Computational Genomics Midterm
MSCBIO 2070/02-710
Spring 2014

March 26, 2014

This exam has 9 questions, for a total of 100 points.

Name: _____

Instructions:

- Write clearly. Unless stated otherwise, make sure to write down your steps for the calculations/derivations, do NOT just write a number.
- If you need more room to work out your answer to a question, use the back of the page. Make sure to indicate that we should check the back of the page for the rest of your answer.
- This exam is open book. 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	HMM	15	
2	Phylogenetics	10	
3	Evolutionary Trees	10	
4	Motif Discovery	12	
5	Normalization	12	
6	Hypothesis Testing	12	
7	Multiple Hypothesis Testing	10	
8	Clustering	9	
9	Feature Selection in Classification	10	

Name: _____

1. (15 points) **HMM**

   Assume we have a DNA sequence that begins in an **exon**, contains one **5′ splice site** and ends in an **intron**. In a given sequence, the problem is to identify where the switch from exon to intron occurred, that is identify where the 5′ splice site is.

   Say that the

   - exons have a uniform base composition on average,
   - introns are A/T rich but have non-zero probability for C/G bases
   - 5′ splice site consensus nucleotide is almost always a G, but sometimes an A.

   Note, the information provided here is not complete, so you have a choice in selecting these numbers, as long as they satisfy the specified constraints.

   (a) *Model set up:* Draw a hidden Markov model diagram for this problem.

   1. Specify the complete set of states.
   2. Specify the emission probabilities. State all the assumptions you made in arriving at these numbers.
   3. Specify the transition probabilities between all states. Assume the Markov chain when it enters an exon or intron state remains there with a probability of 0.9

   Answer:
   1. besides E, 5, and I, dont forget start and end nodes. So from start node go to E, then go to 5′ and then to I and finally to end node.
   2. at E, emission probabilities are all 0.25 because of the uniform base composition. at I, we will set emission probabilities to be say 0.4 each for A and T and 0.1 for C and G because introns are A/T rich. At 5′ splice site, set it to be 0.99 for G and 0.01 for A.
   3. from start to E, it will be 1. At E, the self-loop is 0.9 and transition to 5 is 0.1. at 5′ the transition to I will be 1. No self-loop at 5′. At I the self-loop is 0.9 and the transition to end node is 0.1

Name: _____

(b) *Finding the best path:* Consider the DNA sequence below and an example path through the state space:

where E stands for exon, I for intron and 5 for the 5′ splice site. There are potentially many state paths that could generate the same sequence.

For the HMM you have set up and the 22 nucleotide sequence given above, how many paths in the state space have non-zero probability? Explain why.

Answer: There are 14 possible paths. Because 5′ cannot be at the start or end, it has to be somewhere in between. There are 14 places where you will find either a G or A in the 26 nucleotide sequence, so those are the only places you can place the 5′ state. So the total number of paths is NOT infinity.

(c) *Posterior decoding:* In the example path shown above, the 5′ splice site was shown to be located on the fourth G (counting from the left). Knowing the total number of paths and each of their probabilities, how will you determine if the fourth G is the right choice?

Answer: Divide the probability of any given path by the sum of probabilities over all 14 paths. That will give the relative importance of each path and hence their significance.

Name: _____

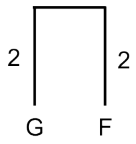
2. (10 points) **Phylogenetics**

Construct a phylogenetic tree using UPGMA for the following distance matrix. Show your steps and specify the resulting branch lengths.

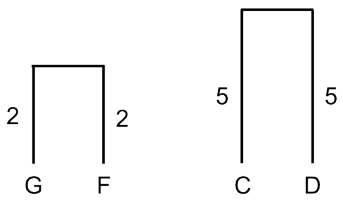
	Cat	Ferret	Dog
Ferret	15		
Dog	10	17	
Gerbil	16	4	18

Answer:

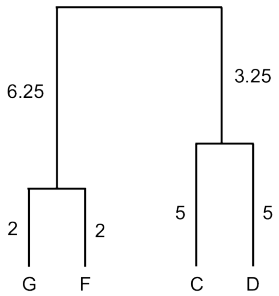
	Cat	Ferret	Dog
Ferret	15		
Dog	10	17	
Gerbil	16	**4**	18



	Cat	GF
GF	15.5	
Dog	**10**	17.5



	CD
GF	**16.5**

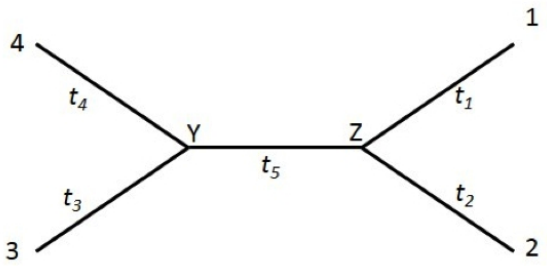


Name: _____

3. (10 points) **Evolutionary Trees**

The probability of the occurrence of a particular evolutionary history as represented by a particular tree can be calculated given a model of evolution. We will assume that the model is time-reversible, which means the likelihood of the model does not change regardless of which time direction is assumed for a branch, and hence the location of the root is unimportant in the calculation.

You will use the tree shown below to calculate the model likelihood using Jukes-Cantor (JC) model of evolution.



This is a tree with four leaves at nodes labeled 1 to 4, with observed sequences at these leaves, but unknown ancestral sequences at the internal nodes Y and Z. The probabilities you calculate depend on the tree topology T. You will consider each branch separately, and the **probability of base i at any position mutating to base j in a time t will be written as $P(j|i,t)$.** You will use Jukes-Cantor (JC) model to evaluate these probabilities.

i. In the JC model with mutation rate parameter α, write down the analytical expressions for the following probabilities:

$P(j|i,t)$ when i is the same as j,
and
$P(j|i,t)$ when i is not the same as j.

Your expressions will use both the rate parameter α and time t.

Answer: We derived these expressions in the class.
$\frac{1}{4}[1+3\exp(-4\alpha t)]$ **and** $\frac{1}{4}[1\exp(-4\alpha t)]$

Name: _____

ii. You can compute the likelihood of **specific bases** x_Y and x_Z occurring at a particular sequence position at internal nodes Y and Z of the tree with topology T and branch lengths t_i . Assume each branch has evolved independently, so you can multiply the probabilities together. Pick x_Z as your root node and expand/simplify the following probability expression:

$$P(x_1, x_2, x_3, x_4, x_Y, x_Z | T, t_1, t_2, t_3, t_4, t_5)$$

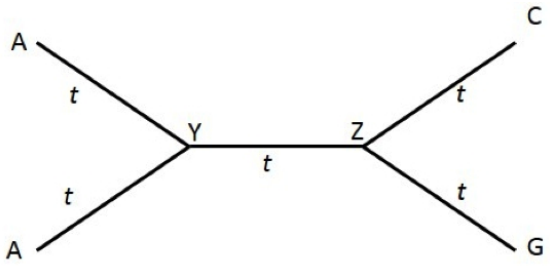
Answer:

$P(x_1, x_2, x_3, x_4, x_Y, x_Z | T, t_1, t_2, t_3, t_4, t_5)$
$= q_{xZ} P(X_1 | X_Z, t_1) P(X_2 | X_Z, t_2) P(X_Y | X_Z, t_5) P(X_4 | X_Y, t_4) P(X_3 | X_Y, t_3)$
where q_{xZ} is the probability of the base at internal node Z being xZ

Name: _____

iii. As an example of this method, you will compute the likelihood of four sequences of one base each, two being A and the other two being C and G, under the simplifying constraint of all five branches having the same length t and using the Jukes-Cantor evolutionary model. Assume the base composition to be equal for all four bases.

Write down the analytical expression for the likelihood of the tree given below:



Note, because the internal nodes Y and Z can have any possible base values for i^{th} position, this must be taken into account.

Use the following notation for the expressions derived from the first problem:
$S = P(j|i;t)$ when i is the same as j
and
$D = P(j|i;t)$ when i is not the same as j
for branch lengths between nodes with identical and different bases respectively.

Answer:
Y and Z can take 16 possible combinations.
$(A,A)(A,C)(A,G)(A,T)(G,A)(G,C),(G,G),(G,T),\ldots$
Every one of these combinations contributes to the likelihood expression. For example plugging in (A,A), we find that there are 3 branches that have same bases on both ends, while two other branches do not. So the contribution to likelihood is: $\frac{1}{4}S^3D^2$, where $\frac{1}{4}$ is because all bases are assumed to be equally present at q_{xZ}. We have to repeat this for all 16 combinations and collect the common terms.

Name: _____

4. (12 points) **Motif Discovery**
In class we discussed two types of matrices for motif discovery. The first, PWM, represents fractions of nucleotides at? each position based on counts of known aligned motifs. The second, PSSM, uses the PWM to compute a score by also taking into account the background distribution of the species being studied.

Assume you are given the following PWM:

	P1	P2	P3	P4	P5
A	0.05	0.3	0.1	0.6	0.8
C	0.8	0.2	0.1	0.2	0.05
G	0.1	0.3	0.7	0.2	0.05
T	0.05	0.2	0.1	0	0.1

As mentioned above, PSSM matrix does not take the genomic background distribution into account. For motifs $M1$ and $M2$ we write $P(M1) > P(M2)$ if the probability that $M1$ is a real motif based on our scoring function is higher than the probability that $M2$ is a real motif. For example, using only the PWM above to score motifs, we have $P(\texttt{CCCGG}) > P(\texttt{GTAGC})$.

For each of the following motif pairs, state what the background distribution should be so that if we use the PSSM derived from the PWM above (by incorporating the background distribution), $P(M1) > P(M2)$. If no background distribution would lead to such relationship briefly explain why.

Note that changing the background distribution can change the scores since if a nucleotide is very rare in a species then even if it appears in only 30% of the motifs (0.3), if we see it in that position it could still lead to a high score. Also note that there may be several different background distributions that will satisfy the questions below, in such cases just state one example of such distribution.

(a) $M1 = \texttt{CCTCC}, M2 = \texttt{GTCGG}$
Answer: A uniform background (25% each nucleotide) would work since even by just using the PWM we have $P(M1) > P(M2)$ and a uniform background would leave the relationship the same.

Name: _____

(b) $M1 = \texttt{CGATG}, M2 = \texttt{GATCC}$

Answer: No background would help here. Since the probability for observing a T in the fourth location is 0, no matter what background we use $P(M1)$ would be 0 and would always be lower than $P(M2)$.

(c) $M1 = \texttt{CCTCC}, M2 = \texttt{GTAGC}$

Answer: We can see that $M1$ has lots of Cs while $M2$ has only 1. On the other hand, $M1$ does not contain any Gs while $M2$ has 2. Even though the first C is much more likely than the first G, if we assume a very skewed GC distribution it would lead to each G being much more surprising than C. So for example $A = 0.1, T = 0.1, C = 0.79$ and $G = 0.01$ would work here. Several other solutions in which $G >> C$ would also work.

Name: _____

5. (12 points) **Normalization**

Assume we have performed a RNA-Seq for two samples from the same species, A and B. We aligned the reads in Seq dataset to the genome and obtained counts for every gene (number of reads mapped to each gene). Let g_1^A and g_2^A be the read counts for genes g_1 and g_2 in experiment A. Denote by $T(g_1^A)$ and $T(g_2^A)$ the normalized values for genes g_1 and g_2 in experiment A.

Assume $g_1^A > g_2^A$. For the following questions choose ALL answers that could be correct. No need to explain your answers.

(a) If we used quantile normalization where for the common values (which we assign to all experiments) we have used the median of each rank then:

i $T(g_1^A) > T(g_2^A)$
ii $T(g_1^A) = T(g_2^A)$
iii $T(g_1^A) < T(g_2^A)$

Answer: (i) and (ii). Since the values are assigned based on rank $T(g_1^A)$ can either be higher than $T(g_2^A)$ (since its ranked higher) or equal if there are ties in the other experiments.

(b) If we used scale factor normalization then:

i $T(g_1^A) > T(g_2^A)$
ii $T(g_1^A) = T(g_2^A)$
iii $T(g_1^A) < T(g_2^A)$

Answer: (i). In this case, because $g_1^A > g_2^A$ any linear transformation (which is what we apply in scale factor normalization) will maintain the same relationship between the two values.

(c) If we used RPKM normalization then:

i $T(g_1^A) > T(g_2^A)$
ii $T(g_1^A) = T(g_2^A)$
iii $T(g_1^A) < T(g_2^A)$

Answer: (i) (i) (iii). In this case it is impossible to tell. For example, if gene 2 is shorter than gene 1 than even though more reads are assigned to gene 1, after RPKM normalization the relationship can be reversed.

Name: _____

For questions d, e, and f, assume that the *total number of reads* in experiment A is the same as the number of reads in experiment B. Let g_1^B and g_2^B be the read counts for genes g_1 and g_2 in experiment B. Similarly, denote by $T(g_1^B)$ and $T(g_2^B)$ the normalized values for genes g_1 and g_2 in experiment B.

Assume $g_1^A > g_1^B$. Again, for the following normalization methods choose ALL answers that could be correct.

(d) If we used invariant set normalization then:

i $T(g_1^A) > T(g_1^B)$
ii $T(g_1^A) = T(g_1^B)$
iii $T(g_1^A) < T(g_1^B)$

Answer: (i) (ii) (iii). It's impossible to tell, it depends on the overall set of values in the two experiments.

(e) If we used RPKM normalization then:

i $T(g_1^A) > T(g_1^B)$
ii $T(g_1^A) = T(g_1^B)$
iii $T(g_1^A) < T(g_1^B)$

Answer: (i). Here, since we know that that total number of reads is the same and that the gene length is the same (its the same gene) RPKM is just a multiplication by a constant which is the same for both experiments and so the relationship is maintained.

Assume we have used scale factor normalization. We know that the variance of experiment A (V^A) is 4 times higher than the overall variance (V). If $g_1^A = 50$, $T(g_1^A) = 75$, and $g_2^A = 100$.

(f) What is the value of $T(g_2^A)$?

i 100
ii 125
iii 150
iv Impossible to tell

Answer: (ii). We know that $\frac{V}{V^A} = 4 \Rightarrow \sqrt{\frac{V}{V^A}} = 2$. In addition, because its the same species in both experiments and the number of reads is

the same, both have the same mean so $M^A = M$. We thus can compute M and we find that $M = 100$ (since $(50-100)/2+100 = 75$). Inserting 100 into the same equation for an original read value of 150 leads to 125.

Name: _____

6. (12 points) **Hypothesis Testing**

In class we discussed using a log likelihood ratio test for hypothesis testing. Such a test is appropriate for nested hypothesis, where H1 is a more detailed version of H0 (in the example we discussed in class, H1 has one extra parameter but both H0 and H1 used the same model). Based on this, for each of the following hypotheses write if they could be tested using a log likelihood ratio test. If the answer is yes, say how many degrees of freedom we have. If the answer is no, briefly explain why.

(a) H0: Gene g_1 in two sets of mircroarray expression experiments (cancer and healthy) is generated from the same Gaussian distribution (same mean and variance in both sets of patients).
H1: Gene g_1 in two sets of mircroarray expression experiments (cancer and healthy) is generated from two different Gaussians with different means and a shared variance $V = 1$.

Answer: No. These are not nested hypotheses. Specifically, the likelihood of H1 can be lower than the likelihood of H0 for example if the real variance is much higher or much lower than 1.

(b) H0: Using a Poisson probabilistic model for read counts in RNA-Seq
H1: Using a Negative Binomial probabilistic model for read counts in RNA-Seq

Answer: Yes. The D.O.F. is 1. We mentioned in class that when the mean equals the variance negative binomial becomes Poisson, so these are nested.

Name: _____

For questions c and d, assume we are clustering expression data from n arrays.

(c) H0: Using k-means with $k = 5$ for clustering gene expression data
H1: Using k-means with $k = 7$ for clustering gene expression data

Answer: No. K-means is not a probabilistic model and so we cannot compute a likelihood for either hypothesis.

(d) H0: Using Gaussian mixtures (diagonal covariance matrix) with $k = 5$ for clustering gene expression data
H1: Using Gaussian mixtures (diagonal covariance matrix) with $k = 7$ for clustering gene expression data

Answer: Yes. The D.O.F. is $4n$ **(**n **additional mean values and** n **additional variance values for each additional cluster).**

Name: _____

7. (10 points) **Multiple Hypothesis Testing**

(a) Assume we are testing 100 genes. We found 5 significant genes with a Bonferroni corrected p-value less than 0.05. What is the FDR (in %) for this set?

Answer: If the corrected p-value is 0.05 **then the actual p-value we used is** $0.05/100 = 0.0005$**. At that p-value we expect to find** $0.005 * 100 = 0.05$ **genes by chance. Since we found 5, the FDR is** $0.05 * 100/5 = 1\%$**.**

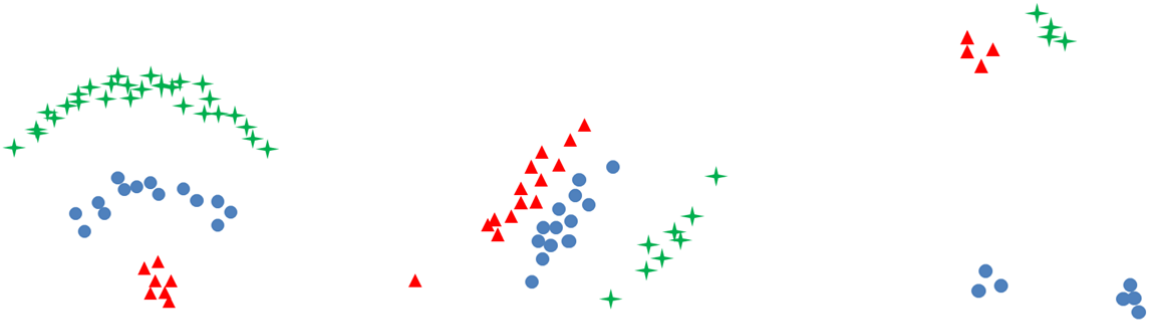
(b) Assume we are testing 100 genes. We found 10 significant genes with a FDR of 1%. What is the Bonferroni corrected p-value that applies to genes in this set?

Answer: If we found 10 genes with a FDR of 1% then for the p-value we used we expected to find by chance 0.1 genes. This corresponds to an (uncorrected) p-value of $0.1/100 = 0.001$**. When correcting using Bonferroni we get** $0.001 * 100 = 0.1$**.**

Name: _____

8. (9 points) **Clustering**

Select the suitable clustering method(s) that can give the following result.



Fill in the following table. Mark T if you think the clustering method is suitable for the figure and mark F if not suitable. No need to explain your answers.

	Figure (a)	Figure (b)	Figure (c)
Gaussian mixtures with full covariance matrix	**F**	**T**	**T**
Gaussian mixtures with diagonal covariance matrix	**F**	**F**	**T**
Hierarchical clustering with single linkage	**T**	**F**	**F**

Name: _____

9. (10 points) **Feature Selection in Classification**

Given a gene expression data set, instead of using all the genes, we want to classify samples (columns) only with a few differentially expressed genes. Here are two strategies:

Strategy (A): First, use all the samples to select top 50 DE genes (similar to what you did in the homework). Then split the samples into training data and test data. Train the classifier with the training data and use the same parameters to compute the test error.

Strategy (B): First, split the samples into training data and test data. Then only use training data to select top 50 DE genes and train the model. Again use test data to compute the test error rate.

Both strategies will use the same classifier. Which strategy is more likely to result in a smaller training error rate? Which one is likely to lead a smaller test error rate? Which method would you prefer to use to classify future (unseen) data? Explain.

Answer: The second strategy will lead to better (smaller) training error rate since the DE genes are only from the training set, while the first will lead to better test error rate since it is 'cheating' by overfitting. We would prefer to use the second strategy because true labels for future data are unknown.