Student ID: _____ Student Name: _____

CompSci 369: Computational Biology

Midterm Test 2008

1. (a) Explain the 'principle of optimality' with respect to dynamic programming.

(***2 marks***)

(b) How many different DNA strings are there of length n?
How many different DNA strands are there of length n?
(Note a string and its reversal are considered the same strand.)

(***2 marks***)

(c) State a fundamental difference between local search and branch-and-bound algorithms with regards to coping with a computational hard biology problem.

(***2 marks***)

(d) Explain the two common types of data (distance and character) used in phylogenic reconstruction.

(***2 marks***)

(e) Compare and contrast the motivation for using the maximum likelihood and maximum parsimony models for predicting evolutionary trees.

(***2 marks***)

2. Consider the problem where we have two strings $X = x_1x_2\cdots x_m$ and $Y = y_1y_2\cdots y_n$ and we want to compute the smallest alignment score where we restrict gaps to length at most 1 except at the ends. Let $D(\texttt{a},\texttt{b}) \geq 0$ denote the cost of matching characters `a` and `b`. Let $\gamma > 0$ denote the cost of matching a gap '`-`' to any character. For example, the alignment X=`aaab-aab-` and Y=`--abb-abb` is okay but the alignment X=`aaab-aab` and Y=`a--bbabb` is not.

   (a) Present a recursive algorithm $f(m, n)$ that solves the subproblems based on aligning prefixes of X and Y. Hint: gaps are not aligned to gaps.

(***5 marks***)

(b) Give an iterative dynamic programming version of your algorithm in part (a).

(***3 marks***)

(c) Explain the running time of your algorithm in terms of Θ notation of a function of m and n.

(***2 marks***)

3. We want to find the best parsimony tree for the taxon represented by S1='ACAG', S2='AATG', S3='GGGT', S4='GAGG' and S5='GGTG'.

   (a) How many parsimony (Steiner) trees do we need to consider in the worst case?

   (***2 marks***)

   (b) Compute the best parsimony score for this specific instance by drawing the tree and labeling the internal nodes (with strings) and edges (with Hamming distances) by using Fitch's algorithm applied with S5 as root.

   (((S1,S2),(S3,S4)),S5)

   (***3 marks***)

4. Let A be the following alignment of two amino acid sequences:

```
x  EHANAWEEG     9
y  -HEAAW--G     6
```

   (a) Using the *Blossum50* scoring matrix provided in annexe and considering a linear gap penalty of 8, what will be the score of A?

   (***1 marks***)

(b) Using the *Blossum50* scoring matrix provided in annexe and considering affine gap-open penalty of 8 and gap-extension penalty of 2, what will be the score of A?

(***1 marks***)

(c) Recall the recurence formula of the Needleman & Wunsch algorithm for global alignment between two sequences x and y, with d the gap penalty and $s()$ the scoring function.

(***2 marks***)

(d) We want to compute a global alignment of two DNA sequences: $x =$ `AGG` and $y =$ `ATAGCG`. The scoring system is as follow: 1 for a match, -1 for a mismatch and -2 for a gap penalty. Fill the corresponding F matrix and trace back all optimal alignments.

(***4 marks***)

5. Two sequences of DNA `alpha` and `beta` are portions of homologous genes and can be aligned as

```
alpha  ACCAACGTCAAGGCCGCCTGGGGTAAGGTT    30
beta   TCTGCCGTTACTGCCCTGTGGGGGAAGGTG    30
```

(a) Write down the expressions of both the observed p-distance and the genetic (evolutionary) d-distance under the Jukes-Cantor model of nucleotide substitution, between `alpha` and `beta`.

(***3 marks***)

(b) Actually, these DNA sequences are parts of the Human genes called α and β globins that are known to be paralogous. Recall the definition of paralogous genes.

(***2 marks***)

(c) Sharks have α and β globins. What can you say about the α globin genes of Humand and Sharks?

(***2 marks***)

(d) Lampreys only have one single type of globin gene. Give two different hypothesis to explain the evolution of globin genes? Based on parsimony, which hypothesis do you consider the most likely?

(***3 marks***)

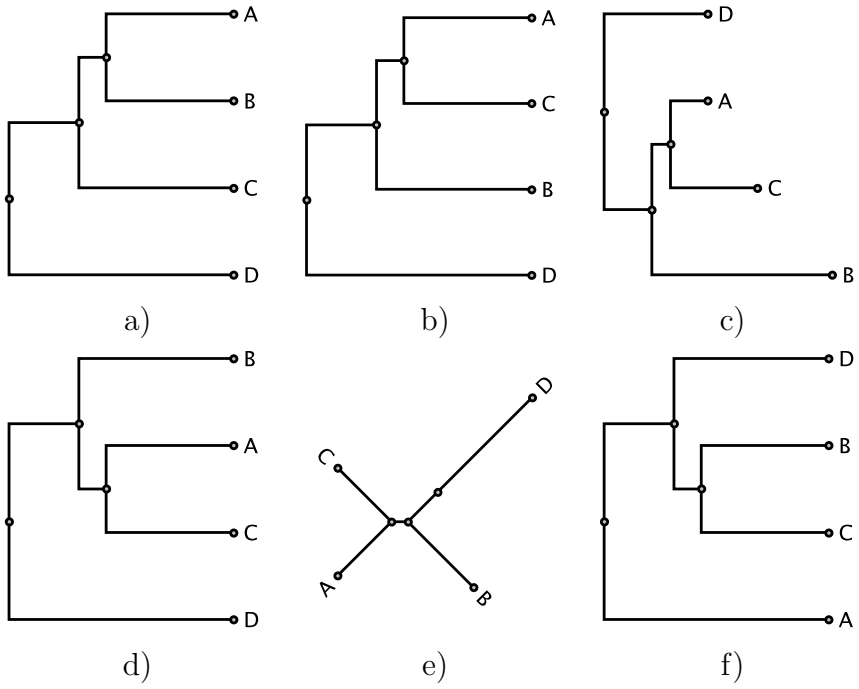
(e) Knowing that Sharks go back in the fossil record for 400 million years ago and that Lampreys are evolutionary related to Sharks. What can you say about the time to the most recent common ancestor of Shark and Lamprey?

(***2 marks***)

6. Using the pairwise distance matrix M between 4 DNA sequences A, B,C and D:

M	A	B	C	D
A	0			
B	8	0		
C	7	9	0	
D	12	14	11	0

Could you tell which of the following trees can be obtained using the UPGMA algorithm?



(***5 marks***)

Annexe

A	5																			
C	-1	13																		
D	-2	-4	8																	
E	-1	-3	2	6																
F	-3	-2	-5	-3	8															
G	0	-3	-1	-3	-4	8														
H	-2	-3	-1	0	-1	-2	10													
I	-1	-2	-4	-4	0	-4	-4	5												
K	-1	-3	-1	1	-4	-2	0	-3	6											
L	-2	-2	-4	-3	1	-4	-3	2	-3	5										
M	-1	-2	-4	-2	0	-3	-1	2	-2	3	7									
N	-1	-2	2	0	-4	0	1	-3	0	-4	-2	7								
P	-1	-4	-1	-1	-4	-2	-2	-3	-1	-4	-3	-2	10							
Q	-1	-3	0	2	-4	-2	1	-3	2	-2	0	0	-1	7						
R	-2	-4	-2	0	-3	-3	0	-4	3	-3	-2	-1	-3	1	7					
S	1	-1	0	-1	-3	0	-1	-3	0	-3	-2	1	-1	0	-1	5				
T	0	-1	-1	-1	-2	-2	-2	-1	-1	-1	-1	0	-1	-1	-1	2	5			
V	0	-1	-4	-3	-1	-4	-4	4	-3	1	1	-3	-3	-3	-3	-2	0	5		
W	-3	-5	-5	-3	1	-3	-3	-3	-3	-2	-1	-4	-4	-1	-3	-4	-3	-3	15	
Y	-2	-3	-3	-2	4	-3	2	-1	-2	-1	0	-2	-3	-1	-1	-2	-2	-1	2	8
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y

— scratch paper; will not be marked —