CMU 10-810 / MSCBIO2070: Computational Genomics, Spring 2007

Midterm ANSWERS

Student Name: *Date: April 5th, 2007*

Advice: Do not spend too much time on a single question! If you cannot solve a question, go to the next one and you can return later, if you have time.

1. (15 points) Suppose that the number of crossovers in a chromosome region *i* follows a Poison distribution:

$$P(k) = e^{-d_i} \frac{d_i^k}{k!}$$

where d_i is the average number of crossovers in region *i*.

(A) (7 points) Assume that crossover is a memoryless point process (i.e., the distributions of crossover numbers in different regions are independent), what is the distribution of the number of crossovers in conjoint region that spans region *i* and *i* + 1 (Hint: You don't have to show derivation, if you know the answer from basic property of Poison distribution. Just write the solution directly. You may want to answer question 1(B) first before answering this one)?

$$P(k_{(i,i+1)}) = e^{-(d_i + d_{i+1})} \frac{(d_i + d_{i+1})^k}{k!}$$

(B) (8 points) What is the average number of crossovers in this region?
(Hint: you may want to answer this first before answering the question in (A)), why this quantity is good for measuring the genetic distance between two loci on a chromosome?

$$E(k_{(i,i+1)}) = d_i + d_{i+1}$$

This quantity is good for measuring the genetic distance between two loci on a chromosome because it is additive, that is, for consecutive loci a, b, c, the average number of crossovers between loci a and c equals to the sum of average crossovers between a and b and between b and c.

2. **Additivity or nonadditivity in QTL analysis**. (12 points) In the graphs below we plot the average phenotype strength of two QTLs measure in F2 resulted from an intercross experiment.

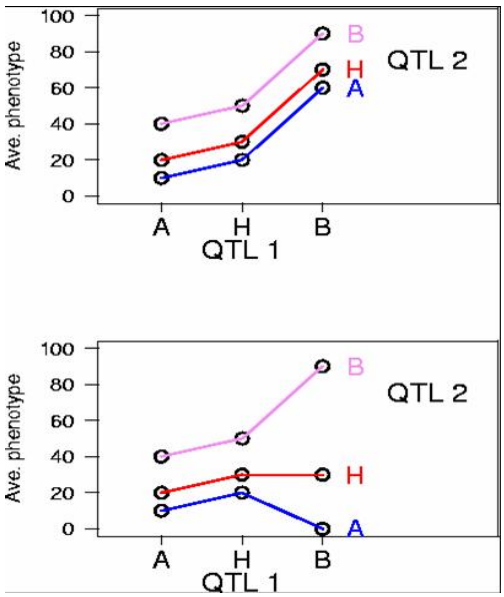
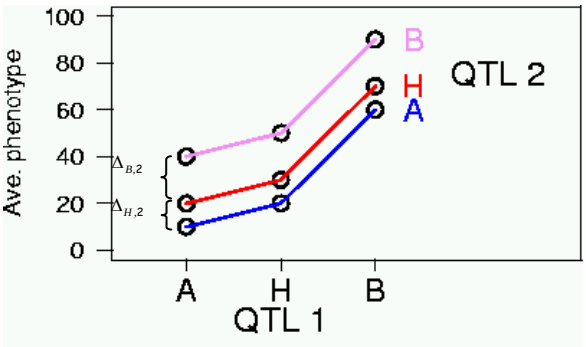


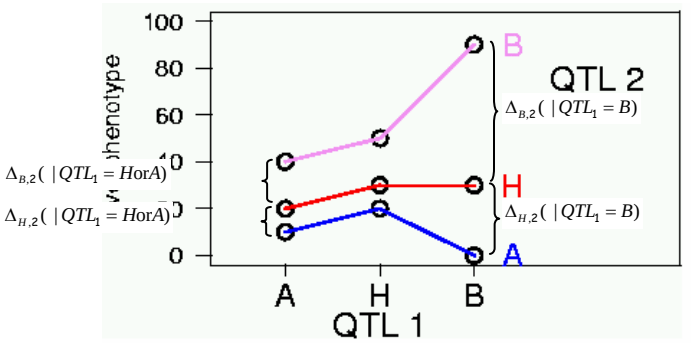
Figure 1. Additivity or non-additivity?

(A) (4 points) What panel in the graph shows an additive phenotype with respect to the 3 states of the 2 QTLs? Which one corresponds to an epistatic phenotype?

A is additive and B is epistatic.

(B) (4 points) What is contribution of the H-type of QTL 2, $\Delta_{H,2}$, on the graph in additive and epistatic cases? What is $\Delta_{B,2}$. Draw both cases on the graph.





Note that here we use H as the reference point of B. The reference point of B can be also be set at A.

(C) (4 points) Give a very brief explanation (e.g., one sentence or one equation) of why epistasis arises.

Because the contributions of one QTL to the phenotype level depend on the states of the other QTL. For example, in figure b above, the magnitude of $\Delta_{H,2}$ depends on the genotype of QTL1.

3. (8 points) Haplotype inference. Let 1 denote genotype 0/0, 1 denote genotype 1/1, and 2 denote genotype 0/1. What is the most parsimonious haplotypes corresponding to the following three genotypes?

```
2 1 2 2 2 2 0 2
0 2 1 1 1 1 2 2
1 1 0 0 0 0 0 1
```

01 1 111 00
11 0 000 01

01 1 111 00
00 1 111 11

11 0 000 01
11 0 000 01

4. Molecular Evolution. (10 points)

(A) (5 points) Why it is necessary to adjust observed percent differences on aligned sequence sites?

Because the observed percentage difference in aligned sequence does not take into consideration multiple back-substitutions that would result in no observable sequence change. Thus the observed percentage difference does not reflect the actual number of substitutions separating the two aligned sequences, and is thus non-linear to the divergence time under a time-homogeneous substitution model.

(B) (5 points) In a time-reversible substitution model, the probability of evolving from the ancestor to the current species equals to the probability of evolving back from the current species to the ancestor. The bellow tree is build based on some molecular substitution model from a pairwise sequence alignment of orang and human sequences. Can you tell whether it is from a reversible model or an irrigable model? Give an example of your guessed model? What do you expect see if using the improbable model?

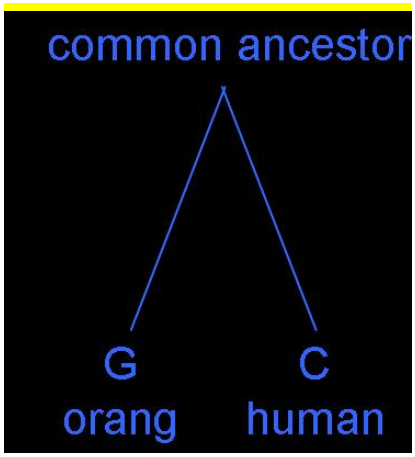


Figure 2. Tree.

It is from a reversible model, such as a Jukes-Canter model. If the model is irreversible, then the distance from the common ancestor to the two leaves can be different because the distances from orang to human and from human to orang can be different, which causes asymmetry.

5. Sequence analysis (10 points).

(A) (2 points) Find the best ungapped alignment of the following two sequences using matrix A BLAST matrix):

ATTCCG
ATCC

Matrix A:

```
   A  T  C  G
A  5 -4 -4 -4
T -4  5 -4 -4
C -4 -4  5 -4
G -4 -4 -4  5
```

Ans.:

```
ATTCCG
|| |          Score: 11
ATCC
```

or

```
ATTCCG
  |||          Score: 11
 ATCC
```

(B) (2 points) Find the best ungapped alignment using matrix B. Unlike with matrix A, in matrix B, high scores denote "bad" alignment.

Matrix B:

```
   A  T  C  G
A  0  5  5  1
T  5  0  1  5
C  5  1  0  5
G  1  5  5  0
```

Ans.:

```
ATTCCG
|| |          Score: 1
ATCC
```

(C) (3 points) Find the best gapped alignment for both matrix A and B, where gap initiation/extension penalties are set to -6 / -2 (matrix A) and 4 / 2 (matrix B), respectively.

Ans.:

```
Matrix A:
ATTCCG
|| ||          Score: 14
AT-CC
```

or

```
ATTCCG
| |||          Score: 14
A-TCC
```

```
Matrix B:
ATTCCG
|| |           Score: 1
ATCC
```

or

```
ATTCCG
|| ||          Score: 4
AT-CC
```

or

```
ATTCCG
| |||          Score: 4
A-TCC
```

(D) (3 points) Assuming that the %GC is 0.5 in the (background) genome, calculate the probability that you will get three or more nucleotides aligned *by chance* given the above sequence lengths.

Ans.:

Given equiprobable background distribution, the probability of matching k nucleotides out of N by chance is binomial with p=0.25. The tail probability of $k \geq 3$ is :

$$P(X \geq 3) = \sum_{k=3}^{4} \binom{4}{k} \cdot 0.25^{k} \cdot 0.75^{(4-k)} = 0.051$$

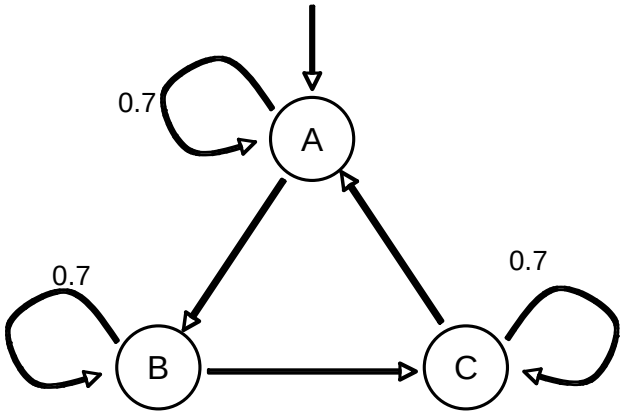
6. HMMs (12 points).

(A) (3 points) What are the differences between the Viterbi and the Baum-Welch algorithms?

Ans.:

Viterbi is a dynamic programming algorithm (similar to sequence alignment algorithms) where the most probable path of an HMM is calculated from the data (i.e., model parameters and emitted data). Baum-Welch is an iterative EM algorithm for calculating the maximum likelihood parameters of an HMM from the data.

(B) (9 points) In the following HMM, each state can generate either *exonic* (E) or *intronic* (I) sequence but not both.



Suppose we observe the following sequence of events: IEEEI.

i. What is the most probable state that generated the third "E"?

Ans.:

State A emits « I », hence state B emits « E ». The question is what state C emits.

a. P (Y_4 = B) = 1.0 * 0.3 * 0.7^2 * 0.3 = 0.0441

b. $P(Y_4 = C) = 1.0 * 0.3 * 0.7 * 0.3^2 = 0.0189$
So, state B is more probable.

ii. What is the probability of seeing this set of states given that the state you specified in A is correct?

Ans.:
If the third "E" is emitted by state B, then the total probability of the observed states is:
$P(\mathbf{Y} = ABBBC) = 1.0 * 0.3 * 0.7^2 * 0.3 = 0.0441$

iii. What is the probability of seeing this set if the state you specified in A is incorrect?

Ans.:
If the third "E" is emitted by state C, then the total probability of the observed states is:
$P(\mathbf{Y} = ABBCA) = 1.0 * 0.3 * 0.7 * 0.3^2 = 0.0189$
Note that $P(\mathbf{Y} = ABBCA) = P(\mathbf{Y} = ABCCA)$.

7. Platforms, normalizations and FDR (12 points).

(A) (8 points) Instead of adjusting the mean and variance we are interested in using a different normalization scheme: Rank adjustment. Given two microarray results we order the values for the first microarray and for the second microarray. We then assign the value of the top gene in microarray 1 to the top gene in microarray 2 etc. until we assign values for all genes (while the values would be the same the order of the genes between the two arrays will obviously be different). Lets call this rank normalization.

A1: The assumptions needed for rank normalization are (circle one)

Stronger Weaker Same Impossible to compare

when compared to the assumptions used for used for normalizing by adjusting the mean and variance (as we discussed in class).

Answer: Stronger

A2: Briefly explain your answer to B1:

Answer: Rank normalization naturally leads to the same mean and same variance. But the assumptions are much stronger since all other moments are also the same and the same values are present for all experiments.

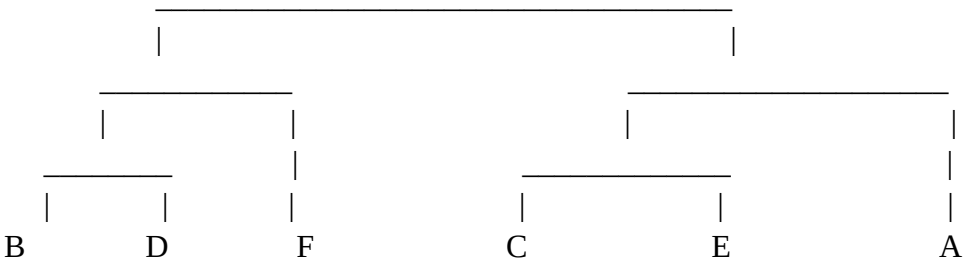
(B) (4 points) False discovery rate (FDR):
We are testing 6000 yeast genes between two different conditions. What p-value would lead to a FDR of 0.3 if we identified 200 genes as differentially expressed at the p-value?

Answer: 0.01. A FDR of 03 implies that 0.3*200 = 60 genes are false positives. Thus, we would expect 60 of the 6000 to be detected at this p-value which leads to 0.01.

8. Clustering (10 points).

(A) (5 points) Generate a hierarchical clustering tree below using complete linkage for the **similarity** matrix below:

	A	B	C	D	E	F
A	1	0.6	0.2	0.1	0.5	0.2
B		1	0	0.8	0.4	0.3
C			1	0.5	0.7	0.6
D				1	0.3	0.4
E					1	0.1
F						1



(B) (5 points) Present the optimally ordered tree. What is the value of the optimal order?

Answer: See above. The value is 3.

9. Classification (11 points).
We are interested in choosing genes for a Naïve Bayes classifier between cancer and healthy cells.

(A) (3 points) (**YES or NO**) Assume that we have two genes, G1 and G2. Both G1 and G2 have the same mean and variance (μ_h and σ^2_h) for the healthy cells and the same mean and variance (μ_c and σ^2_c) for the cancer cells (but the means and variances for the two classes are different, that is $\mu_h \neq \mu_c$ and $\sigma^2_h \neq \sigma^2_c$). Based on the criteria defined in class, could it be that we will prefer one over the other?

Answer: NO. For the log likelihood ratio the only thing that counts is the difference between the empirical means and variances. Since these are the same for both genes both will lead to the same value.

(B) (3 points) (**YES or NO**) What would be your answer to B be if we change the criteria to: We are interested in a gene that (using the Naïve Bayes classifier) maximizes the number of accurately classified samples in the training data.

Answer: YES. As an example consider the following two genes in two classes:
G1: Class 1 {0, 0, 0, 4} Class 2 {4, 4, 4.2, 4.2}
G2: Class 1 {-0.73, -0.73, 2.73, 2.73} Class 2 {4, 4, 4.2, 4.2}

Both have the same mean (1) and variance (12). But the first will lead to one misclassification (the 4 in class 1) and the second to correct classification.

(C) (3 points) For each of the two figures below, can a Naïve Bayes Classifier correctly classify the points

Figure 1: **YES** **NO** (circle one)

Answer: YES. While the mean is the same the variances are clearly different so a Naïve Bayes classifier will accurately separate the two classes.

Figure 2: **YES** **NO** (circle one)

Answer: NO. Both mean and variances are the same in both directions so there is no way to separate the two.

(D) (2 points) Same as C for a full Bayes classifier

Figure 1: **YES** **NO** (circle one)

Answer: YES. It can do as least as well as a Naïve Bayes classifier.

Figure 2: **YES** **NO** (circle one)

Answer: YES. While both mean and variances are the same in both directions, the covariance's are different. The genes in class one have correlated values whereas the genes in class two are anit-correlated.

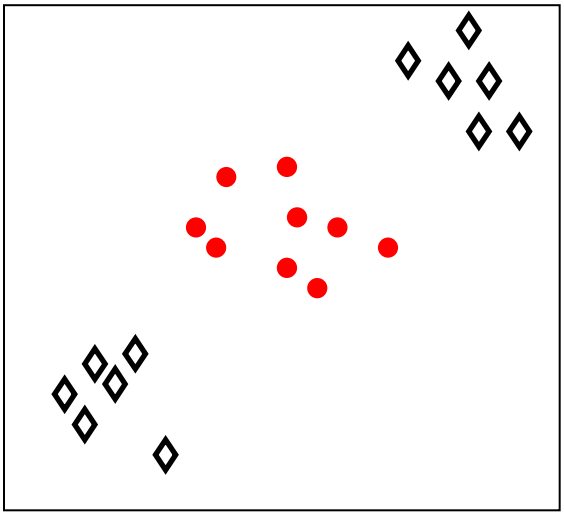


Figure 1

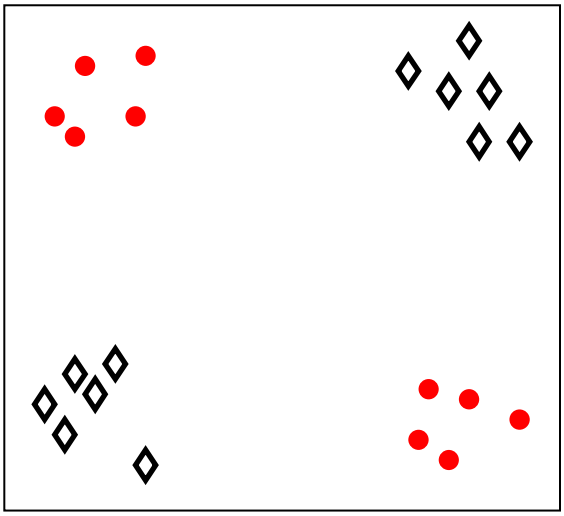


Figure 2