Name: _____

SID: _____

Topics in Computational Biology and Genomics
Plant & Microbial Biology / Molecular & Cell Biology / Bioengineering c146/c246

Spring 2003

MIDTERM EXAM

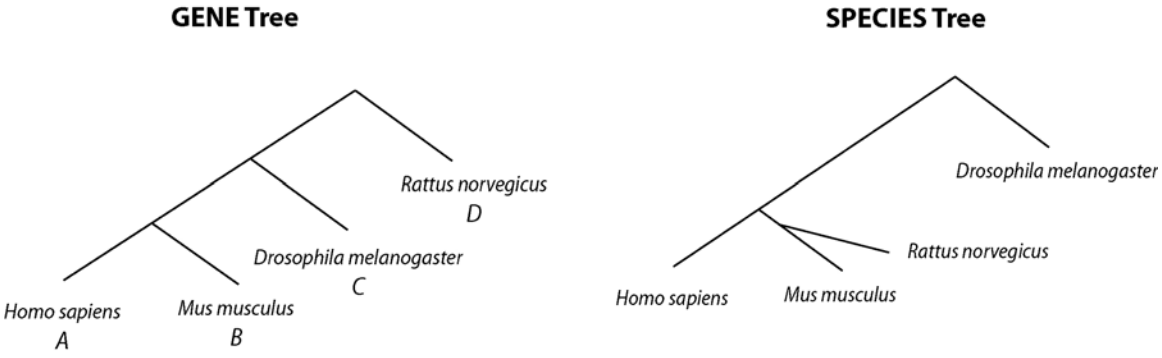
- Concise answers and clear phrases acceptable
- Complete sentences not essential
- Answers can be in functional form (*e.g.*, $\frac{1}{\sqrt{2}}$ instead of 0.707)

1	_____ /	5
2	_____ /	5
3	_____ /	10
4	_____ /	10
5	_____ /	10
6	_____ /	10
7	_____ /	10
8	_____ /	10
9	_____ /	10
10	_____ /	5
11	_____ /	15
Extra credit	_____ /	5
TOTAL	_____ /	100

NAME: _____

1. 5 points

The tree on the left depicts the evolutionary history of a family of 4 proteins found in 4 species; the tree on the right depicts the evolutionary history of the species.



(A) What is the most likely evolutionary relationship between proteins A and B? Why?

(B) What is the most likely evolutionary relationship between proteins B and D? Why?

2. 5 points

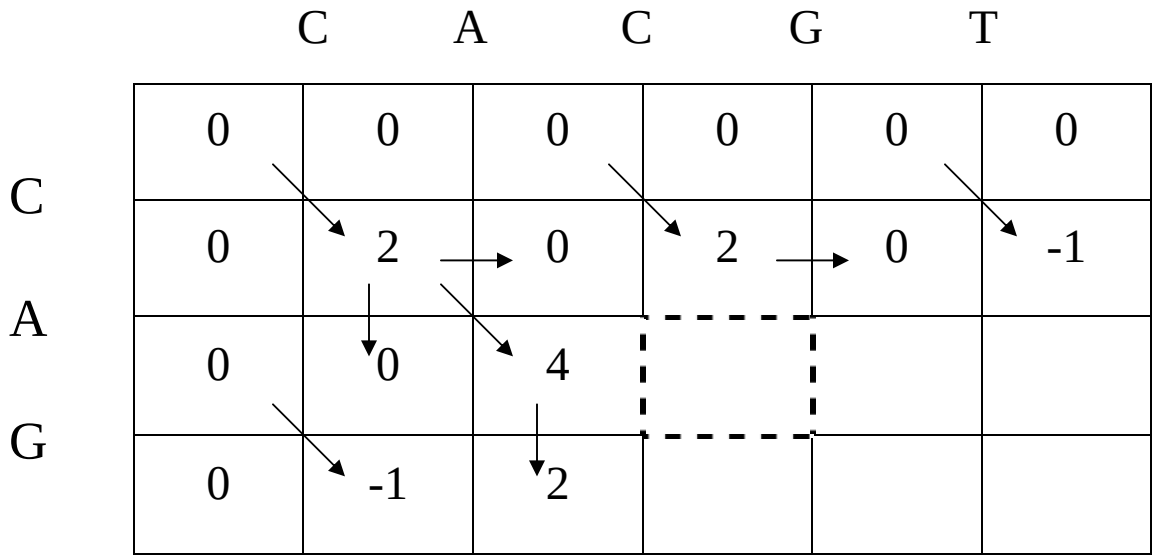
What is the theoretical basis that allows us to infer that proteins are orthologous from their sequences?

NAME: _____

3. 10 points

(A) (6 points) Provide the dynamic programming matrix entries and traceback vectors for the semi-global alignment (no end gap penalties for either sequence) of the sequences below. Use the following scores:

Match	+2
Mismatch	–1
Gap (fixed)	–2



(B) (2 points) What is the resulting alignment and score?

(C) (2 points) What alignment does the value in the cell with dotted borders represent?

NAME: _____

4. 10 points

In the procedure to construct PAM matrices, how does matrix multiplication incorporate a Markov model of evolution? (*Recall*: M = A × B, where $m_{ij} = \Sigma_k \, a_{ik} \times b_{kj}$)

5. 10 points

(A) (5 points) Write the recursion relations required for dynamic programming alignment with *generalized* gap penalties.

(B) (5 points) What is the time complexity (in big-oh notation) for sequence alignment with *generalized* gap penalties? for sequence alignment with affine gap penalties?

NAME: _____

6. 10 points

A BLAST database search is performed with a query sequence of length 16 (2^4) and an effective database size of 65,536 (2^{16}). The raw score of an alignment is 25, using a matrix with K = 2 and $\lambda = \ln 2 = 0.693$. What is the BLAST E-value (you may leave your answer in base 2)?

7. 10 points

Suppose you want to perform a database search for extremely similar hits to your query sequence. Name 3 BLAST parameters that could be changed to make the search run faster. Explain your reasons for each change in a few words.

NAME: _____

8. 10 points

(A) (5 points) What are the key reasons that rigorous traditional sequence alignment by dynamic programming becomes intractable with multiple sequences?

(B) (5 points) What are advantages of full dynamic programming over other multiple alignment approaches?

9. 10 points

(A) (5 points) How do CLUSTALW and CLUSTALV differ? What effects do these differences have on speed and alignment quality?

(B) (5 points) How do CLUSTALW and T-COFFEE differ? What effects do these differences have on speed and alignment quality?

10. 5 points

PAM matrices are constructed using the formula given to the right. What is the scaling factor, λ, associated with a PAM matrix?

$$S_{ij} = 3\log_{10}\frac{q_{ij}}{p_i p_j}$$

Note: $\log_{10} e \approx 0.434$ $\log_{10} 3 \approx 0.477$ $\log_{10} \pi \approx 0.497$

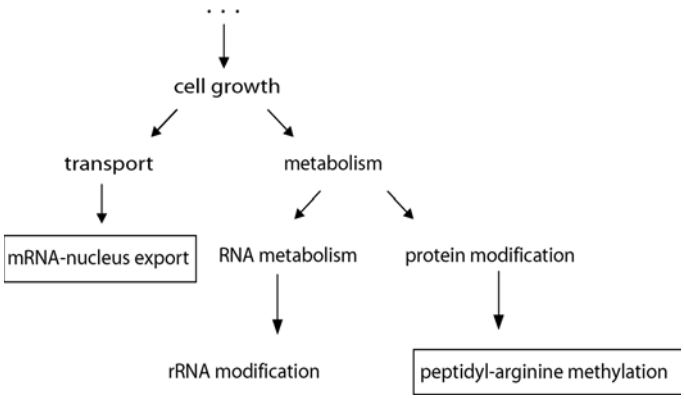
$$\log_b a = \frac{\log_{10} b}{\log_{10} a}$$

NAME: _____

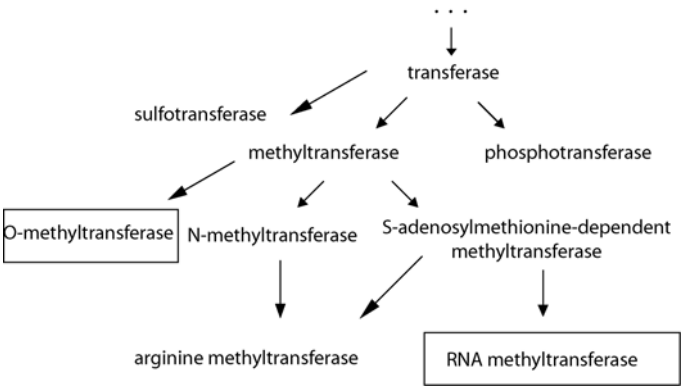
11. 15 points

A BLAST search of an unannotated query protein yielded two weak but significant hits. These hits have biochemically-verified annotations for the boxed Gene Ontology terms:

Biological Process



Molecular Function



(A) (5 points) What can you deduce about the *molecular function* of the query protein?

(B) (5 points) What can you deduce about the *biological process* of the query protein?

(C) (5 points) Describe a method that could help annotate the biological process of the query protein.

Extra credit: (5 points – all or none. Use the reverse side of the page.)
MAFFT is claimed to be a quick and accurate multiple alignment algorithm.

(A) What features of MAFFT make it quick (just saying FFT is not enough)?
(B) What features of MAFFT make it accurate?
(C) What is a fundamental limitation in its optimization function that prevents MAFFT from making a biologically optimal alignment? Why?