Name: _____

Comparative Bioinformatics Midterm II
Fall 2003

Objective Answer, part I: For each of the following, select the single best answer or completion of the phrase. (5 points each)

1. Which of the following is a valid reason to sequence the dog genome?

a. Craig Venter really, really likes his Welsh Corgis.
b. The dog genome is more compact than any vertebrate genome other than that of Fugu (pufferfish).
c. More genetic diseases have been identified in dogs than in any animal other than humans.
d. The dog genome shares closer synteny with the human genome than with any common laboratory animal other than chimpanzees and bonobos.

2. What is an advantage of sequencing the dog genome to only 1.5x coverage?

a. This is enough coverage to permit complete contig assembly and closing of the genome.
b. It will probably be desirable to sequence the genome of many breeds of dog, so this gives enough coverage to get useful information at a low cost per breed.
c. The publicly funded genome project has already published a dog genome, so there really isn't any point in doing any more than this much coverage.
d. All of the above.

3. What is an advantage of an approach to multiple sequence alignment based on the Smith-Waterman algorithm, but extended to an n-dimensional matrix?

a. It requires very little computer memory.
b. It doesn't require very much calculation.
c. It is an exact algorithm.
d. It performs well with large numbers of sequences.

4. In a Markov process…

a. The system is defined by a set of states, the allowed transitions that can occur among these states.
b. Probabilities are associated either with the states, or with the allowed transitions.
c. The states into which the system can change are dependent only on the state the system is in.
d. All of the above.

5. Orthologs are homologous copies of a gene that are related by…

a. phylogenetic descent.
b. gene duplication.
c. at least 85% identity.
d. none of the above.

Name: _____

Objective Answer, part II: For each of the analytical methods listed below, indicate which of the following assumptions apply to that analytical method in the context of DNA sequence analysis. (25 points)

a. Sequences are descended from a single common ancestor and are related by a dichotomously branching tree.
b. Character-state reversal is rare.
c. The tree that requires the smallest number of character-state changes is the best tree.
d. All nucleotides occur with equal frequency.
e. The four nucleotides occur with different frequencies.
f. All substitutions are equally likely.
g. Transitions and transversions occur at different rates.
h. Each nucleotide substitution can occur at a distinctive rate.
i. The phylogenetic tree can be reconstructed from a corresponding set of pairwise distances.

6. Parsimony

a b c d e f g h i

7. Jukes-Cantor Distances (just the calculation of the distance matrix)

a b c d e f g h i

8. Kimura Two-Parameter Distances (just the calculation of the distance matrix)

a b c d e f g h i

9. Fitch-Margoliash (sum-of-errors distance) with Felsenstein 1984 (F84) distances

a b c d e f g h i

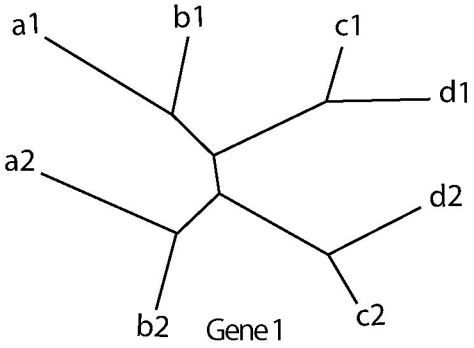
10. Maximum Likelihood with a General Time Reversible model

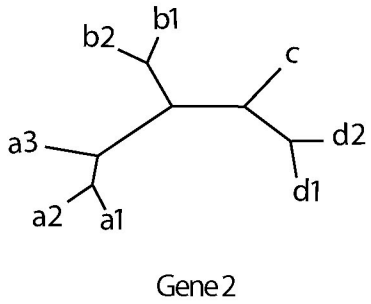
a b c d e f g h i

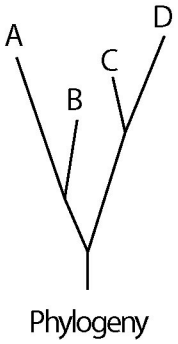
Name: _____

Short Answer. (5 pts each)

11-13: The following phylogenetic trees represent the results of analyses of DNA sequences of two gene families isolated from four organisms, A B C & D which are known to be related by the phylogeny shown at right. The first letter of the sequence label indicates which organism the sequence was determined from, so a1 and a2 were both isolated from organism A, b1 and b2 from organism B, and so forth.







11. Organism A has two copies of both Gene 1 and Gene 2. For which of these genes do you feel it more likely that there is a functional distinction between different copies? Why?

12. Which of the two genes seems more likely to be undergoing concerted evolution? Why?

13. Draw the phylogenetic tree that one would obtain if (in the absence of genomic data from all of the organisms) one performed a phylogenetic analysis using only sequences a1, b2, c1, and d2 from Gene 1. To what erroneous conclusion would this analysis lead?

Name: _____

For each of the scenarios described below, (1) indicate what analytical method you would recommend applying to the problem, and (2) explain why you recommend this approach.

14. Chau is studying the rapid radiation of insect diversity that occurred in the Carboniferous Period. To address this topic she has determined the sequence of 5 protein-coding genes from 60 representative insects, and she plans to use careful interpretation of the reconstructed branch lengths and branching patterns to interpret this difficult phylogenetic problem. She has come to you to ask advice on how to proceed. What do you advise?

15. Kevin has determined the1kb sequence of a gene that encodes a viral coat protein from 800 isolates of tissue from animals in a Chinese market. He would like to compare these sequences to a homologous sequence isolated from a patient who is gravely ill with an uncharacterized disease in the hope that this will identify the source of the infection. Time is of the essence, because this information might help guide treatment of the patient, and could be used to limit the spread of the infection. What phylogenetic method should Kevin use to analyze these data?

16. Radhika is studying the history of ancient manuscripts, and has received a grant to study the twenty ancient copies of a manuscript that date to roughly the fourth century AD. She would like to reconstruct the history of how these manuscripts were transcribed. How should she proceed?

Name: _____

Essay. (20 points)

17. Parsimony is known to be inconsistent (i.e., to converge on an incorrect solution) under some analytical conditions, particularly when branch-lengths are unequal. Describe such conditions, and explain how model-based methods can perform better with the same data. Also address when model-based methods might be expected to be inconsistent.