BioE/MCB/PMB C146/246, Spring 2005

Problem Set 2

Due 31 January 2005 by **midnight** to lfl@compbio.berkeley.edu

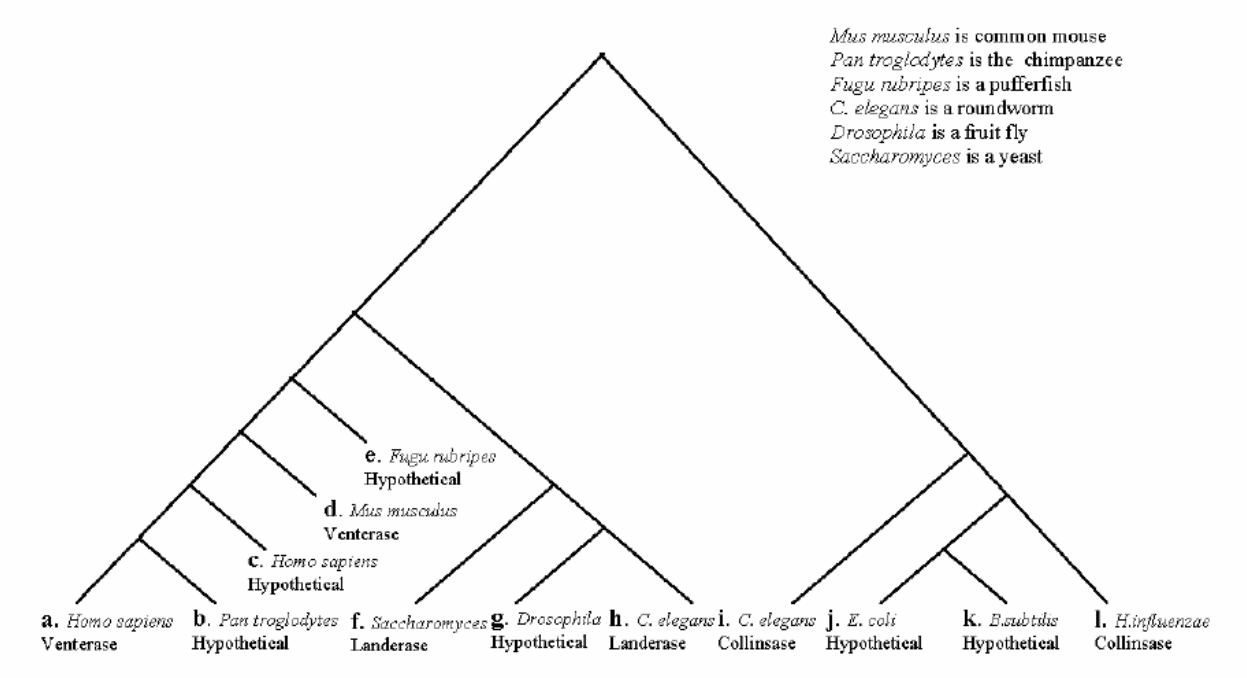
1) Please attach dynamic programming matrices for 2A, 2B, 2C, 3A, 3B in your mail as TAB-DELIMITED files: *jrandom*_ps2_2A.txt, *jrandom*_ps2_2B.txt, etc.
2) Attach your code for Question 3. In the text of your email, please include instructions on how to run your program.

1. 20 points

Using the following tree depicting the evolutionary history of a family of 12 genes found in 10 species, describe the most likely evolutionary relationship between the following genes. Where functional information is available, it is given.

A. a and each of c, d, e
B. j and each of i, k, l
C. f and each of j, k, l

Hint: For this problem, you need to know the evolutionary relationships among these species. Look this up if you don't know them already.



2. 25 points

Align the sequences

EEHGWAGAEH and EAEHWAP

using the following scoring matrix and a fixed per-position gap penalty of –8 (you can download this as SCORE_MATRIX from the course website),

A	Ala	4																		
R	Arg	-1	5																	
N	Asn	-2	0	6																
D	Asp	-2	-2	1	6															
C	Cys	0	-3	-3	-3	9														
Q	Gln	-1	1	0	0	-3	5													
E	Glu	-1	0	0	2	-4	2	5												
G	Gly	0	-2	0	-1	-3	-2	-2	6											
H	His	-2	0	1	-1	-3	0	0	-2	8										
I	Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4									
L	Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4								
K	Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5							
M	Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5						
F	Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6					
P	Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7				
S	Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4			
T	Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5		
W	Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	
Y	Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7
V	Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

A. Globally
B. Locally
C. Globally with no end gap penalties.

For all alignments, provide the complete dynamic programming matrix (including traceback of how you found the alignment), as well as the score. You may write a program to assist you (see Question 3), or you may do the alignment by hand.

D. Use SSEARCH to align these sequences and compare its alignment with your result.

* SSEARCH can be run on the Internet from http://workbench.sdsc.edu/ (Registration is free.)

1) Click on the *Protein Tools* button.
2) Choose *Add New Protein Sequence*. Push *Run*
3) Enter the first protein sequence. Push *Save*.
4) Repeat step 3 for the second protein sequence.
5) Check the box next to one of the sequences.
6) Click on "SSEARCH" in the list. Push *Run*.
7) Click "Choose One or More Sequences"
8) Use the "blosum62" matrix and set gap penalties

3. 30 points

Pairwise sequence alignment

Write a program that implements global DNA alignment[♦] using the following parameters:

Identity	+4
Transition	–2
Transversion	–4
Gap	–8 *fixed*

Given the sequences

```
GCCCCGGTATATGCGTTAT
CTATGGGCGCGT
```

A. (10 points) Align the two sequences and report their scores.
B. (10 points) Revise the algorithm from part B to produce a local alignment.
C. (10 points) Submit your code, using good style and documentation, along with instructions on how to run it.

4. 15 points

Give an example of when each of the following alignment strategies would be appropriate:

A. global
B. local
C. global with no end gap penalties

♦ Although only a DNA alignment program is required for this problem set, you will benefit by making your code as general as possible. Design your program in a modular way such that you could easily change it to perform protein sequence alignments, use generalized or affine gap penalties, or read in a score matrix.

5. 10 points

A. The following two sequences can be downloaded from the course website. Look at the two sequences using a dot plot viewer (many are available on the Internet). What do you infer from them?

```
>SEQUENCE1
TWWFVRIQWICHTVLCFQIPGGIWNHMHVGWEYQHECHKVYQKMSTKHGTY
WSRFSEVWIPAMTYREWGHEREMTAAYLMTFQYWVRPQFADYNQYEFPKFM
SGGSYNLGAYREPMGTQRFFFWEQQPGKKPYAWEGATAHYDCTIVVSDWGT
NIHSHTDHKWHEPWEAMRPLPLMELDNDLSMRNRVRTIHVSDWGTNIVSET
DHFWHEPWEADGTKIMGAYYCVPGRHALALVYMTIHVSDWVTNIHSHTDHA
EHEPWEAQVWFRMRCCKAGTMPRLKGAYARHKMGRWTICTIHVSDWGTNIH
SMSDHKWHEPWEAPAISEAMQSKVIIHVSDWVTNIHSHTDHKHHEPWEASM
PLFKYAAKCYACKFDCTIHVSDWGTNIQSHTDWKWHEPWEA

>SEQUENCE2
TQNRAFKMCPQDNEWTSQNWCTSHFCASKIDGIWCQIHVGWEYKHECGKVY
QKMSTKHPTYWSTFSEVWITAETYLEWGHWREMWAAYLMTVQVWVRPQFRD
VNQYEFPKFRSGGDYALGAYREPMGTQRIMAPLIAEPKFTNTTIHVSDWGT
NIWSHTDHKWHEPWEYGQTQEPGIKNFKGHKQHDWQQTTIHVSDWGTNIHS
HTDHKWHEPWEAMHDIHFWMMYYSETIHTSDQGTNIHGHTDHKWHEPWLAR
DLSFECDFAMHGHMTYPYTDKTIWVSDWGTNVHSHTDHRWHEDWEA
```

B. Why might you look at a protein against itself in a dotplot?