CS284A: Algorithms for Molecular Biology Assignment #3

Exercise 1

Perform pair-wise alignment between the following two sequences:
Seq A: `ACCGCGCATGCC` and Seq B: `ACCGCATAGCA`.
Use the following scoring scheme: match = 2, mismatch = -1, and gap = -2.

1. Perform Needleman-Wunsch global alignment by filling in the following dynamic programming matrix. Report the best alignment score and pair-wise alignment.

	-	A	C	C	G	C	G	C	A	T	G	C	C
-													
A													
C													
C													
G													
C													
A													
T													
A													
G													
C													
A													

2. Perform Smith-Waterman local alignment by filling in the following dynamic programming matrix. Report the best alignment score and pair-wise alignment.

	-	A	C	C	G	C	G	C	A	T	G	C	C
-													
A													
C													
C													
G													
C													
A													
T													
A													
G													
C													
A													

Exercise 2

Suppose we have derived the following dinucleotide counts from two sets of human DNA sequences. The (+) set corresponds to DNA sequences from CpG islands and the (-) set corresponds to DNA sequences that are not from CpG islands.

Dinucleotide	(+) Set Count	(-) Set Count
AA	180	300
AC	274	205
AG	426	285
AT	120	210
CA	170	322
CC	368	298
CG	274	78
CT	188	302
GA	161	248
GC	339	246
GG	375	298
GT	125	208
TA	79	177
TC	355	239
TG	384	292
TT	182	292

1. Draw 2 separate Markov chains, one Markov chain (MC+) to represent DNA sequences from CpG islands and a second Markov chain (MC-) to represent DNA sequences that are not from CpG islands.

2. Find the maximum likelihood estimates for transition probabilities in both Markov chains (MC+ and MC-). Fill in the transition probabilities in the following table:

MC+	A	C	G	T
A				
C				
G				
T				

MC-	A	C	G	T
A				
C				
G				
T				

3. Suppose we want to use these models for discrimination of the following two DNA

sequences seq1=GCAC and seq2 = GCTC.

Calculate the log-likelihood ratios S(seq1) and S(seq2) for these two DNA sequences ($\log_2$ based, ratio between MC+ and MC-). Which sequence do you think is from a CpG island? Explain.