Due on Monday, December 2nd

Name:
Are you a graduate or undergraduate student? Please circle one.

Bioinformatics Take Home Test #8

(This is an open book exam based on the honors system -- you can use notes, lecture notes, online manuals, and text books.
Teamwork is not allowed on the exams, write down your own answers, do not cut and paste from webpages.
If your answer uses a citation, give the source of the quoted text.)

Notes on Formatting Quizzes: Please make sure each answer is only on one page, by using page breaks. Splitting an answer onto two pages tend to lead to grading errors.

Please do not write or type in font smaller than 12 point or write in cursive.

If you submit your quiz via email, please remove the instructions and extras (blank lines, alternative answers for multiple choice questions) from your document, so that only your answers, a minimal amount white space, and optionally the questions, are left.

1. 1pt **True/False** HMMER is a suit of search program similar to psiblast to find distant homologs to a query sequence.

2. 1pt **True/False** It is NOT possible to align nucleotide sequences based on a protein alignment.

3. 1pt **True/False** Amino acids on the outside of virus capsids interact with the immune system and are therefore often are under positive selection, to evade the immune system.

4. 1pt **True/False** The chances of attaining false negatives when performing a PSI blast are decreased as compared to a normal blast search.

5. 1pt **True/False** PSSMs ARE corruptible, therefor the E-value of a match obtained in a later iteration of a PSIBlast search is NOT a good measure for obtaining a match of this quality due to chance.

6. 1pt **True/False** You should NOT run psi-blast for more than 5 iterations to minimize corruption of the PSSM.

7. 1pt **True/False** A PSI-Blast search is done using an ATPsynthase catalytic subunit as query. In the 5th iteration a match to myosin with an E-value of 10^-15 is reported. This demonstrates that at least a portion of ATPsynthase catalytic subunit is homologous to part of the myosin molecule.

8. 1pt **True/False** Turning on the filter for low complexity slows down corruption of a PSSM.

9. 1pt **True/False** Because PSI blast builds a Scoring Matrix from many homologous sequences, low complexity sequences pose a problem in PSI blast, by corrupting the PSSM.

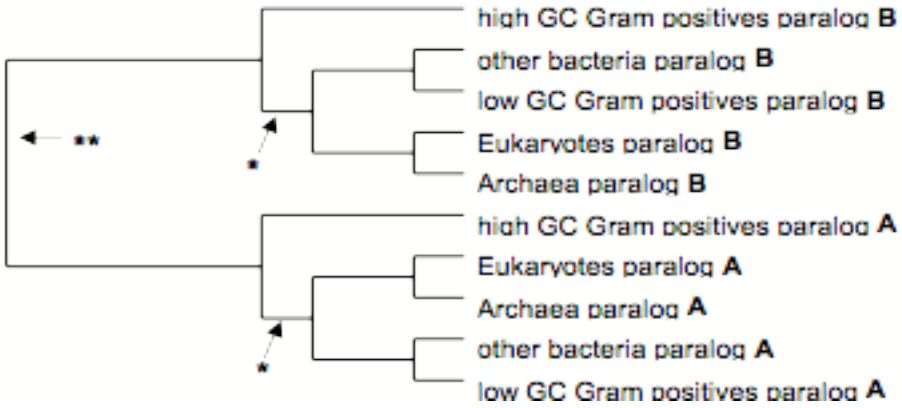
10. 1pt Regarding types of error in a normal BLAST search, which is correct:
    A) Unrecognized false positives are a serious problem in identifying homologs
    B) False negatives occur frequently.
    C) Most false negatives can be identified, if one turns on the filter for low complexity

11. 1pt In applying the Bayesian framework to the analysis of molecular data, which of the following are true?
    A) The probability of the model given the data is assessed
    B) The probability of the data given the model is assessed
    C) This is the same as maximum likelihood analysis
    D) Both a) and c)
    E) Both b) and c)

12. 1pt In applying the Maximum Likelihood principle to the analysis of molecular data, which of the following are true?
    A) The probability of the model given the data is assessed
    B) The probability of the data given the model is assessed
    C) This is the same as Bayesian framework analysis
    D) Both a) and c)
    E) Both b) and c)

13. 1pt What outgroup could you use to root the tree of life?
    A) Giardia
    B) Any randomly chosen taxon on the tree
    C) Viruses
    D) Genes derived from ancient gene duplication
    E) A stone
    F) There is no way to choose an outgroup for the tree of life

14. 1pt Analyzing a gene family you find that two paralogs each are present in all three domains of life. The two groups of paralogs are joined by a branch that connects the bacterial domains



Ignoring horizontal gene transfer, this tree would suggest:
A) that the bacteria are a monophyletic (and not a para- or polyphyletic) group
B) that the last common ancestor of the two paralogous types of genes existed in a bacterium; and that ignoring horizontal gene transfer the most recent common ancestor of the three domains of life was a bacterium.
C) that the last common ancestor is placed into the tree of life between the bacteria on one side and the archaea and eukaryotes on the other.

15. 1pt **True/False** If the branches in the tree in the previous question that are indicated by *s have only 55% bootstrap support, (even though the branch indicated by ** is strongly supported), this would indicate that the conclusion drawn in question (10) is not strongly supported.

16. 1pt **True/False** If the branches in the above tree that are indicated by **s have only 55% bootstrap support, (even though the branch indicated by * is strongly supported), this would indicate that the conclusion drawn in question (10) is not strongly supported.

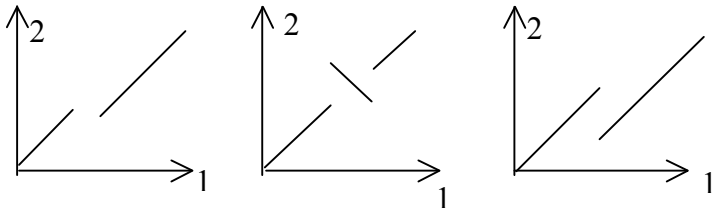
17. 1pt When considering data obtained from flipping one coin three times and obtaining all heads, what will Bayesian Inference calculate? (Consider that there are three models possible for this coin toss and that the first one is twice as likely as the other two: 1. A fair coin model. 2. A coin with both sides heads. And 3. A coin with both sides tails.)
    a. The probability of obtaining all heads, averaged over all possible models, with respect to the prior of how likely each model is; (i.e. ((.5)^3 * .5) + (.25 * 1.0) + (.25 * 0))
    b. The probability of obtaining all heads, given the model that maximizes this probability (i.e. 100% and it will always chose the second model)
    c. The probability of obtaining all heads when using a fair coin (i.e. (.5)^3 * .5))
    d. Bayesian Inference is not applicable to coin toss data, only nucleotide or amino acid sequence data can be used.

18. 1pt When considering data obtained from flipping one coin three times and obtaining all heads, what will maximum likelihood calculate? (Consider that there are three models possible for this coin toss and that the first one is twice as likely as the other two: 1. A fair coin model. 2. A coin with both sides heads. And 3. A coin with both sides tails.)
    a. The probability of obtaining all heads, averaged over all possible models (i.e. ((.5)^3 * .5) + (.25 * 1.0) + (.25 * 0))
    b. The probability of obtaining all heads, given the model that maximizes this probability (i.e. 100% and it will always chose the second model)
    c. The probability of obtaining all heads when using a fair coin (i.e. (.5)^3 * .5))
    d. The probability of obtaining all heads, without considering possible models. This is possible because a robot is used to explore probability space.
    e. Maximum likelihood is not applicable to coin toss data, only nucleotide or amino acid sequence data can be used.

19. 1pt When considering the above coin toss experiment in a Bayesian framework, which other quantity do you need to know to arrive at a conclusion?
    a. the posterior probability
    b. the prior probability
    c. the joint probability of the three possible models

20. 1 pt Your health care provider performed a test for a rare genetic disease on you. The test gives a false positive result only in 1 out of a 1000 cases. The rate of false negatives is zero (if you have the disease, the test will detect it). About 1 in a million Americans have the disease. Your test is returned positive. What is the probability that you actually have the disease? (You might get partial credit, if you show your reasoning.)

21. 3 pt You compare two genomes from closely related organisms using a Genome plot. Which processes that could have given rise to the following:
(Genome 1 is on the ordinate (X), Genome 2 is on the abscissa (Y))



I) II) III)

Panel I)

A) Genome 1 has an insertion that is not present in genome 2, or genome 2 had a deletion.
B) Genome 2 has an insertion that is not present in genome 1, or genome 1 had a deletion.
C) Genome 1 possesses a region that is duplicated.
D) Genome 1 or genome 2 underwent an inversion

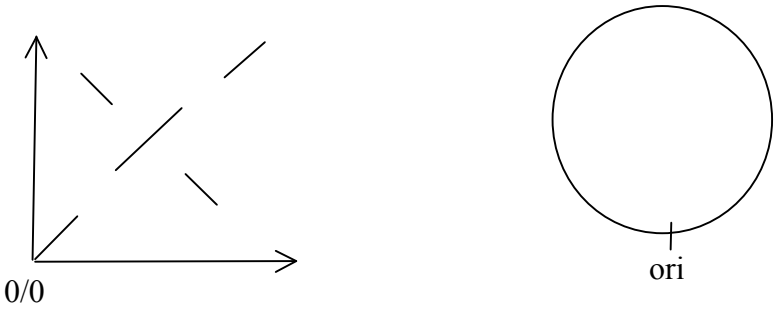
Panel II:

A) Genome 1 has an insertion that is not present in genome 2, or genome 2 had a deletion.
B) Genome 2 has an insertion that is not present in genome 1, or genome 1 had a deletion.
C) Genome 1 possesses a region that is duplicated.
D) Genome 1 or genome 2 underwent an inversion.

Panel III:

A) Genome 1 has an insertion that is not present in genome 2, or genome 2 had a deletion.
B) Genome 2 has an insertion that is not present in genome 1, or genome 1 had a deletion.
C) Genome 1 possesses a region that is duplicated.
D) Genome 1 or genome 2 underwent an inversion.

22. 1 pt Assuming that the origin of replication is at (0/0) in the coordinate system, and that you compare two closely related genomes. The organisms in question have circular genomes. In the sketched genomes on the right, indicate where the rearrangement event(s) took place that could have given rise to the genome plot on the left



For Graduate Students:

23. 2pts You want to find all copies of a transposase gene in a particular microbial genome. A blastp search of the annotated genome resulted in 12 significant hits. A PSI-blast search of the annotated genome using a PSSM calculated from first searching nr for 5 generations resulted in 16 significant matches. A PSI-blast search of the 6 frame translation of the genome gives 42 significant matches.
Explain why there are additional matches obtained in PSI blast searches? Which approach also finds decaying transposase genes in the genome?

24. 2pt – Why is a Bayesian consideration advantageous, as opposed to the Maximum Likelihood estimate, of the coin toss example used in previous questions?