Practice final for Math 5233.

- What is the difference between a transversion and a transition between nucleotides? What is the significance of this for bioinformatics?

- Using a scoring system of $+3$ for a match, -2 for a mismatch, and -1 for a gap, use the Needleman-Wunsch algorithm to fill in a dynamic programming table and find the best local alignment between the sequences $AACTG$ and $ACCG$. If there are ties, list all of the best alignments.

- What are some of the possible problems with using the BLOSUM scoring matrices to identify proteins in Plasmodium falciparum?

- What are some of the advantages of the T-Coffee algorithm over Clustal?

- What is the apicoplast in Plasmodium falciparum and why is it interesting?

- The PROSITE pattern for G6PD (glucose-6-phosphate dehydrogenase) is `D - H - [YF] - L - G - K - [EQK]`.
  How many 7-amino acid sequences match this pattern? How many 7-amino acid sequences are there? If you found a match to this pattern in the Macaca mulatta (rhesus monkey) genome, how confident would you be that it was orthologous to G6PD? What could you do to check?

- Using the complete mitochondrial genomes, the Jukes-Cantor distances in the table below were calculated. Use the UPGMA algorithm to construct a possible phylogeny for these 5 species.

	Chimp.	Gorilla	Orang.	Human	Pigmy Chimp
Chimp.	0	80	120	65	30
Gorilla	80	0	120	80	80
Orangutan	120	120	0	120	120
Human	65	80	120	0	65
Pygmy Chimp.	30	80	120	65	0

- Why is the quality of multiple sequence alignment important in molecular phylogeny?

- Explain why the statement "protein A and protein B are 50 percent homologous" is incorrect.

- Suppose the entries of a scoring matrix are computed from the formula $s(a,b) = \log_{10} \frac{p_{ab}}{f_a f_b}$. How would you interpret an entry equal to -5?

- What is the information content in bits of the sequence $A-A-A-C-G-T$ if the background frequencies of the nucleotides are equal?

- What is a greedy algorithm?