Name: ___________________________

Comparative Bioinformatics Midterm II
Fall 2004

Objective Answer, part I: For each of the following, select the single best answer or completion of the phrase. (3 points each)

1. *Deinococcus radiodurans* has been isolated from:

a. Radiation "sterilized" canned meat.
b. The surface of Mars.
c. The inside of an autoclave.
d. All of the above.

2. NP-Hard (nonparametric polynomial) problems are a class of problems for which…

a. exact solutions become very difficult as the size of the problem increases.
b. it is possible to calculate a goodness-of-fit score for a candidate solution.
c. heuristic algorithms are widely used to find approximate solutions.
d. all of the above.

3. Clusters of Orthologous Groups are identified by…

a. reciprocal best hits from "all-against-all" BLAST analyses of genomes.
b. phylogenetic analysis of sequences excluded by the "triangle criterion."
c. Hierarchical clustering and "synteny-mapping" of chromosomes.
d. none of the above.

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. Paralogs 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: The groups Archaea, Bacteria, and Eukarya, as well as combinations of these, are listed below.

a. Transcription start site with TATA motif (TATA box)
b. Formylmethionine as first amino acid in new protein
c. Circular chromosomes
d. Protein-coding genes organized into operons
e. Abundant introns
f. DNA bound to histones or histone-like proteins
g. Several distinct RNA polymerases that require transcription factors
h. Nucleus
i. Ether-linked phospholipids

For each of the following sets of organisms, circle those of the characteristics shown above that apply to all (or many) of the organisms in that set. (10 points)

6. Archaea

a b c d e f g h i

7. Bacteria

a b c d e f g h i

8. Eukarya

a b c d e f g h i

9. Archaea *and* Bacteria

a b c d e f g h i

10. Archaea *and* Eukarya

a b c d e f g h i

Name: ________________________

Short Answer. (5 pts each)

11. In principle, one could perform multiple sequence alignment by extending the Smith-Waterman algorithm to create an n-dimensional matrix, where n corresponds to the number of sequences to be aligned. In fact such an algorithm has been implemented, and provides an exact solution to multiple sequence alignment. Why is this software tool unlikely to find widespread use? 5 points.

12. Describe the cluster approach to multiple sequence alignment (e.g., as implemented by clustalw or pileup). 5 points.

13. What is a consensus sequence? Provide an example. 5 points.

Name: ___________________________

For each of the scenarios described below, (1) indicate what type of analysis you would recommend as the next step, and (2) explain why you recommend this approach (5 points each).

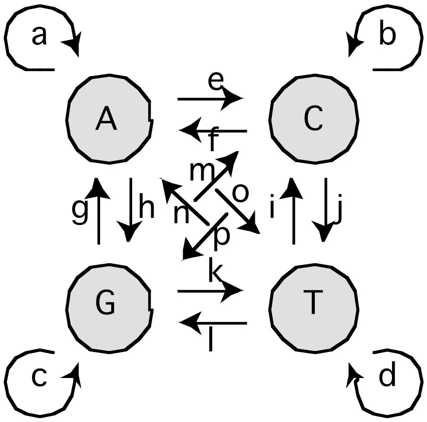
14. Kevin has determined 20,000 EST sequences from five different wombat tissues. He would like to make tentative inferences of homology for each of these sequences.

15. Julia has completed the genome sequence of five different species of nematode, has assembled each of the genomes, and has identified a large number of open reading frames. She has used BLAST analysis to search each ORF against all of the genomes.

16. Arya has determined the gene encoding a highly variable surface antigenic protein from 75 isolates of *Bacillus anthracis*, and would like to understand how these strains are related to each other.

17. We have obtained the following DNA sequence, and would like to determine whether or not it is derived from the *Plasmodium* genome. We can represent the sequence with a Markov model showing the frequencies with which individual nucleotides follow each other. (25 points)

```
QUERY: AACCTGGTTG ATCCTGCCAG TAGTCATATG CTTGTCTCAA AGATTAAGCC
```



a) What dinucleotide pair does transformation "j" represent?

b) How often is this dinucleotide pair observed in the sequence above?

c) What are the observed frequencies of *each* transformation?

a: b: c: d:
e: f: g: h: i: j: k: l:
m: n: o: p:

d) The Plasmodium genome is 80% A+T. From this we can estimate some relationships among expected transformation frequencies in *Plasmodium*. Provide a *numerical* estimate for each of the following values:

a+e+h+o = d+l+i+n = ? ________

c+g+k+m = b+f+j+p = ? ________

e) What assumptions did you use to make your inference in (d)?

f) [optional] Is the sequence likely to have been derived from *Plasmodium*? Why?

Name: ___________________________

18. Assume that you have sequence data from four animals, Albatross, Beluga Whale, Cat, and Dugong (10 points).

a. Draw the three possible unrooted trees that can interrelate these taxa.

b. Draw the three corresponding ROOTED trees that assume that the Albatross is the outgroup.

You may wish to refer to the following algorithm (described by Swofford and Maddison, 1987), which can be used to determine the length of a tree under the parsimony criterion:

```
1.Assign a state to each terminal node
(2). Visit first internal node
        A. is the intersection of states non-empty?
                I. Yes: set internal state to this.
                II.  Else:
                        a. set the state to the smallest set containing the states of the daughter
                        nodes
                        b. increase the tree length by 1.
3. Are you at the root of the tree?
        A. No: go to 2.
        B. Yes: go to 4.
(4). Is the state at this node the same as the outgroup state?
        A. Yes: Proceed to the next character
        B.  Else: Add one to the length of the tree; proceed to next character
```

```
          123456789
Albatross GACCGTATA
Beluga    GTTCCTCTC
Cat       GTTCCCTTC
Dugong    GTCCGCTTA
```

19. Use the Swofford and Maddison algorithm (above) to (a) determine which of the three rooted trees is most parsimonious, and (b) show the node assignments of character states for characters 5 and 6 on the most parsimonious tree (10 points).