Name:
Due on Wensday, December 7th

Bioinformatics Take Home Exam #9

Pick one most correct answer, unless stated otherwise!

1. What process brought 2 divergent chlorophylls into the ancestor of the cyanobacteria, which enabled oxygen-producing photosynthesis?
   A. Gene duplication
   B. Genome duplication
   C. Meiotic hybridization between species
   D. Horizontal gene transfer
   E. Long branch attraction

2. What happened to allow some Asian people to be able to digest algae?
   A. Gene duplication, followed by neofunctionalization
   B. HGT from a red and brown algal parasites to symbionts in the human gut
   C. HGT to humans from the red algae
   D. Humans gained a new symbiont in their gut from the red algae, that can digest algae
   E. None of the above

3. Which of the following is true regarding HGT?
   A. It is a process through which genes enter a genome, without being inherited parentally
   B. It can lead to important biological innovations
   C. A transferred gene can be inherited parentally, so that a clade of organisms all share the same inherited ancient HGT.
   D. It is more common in Bacteria than in humans.
   E. All of the above.

4. Which processes allow favorable genetic changes to be combined into the same individual, speeding up the rate of evolution?
   A. Gene duplication and neofunctionalization
   B. Genetic drift
   C. Punctuated equilibrium
   D. Sex and HGT
   E. None of the above.

5. Why is PSI-BLAST more prone to false positives than normal blast, especially as the number of iterations increases?
   A. As the program narrows down possible homologous proteins, each new protein is more likely to be recognized as homologous simply because there are less proteins remaining in the pool.
   B. With every iteration, there is the possibility that false positives are incorporated into the position specific scoring matrix.

C. The program creates a more comprehensive profile-HMM as more sequences added with each iteration, increasing the likelihood of false positive because more protein domains are being scored.
D. All of the above

6. What does the PSI in PSI-BLAST stand for?
   A. Phylogenetic Sequence Initiative
   B. Phylogeny Stimulated Image
   C. Position-Specific Iterated
   D. Position-Sequence Initiated
   E. None of the above

7. What are false negatives?
   A) non-homologous sequences that are listed as matches
   B) homologous sequences that are not listed as matches
   C) homologous sequences that are listed as matches
   D) non-homologous sequences that are not listed as matches

8. **True/**False PSI Blast has more false positives than normal Blast because of profile corruption.

9. **True/**False **–** The E-value reported as the result of the last iteration of the PSI blast search indicates how many matches to the initial query sequence are expected due to chance alone (i.e. in the absence of shared ancestry)

10. Which program is specifically designed to detect distant relationships between proteins?
    A) Seaview
    B) PSI-BLAST
    C) njplot
    D) Swiss-pdb
    E) None of the above

11. What is a PSSM?

12. Psi-Blast can use:
    a. existing multiple alignment
    b. use RPS-Blast to search a database of PSSMs
    c. Both A & B
    d. Neither A or B

13. **True** or **False -** HMMER is a free and commonly used software package used to identify homologous protein or nucleotide sequences.

14. **True or False -** PSMs and Profile-HMMs are derived from a set of aligned sequences that are thought to be homologous. They have become an important part of many software tools for computational motif discovery.

15. Which computing program uses MCMC algorithms?
    A. MrBayes
    B. Seaview
    C. PSI-BLAST
    D. Njplot
    E. Swiss pdb

16. **True or False -** According to the modern synthesis mutations play little to no role in directing the evolution of organisms; selective processes play the main role in evolution.

17. Why is it not recommended to conduct PSI-BLAST for more than 5 iterations?

18. What can one do to minimize the possibility to corrupt a PSSM with non-homologous sequences?
    A) Use a smaller databank
    B) Use a smaller cut-off value for inclusion of database hits in the next PSSM
    C) Filter the query for regions of low complexity
    D) Limit the number of iterations
    E) B-D
    F) All

19. **True or False -** There are fewer false negatives with PSI blast than with normal blast.

20. **True or False -** MrBayes is extremely reliable in predicting the correct tree. As a result, the support values that the program produce are conservative and should be considered more reliable than bootstral support values.

21. **True or False -** psiBLAST is the algorithm that HMMER uses to find distant homologs to a query sequence.

22. Which program can align nucleotide sequences based on a protein alignment?
A. MrBayes
B. Seaview
C. psiBLAST
D. Clustalo
E. Cluster

23. Not doing which of the following commands at the beginning of a session on the cluster is considered rude?

A. ls
B qlogin
C qstat
D cd
E qdel

24. Why might amino acids on the outside of virus capsids be under positive selection?

A. They interact with the immune system and need to change to evade recognition and capture
B. These positions are under strong selection to maintain function, because they are important to the virus
C. They interact with the host DNA and need to change as the host evolves
D. Binding of host antibodies triggers mutations in the virus
E. All the above

25. What allowed Walter Fitch to beat of the CDC in picking the strain of flu to vaccinate, year after year, until the CDC finally started doing things his way?

A. He had a huge team of researchers on the project, while the CDC just had one retired professor on the project
B. He had enormous computing power at his disposal, while the CDC was using a pocket calculator
C. He had new, modern laboratory equipment, allowing him more to obtain more accurate sequences than the CDC
D. He had decades of knowledge of evolution and with it was able to develop the novel idea of using an evolutionary pattern he spotted in flu phylogenies to predict next years' epidemic
E. A, B, and C

26. The Bayesian framework tries to calculate

A) the probability of the model given the data
B) the probability of the data given the model
C) This is the same as maximum likelihood analysis
D) Both a) and c)
E) Both b) and c)

27. In Maximum Likelihood

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)

28. When considering data obtained from flipping one coin four times and obtaining all tails, what will the maximum likelihood approach calculate? (Consider that there are three models possible for this coin toss: 1. A fair coin model. 2. A coin with both sides heads. And 3. A coin with both sides tails. Priors are 1. 99.8%, 2. 0.1%, 3. 0.1%)

A. The probability of obtaining all tails, averaged over all possible models (i.e. ((.5)^4 * 0.998) + (0 * 0.001) + (1.0 * 0.001))

B. The probability of obtaining all tails, given the model that maximizes this probability (i.e. 100% and it will always choose the third model)

C. The probability of obtaining all tails when using a fair coin (i.e. (.5)^3 * 0.998))

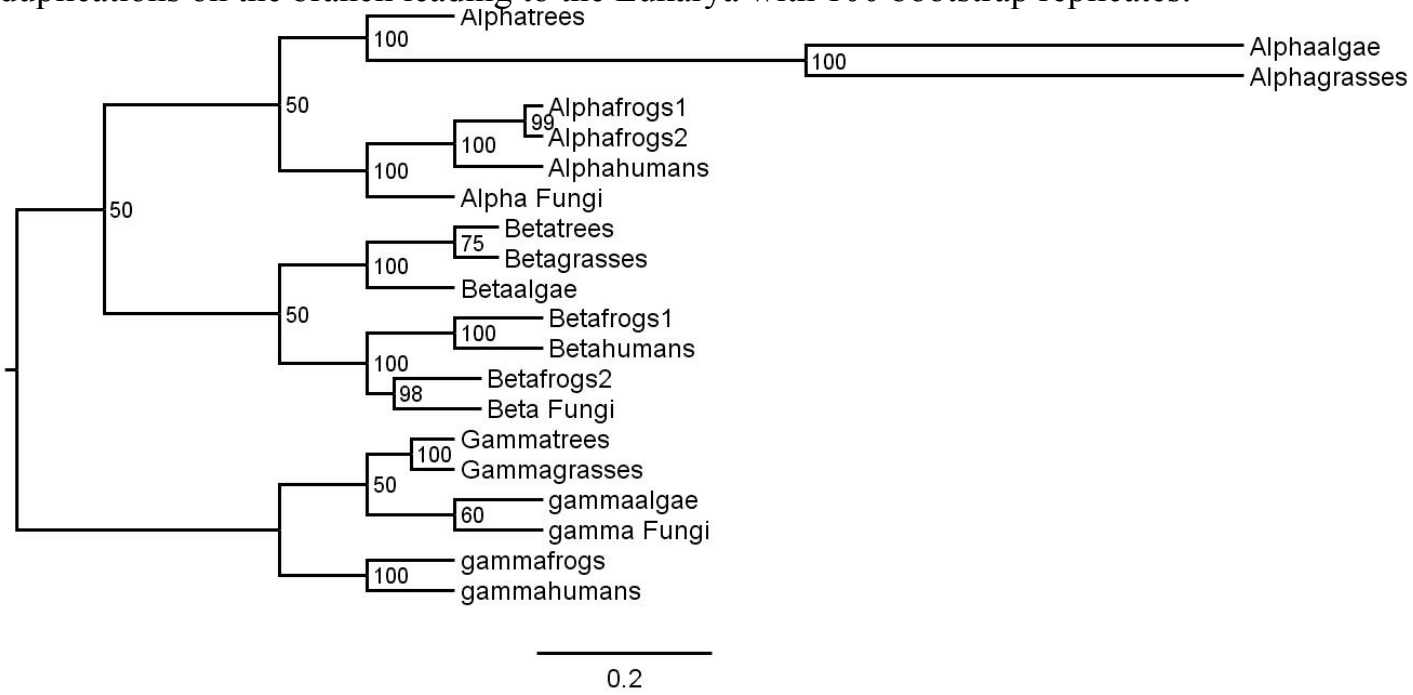
D. The probability of obtaining all tails, without considering possible models. This is possible because a robot is used to explore probability space.

E. Maximum likelihood is not applicable to coin toss data, only nucleotide or amino acid sequence data can be used.

29. What is NOT considered an example of horizontal gene transfer producing new pathways in organisms?

a) The evolution of archaeal ATP synthase
b) Oxygen-producing photosynthesis
c) Acetoclastic methanogenesis
d) Acetyl-CoA assimilation
e) All are examples

For following questions use this diagram showing three genes related by ancient gene duplications on the branch leading to the Eukarya with 100 bootstrap replicates:



30. The following is a list of phenomena occurring that makes this tree deviate from the species tree. Match them to the deviations below:

i. In-paralogs resulting from a recent gene duplication
ii. Horizontal gene transfer
ii. Long branch attraction
iv. Difficult to tell, but could be a result of lack of resolution
    A. Beta2frogs grouping with Betafungi
    B. Alphaalgae grouping with Alphagrasses
    C. Gammaalgae grouping with Gammafungi
    D. Two copies of the alpha gene being present in frogs

31. Might whole genome duplication have played a role in forming this tree (Yes or No)? If yes, how many rounds is most likely to have occurred?

32. Is it possible that gammaalgae groups with the 2 gamma plant sequences in 50% of bootstrap samples? If not, what is the maximum number of samples?

33. According to the Alpha and Beta paralogs, where is the root within the Eukarya?

34. Which value of non-synonymous/synonymous rate ratio (+omega or dN/dS) would you expect for a protein-coding gene that encodes an enzyme vital for photosynthesis?
A) 0.01 or smaller
B) about 1.0
C) 1.2 or larger

35. **True or False -** The 60 means that 60% of the time, the two gamma plant sequences (trees and grasses) group with the two gamma animal sequences and all of the alpha and beta sequences.

36. **True or False -** Constructive Neutral Evolution (CNE) emphasizes adaptive mechanisms of evolution

37. Which of the following are not considered methods to detect positive selection?
a. Selective Sweeps
b. Have few alleles present in a population
c. High dN
d. dN/dS larger than 1
e. SNPs in the allele are not in linkage equilibrium
f. all of the above
g) a, c and d

38. **True or False -** the car trunk analogy illustrates that genes can be under purifying selection without increasing the fitness of the individuals carrying the gene.

39. Mitochondial Eve lived:
A) 3.2-4.2 million years ago
B) 750,000 years ago
C) 166-249 thousand years ago
D) 10 thousand years ago
E) 90-100 thousand years ago

Extra Credit

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.

1) Explain why there are additional matches obtained in PSI blast searches?

2) Why is a Bayesian consideration advantageous, as opposed to the Maximum Likelihood estimate, for the coin toss example used in previous question?

3) Molecular evolution of which macromolecule is important in flu vaccine development? (more than one may be correct)

ATPsynthase
DNA
Hemagglutinin
Inteins
Lipid bilayer
Neuraminidase

4) Darwin considered evolution as a slow, rigorous, and gradual process. Describe 3 processes that might lead to the rapid/increased evolution of an organism. Explain your reasoning.