Name:

Bioinformatics Take Home Test #7

Due Date Wednesday 11/18/2013 before class

(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.)

Make sure each answer is only on one page, by using page breaks. Splitting an answer onto two pages leads to grading errors.

Do not write or type in font smaller than 12 point or write in cursive. Doing so will lose you 2 pts.

If you have an emergency and cannot submit a quiz in person, email it in by the start of class on the due date. If you do so, first 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.

Note on Late Quizzes: Late quizzes are an inconvenience and cannot be accepted at all after the answers have been released. If your quiz is submitted within the first 12 hours after the deadline, you will receive 5% off. Each additional 12 hours is an additional 5% off, up until the graded quizzes are returned or the answers released.

All questions worth 1pt unless otherwise stated.

1. Submit your question for Quiz 7 by 5pm on Friday

2. I slice an alignment up by column, put each column in a hat, and pick a column from the hat at random, writing it into a new dataset. I put the column back in the hat and randomly pick another, repeating this process until I have the same number of columns as the original dataset. What have I created?

A. Crap
B. A jackknife sample
C. A single nonparametric bootstrap sample
D. A phylogenetic tree
E. A single parametric bootstrap sample

3. . If evolutionary parameters are estimated from a dataset and associated tree and those parameters are then used to simulate a new dataset, what is this dataset?

A. Crap
B. A jackknife sample
C. A single nonparametric bootstrap sample
D. A phylogenetic tree
E. A single parametric bootstrap sample

4. True/False Parsimony does a better job handling gaps than Neighbor Joining, but Neighbor Joining does better with long branches.

5. Which of the following are reasons a gene tree may not match the species tree?
A. Incomplete lineage sorting
B. Horizontal gene transfer
C. Insufficient phylogenetic signal
D. Long branch attraction
E. All of the above

6. Which of the following is a tree reconstruction artifact?
A. Incomplete lineage sorting
B. Horizontal gene transfer
C. Insufficient phylogenetic signal
D. Long branch attraction
E. All of the above

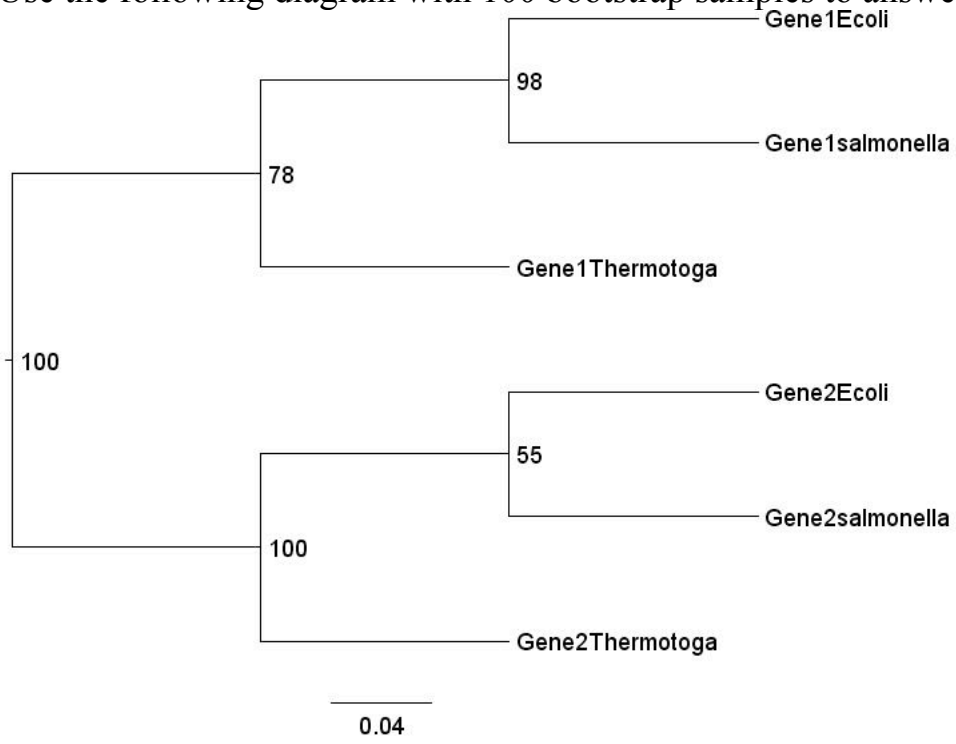
7. Long Branch Attraction is caused by which of the following?
A. Homoplasies resulting from the long branches independently acquiring the same substitution.
B. Alignment programs misalignment sequences to maximize similarity
C. Tree building programs underestimating the number of substitutions occurring
D. All of the above.
E. None of the above.

8. Showing Branch lengths on a tree gives what sort of information?
A. How much evolution has occurred along a branch
B. An indication if long branches might be a problem
C. An indication that a strictly bifurcating tree is not the best model, due to a polytomy
D. All of the above
E. None of the above

9. Which of the following is true with respect to the outgroup of a phylogenetic tree?
A. It represents the ancestor of the ingroup.
B. It is used to inform upon the ancestral state of a given character.
C. It can be a random sequence
D. It does not matter how distantly related it is to the ingroup; distantly related outgroups are just as informative as closely related ones.

10. Which of the following is a factor when calculating the bootstrap support of a branch in a phylogenetic tree?
A. Branch lengths
B. The number of times a split is recovered in a set of bootstrap samples
C. The Gamma parameter and among site rate variation
D. Lineage sorting
E. All of the above

Use the following diagram with 100 bootstrap samples to answer questions 11-28:



11. What does the line at the bottom, Labeled with 0.04 represent?
A. This is meaningless, other than to separate the estimate of 0.04 from the tree
B. The value of the gamma shape parameter estimated for this tree
C. The percent of bootstrap samples expected to produce that split if the sequences were random
D. The scale bar, used to indicate branch length in the tree

12. What does the number 0.04 refer to?
A. The percent of bootstrap samples expected to produce that split if the sequences were random
B. The average number of substitutions per site in a branch of that length
C. The degree to which among site rate variation occurs in the dataset
D. The percent of long branch attraction occurring in the tree
E. None of the above

13. True/False The number 55 indicates that in 55% of the bootstrapped samples Gene2Salmonella groups with Gene2Thermotoga.

14. True/False The number 98 indicates that in 98% of the bootstrapped samples Gene1Salmonella groups with Gene1Ecoli.

15. True/False The number 78 indicates that in 78% of the bootstrapped samples Gene1Thermotoga groups with Gene1Salmonella to the exclusion of Gene1Ecoli.

16. True/False The number 55 indicates that in 55% of the bootstrapped samples Gene2Salmonella groups with the three Gene 2 sequences.

17. True/False The number 78 indicates that in 78% of the bootstrapped samples the three Gene2 sequences group together.

Use the following sequence labels for the following bifurcation table in question 18 and 19: 1. Gene1Ecoli 2. Gene1Salmonella 3. Gene1Thermotoga 4. Gene2Ecoli 5. Gene2Salmonella 6. Gene12Thermotoga.

18. Which bifurcations are represented in the above tree?
A) ***…
B) ..****
C) ….**
D) …***
E) All of the above

19. Which of the following bifurcations is incompatible with the above tree?
A) .*…*
B) *….*
C) ..*.**
D)..**..
E) All of the above

20. True/False It is possible that in 22% of bootstrap trees Gene2 groups within the Gene1 sequences.

21. True/False It is possible that in 45% of bootstrap trees Gene2Thermotoga groups the Gene2Ecoli sequence to the exclusion of Gene2Salmonella.

22. True/False Gene2 can be used to root Gene1 and vice versa.

23. True/False It is possible that in 20% of bootstrap trees, Gene1 groups within Gene2.

24. True/False It is not possible that the grouping of the three Gene2 sequences together is due to Long Branch Attraction, because the bootstrap support is 100%.

25. Why is the value 100 on the tree twice?
A. This is an artifact of the tree drawing program, which has taken the value associated with the root and written it in two places.
B. Because there are 2 nodes supported with 100% bootstrap support
C. Because of Long Branch Attraction
D. A, B, and C cannot be discriminated between
E. All of the above

26. It is possible the in 2% of bootstrap samples the three Gene2 sequences group with Gene1Ecoli.

27. The topology of Gene 1 compared to the topology of Gene2 indicates that which of the following has most likely occurred in these sequences?
A. Incomplete lineage sorting
B. Long Branch Attraction
C. Long Branch Repulsion
D. Vertical inheritance only
E. Horizontal Gene Transfer

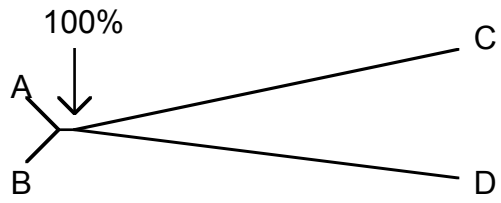
28. It is possible the in 20% of bootstrap samples the Gene2Thermotoga sequence group with Gene1Salmonella.

29. What is a Xenolog?
A. A homolog resulting from gene duplication.
B. A gene with the same function as another, but evolved independently.
C. A homolog resulting from the hybridization of two species.
D. A homolog resulting from Horizontal Gene Transfer.
E. A homolog resulting from a speciation event.

30. In the PHYLIP package, Trees written onto "outtree" are in the:
A. Newick format
B. Mathlab Format
C. Java Format
D. C++ Format
E. PHYLIP Format

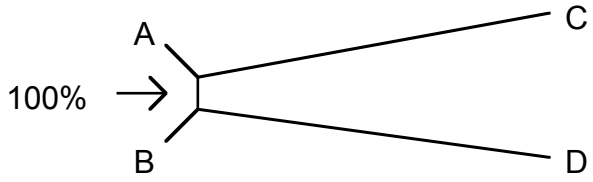
31. Parsimony aims to build the tree that?
A) under which the data set (e.g., aligned sequences) is most probable.
B) is most probable given the data.
C) explains the evolutionary history that gave rise to the aligned sequences with the least number of substitution events.
D) is in the best possible agreement with the observed number of substitutions observed between the sequences.

32. When analyzing a quartet of putatively orthologous sequences, the maximum parsimony tree looks like this, with the central branch having 100% bootstrap support:



A. This tree groups the two long branches together. The possibility exists that this result might represent a long branch attraction artifact.
B. The central branch is so strongly supported that one can exclude a long branch attraction artifact. (LBA is a statistical phenomenon and never reaches 100% bootstrap support.)
C. Maximum parsimony is not subject to the long branch attraction artifact, rather it always has the tendency to group the long with the short branches. Therefore, the finding that A and B group together is reliable.

33. When analyzing a quartet of putatively orthologous sequences, the maximum parsimony tree looks like this, with the central branch having 100% bootstrap support:



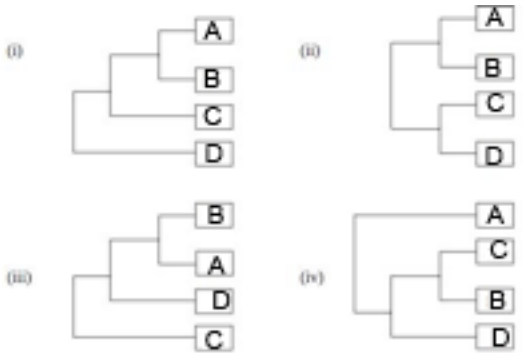
A) Maximum parsimony is not subject to the long branch attraction artifact, but always has the tendency to group the long with the short branches (aka long branch repulsion). Therefore, the finding that A and C group together is unreliable.
B) This tree does not group the two long branches together, indicating that the result is not due to long branch attraction.
C) The central branch is so strongly supported that one can exclude any artifact that might occur during phylogenetic reconstruction. (The artifacts caused by long branches are a statistical phenomenon and never reached 100% bootstrap support.)

34. Neighbor joining calculates trees by?
A. minimizing the number of substitution events.
B. maximizing the probability of a model, given the data.
C. By using pairwise comparisons to determine the nearest neighbor and then collapsing a node and recalculating all of the pairs for the node.
D. maximizing the probability of a dataset, given the model.
E. programing a little robot to walk around in tree space.

35. True/False Phylogenetic reconstruction using Markov chain Monte Carlo sampling aims to find the phylogenetic tree that is most probable given the data by walking around in tree space with a biased walk and sampling the trees.

36. Maximum likelihood aims to build the tree that?
A) that is most probable given the data.
B) that explains the evolutionary history that gave rise to the aligned sequences with the least number of substitution events.
C) that is in the best possible agreement with the observed number of substitutions observed between the sequences.
D) the data set (e.g., aligned sequences) is most probable.

37. Which of the following rooted trees does NOT have an identical topology when considered as unrooted?



A) Tree (i)
B) Tree (ii)
C) Tree (iii)
D) Tree (iv)
E) All trees are identical in their topology

Extra credit:
1. Max of 2pts Models used to describe sequence evolution frequently use the Gamma distribution, using the alpha parameter.
A. What is the name of the process often described by the Gamma distribution?

B. Why is the Gamma distribution more useful than the normal distribution?

2. 1pt Often it is said that humans are the highest life form, because we are more highly evolved than all the other animals. Is this correct? Can *any* species be more evolved than another?