	UCU Ser, UCC Ser, UCA Ser, UCG Ser	UAU Tyr, UAC Tyr, UAA Stop, UAG Stop	UGU Cys, UGC Cys, UGA Stop, UGG Trp	U C A G
C	CUU Leu, CUC Leu, CUA Leu, CUG Leu	CCU Pro, CCC Pro, CCA Pro, CCG Pro	CAU His, CAC His, CAA Gln, CAG Gln	CGU Arg, CGC Arg, CGA Arg, CGG Arg	U C A G
A	AUU Ile, AUC Ile, AUA Ile, AUG Met	ACU Thr, ACC Thr, ACA Thr, ACG Thr	AAU Asn, AAC Asn, AAA Lys, AAG Lys	AGU Ser, AGC Ser, AGA Arg, AGG Arg	U C A G
G	GUU Val, GUC Val, GUA Val, GUG Val	GCU Ala, GCC Ala, GCA Ala, GCG Ala	GAU Asp, GAC Asp, GAA Glu, GAG Glu	GGU Gly, GGC Gly, GGA Gly, GGG Gly	U C A G

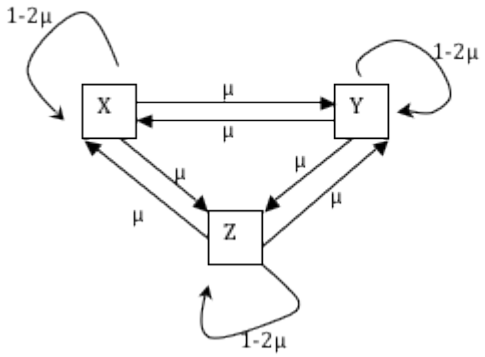
2. **Tree Construction (5 points)**
   In HW2, you have constructed a phylogenetic tree for using UPGMA. Now let's consider using Neighbor Joining:

	Homo sapiens	Homo neanderthalensis	Pan troglodytes
Homo sapiens	0	1	5
Homo neanderthalensis	1	0	6
Pan troglodytes	5	6	0

   2.1. According to NJ algorithm , which two species will you consider to join together first? Show calculations (3 points).

2.2. Show an example of a distance matrix with three species (A, B and C) that will lead to the same unrooted tree no matter which method (UPGMA/NJ) is used (2 points)

3. (15 points) Your colleagues in NASA′s wet lab discovered a group of new forms of life on the moon Pandora. These organisms have simple genetic systems which are based on only three nucleotides X, Y and Z. They managed to obtain sequences of two homologous genes from two such organisms, which are believed to have a common ancestor. They ask you to derive a correction for the observed distance between these two genes. You hypothesize that the substitution rates from any nucleotide to another are the same, and you denote it by μ. You draw the following diagram to understand the transitions between X, Y and Z:



a. Write down an expression for K(t), the expected number of substitutions per site between the two sequences, in terms of μ and t (2 points).

You managed to derive the following probability transition matrix P(t) for this model:

$$P(t) = \begin{bmatrix} r(t) & s(t) & s(t) \\ s(t) & r(t) & s(t) \\ s(t) & s(t) & r(t) \end{bmatrix}$$

in which:

$r(t) = \frac{1}{3}(1 + 2e^{-3\mu t})$
$s(t) = \frac{1}{3}(1 - e^{-3\mu t})$

b. Does this model have a stationary distribution? If so, find it. If not, explain why (1 point).

c. Consider one site in the two aligned sequences. What is the probability that the two nucleotides at this site differ at time t? (Remember that they shared a common ancestor at t = 0.) Express it in terms of s(t) and r(t). (4 points)

d. Denoting the probability you calculated above by U and the length of the two sequences by L, derive an expression for the corrected distance between the two sequences in terms of U and L. (6 points)

e. How would you estimate the substitution rate μ, given that the sequences are separated from their common ancestor t years ago? (2 points)

5 Motif Discovery and MEME (10 points)

1. (2 points)Your friend made a PSSM based on log-likelihood ratios from fruitfly for the motif that resembles ATGCT, and asked to find similar motifs in a genome that only you have access to (let's just say you sequenced the legendary Bigfoot genome). Do you think it is OK to use this PSSM for finding motifs in the Bigfoot genome? Explain your answer.

2. Suppose at E-step, we have:

   The parameter : $\lambda = 0.4$, $\theta = \{\theta_1, \theta_2\}$, where $\theta_1 =$

A:	0.2	0.3	0.1	0.1	0.6
C:	0.5	0.1	0.1	0.1	0.1
G:	0.1	0.1	0.7	0.1	0.1
T:	0.1	0.5	0.1	0.7	0.2

   and θ_2 = [1/4 1/4 1/4 1/4] for A,C,G,T.

   Now suppose a string K = "CTATA"
   (3 points) Calculate E(Z_{K1}) , which E [P(string K is a motif)]

   (3 points) Calculate E(Z_{K2}) , which E [P(string K is from background)]

3. What are the possible pitfalls and disadvantages of using EM (expectation-maximization) for motif discovery? In other words, what kind of situation MEME can not handle well? Pick two to explain (2 points)

6 Hypothesis Testing (12 pts)

1. Assume we are studying cancer vs. healthy individuals using microarrays. We have 20 cancer patients and 10 healthy individuals in our sample set. We are interested in performing randomization tests to determine if the differences in mean values we observe for each gene are significant. Write a formula for the number of randomization tests that can be performed for this data (no need to provide the actual number, only a formula that when evaluated would give this number).

2. For another cancer vs. healthy experiment assume that we have performed 500 permutation tests for each gene in our study. We use these to compute p-values for every gene we test. Out of the 1000 genes we tested, we identified 100 as significant. What is the minimum FDR for this set?

3. We tested n genes and identified 40 with a p-value of 0.005. We are told that the FDR for this set is 20%. What is n?
   a. 8000 b. 4400 c. 1600 d. 800 e. impossible to tell

4. We tested n genes using a Bonferroni corrected p-value of 0.001 and identified 50 as differentially expressed. What is uncorrected p-value we started with?
   a. 0.05 b. 0.01 c. 0.005 d. 0.001 e. impossible to tell

7 Normalization (12 pts)

For each of the following experiments select the answer(s) that would provide a good normalization for the experiment described. You can select multiple answers for each question but you will be penalized for every wrong answer you selected.

1. Over crowding: We are performing experiments in which we compare wild type e. coli. to e. coli's that are overcrowded (that is, in which we are trying to express as many genes as possible to their highest levels).
   a. Quintile b. Same mean / variance c. dChip d. non of the answers is appropriate for this experiment.

2. We are comparing experiments from two different types of arrays where one set has twice as many genes printed on it as the other (randomly selected).
   a. Quintile b. Same mean / variance c. dChip d. non of the answers is appropriate for this experiment.

3. We are in a hurry and would like to use the fastest method described in class
   a. Quintile b. Same mean / variance c. dChip d. non of the answers is appropriate for this experiment.

4. We are comparing samples from different tissues of the same individual.
   a. Quintile b. Same mean / variance c. dChip d. non of the answers is appropriate for this experiment.

8 Clustering & Classification (10 pts)

In gene clustering, our goal is to identify groups of co-expressed genes by mining gene-expression data.

1. (5 points) In the Figure below, you see the cluster assignments using three different cluster similarity measures. Circle the correct cluster similarity measure/measures corresponding to each figure:
   Figure 8.1.a: A. Single-Link B. Complete-Link C. Average-Link
   Figure 8.1.b: A. Single-Link B. Complete-Link C. Average-Link
   Figure 8.1.c: A. Single-Link B. Complete-Link C. Average-Link

Figure 8.1.a

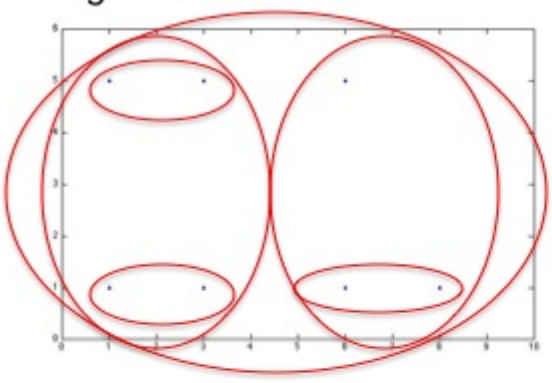


Figure 8.1.b

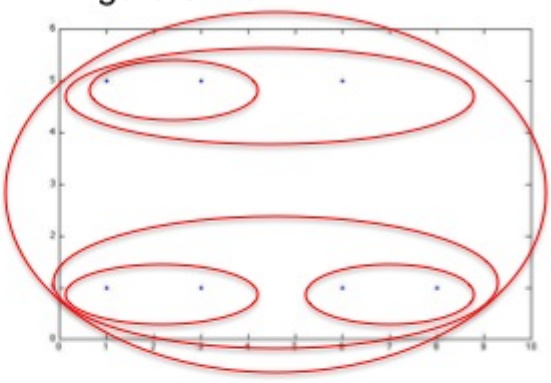
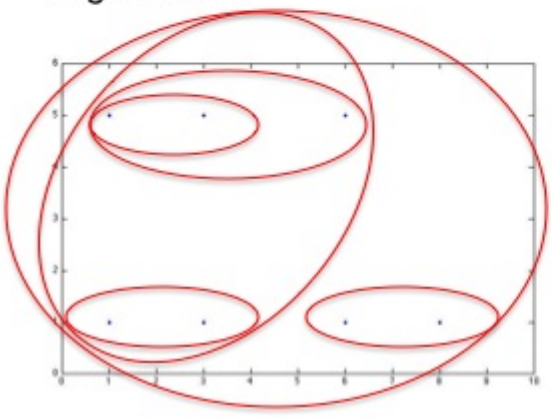


Figure 8.1.c



2. (5 points) Suppose we are using Gaussian Naive Bayes to classify genes into one of two classes: cancer and healthy, based on their expression values in a stress response experiment. Assume we trained a Gaussian Naive Bayes classifier using two different

datasets of known cancer and healthy genes. Dataset one contains C1 cancer genes and H1 healthy genes and dataset two contains C2 cancer genes and H2 healthy genes. Surprisingly, we noticed that the classification models obtained for from these two sets (M1 and M2 respectfully) had exactly the same parameters for mean and variance in the two classes. In other words both had the same mean for healthy genes and the same variances for healthy genes and also the same mean and variances for cancer genes (of course, the means and variances differed between healthy and cancer genes in both models). We next obtained expression values for a new gene G that was not included in either of the training datasets. When classifying G using M1 we determined that it was a cancer gene. However, when we used M2 we determined that it is a healthy gene. Which of the following answers is correct (you may circle more than one, but would be penalized for any wrong circles).

a. C1 > C2 b. H2 > H1 c. C1+ H1> C2+H2 d. C1/ C2 > H1/ H2 e. None of these has to be correct