BioE/MCB/PMB C146/246, Spring 2005

Problem Set 6

Due **Tuesday**, 8 March 2005 by **11 am** to lfl@compbio.berkeley.edu. Turn in a hard copy in class. NOTE: Solutions will be available Tuesday evening. Thus, we can not accept late problem sets turned in after 6pm.

1.   10 points

Given two sets of 5 sequences of length 500, sometimes MSA will run slowly and sometimes it will run very slowly. What features of the sequences and their alignments would cause this to happen?

2.   10 points

What is the meaning of "once a gap, always a gap"? Name three multiple alignment programs that have this "feature" and three that do not.

3.   10 points

Name and justify three heuristics used by CLUSTALW.

4.   15 points

(A) How are trees used by CLUSTALW in conducting multiple sequence alignment?

(B) Perform a CLUSTALW alignment on the orthologous sequences in `hmg.fasta` (data on the course website). Does the CLUSTALW output tree agree with the species tree?

5.   15 points

The following four sequences are members of the `Arthro_defensin` protein family (found in Pfam, http://www.sanger.ac.uk/Software/Pfam/). Perform pairwise alignments (global alignment with **no end gap penalty**) between the first sequence and each of the other sequences, and assemble a master-slave alignment. Compare your alignment with the Pfam alignment.

```
>DEFI_APIME/53-82
GQVNDSACAANCLSLGKAGGHCEKVGCICR

>DEFI_AESCY/1-37
GFGCPLDQMQCHRHCQTITGRSGGYCSGPLKLTCTCY
```

```
>DEFA_ZOPAT/10-43
IAGTKLNSAACGAHCLALGRRGGYCNSKSVCVCR

>SAPC_SARPE/10-39
GVQHSACALHCVFRGNRGGYCTGKGICVCR
```

6.   20 points

The following alignment can be downloaded from the course website.

```
ATTDEWKKQRKDSHKEVERRRRENINTAINVLSDLLPVRESSKAAILACAAEYIQKLKETDEAN
HGSEEWHRQRRENHKEVERKRRESINTGIRELARLIPTTDTNKAQILQRAVEYIKRLKENENNN
TGSTAWKQQRKESHKEVERRRRQNINTAIEKLSDLLPVKETSKAAILSRAAEYIQKMKETETAN

IEKWTLQKLLSEQNASQLASANEKLQEELGNAYKEIEYMKRVLRK----------
IEKWTLEKLLTEQAVSELSASNEKLKHELESAYREIEQLKRGKK----------
IEKWTLQKLLGEQQVSSLTSANDKLEQELSKAYKNLQELKKKLKEAGIEDPTEEE
```

Ignoring any columns with gaps,
(A) Calculate the minimum entropy score for this alignment.

(B) Calculate the sum of pairs score for this alignment, based on BLOSUM62.

7.   15 points

(A) Describe the idea of "consistency" as used in ProbCons. What other alignment program that we discussed uses a similar concept in building its alignment?

(B) What general alignment strategies are used in ProbCons? What general alignment strategies are used in MUSCLE?

8.   5 points – for taking the midterm (March 10).