Practice final for Math 5233. This should give a sense of the sort of questions that will be on the exam. You are allowed to use your notes for the exam, but not the textbook or printed articles.

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

(2) Using a scoring system of $+3$ for a match, -2 for a mismatch, and -1 for a gap, use the Smith-Waterman 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.

(3) Calculate the score (in bits) of matching the sequence ACGG with a PSSM derived from the aligned sequences

```
A C G G
A C A G
A C A A
T C A A
```

assuming a uniform background distribution of nucleotides.

(4) What are G-protein coupled receptors? How can they be recognized from their amino acid sequences?

(5) 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?

(6) 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

FIGURE 1.

(7) What are the assumptions of the Jukes-Cantor model? Do you think that sites in the multiple DNA sequence alignment in Figure 1 satisfy those assumptions? Why or why not?

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

(9) Suppose the entries of a DNA positional scoring matrix are computed from the formula $s(a, b) = \log_2(4f_{ab})$, where f_{ab} are empirical frequencies of DNA substituions (a is the column, b is the nucleotide). How would you interpret an entry equal to -5? What assumptions are being made about the sequence?

(10) True or False: it does not matter which program you use to compute a maximum parsimony phylogenetic tree - you will get the same answer as long as you use the same data, since there is a unique optimal solution. Explain your answer.

(11) What are the characteristics of a dynamic programming algorithm?