Midterm Exam for 02-710/10-810/MSCBIO2070 Computational Genomics 2013

Name:

Instructions:

- Write clearly. 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 clearly mark on the front of the page if we are to look at what's on the back.
- This exam is open book. Calculators are allowed, but no computers, PDAs, or other communication devices.
- You have 1 hour and 20 minutes. Good luck!

No.	Topic	Max Score	Your Score
1	Biological Background	14	
2	Sequencing and Analysis	23	
3	Substitution Model	23	
4	Differential Expression	12	
5	Normalization	16	
6	Clustering	12	
	Total	100	

1. Biological Background (14 Points)

1. [8 Points] List the names of four human genes and describe in one sentence their main function/significance in human biology.

depends on the answer. but please do not reward them if they quote famous genes from other species that are not in humans

2. [3 Points] which of the following is correct? [**Note: -1 point for wrong answers**]

   A) Deamination or methylation of protein residues A, T, G, and C yields the four bases that make up DNA
   B) The central dogma states that proteins make DNA, which makes RNA
   C) The term cloverleaf, used for tRNAs relate to the secondary structure of tRNA
   D) The total number of all transcribed sequences in human genome is approximately 25,000
   E) RNA extraction from ancient fossils is now being recognized as an easier way to study evolution.
   F) Molecular clock hypothesis implies that evolutionary rate of different proteins is a constant.
   G) None of the above

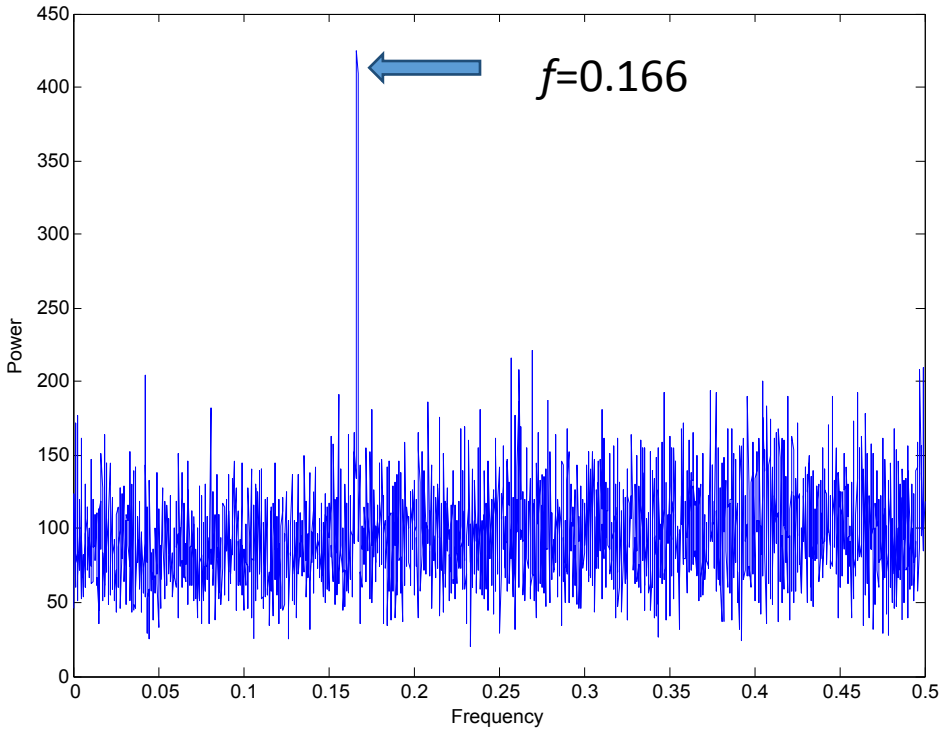
C is correct

3. [3 Points] 10 cycles of standard PCR is performed on a starting set of ~1000 double stranded DNA molecules. Assuming that each cycle of the PCR is 100% efficient, ie all molecules present in a given cycle is amplified without any error or deviation from theory, how many molecules will be made after 10 cycles?

1000-> 2000 in second cycle-> 4000 in 3rd and so on => 1000*2^10 after 10 cycles

2. Sequencing and Analysis (23 Points)

1. [5 Points] a. The curiosity rover found life on mars. The alien has similar DNA to earthlings, but it is unknown if the alien DNA codes for proteins in the same way as earth DNA. To test this, you perform a Fourier transform on the alien DNA sequence (as you saw in class lecture) and get the following power plot. What does this suggest about how translation works in the alien cell?



1/0.166=6 nts. the codons seem to be 6 nts long.

2. [4 Points] You want to see if the ~8K long alien mitochondrial genome aligns with the ~16K long human mitochondrial genome using Needleman-Wunsch algorithm. Assuming that aligning two sequences of length 4K each, would take 1 second, how long will the alignment take?

   algo is O(NM) => 4*4 eq 1 sec, 16*8 eq to 8 sec.

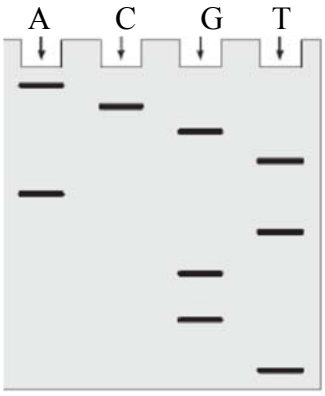
3. [6 Points] You are tasked to perform a global sequence alignment of two sequences X and Y. However, you know that Y is only a partial sequence and hence you do not want to penalize for any gaps that precede Y or occur after the last residue in Y. See illustration below for an example. How will you achieve this based on the basic Needleman-Wunsch algorithm? Hint: your answer should state clearly the array initialization and where to trace back from and where to end the trace back. Assume residues of X are arranged in columns in the alignment matrix, F.

   Example: if X=AGGGATCACGGC, Y= GGACCG, we want to get an alignment like this using Needleman-Wunsch algorithm.

   ```
   X=  GGATCACG
   Y=  GGAC--CG
   ```

X indices are j, Y indices are i. gaps before Y should not be penalized but of X should be.
F(0,0)=0, F(0,j)=0 for all j, F(i,0)=-i*go;
for not penalizing gaps after Y, the traceback will need to start from the best score in the last row ie traceback from argmax {F(length(Y),j)} for j=0..length(X)
Traceback will end as soon as you reach row 0.

4. [8 Points] An mRNA (sequence R) is amplified by 30 cycles of PCR. DNA sequencing was performed using the sanger method using the DNA strand (sequence D) that contained a stretch of over 20 Ts. the following patterns on the gel was obtained for the sequencing products. What is the best inference for the sequences of D and R (assume that the consecutive T's are related to polyadenylation seen in mRNAs) ? **Label the 5' and 3' ends of your sequences**.

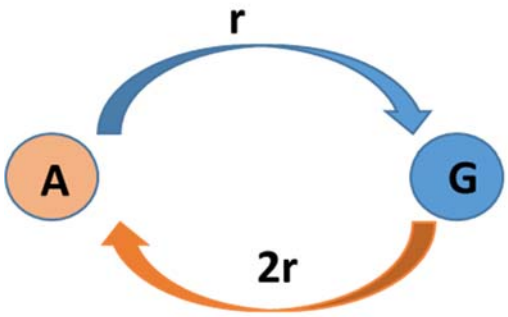


The sequencing product is 5'-TGGTATGCA-3' so the DNA strand used for sequencing is 3'-ACCATACGT-5' but this sequence is reverse complementary to the RNA sequence because RNA sequence has the polyA and the DNA has the polyT. also RNA has U instead of T.

which means: R= 5'-UGGUAUGCA-3'
D= 3'-ACCATACGT-5' OR 5'-TGCATACCA-3'

3. Substitution Model (23 Points)

1. [20 Points] Assume proteins carry only two amino acids, denoted as A and G. with the substitution rates shown in figure between the two residues. Assuming a first order Markov model, derive the probability transition matrix P(t) for this model. Show all steps clearly.



P(t)=

$$M(t) = \begin{array}{c} \begin{matrix} A & \quad G \end{matrix} \\ \begin{pmatrix} 1-f(t) & f(t) \\ 2f(t) & 1-2f(t) \end{pmatrix} \end{array} \quad \text{where } f'(0) = r$$

$$M'(t) = M(t).M'(0)$$

$$\Rightarrow f'(t) = r - 3rf(t)$$

$$\Rightarrow f = \frac{1-e^{-3rt}}{3} \quad \text{since } f(0) \text{ is } 0$$

X

2. [3 Points] now consider position 2 in the following sequences and assume that the ancestral residue at that position was a G (as shown). How many total mutations are expected to occur between time 0 and Δt? express your answer in terms of r and time.

```
sequence 1: AGGGTA
sequence 2: GAGGAA
Position  : 123456
```

2r*delta-t

3. [3 Points] Now assume the ancestral residue at position 3 is A. How many total mutations are expected to occur between time 0 and Δt?

r*delta-t

4. Differential Expression (12 points)

In the following two questions we will compare two separate experiments that were performed to identify differentially expressed (DE) genes. The first was performed for condition A and the second for condition B. Let NA and NB be the total number of genes measured in experiment A and B respectfully, PA and PB be the p-values used in the analysis of each condition, GA and GB the number of DE genes identified for each condition and FA and FB the FDR for the DE gene lists in the two conditions.

1. [2 Points] If NA>NB, PA=PB and GA=GB then:

   A. FA > FB
   B. FB > FA
   C. FA = FB
   D. Impossible to tell

Answer: A. Since NA>NB we can expect more genes for the same p-value and so the FDR for A is higher.

2. [2 Points] If NA > NB, PA > PB and GA=GB then:

   A. FA > FB
   B. FB > FA
   C. FA = FB
   D. Impossible to tell

Answer: A. Here both NA>NB and the p-value for A is higher so we expect even more genes under that p-value so FA>FB.

3. [2 Points] If NA > NB, PB > PA and GA=GB then:

   A. FA > FB
   B. FB > FA
   C. FA = FB
   D. Impossible to tell

Answer: D. Since PB>PA its impossible to know how many we would expect at random for the two experiments.

4. [2 Points] Assume that PA and PB are the Bonferroni corrected p-values (for each condition) for an initial p-value of 0.01. If GA > GB, NB > NA and FA > FB then
   A. PA > PB
   B. PB > PA
   C. PA = PB
   D. Impossible to tell

Answer: A. The only thing that matters for the correction are NA and NB. Since NB>NA the corrected p-value for B would be lower.

5. [4 Points] Instead of using Bonferroni we decided to perform randomization tests (separately for each condition). Experiment A had 5 healthy arrays and 5 cancer arrays. What is the lowest p-value we can achieve for a gene in this experiment using randomization tests?

Answer: $\frac{1}{\binom{10}{5}} = \frac{1}{252}$

5. Normalization (16 Points)

1. [4 Points] It was recently shown that some cancer cells have as much as three times more total mRNA than healthy cells. Assume we have performed RNA-Seq experiments in healthy and cancer patients. Is this finding a problem for the following two normalization methods (circle the correct answer and briefly explain):

   A. Quintile normalization: Yes No

   Answer: Yes, it's a problem. Quantile assumes that the total levels are same and so is not appropriate for this experiment.

   B. RPKM: Yes No

   Yes, it's a problem. The values are normalized per million reads and so the underline assumption is still that the overall RNA levels are this same.

2. [4 Points] Several genes are alternatively spliced meaning that in some conditions they use one subset of exons and in another they use a different subset, or several different subsets. For example, a gene with exons A,B,C and D may give rise to several different proteins, for example: ABD and ACD. Assume we only care about the total number of transcripts for a gene (regardless of hat splice variant is used). Explain why alternative splicing may cause a problem when using microarrays (Hint: think of how they are designed).

   Answer: Depending on what probes we select, we may not represent all exons and so if a gene is alternatively splices and the remaining exons are not on the array we will lose the ability to infer that the genes is expressed when using microarrays.

3. [8 points] Instead of microarrays we are using RNA seq. We know that gene G has 4 potential splice variants: ABC, ABD, ACD and BCD. Assume that read counts are accurate. Explain how we can determine the expression levels (in RPKM) of each of these variants in a specific experimental condition (you can assume that reads can be mapped to exons to determine RPKMs but not to larger units so its impossible to tell for a given read assigned to exon A if it was generated by ABC or ABD). Be specific.

We need to solve a set of linear equations for this. As the question notes, we can assume that we have RPKM reads for the specific exons. We now need to distribute them across the splice variants. Let RA, RB, RC and RD be the RPKM values for the 4 exons, respectively. Let R1, R2, R3 and R4 be the (unknown) RPKM for the 4 splice variants (ABC, ABD, ACD and BCD). Then the following 4 equations must hold:
(R1+R2+R3)/RA = 1
(R1+R2+R4)/RB= 1
(R1+R3+R4)/RC = 1
(R2+R3+R4)/RA = 1
To see why this is the case, note that R1, R2 and R3 all contain exon A and so together they account for all reads from this exon. By solving these equations we can obtain the values for R1, R2, R3 and R4.

6. Clustering (12 Points)

1. [4 Points] The blue points in Figure X represent cluster centers (for both k-means and Gaussian mixtures). The red points are two genes in our study (each gene has two measurements which are reflected by the x and y values that determine the point location in the figure). There are several other genes in our study, but they are not shown in the picture (though, of course, they were used to determine the cluster centers).Using the following options:

   A. The two genes belong to the same cluster
   B. The two genes belong to different clusters
   C. Impossible to tell

   For each of the clustering methods discussed in class chose the correct letter from the options above:

   i. Hierarchical clustering using Pearson correlation (we define a cluster by splitting the tree at the 3rd level from the root so that we have 4 total clusters) – C. Depends on the other points and the linkage method used.

   ii. K-means – A

   iii. Gaussian mixture (complete covariance matrix) – C. It depends on the variance for each cluster.

2. [6 Points] We performed experiments in which we tested 50 biological human samples using microarrays (the arrays profile thousands of genes). For two genes, A and B we have 10 missing values for each across all samples (the missing values may come from different samples, so gene A may have a missing value in sample 1 whereas gene B has a value for sample 1). For each of the three clustering methods state if such data can be used by the clustering method and if so how.

    i. Hierarchical clustering using Pearson correlation (we define a cluster by splitting the tree at the 3rd level from the root so that we have 4 total clusters) – Yes. We can compute the Pearson correlation using the set of points shared by the two genes.

    ii. K-means – – Yes. Compute the distance for each cluster center for the observed values.

    iii. Gaussian mixture (complete covariance matrix) – No. Given the full covariance matrix, we cannot determine the likelihood in case of missing values.

3. [4 Points] Assume that apart from the missing values, all values that are measured for both A and B are the same (so if A has a value for sample i and B also has a value for that sample, their values are the same for that sample, however, it could be that A has a value for i and B does not have a value for that sample). For the methods for which you answered 'Yes' in 2, chose one of the three letters in 1 to state whether A and B will be in the same cluster. If you answered 'No' in 2 for any of these methods just write 'No' again.

    i. Hierarchical clustering using Pearson correlation (we define a cluster by splitting the tree at the 3rd level from the root so that we have 4 total clusters) – A. The correlation is 1.

    ii. K-means – C. Depending on the values present independently for each gene, they can be assigned to different clusters.

    iii. Gaussian mixture (complete covariance matrix) – No.