CBB752; Quiz2; Name:

1. List two reasons why modeling (in general) is important.
2. List two reasons why modeling (in general) can be challenging.
3. Give an example of how biomolecular modeling is used or could someday be used to improve patient health.
4. Why is dimensionality reduction useful?
5. What is the difference between supervised and unsupervised learning? When would you use one versus the other?
6. What is the kernel trick in SVM? What are its advantages and pitfalls?
7. What makes a good decision tree rule?
8. Describe the main steps of Expectation Maximization for predicting transcription factor binding motifs and their instances.
9. What is the special statistical interpretation of the *first* principle component as determined by Principle Component Analysis?
10. What is meant by the *ruggedness* of a protein’s folding energy landscape? Are diseases of misfolded proteins more likely to arise from proteins with rugged or with smooth energy landscapes? Explain.
11. The radius of gyration Rg of a protein with N residues is approximated by Rg ~ Nν. Rank the following settings from smallest to largest values of ν: a) fully extended chain, b) random walk, and c) self-avoiding random walk
12. Below, a Ramachandran plot based on the original theoretical calculations is shown. Describe roughly what is the difference between the yellow and red regions of the plot?



1. Which of the following observations about the mutational properties of a gene from a *germline* genome sequencing effort of *healthy* humans like the Exome Aggregation Consortium provides the best evidence that the gene is essential for human survival and reproduction?
2. Low synonymous [aka silent] mutation rate
3. Low loss of function [aka premature stop] mutation rate
4. Low ratio of the loss of function mutation rate to synonymous mutation rate
5. High synonymous [aka silent] mutation rate
6. High loss of function [aka premature stop] mutation rate
7. High ratio of the loss of function mutation rate to synonymous mutation rate
8. The Lennard-Jones potential is written below. What physical force/setting accounts for the repulsive potential in the Lennard-Jones potential? The attractive force? What is the difference in the definitions of r and sigma?



1. Suppose a typical base in a whole genome sequencing experiment is overlapped by 30 reads (half of which cover the maternal chromosome, half the paternal chromosome) and that the reads are 90 bp long, uniformly distributed throughout the genome, and single-ended (i.e. not paired-end). How many split reads would you expect could contribute to the identification of a 3,000 bp heterozygous deletion within a fully mappable region? Assume that split reads are eligible so long as at least one nucleotide is present from both the portion 5’ and 3’ to the portions. Explain your reasoning for partial credit.
2. 15
3. 30
4. 90
5. 900
6. 1500
7. 1800
8. 3000
9. 90000

Hint: it may help to consider a simple diagram like the scaled-down version below, where the sequence TTT has been deleted from the paternal chromosome. Notice that in this diagram, although all bases are covered by six reads, only some bases are overlapped by the 5’ end of a read).

Maternal chromosome

Maternal: 

Paternal:

Paternal chromosome

1. Which SV detection methods could detect the existence of the (red) insertion that has occurred in Sequence 2, with respect to Sequence 1? Select all that apply:
Hint: The sequence is repetitive. Explain your reasoning for partial credit.
2. Split reads with 75-bp long reads
3. Read depth analysis
4. Paired-end read analysis, with a 300 bp insert size and 75 bp-long reads

Sequence 1:

1000 ACTGACGCTA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA

1100 TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA

1200 TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA

1300 TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA

1400 TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA

1500 TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA

1600 TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA

1700 TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA

1800 TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA GTAGACAGTTC

Sequence 2:

1000 ACTGACGCTA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA

1100 TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA

1200 TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA

[insert1]TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA

1300 TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA

1400 TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA

1500 TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA

1600 TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA

1700 TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA

1800 TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA TAGCAGCGAA GTAGACAGTTC