Study questions

These study problems are intended to help you to review for the final exam. This is not an exhaustive list of the topics covered in the class and there is no guarantee that these questions are representative of the questions on the final exam. You should also review your class notes, the syllabus and your home work assignments.

1. Pairwise sequence alignment

   (a) Compare, dynamic programming for *local*, *global* and *semiglobal* alignment in terms of when they should be used; differences in initialization and termination condition; recursion relation, which scoring functions are suitable.

2. Multiple sequence alignment

   (a) What are the differences between the exact dynamic programming for multiple sequence alignment and the progressive alignment heuristic?

   (b) Score the following multiple sequence alignment using sum-of-pairs. Compare the scores and explain why they differ?

```
A
G
G
A
T
```

3. Phylogeny reconstruction

   (a) Concepts:

      i. Taxa, leaf nodes, internal nodes, root, branch lengths.
      ii. Rooted versus unrooted trees; what methods are used to root a tree? For each method, under what circumstances can they be applied?
      iii. Gene trees versus species trees
      iv. How many trees with k leaves?
      v. How to enumerate trees and/or traverse tree space.
      vi. Models of sequence substitution, what are they how are they used?
      vii. Character data: characters and character states
      viii. Distance data: how distances are derived from sequence data; why it is necesary to correct such distances; additive and ultrametric distances.

   (b) For each of the phylogeny reconstruction methods listed below,

      i. Which types of information can they infer?
      ii. If a method cannot infer all of the types listed, what biological or algorithmic assumptions impose these limitations?
      iii. Which reconstruction methods are appropriate for each of the types of data listed above? Why?

Phylogeny reconstruction methods

- UPGMA
- Neighbor Joining
- Maximum parsimony
- Maximum likelihood

Information inferred by such methods

- unrooted tree topology
- rooted tree topology
- branch lengths,
- labels on internal nodes

Sequences

- nucleotide
- amino acid

- evolving rapidly
- under selective pressure
- evolving at a constant rate in all lineages

4. Review Fitch's algorithm.

   (a) What does it do?

   (b) How does it work?

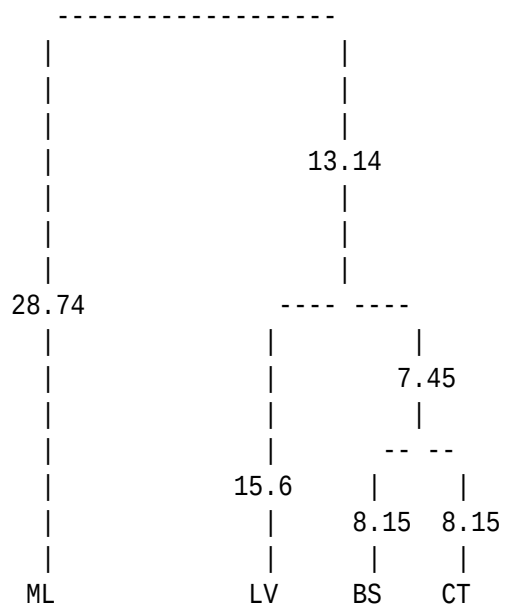
   (c) What can't it do?

5. Tree reconstruction

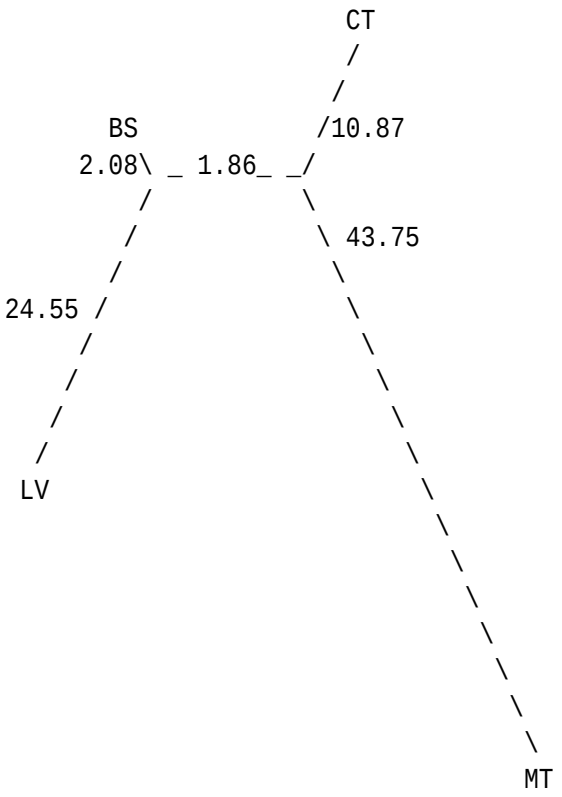
   Below you see trees constructed using 5S small subunit ribosomal RNA genes from four bacterial species:

```
M10815 Bacillus subtilis (BS)
X02713 Lactobacillus viridescene (LV)
M58416 Clostridium tyrobutyicum (CT)
K02683 Micrococcus luteus (ML)
```

UPGMA Tree:



Neighbor-Joining tree:



(a) Are the topologies the same? If not, how do they differ?

(b) In which tree , is the distance from *Lactobacillis viridescens* to *Micrococcus luteus* most distorted with respect to the original distance matrix derived from the multiple sequence alignment?

```
DISTANCE MATRIX FROM OUTPUT

Key for column and row indices:

    1 BSubtilis
    2 CTyro
    3 LViridescens
    4 MLuteus

Matrix 1: Part 1

               1         2         3          4
 -------------------------------------------------
|    1    |   0.00      16.31     26.63     46.18
|    2    |              0.00     35.77     54.62
|    3    |                        0.00     71.66
|    4    |                                  0.00



Distance derived from the data: 71.66
Distance derived from the NJ tree: 43.75 + 1.86 + 24.55 = 70.16
Distance derived from the UPGMA: 15.60 + 13.14 + 28.74  = 57.48
```

(c) Why might UPGMA give a different topology than NJ for these sequences?

6. Short questions

   (a) Using a PSSM instead of a single query sequence in a database search can result in improved retrieval of distantly related motifs. What might this be the case?

   (b) Explain why we use pseudocounts to correct for the zero frequency case when building profiles.

   (c) What are the properties that HMM's can capture that PSSM's cannot?

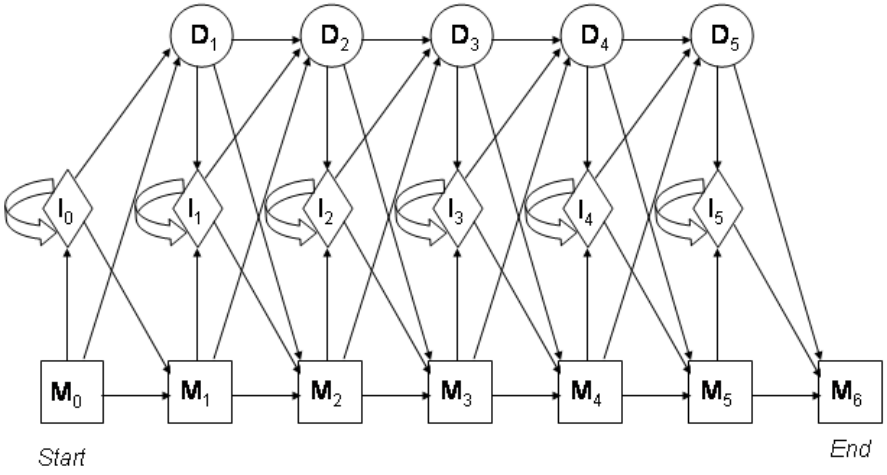
   (d) How would you use an HMM to calculate a global multiple alignment?

   (e) What is meant by "unlabeled" data? By "labeled" data?

   (f) Generally, it is believed that the BLOSUM matrices give better results than the PAM matrices. Nevertheless, PAM30 and PAM70 are still used when searching for very closely related sequences. Why?

   (g) What is model surgery? How would you determine whether model surgery is required? If it is required, what steps would you take to modify your HMM?

   (h) Consider the profile HMM shown below:



   i. M_0 and M_6 are silent start and end states. Since these states do not emit a symbol, what purpose do they serve?

   ii. What is the purpose of the I_0 state?

7. Define an HMM $\mathcal{H}$ with three states $\{A,B,C\}$, and alphabet $\{1,2,3\}$, initial state probabilities $\pi_A = 1, \pi_B = 0$ and $\pi_C = 0$ and the following transition and emission probabilities:

	A	B	C	1	2	3
A	0.25	0.25	0.5	0.5	0.25	0.25
B	0.5	0.00	0.5	0.5	0.0	0.5
C	0.33	0.33	0.33	0.0	0.5	0.5

   (a) Draw the state diagram of this HMM and show the transition probabilities.

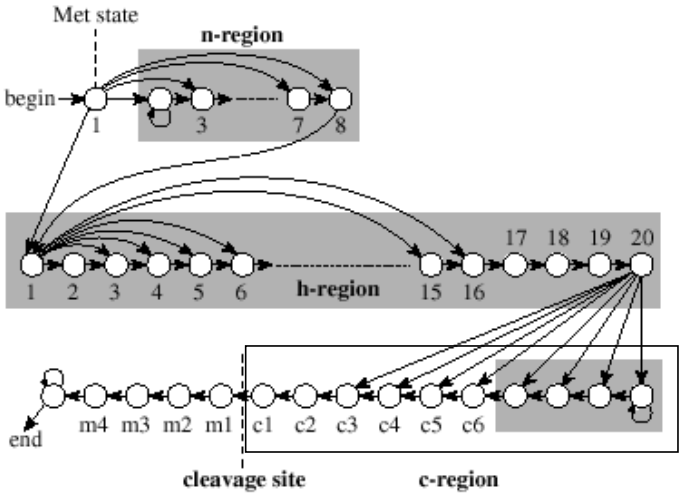
   (b) Give all the possible state paths for the sequence $O = 1,3,1$.

   (c) What is $P(O|\mathcal{H})$?

   (d) What is the most probable path? Give the probability of O for this path.

   (e) For this HMM, would the Viterbi algorithm be a good approximation for Forward algorithm. Why or why not?

8. Signal peptides control the entry of proteins into the secretory pathways. These peptides are typically less than 40 residues long, contain a charged N-terminal region, a central stretch of hydrophobic residues and a cleavage site preceded by three to seven polar residues. The following HMM to recognize signal peptides, proposed by Nielsen and Krogh, contains an n-region module, an h-region module and a c-region module:



(a) Which of the three modules, if any, recognize subsequences with geometric length distributions?

(b) What is the minimum length of the n-region sequences this HMM will recognize?

(c) What is the minimum length of the h-region sequences this HMM will recognize?

(d) What is the maximum length of the h-region sequences this HMM will recognize?

9. What are the similarities and differences

   - between the Jukes Cantor and the Kimura 2 Parameter models?

   - between Jukes Cantor and PAM?

   - between PAM and BLOSUM?

10. BLAST

    What is the impact on

    - the speed of the heuristic
    - the number of false negatives
    - the number of false positives

    of the following changes in BLAST parameters

    (a) increase/decrease w

    (b) increase/decrease T

    (c) increase/decrease A

    (d) increase/decrease

    (e) increase/decrease S

11. Give an expression for the probability of a path of length l through the HMM shown below, where q_1 is the start state and q_5 is the end state. (Assume an alphabet with one symbol and all states q_i emit that symbol with probability equal to one.)