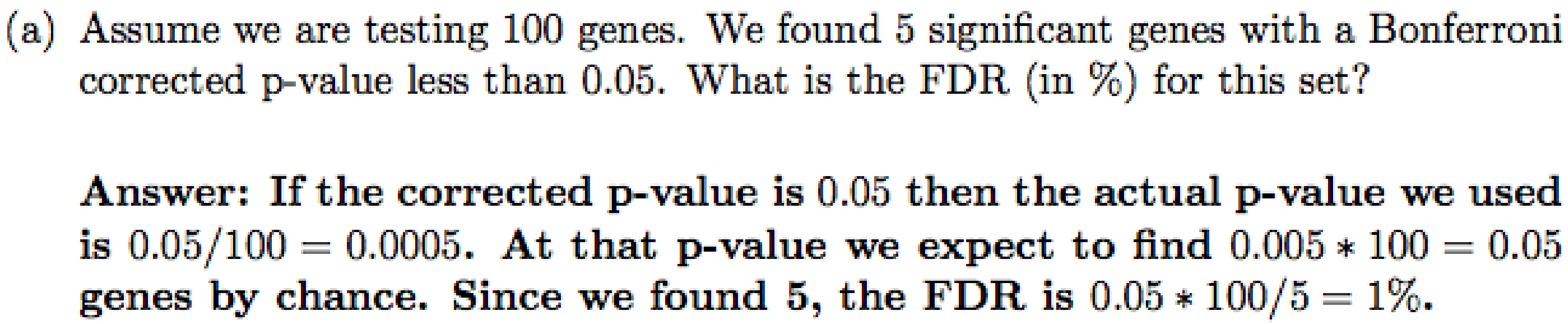
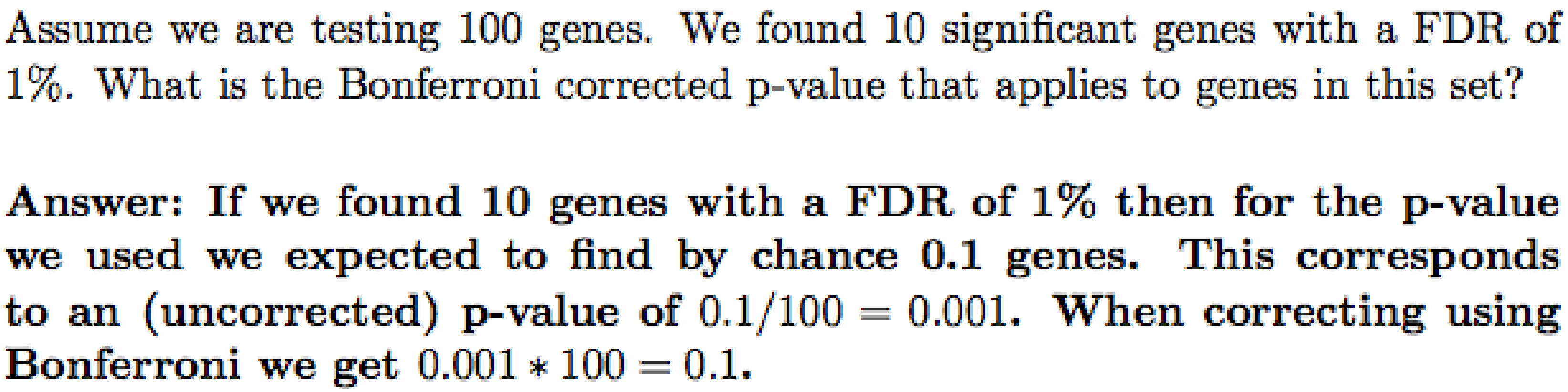
Review Notes

- Material covered in different years is slightly different
- You are only responsible for the material discussed in the lectures so far



(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** $0.05 * 100/5 = 1\%$**.**



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** $0.001 * 100 = 0.1$**.**

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

   Answer: d. Given this FDR we expect 8 genes to pass the p-value by chance. Since the p-value is 0.005 we have n*0.005 = 8 -> n = 1600

4. We tested n genes using a bonforonnie 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

Answer: e. The corrected p-value is a function of the initial p-value and n (number of genes). Since we do not know both of these it is impossible to tell what is p.

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.

Genome Assembly

9. (1.5 points) Eulerian approach to assemble a sequence from its k-mers using de Bruijin graph

Assume k = 3 for the sequence: ATAGCCTAGCAAT

A) (0.5 point) Write down the set of k-mers. Draw the di Bruijin graph for the set of k-mers where each vertex corresponds to k-1 mer and edges correspond to the k-mers. Do not allow multiple edges between two vertices.

B) (0.5 point) Identify an Eulerian path and the corresponding string from the di Bruijin graph.

C) (0.5 point) Is there an Eulerian path in the graph that corresponds to the original string? Can you say why?

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

Alignment

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

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

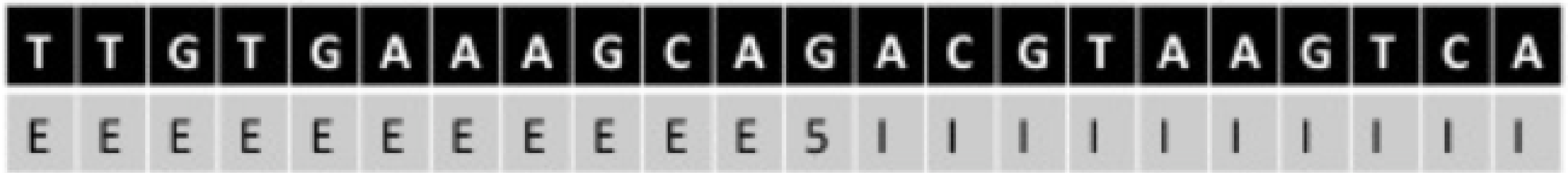
Answer:

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

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



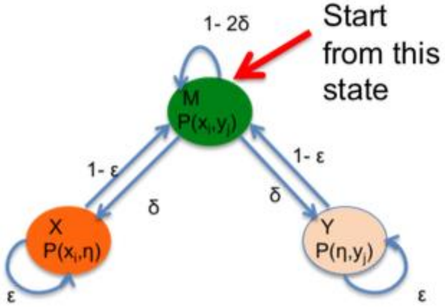
T	T	G	T	G	A	A	A	G	C	A	G	A	C	G	T	A	A	G	T	C	A
E	E	E	E	E	E	E	E	E	E	E	5	I	I	I	I	I	I	I	I	I	I

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.

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.

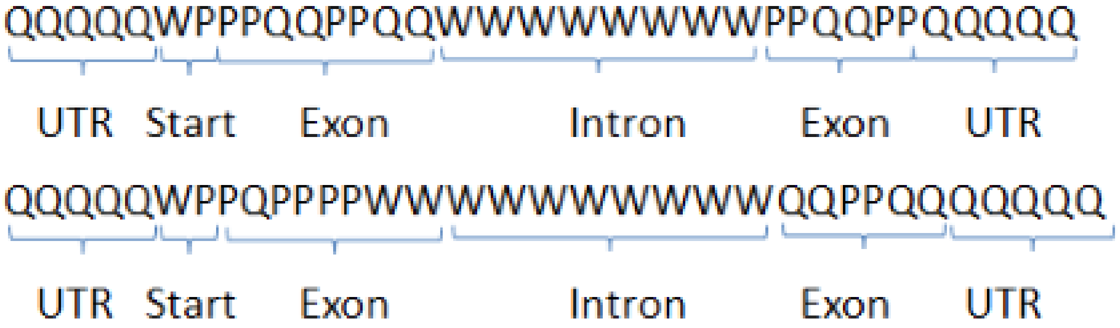


4. (3 points) In the above model can you explain why the transition between X and Y are not needed?

7. (1 point) You are a famed exobiologist studying the newly discovered lifeforms on Mars. Due to their highly divergent evolution from Earth species, existing gene finding tools are completely useless. Using the following facts and annotated Martian genes, build an HMM-based gene model. Show your states, as well as emission and transition probabilities.

Fact 1: The genome uses 3 symbols: QWP.
Fact 2: Like Earth eukaryotes, there have both exons and introns

Genes:


QQQQQWPPPQQPPQQWWWWWWWWPPQQPPQQQQQ

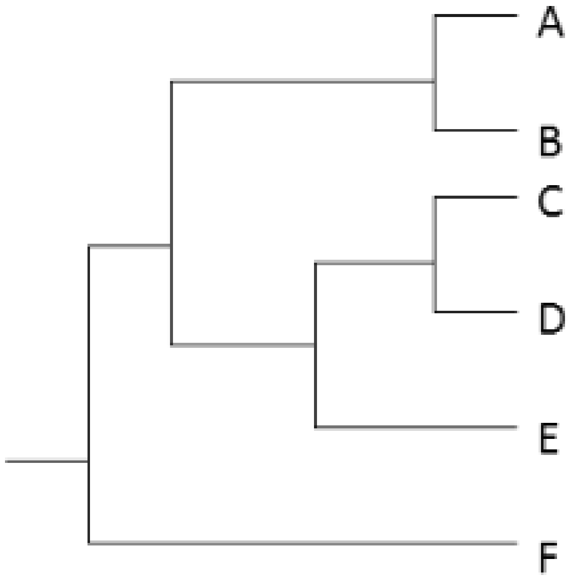
UTR Start Exon Intron Exon UTR

QQQQQWPPQPPPPWWWWWWWWWWQQPPQQQQQQQ

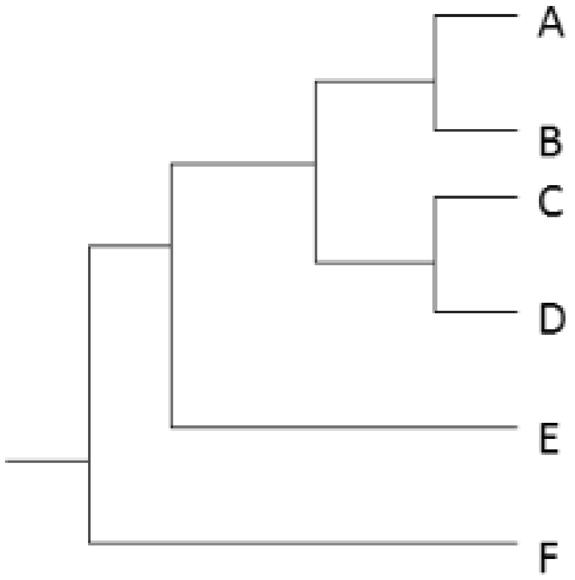
UTR Start Exon Intron Exon UTR

Phylogeny

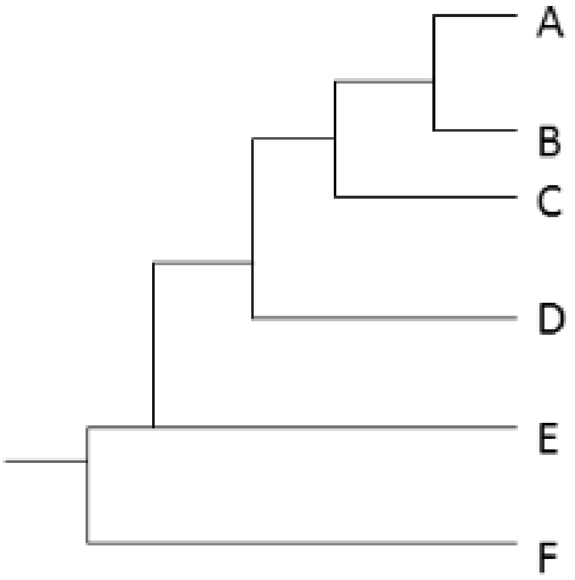
5. [**3 points**] GENE1 is found in species A and B, but not in C, D, E, and F. GENE2 is found in species C and D, but not in the other species. Which of the following rooted phylogenetic trees is supported by these findings? There may be more than one tree that could fit this data.



1.



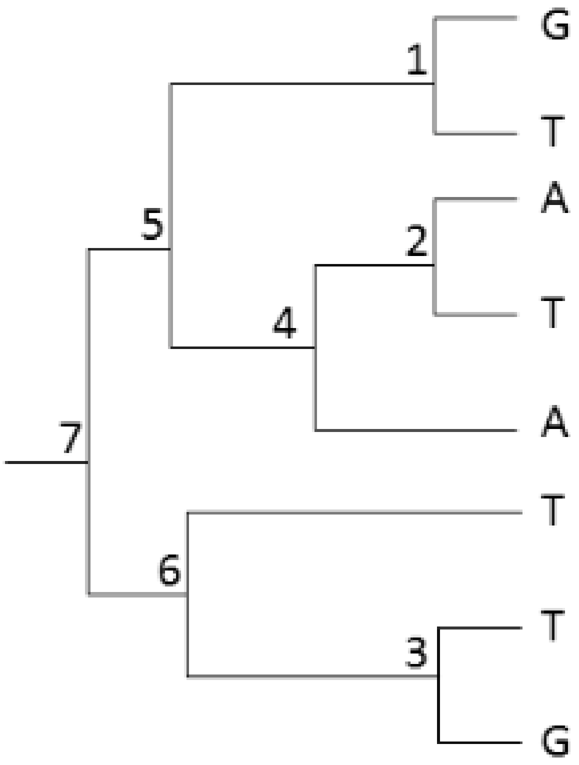
2.



3.

Answer: Trees 1 and 2.

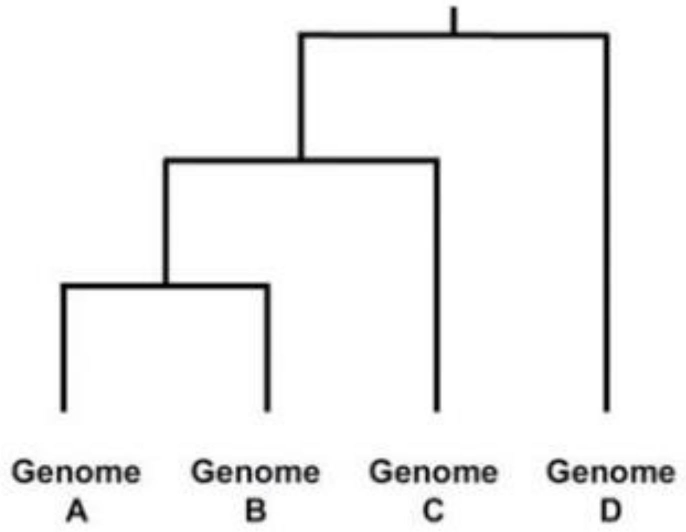
6. **[3 points]** Using maximum parsimony, reconstruct the ancestral nucleotide at the internal nodes of the following tree by labeling the ancestors at each node, 1-7. If more than one nucleotide is possible, indicate which they are.



Answer:
1. G/T
2. A/T
3. T/G
4. A
5. G/T/A
6. T
7. T

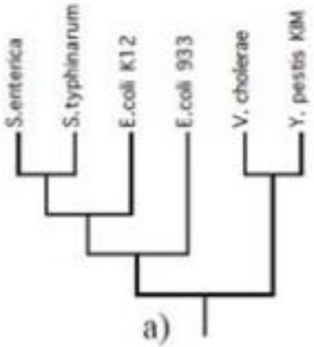
2.2. Show an example of a distance matrix with three species (A, B and C) that will lead to the same unrooted tree no matter which method (UPGMA/NJ) is used (2 points)

10. (1.5 point) In the figure below, each row in the table lists a set of four bacterial taxa whose evolutionary relationship follows the topology of the tree. Each row can be interpreted as a four taxon tree.

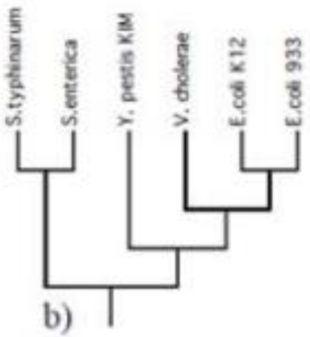


E. coli 933	*E. coli K12*	*S. typhimurium*	*Y. pestis KIM*
S. enterica	*S. typhimurium*	*E. coli K12*	*Y. pestis KIM*
E. coli K12	*S. typhimurium*	*Y. pestis KIM*	*V. cholerae*

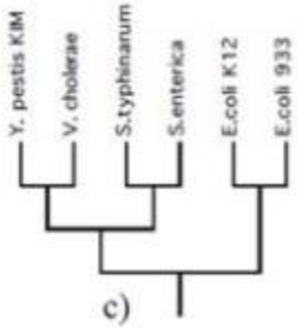
Which of the four trees below is compatible with the information provided in all of the rows of the table?



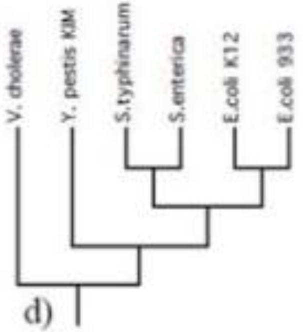
a)



b)



c)

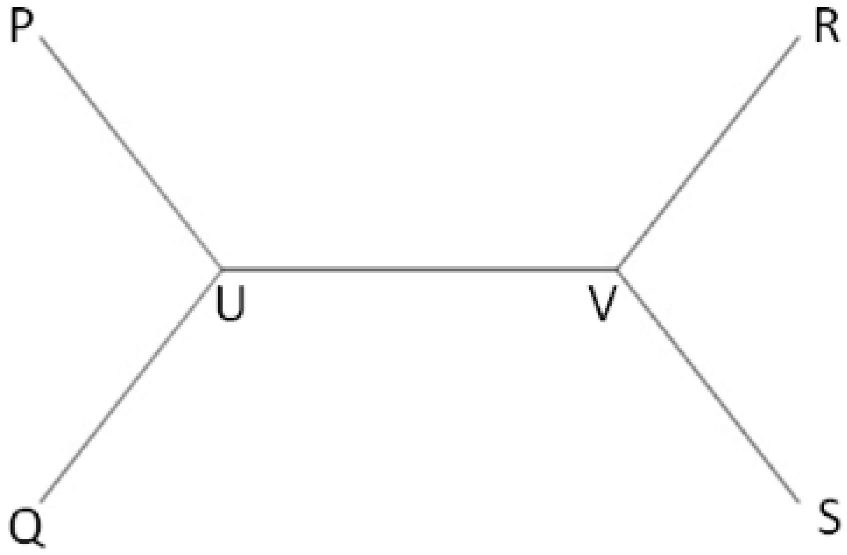


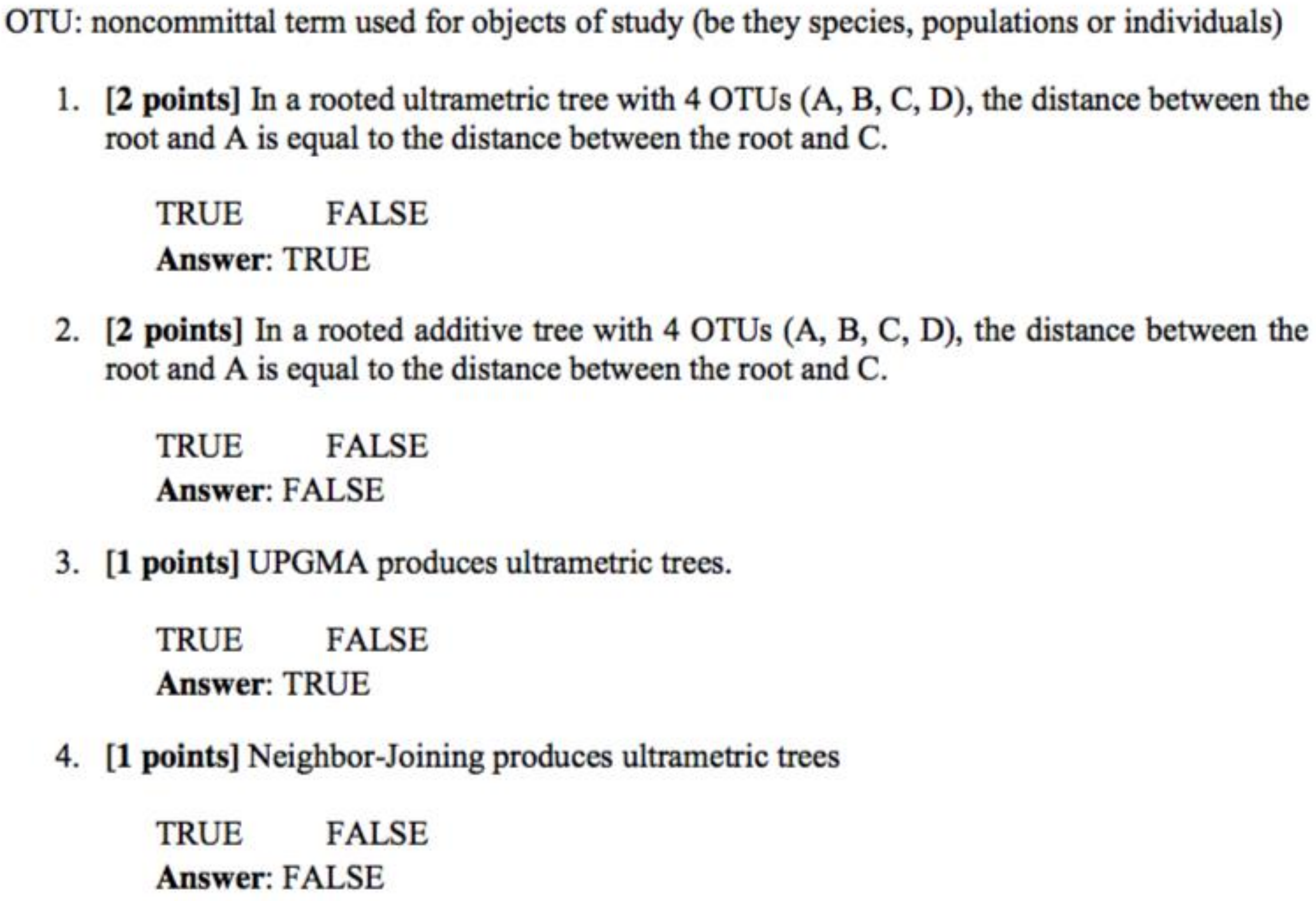
d)

11. (1 point) Most parsimonious tree: Suppose we are given the following four sequences:

```
1: G  A  G  T  G  A
2: A  T  A  T  C  A
3: C  C  G  T  G  G
4: A  G  A  T  C  G
```

We are trying to compute the most parsimonious unrooted phylogenetic tree showing the evolutionary relationship amongst these four sequences. Using the tree structure given below, place the taxon in four leaves (P, Q, R and S), so that it results in a maximum parsimonious tree with minimum number of mutations (assume no insertions/deletions in sequences). Infer the two ancestral sequences (U and V) and show the total number of mutations along each edge that is needed to explain the tree.





OTU: noncommittal term used for objects of study (be they species, populations or individuals)

1. **[2 points]** In a rooted ultrametric tree with 4 OTUs (A, B, C, D), the distance between the root and A is equal to the distance between the root and C.

   TRUE FALSE

   Answer: TRUE

2. **[2 points]** In a rooted additive tree with 4 OTUs (A, B, C, D), the distance between the root and A is equal to the distance between the root and C.

   TRUE FALSE

   Answer: FALSE

3. **[1 points]** UPGMA produces ultrametric trees.

   TRUE FALSE

   Answer: TRUE

4. **[1 points]** Neighbor-Joining produces ultrametric trees

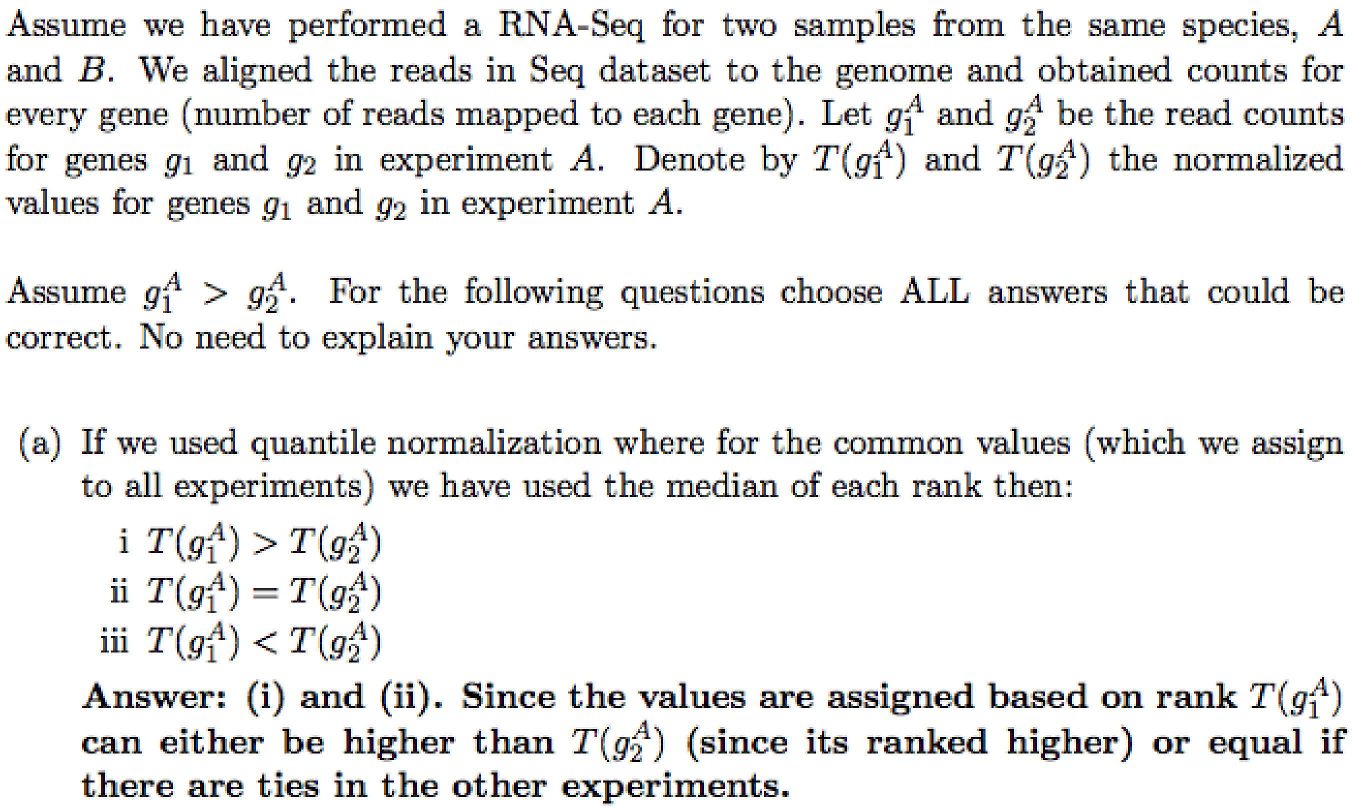
   TRUE FALSE

   Answer: FALSE

Gene Expression/RNAseq

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.



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.

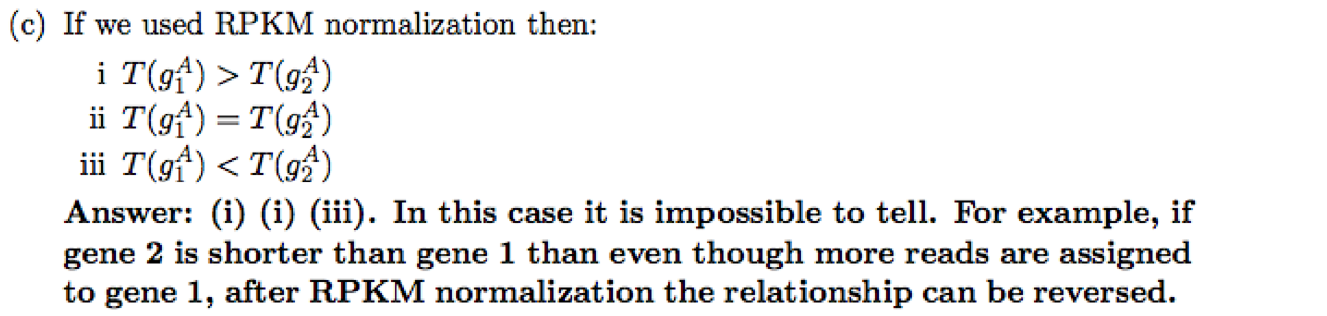
(a) If we used quantile normalization where for the common values (which we assign to all experiments) we have used the median of each rank then:

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.



(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.

Clustering

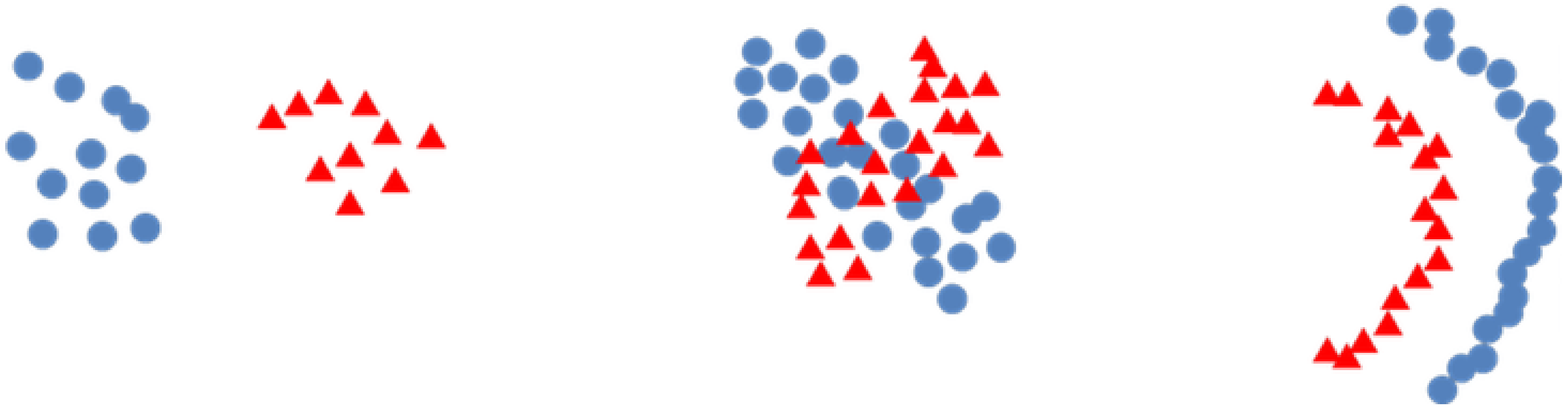


Figure: Three clustering results for the question below.

Select all the clustering method(s) that will lead to the results in the Figure above. Fill in the table below by marking T if the clustering method can lead to these results and F if it cannot.

	Figure (a)	Figure (b)	Figure (c)
Gaussian mixture model	**T**	**F**	**F**
k-means	**T**	**F**	**F**
Hierarchical clustering with single linkage	**T**	**F**	**T**

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

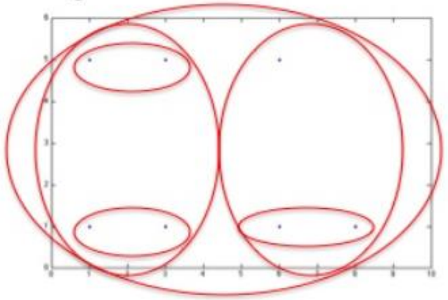


Figure 8.1.b

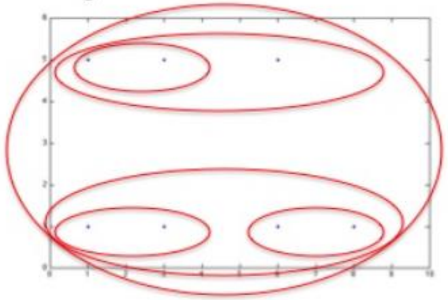
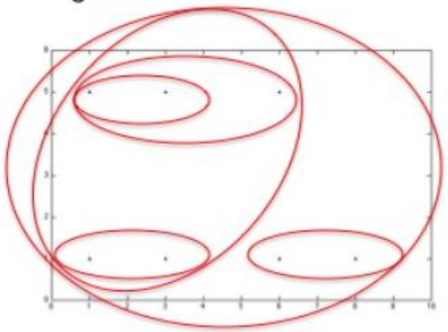
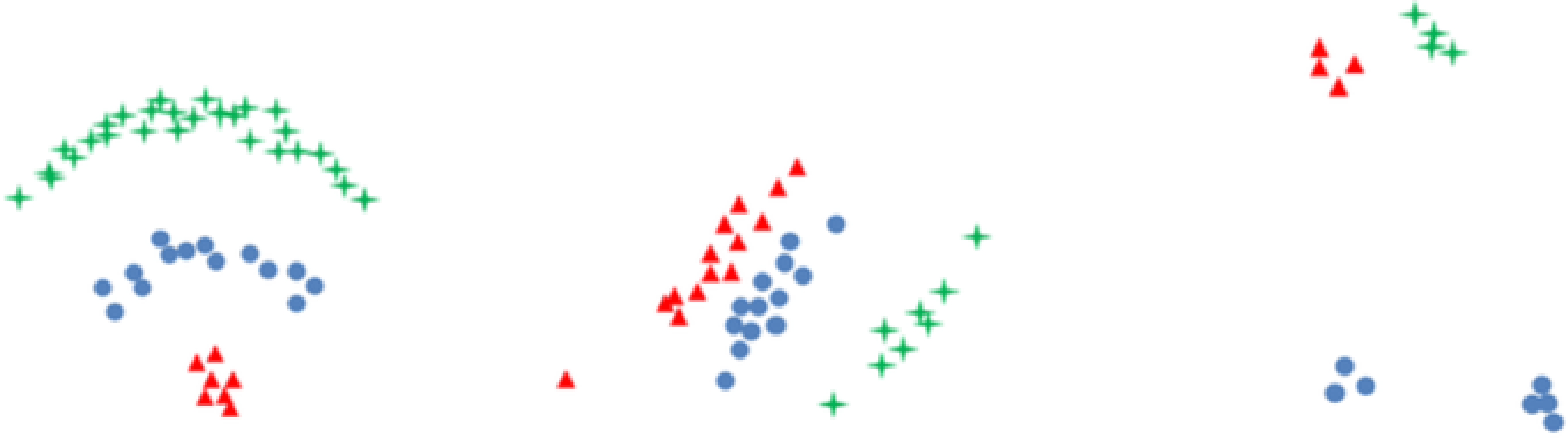


Figure 8.1.c



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



(a) (b) (c)

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**

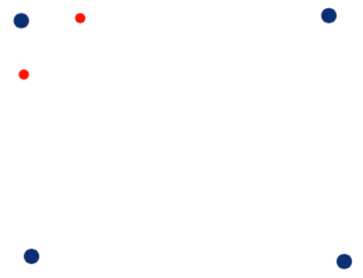
2. [6 Points] We performed experiments in which we tested 50 biological human samples using microarrays (the arrays profile thousands of genes). For two genes, A and B we have 10 missing values for each across all samples (the missing values may come from different samples, so gene A may have a missing 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.

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.

Other

13. (1pt) The genotypes of father, mother, and their two children are given as below. 1 and 0 represent two alleles, and X represents missing values.

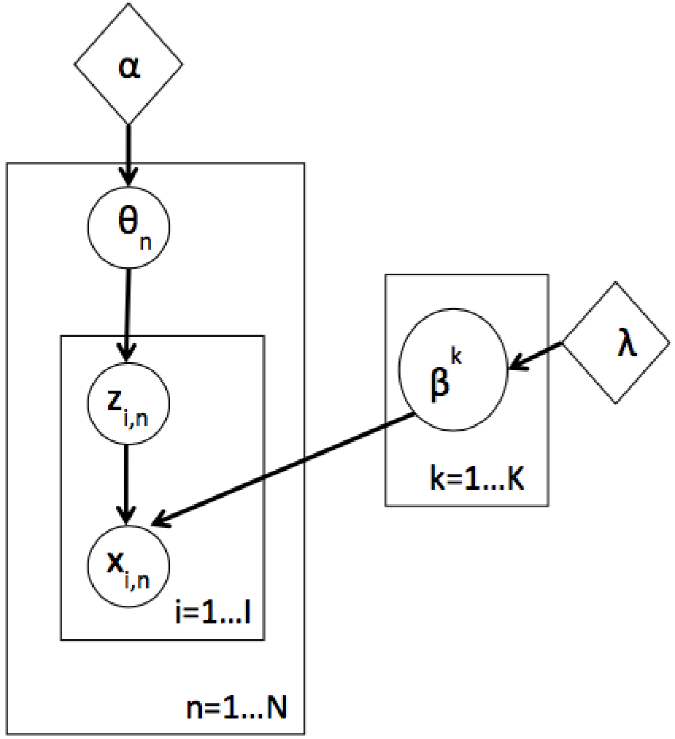
Father	0 1 0 0 0 0 0 0 0
	0 1 1 1 1 1 1 1 0
Mother	1 0 0 0 1 0 1 0 0
	0 0 1 0 0 1 0 0 0
Child 1	0 1 0 0 1 X 1 1 0
	1 0 0 0 0 X 0 0 0
Child 2	0 1 1 1 0 0 0 0 0
	1 0 0 0 1 0 1 0 0

(a) (0.5 point) Where are the recombination sites?

(b) (0.5 point) Can you infer the missing genotypes of the loci marked as ‘X’?

14. (1.5+1.5=3 points) Assume that we fit an admixture model (as shown below) to the genotype data collected from N individuals at I loci (i.e., I genetic markers), with the number of populations K=3 for Asian, Caucasian, and African populations. We use the MCMC sampling algorithm to learn the unknown parameters. After running the sampling algorithm for a large number of iterations until convergence, we summarize the parameter estimate as the single sample with the highest posterior probability. Given this sample, describe how you would answer the following questions.

(a) (1.5 points) Does the region of the genome that contains *LCT* gene exhibit a certain population structure? (Assume that there are 5 genetic markers covering the region of *LCT* gene.)



α, λ: parameters of the prior distribution
$x_{i,n}$: genotype of the nth individual at the ith locus
$z_{i,n}$: the population label of the ith locus of the nth individual
θ_n: a vector of length K for the proportions of populations in the nth individual's genome
β^k: a vector of length I for the allele frequencies in the kth population

(b) (1.5 points) Does an individual who identified himself as Asian have any African ancestry in his genome?

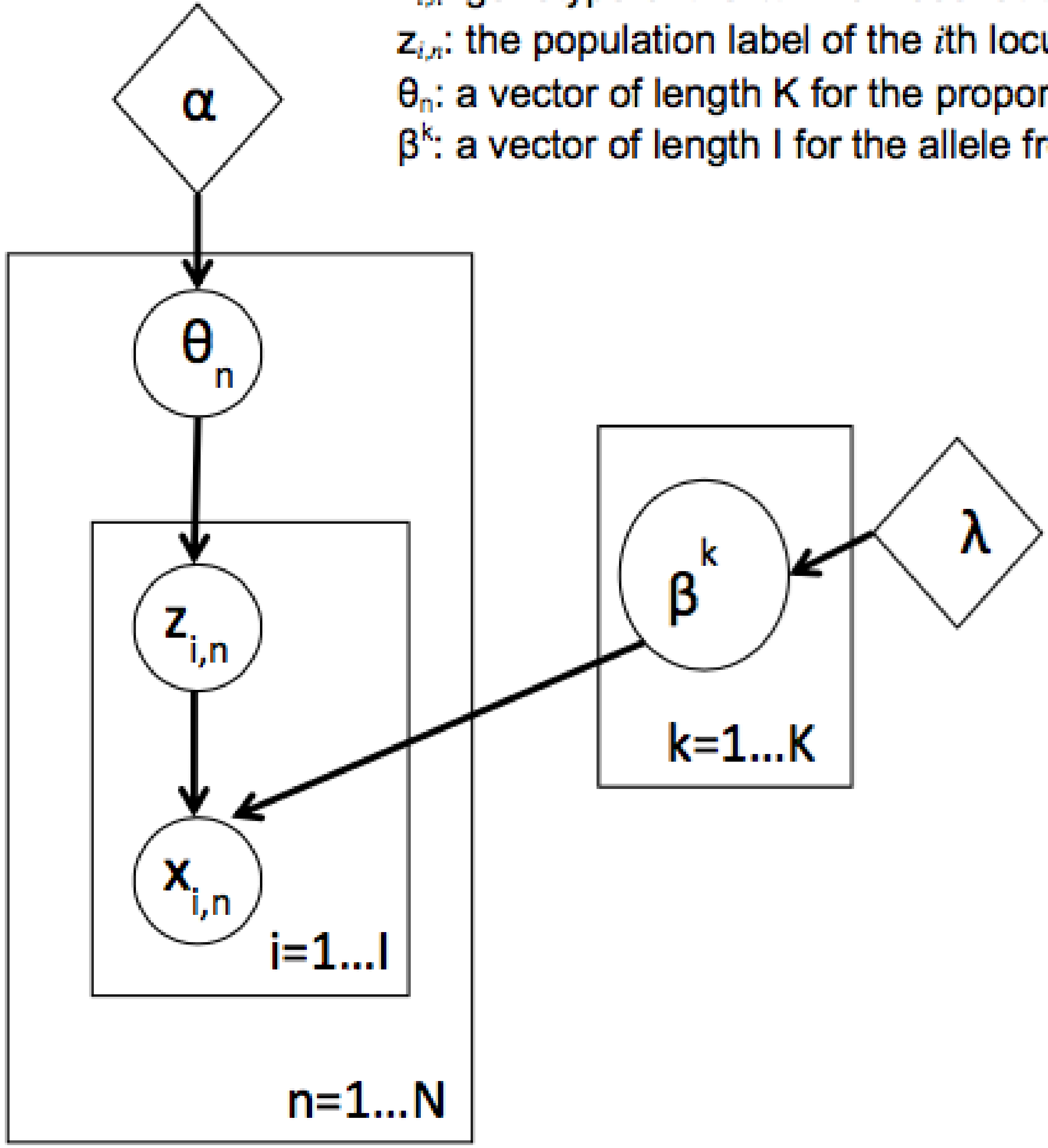
α, λ: parameters of the prior distribution

$x_{i,n}$: genotype of the nth individual at the ith locus

$z_{i,n}$: the population label of the ith locus of the nth individual

θ_n: a vector of length K for the proportions of populations in the nth individual's genome

β^k: a vector of length I for the allele frequencies in the kth population



(b) (1.5 points) Does an individual who identified himself as Asian have any African ancestry in his genome?

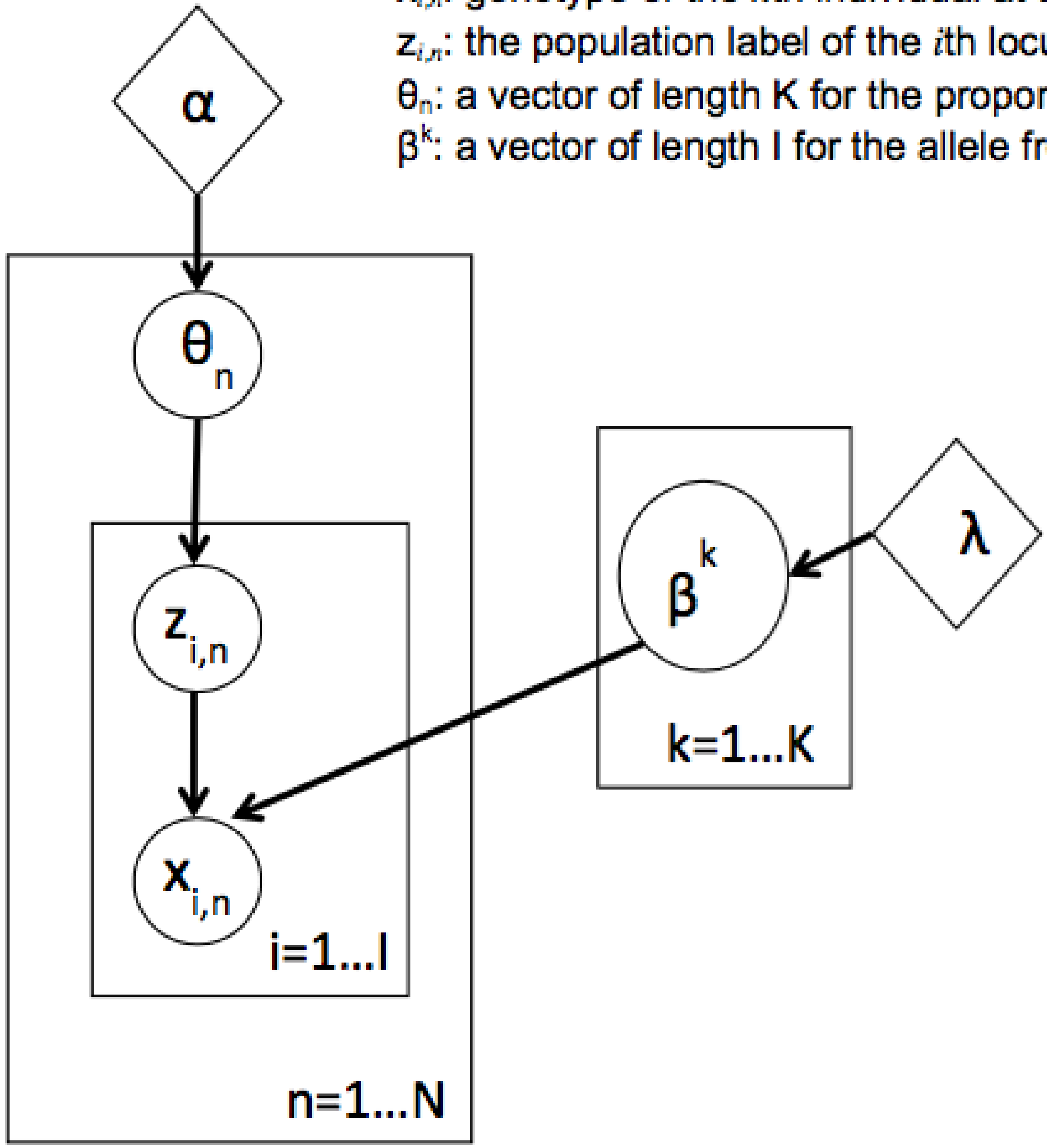
α, λ: parameters of the prior distribution

$x_{i,n}$: genotype of the nth individual at the ith locus

$z_{i,n}$: the population label of the ith locus of the nth individual

θ_n: a vector of length K for the proportions of populations in the nth individual's genome

β^k: a vector of length I for the allele frequencies in the kth population



Population Genetics

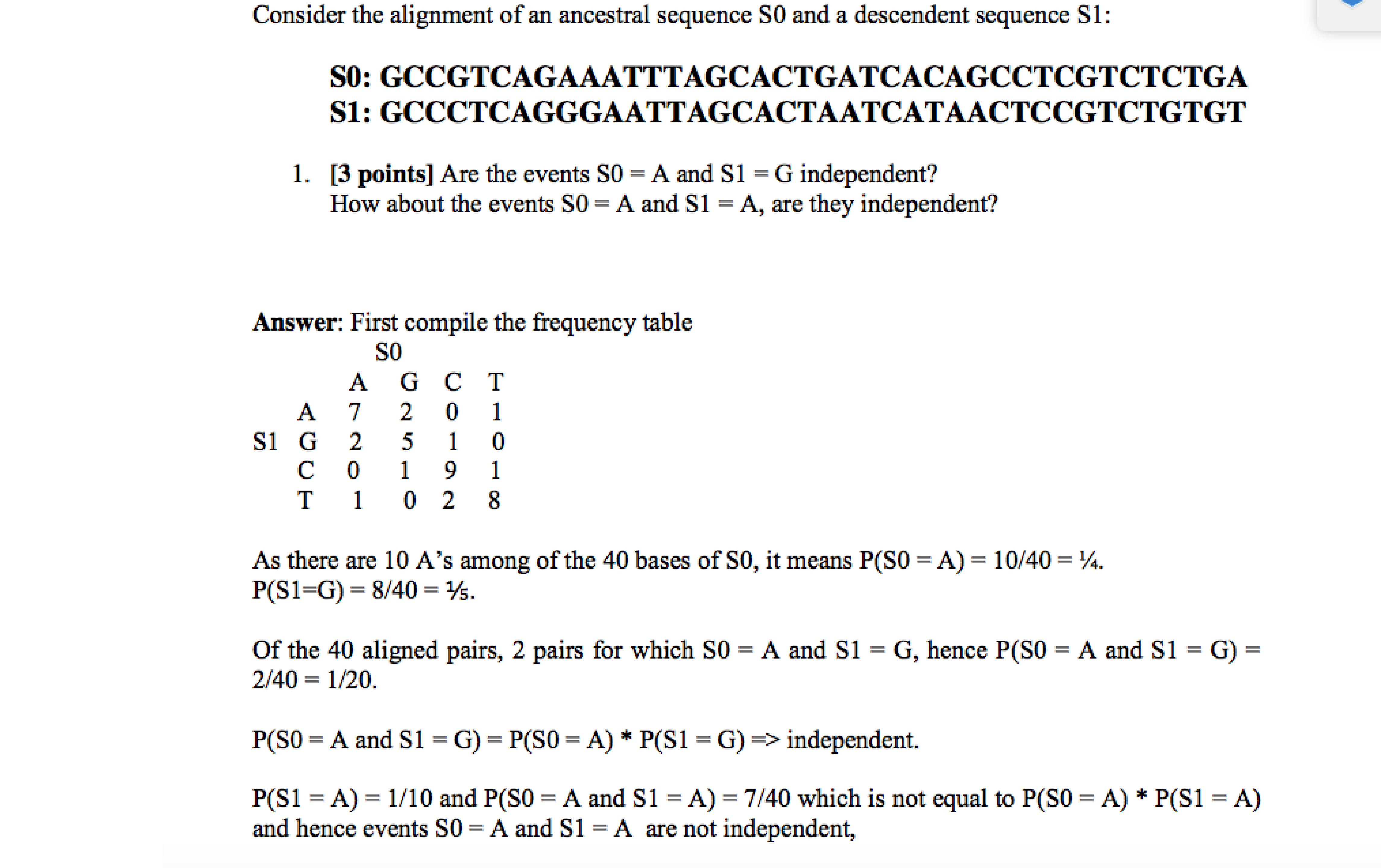
4. (1 point)

A) What is the role of methylation in gene expression regulation?

B) Where do you find CpG islands the most – your choices are: upstream regions of transcription start site; downstream of 3' UTR; near the first intron; at exon-intron locations.

2. **[6 points]** Let's make up a DNA score matrix where we want to optimize the matrix for finding 88% identity elements. Assume all mismatches are equally probable and the composition of both alignments and background sequences is uniform at 25% for each nucleotide. Assuming $\lambda = 0.25$, what is the score you will assign for a match (such as AA, GG, CC, TT) and what is the score you will assign for a mismatch (such as AG, CT and so on) (hint: round up the scores where convenient).

Answer:
Match probability: set p_AA and so on = 0.22
Mismatch probability: set p_AG and so on = 0.01 for each of 12 mismatches
Background probability = 0.25
Match score = ¼ log (0.22/(0.25^2)) = ~5
Mismatch score = ¼ log(0.01/(0.25^2)) = ~ (-7)



Consider the alignment of an ancestral sequence S0 and a descendent sequence S1:

S0: GCCGTCAGAAATTTAGCACTGATCACAGCCTCGTCTCTGA
S1: GCCCTCAGGGAATTAGCACTAATCATAACTCCGTCTGTGT

1. **[3 points]** Are the events S0 = A and S1 = G independent?
   How about the events S0 = A and S1 = A, are they independent?

Answer: First compile the frequency table

		S0			
		A	G	C	T
	A	7	2	0	1
S1	G	2	5	1	0
	C	0	1	9	1
	T	1	0	2	8

As there are 10 A's among of the 40 bases of S0, it means P(S0 = A) = 10/40 = ¼.
P(S1=G) = 8/40 = ⅕.

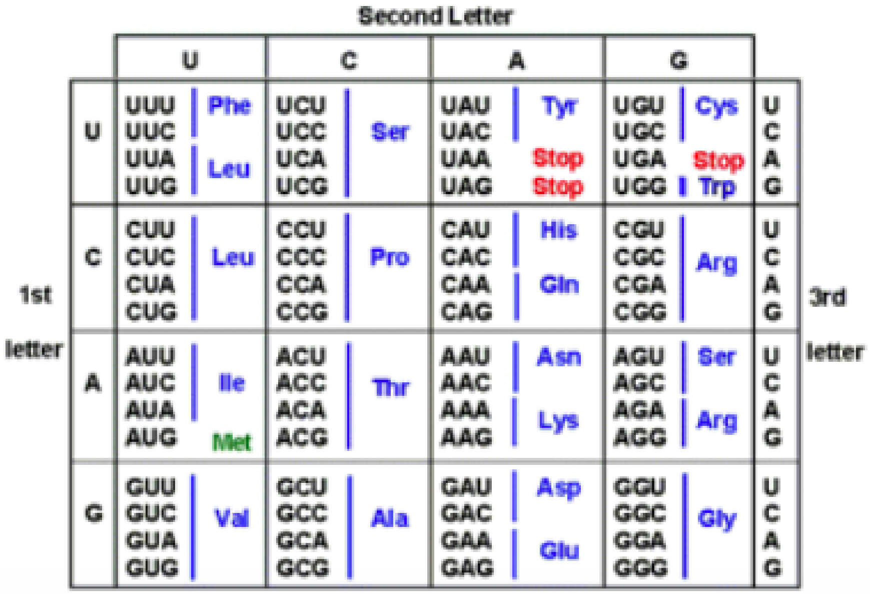
Of the 40 aligned pairs, 2 pairs for which S0 = A and S1 = G, hence P(S0 = A and S1 = G) = 2/40 = 1/20.

P(S0 = A and S1 = G) = P(S0 = A) * P(S1 = G) => independent.

P(S1 = A) = 1/10 and P(S0 = A and S1 = A) = 7/40 which is not equal to P(S0 = A) * P(S1 = A) and hence events S0 = A and S1 = A are not independent,

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:



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