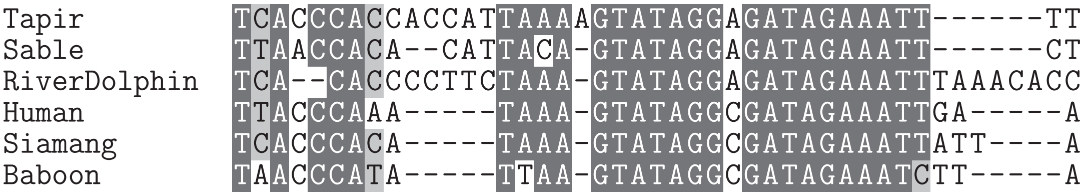
Math 5233 practice exam. The actual exam will be somewhat shorter than this (5-6 questions). You may use notes, including worksheets, but not the text. It may be helpful to have a calculator as well.

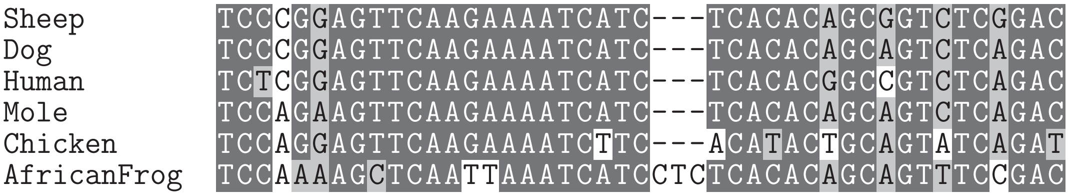
(1) (a) What are mitochondria?
(b) Why are they important for evolutionary studies?
(c) Why might a scoring matrix derived from mitochondrial sequences differ from one derived from nuclear sequences?

(2) Decide whether each of the alignments below are from protein-coding regions of DNA or not. Explain your reasoning.



```
Tapir         TCACCCACCACCATTAAAAGTATAGGAGATAGAAATT------TT
Sable         TTAACCACA--CATTACA-GTATAGGAGATAGAAATT------CT
RiverDolphin  TCA--CACCCCTTCTAAA-GTATAGGAGATAGAAATTTAAACACC
Human         TTACCCAAA-----TAAA-GTATAGGCGATAGAAATTGA-----A
Siamang       TCACCCACA-----TAAA-GTATAGGCGATAGAAATTATT----A
Baboon        TAACCCATA-----TTAA-GTATAGGCGATAGAAATCTT-----A
```

Figure 1. A



```
Sheep         TCCCGGAGTTCAAGAAAATCATC---TCACACAGCGGTCTCGGAC
Dog           TCCCGGAGTTCAAGAAAATCATC---TCACACAGCAGTCTCAGAC
Human         TCTCGGAGTTCAAGAAAATCATC---TCACACGGCCGTCTCAGAC
Mole          TCCAGAAGTTCAAGAAAATCATC---TCACACAGCAGTCTCAGAC
Chicken       TCCAGGAGTTCAAGAAAATCTTC---ACATACTGCAGTATCAGAT
AfricanFrog   TCCAAAAGCTCAATTAAATCATCCTCTCACACAGCAGTTTCCGAC
```

Figure 2. B

(3) Describe how a progressive pairwise alignment of the sequences ACCCGC, ACCGC, ACTCTC, and ACGCGC (in that order) could generate a sub-optimal multiple alignment.

(4) Using a scoring system of +5 for a match, -3 for a mismatch, and -4 for a gap, use the Smith-Waterman algorithm to fill in a dynamic programming table and find the best local alignment between the sequences TAGCA and TAAGCG. If there are ties, list all of the best alignments.

(5) Suppose two different types of nucleotide sequences are used to study the evolutionary distances between humans, chimps (Pan troglodytes), and mice (Mus musculus). The first sequence type is mRNA of the H4 histone protein, and the fraction of sites that differ between the species are: human-mouse: 40/356, human-chimp: 5/356,

and chimp-mouse: 43/364. The second sequence type is the D-loop region of the mitochondrion, and for that the fractional differences are: human-mouse: 527/1092, human-chimp: 138/1107, and chimp-mouse: 531/1088.

(a) What would you conclude from each type of sequence if you used the Jukes-Cantor model?

(b) What are some possible problems with using each type of sequence with the Jukes-Cantor model? What sort of third sequence type would be good to use to check your results?

(6) Suppose that P and Q are probability distributions on N items - that is, $P = \{p_1, \dots, p_N\}$ and $Q = \{q_1, \dots, q_N\}$ are both lists of N numbers between 0 and 1, which each sum to 1.

(a) Show that the difference in entropy

$$H_Q - H_P = -\sum_{i=1}^{N} q_i \log q_i + \sum_{i=1}^{N} p_i \log p_i$$

is equal to the relative entropy of P relative to Q

$$H(P||Q) = \sum_{i=1}^{N} p_i \log(p_i/q_i)$$

if Q is the uniform distribution (each item has a $1/N$ probability).

(b) If $N = 2$, for a fixed distribution P for which distributions Q will $H_Q - H_P = H(P||Q)$?

(c) Extra Credit: same as the previous question, for $N = 3$. Describe the situation as completely as possible.

(7) Suppose we view each of the approximately 30,000 proteins in the human genome as a message from the genome to the cell, and suppose that their frequency of expression follows Zipf's law: the most frequent protein is twice as likely as the second-most frequent protein, three times as likely as the third-most frequent protein, and so on.

(a) What is the entropy of this distribution?

(b) How does it compare with the entropy of the uniform distribution for 30,000 items?

(8) For the DNA scoring matrix which has $S_{ii} = 3$ and $S_{ij} = -5$ for $i \neq j$, if you assume a uniform background distribution of nucleotides $P_A = P_C = P_G = P_T$ calculate the implied transition matrix Q_{ij} and the (relative) entropy $\sum_{i,j} Q_{ij} \log_2(\frac{Q_{ij}}{P_i P_j})$.