

10-810/MSCBIO 2070: Computational Genomics, Spring 2008

Midterm Questions and Solutions

- There are 7 questions in this exam (10 pages including this cover sheet).
- Questions are not equally difficult.
- 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. Computers, PDAs, cell phones are not allowed.
- You have 1 hour and 20 minutes. Good luck!

| |
|--------|
| Name: |
| Email: |

| Question | Topic | Max Score | Score |
|----------|----------------------------------|-----------|-------|
| 1 | HMM | 20 | |
| 2 | Sequence Alignment/Tree Building | 15 | |
| 3 | Genome Assembly | 15 | |
| 4 | DNA signals/Gene finding | 10 | |
| 5 | Normalization Methods | 15 | |
| 6 | Multiple Hypotheses Testing | 15 | |
| 7 | Classification | 10 | |
| | Total | 100 | |

1. Introduction to Statistical Learning - HMM (20 pts)

1.1 (6 pts) Here are the names of some algorithms introduced in class. For each problem, select the corresponding algorithm.

- (A) Forward algorithm
- (B) Backward algorithm
- (C) Viterbi algorithm
- (D) Baum-Welch algorithm

Problem 1: Given the observation sequence and a HMM, efficiently compute the probability of the observation sequence.

Ans.: A,B

Problem 2: Given the observation sequence and a HMM, choose a corresponding state sequence which best explains the observation.

Ans.: C

Problem 3: Adjust model parameters to maximize the probability of observation given a HMM

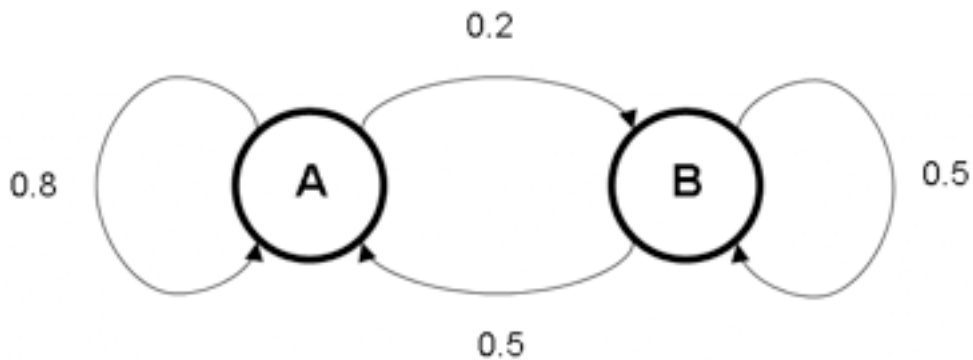
Ans.: D

1.2 (4 pts) Which of the following problem(s) could be done with HMM?

- (A) CpG island identification
- (B) Protein function prediction
- (C) Pairwise alignment of DNA sequences
- (D) Binding site prediction

Ans.: A,C,D (B is not on sequence level, in another word, it's not a segmentation problem)

1.3 (10 pts) In the following HMM model, each state ('A' or 'B') could generate either 'S' or 'T' sequence.



The emission probability table is:

| | S | T |
|---|-----|-----|
| A | 0.6 | 0.4 |
| B | 0.7 | 0.3 |

The initial probability table is:

| | Initial prob |
|---|--------------|
| A | 1 |
| B | 0 |

- 1) What is the probability that the emission sequence generated by this HMM is 'TSS'?

Ans.: Using forward algorithm: $a_1(A)$ means the probability of having state 'A' at the first position; $t(AA)$ is the transition probability from 'A' to 'A'; $b(S|A)$ is the emission probability of observing 'S' given state 'A'.

$$a_1(A) = b(T|A) = 0.4$$

$$a_1(B) = 0$$

$$a_2(A) = b(S|A) \cdot a_1(A) \cdot t(AA) = 0.1920$$

$$a_2(B) = b(S|B) \cdot a_1(B) \cdot t(AB) = 0.0560$$

$$a_3(A) = b(S|A) \cdot (a_2(A) \cdot t(AA) + a_2(B) \cdot t(BA)) = 0.1090$$

$$a_3(B) = b(S|B) \cdot (a_2(A) \cdot t(AB) + a_2(B) \cdot t(BB)) = 0.0465$$

$$P(TSS) = a_3(A) + a_3(B) \approx 0.155$$

or

$$\begin{aligned}
 P(TSS) &= P(TSSIAAA) + P(TSSIAAB) + P(TSSIABA) + P(TSSIABB) = \\
 &= 0.4 \cdot (0.8 \cdot 0.6 \cdot 0.8 \cdot 0.6 + 0.8 \cdot 0.6 \cdot 0.2 \cdot 0.7 + 0.2 \cdot 0.7 \cdot 0.5 \cdot 0.6 + 0.2 \cdot 0.7 \cdot 0.5 \cdot 0.7) \\
 &= 0.1554
 \end{aligned}$$

2) What is the most probable state that generates the first 'T'?

Ans.: 'A'

Because the initial probability of 'A' is '1', and $t(AA) > t(AB)$, the HMM would like to stay at 'A'. (Enough for full credits)

If you use Viterbi algorithm, for a observed sequence 'SS.....SST' (with N 'S'), the state that gave 'T' is 'A'.

2. Sequence alignment & tree building (15 pts)

2.1 (8 pts) Locally align the sequences ATGAAA and TGCCAA. Use (match) $M=2$, (mismatch) $m=-1$, and an affine gap scoring function where the first gap in a sequence scores -3 and contiguous extensions score -1. That is, a gap of length k will score, in sum, $g=-2-k$.

Provide the dynamic programming matrix, and write all optimal alignments.

| A | T | G | A | A | A | |
|---|---|---|---|---|----|----|
| T | 0 | 2 | 0 | 0 | 0 | 0 |
| G | 0 | 0 | 4 | 1 | 0 | 0 |
| C | 0 | 0 | 1 | 3 | 0 | 0 |
| C | 0 | 0 | 0 | 0 | 2 | 0 |
| A | 2 | 0 | 0 | 2 | 2 | 4c |
| A | 2 | 1 | 0 | 2 | 4a | 4b |

Three alignments of score 4:

Ans.:

(1) atgAAa
 tgccAA

(2) aTGAAA
 TGCCA

(3) atgaAA
 tgccAA

2.2 (7 pts) Reconstruct the tree using UPGMA from the following distance matrix. Label the distances of internal edges. (More writing space in next page)

| | B | C | D | E |
|---|---|----|----|----|
| A | 2 | 10 | 10 | 10 |
| B | | 10 | 10 | 10 |
| C | | | 4 | 4 |
| D | | | | 2 |

Ans.: In newick format:

((A:1,B:1):4,(C:2,(D:1,E:1):1):3)

3. Genome assembly (15 pts)

3.1 (7 pts) You are tasked with assembling a genome of size $G=1 \times 10^8$ from a set of lambda clones. If 150000 clones each of average length 20kb are provided, what is the expected fraction of bases in the genome covered by at least one clone fragment?

Ans.: $P(\text{covered by } \geq 1) = 1 - P(\text{not covered by } n)$
 $= 1 - (1 - L/G)^n = 1 - 10^{-14}$

3.2 (8 pts) Assume that you have a genome of size $G=3 \times 10^9$ bp covered by a library of plasmid clones of average length $L=2$ kb, of which we sequence $n=500$ bases on both ends. What should the minimum detectable overlap between clones be so that by sequencing $N=10,000,000$ of the clones we get $E=45$ expected number of islands?

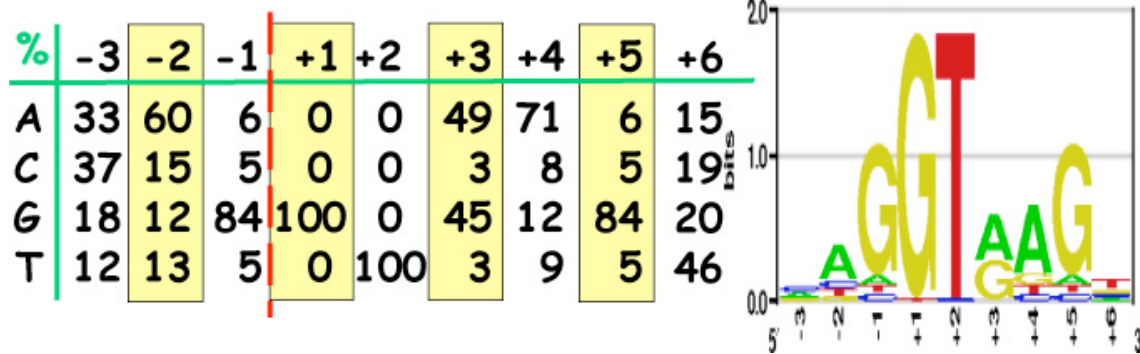
Note: There was a typo in the average read length above, so full credit was given to those that worked a solution on the basis of the Lander-Waterman statistics and ½ credit to those who didn't.

4. DNA signals / gene finding (10 pts)

4.1 (5 pts) What is the most important assumption of the PSSM DNA-binding models? How well does it hold for real proteins? How could someone model DNA-binding sites if this assumption brakes?

Ans.: From the notes in the class: the most important assumption is the “additivity assumption” which does not hold overall, but it was found to be pretty good for practical purposes in the cases presented in the class (C2H2 zinc finger proteins and Mnt. If additivity brakes, then one can model the sites with higher order models (e.g., Markov models for consecutive nucleotide biases, or other base-dependence models for long-term biases). Although it was only casually mentioned in the class, one can think alternative ways for modeling problematic data, such as suffix trees or dictionary-based approaches on known sequences.

4.2 (5 pts) Describe briefly the MDD method that *Genscan* uses to model splice sites. Assuming that the motif represented by the LOGO and the weight matrix below are derived from 1,000 sequences (splice sites), draw the flowchart of the MDD partitions and provide the expected number of sequences in each step.



Ans.: For the MDD description, see class notes. The difference between this model and the one presented in the class is that here positions -1 and +5 have the same frequencies, making difficult to split. A simple solution is to choose one of the two positions (randomly) to model first with the MDD approach, followed by the other position. The number of sequences remaining in each step can be calculated from the percentages of the most dominant base in the corresponding position.

5. Normalization methods (15 pts)

In the following question use these letters to denote the three normalization methods we discussed in class by:

- A – Mean and variance normalization
- B – Quintile (ranking) normalization
- C – Invariant genes normalization

Below we present three alternative normalization methods. For each of these additional methods determine which of the methods learned in class make stronger or weaker *assumptions*. If there the assumptions that cannot be directly compared chose ‘cannot be compared’. Note that some categories may be empty for some of the new methods.

5.1 (5 pts) Normalizing by assigning genes in each array to the same predefined Gaussian distribution.

Methods that make stronger assumptions than this: [None](#)

Methods that make weaker assumptions than this: [A,B](#)

Methods that cannot be compared: [C](#)

When normalizing using a predefined distribution for all arrays we arrive at arrays with the same values (as in B) and the same mean and variance (as in A). In addition, unlike B we do not use any values from the measurements for the global values (just the ranking) and rely on the predefined distribution for the values we end up with. The assumptions made by C are not directly comparable (though we accepted answers that said that C had weaker assumptions as well).

5.2(5 pts) Normalizing by only using the spike in controls (which are added to each of the samples and contains genes from a different species).

Methods that make stronger assumptions than this: [A,B,C](#)

Methods that make weaker assumptions than this: [None](#)

Methods that cannot be compared: [None](#)

Spike controls require very little assumptions regarding the actual similarities between the global or partial values and ranking of the real experimental samples. Unlike A, B and C when using spike controls we rely on the procedure used to carry out the experiments which is controlled by the experimentalist.

5.3 (5 pts) Normalizing by equating the mean, variance and third moment of the distribution.

Methods that make stronger assumptions than this: [B](#)

Methods that make weaker assumptions than this: [A](#)

Methods that cannot be compared: [C](#)

For B, since all values are the same all moments are equal as well so B is stronger. A is weaker since it only equates the mean and variance. Again, C cannot be directly compared since C looks at specific gene rankings whereas this assumptions looks at global mRNA quantities.

6. Multiple hypotheses testing (15 pts)

We carried out microarray experiments to compare cancer and healthy cells. For each of the 2000 genes we tested we computed the log likelihood ratio score and used the chi-square distribution to assign p-values for these scores. The table below contains the scores and their corresponding p-values.

Since we tested 2000 genes, we can use the Bonferroni correction to correct for multiple hypothesis testing. Another way is to use randomization tests. We have carried out 1000 such tests (by randomizing the labels of the experiments). In the table below we also present the number of randomization experiments containing at least one gene with a score above each of the three scores:

| | | | |
|---|------|------|-------|
| score | 10 | 20 | 30 |
| p-value | 0.05 | 0.01 | 0.001 |
| # of times we found a gene with a lower p-value | 300 | 100 | 40 |

6.1 (3 pts) What is the corrected p-value required for an original p-value of 0.1 according to the Bonferroni correction?

Ans.: 0.1/2000

6.2 (3 pts) What is the corrected p-value required for an original p-value of 0.1 according to the randomization correction method?

Ans.: 0.01. As the table above indicates, there were 100 out of 1000 randomization runs that led to a value of 20 or higher for at least one gene. Thus, for a p-value of 0.1 we need the uncorrected p-value that corresponds to a score of 20, which is 0.01.

6.3 (3 pts) Is there a method (Bonferroni or randomization) that is stricter (leads to lower corrected p-values) for *all* three randomization corrected p-values that can be derived from the table above?

Ans.: Yes. In all cases the randomization correction leads to higher (less strict) p-values.

6.4 (3 pts) Assume we have identified 200 genes with a p-value < 0.01 . What is the false discovery rate (FDR)?

Ans.: Since we started with 2000 genes we would expect to see 20 genes with this p-value. Thus, the FDR is $20 / 200 = 10\%$.

6.5 (3 pts) What is the FDR if we identify 10 genes with the *Bonferroni corrected* p-value for the original p-value of 0.1 (the p-value in your answer to a)?

Ans.: The Bonferroni corrected p-value for 0.1 is $0.1 / 2000$. Thus, the expected number of genes at this p-value is 0.1 ($2000 * (0.1 / 2000)$). Since we identified 10 genes the FDR is 1%.

7. Classification (10 pts)

7.1 (6 pts) Suppose you have the following training set with three genes A, B, C taking Boolean values of 1 for upregulated and 0 for downregulated. The labels are 0 for normal and 1 for cancer.

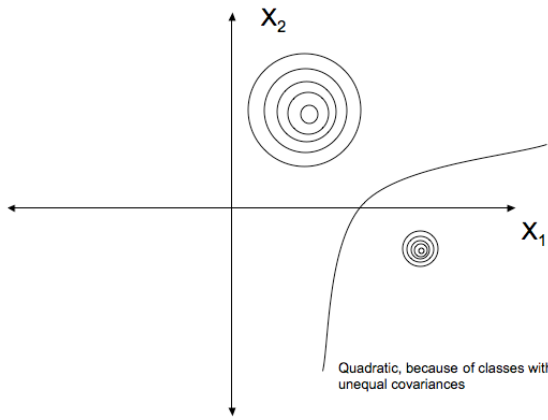
| A | B | C | Class |
|---|---|---|-------|
| 1 | 0 | 0 | 0 |
| 0 | 1 | 1 | 0 |
| 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 |
| 0 | 0 | 1 | 1 |
| 0 | 1 | 0 | 1 |
| 1 | 1 | 0 | 1 |

We would like to predict cancer vs normal using a Naïve Bayes classifier. After learning is complete what would be the predicted probability $P(\text{Class} = \text{normal} \mid A = 0, B = 1, C = 0)$?

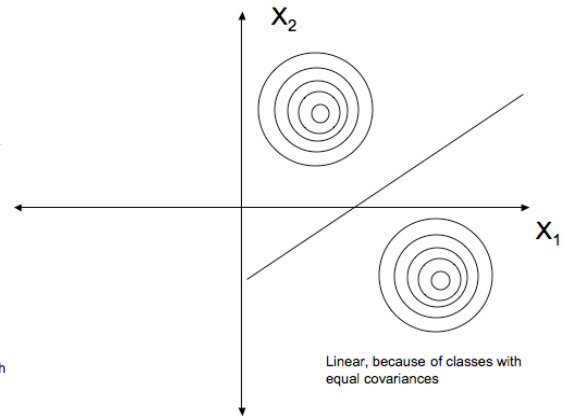
Ans.: $P(\text{class} = 0 = \text{normal} \mid A = 0, B = 1, C = 0)$

$$\begin{aligned}
 &= \frac{P(\text{class} = 0)P(A = 0 \mid \text{class} = 0)P(B = 1 \mid \text{class} = 0)P(C = 0 \mid \text{class} = 0)}{P(A = 0, B = 1, C = 0)} \\
 &= \frac{P(\text{class} = 0)P(A = 0 \mid \text{class} = 0)P(B = 1 \mid \text{class} = 0)P(C = 0 \mid \text{class} = 0)}{P(\text{class} = 0)P(A = 0, B = 1, C = 0 \mid \text{class} = 0) + P(\text{class} = 1)P(A = 0, B = 1, C = 0 \mid \text{class} = 1)} \\
 &= \frac{8}{35} \\
 &= 0.229
 \end{aligned}$$

7.2 (4 pts) On plotting data points in feature space of gene X1 and gene X2 shows two Gaussian clusters with contours shown below. Assuming a Bayes classifier, draw the decision boundary for the two examples and label what kind of boundary you get.



(A)



(B)