Student ID: _____ Student Name: _____

CompSci 369: Computational Biology
Midterm Test 2008

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

The optimal solution to a problem is based only on optimal solutions to its sub-problems.

(***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.)

4^n and $(4^n + 4^{\lceil \frac{n}{2} \rceil})/2$. Need to consider how many palindrones for the 2nd case.

(***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.

The local search algorithm may not find an optimal answer whereas the branch-and-bound eventually will.

(***2 marks***)

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

Distance data, usually stored in an n-by-n matrix, gives pairwise genetic distances. Character data, usually stored as a binary array or bit vector, indicates which traits hold for each taxon.

(***2 marks***)

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

The maximum likelihood model is based on how likely (using probability and statistics) a tree evolved. The maximum parsimony model uses a metric that minimizes the total tree distances between all evolutionary branches (i.e. least number of evolutionary changes).

(***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.

$$
\begin{aligned}
f(m,n) &= \min(f_1(m,n), f_2(m,n)) \\
l &= \min(m,n) \\
f_1(m,n) &= \min\left(\min_{1\leq i\leq l} f_2(m-i,n)+i\gamma, \min_{1\leq i\leq l} f_2(m,n-i)+i\gamma\right) \\
f_2(m,n) &= \begin{cases}
n\gamma & \text{if } m=0 \\
m\gamma & \text{if } n=0 \\
\min\begin{pmatrix}(n-1)\gamma + D(x_1,y_n),\\ (n+1)\gamma\end{pmatrix} & \text{if } m=1 \text{ and } n\geq 1 \\
\min\begin{pmatrix}(m-1)\gamma + D(x_m,y_1),\\ (m+1)\gamma\end{pmatrix} & \text{if } m>1 \text{ and } n=1 \\
\min\begin{pmatrix} f_2(m-1,n-1)+D(x_m,y_n),\\ f_2(m-1,n-2)+\gamma+D(x_m,y_{n-1}),\\ f_2(m-2,n-1)+\gamma+D(x_{m-1},y_n)\\ f_2(m-2,n-2)+2\gamma+D(x_{m-1},y_{n-1})\end{pmatrix} & \text{otherwise}
\end{cases}
\end{aligned}
$$

Note that many other possible answers exist. The key point is that if one has aligned a gap (that is not at the ends) then the next position to the left must be an alignment of two characters, as f_2 enforces.

(***5 marks***)

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

First fill in all of table f_2 by using two nested loops. Then calculate $f_1(m, n)$. Finally, compute and return $f(m, n)$.

(***3 marks***)

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

We need to fill out a table of size $m \times n$ for f_2. As specified above, each entry in the table for f_1 requires $\Theta(l)$ time to compute. However we only need to compute one entry $f_1(m, n)$. Thus the total time, as presented, is $\Theta(m \cdot n + \min(n, m)) = \Theta(mn)$.

(***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?

   $3 \cdot 5 = 15$, which is the same as the number of rooted binary trees with 4 tips (say, the root being S5 and leaves labeled with {S1,S2,S3,S4}).

   (***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)

   Possible internal nodes (from top down) are GATG, AATG and GAGG for a total tree of cost 7.

   (***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?

   12

   (***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?

18

(***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.

$$F(i,i) = \max \left\{ \begin{array}{lll} F(i-1,j-1) & + & s(x_i, y_j) \\ F(i,j-1) & - & d \\ F(i-1,j) & - & d \end{array} \right\} \text{with } F(0,0) = 0$$

(***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.

$F(x,y)$		A	T	A	G	C	G
	0	**-2**	**-4**	-6	-8	-10	-12
A	-2	**1**	**-1**	**-3**	-5	-7	-9
G	-4	-1	0	-2	**-2**	**-4**	-6
G	-6	-3	-2	-1	-1	-3	**-3**

```
x  A--G-G    3
y  ATAGCG    6
```

and

```
x  --AG-G    3
y  ATAGCG    6
```

(***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`.

The length of the alignment is 30 nucleotides and we observe 12 nucleotide substitutions. So the p-distance is simply $12/30 = 0.4$. Under the Jukes-Cantor model, the evolutionary distance is:

$$d = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right)$$

which is about 0.57 substitutions per site. This means that the effective number of substitutions is about 17.

(***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.

All homologous genes share a common evolutionary ancestor but paralogues are derived from gene duplication events within a single species, so they are found in more than one copy in the same genome.

(***2 marks***)

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

The two genes are orthologous.

(***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?

Two hypothesis can be made:

- in the first one, the duplication event occured in an ancestor of Shark and Lamprey and the Lamprey loose one of the paralogues.
- in the second one, the duplication event occured in an ancestor of the Shark only.

The second hypothesis is the most likely to occur.

(***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?

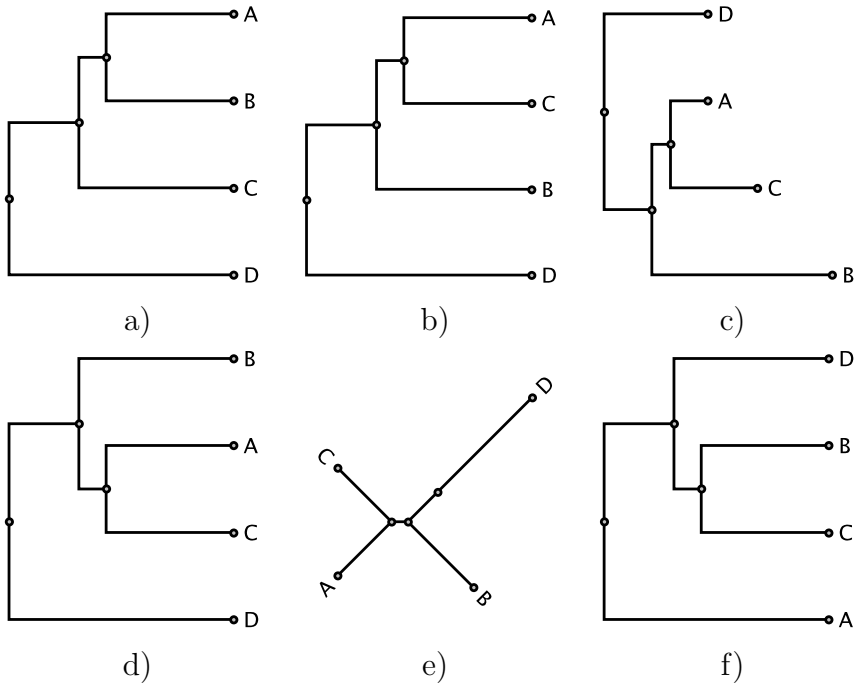
The time to the most recent common ancestor of Sharks and Lampreys is at least 400 million years ago.

(***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?



Trees b and d.

(***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