Homework 1 (due 2/7/05)

1. Find the optimal alignment of the following two sequences:
   AGGCTATCACCTGACCTCCAGGCCGATGCCC
   TAGCTATCACGACCGCGGTCGATTTGCCCGAC

The scoring functions are: match=1, mis-match=-1, and indel=-1.

You are asked to complete the alignment matrix, find the optimal alignment path, and given the optimal alignment score.

2. As we learned during the lecture, the maximal likelihood estimation of the model parameter can be obtained by first writing down the likelihood of the observed data (e.g., DNA sequences) under a candidate model (e.g., an HMM) with parameters of unknown value, and then find the argmax of the likelihood for the parameters. Let $\{Y_{t,n}\}$ denote the hidden variable sequence (t index position in the sequence and n index each sequence, $\{X_{t,n}\}$ denote the observed sequence, $\{A_{i,j}\}$ denote the number of times a transition between state i and j take place, $\{E_{i,k}\}$ denote the number of times state i emit symbol k, and $\{I_i\}$ denote the number of times the initial state of n given sequences are in state i:

   a. Supervised learning: Given both $\{X_{t,n}\}$ and $\{Y_{t,n}\}$, derive the formula for the maximum likelihood estimation of the initial, transition, and emission probabilities of the HMM (as functions of $\{A_{i,j}\}$, $\{E_{i,k}\}$, $\{I_i\}$, which are called the *sufficient statistics* of the parameters, and can be counted from the values of $\{Y_{t,n}\}$ and $\{X_{t,n}\}$. Note that once we get these counts, we do not need to keep the sequence and state data because the counts are *sufficient* for estimating the parameters.).

   b. Unsupervised learning: When only the sequences, i.e, $\{X_{t,n}\}$ are given, but not the underlying state sequence (i.e., $\{Y_{t,n}\}$), the above algorithm breaks down. In an EM algorithm, we replace the actual counts $\{A_{i,j}\}$, $\{E_{i,k}\}$, $\{I_i\}$ with their expectations given the sequences in the so-called "E" step. The expectations are computed as following:

$< A_{i,j} > = \Sigma_{t,n}\, p(Y_{t,n}=i, Y_{t+1,n}=j|X)$
$< E_{i,k} > = \Sigma_{t,n}\, p(Y_{t,n}=i|X)\delta(X_{t,n}=k)$
$< I_i > = \Sigma_n\, p(Y_{0,n}=i|X)$
where $\delta(.)$ is an indicator function which equals to 1 if the argument is true and 0 otherwise.

How these formulas relate to the forward-backward algorithm?

(In the "M" step, we do the ML estimation as in problem "a" to get an *estimator* of the parameter. Then we return to the "E" step the re-compute the $< A_{i,j} >$, $< E_{i,k} >$, and $< I_i >$. Then we do "M" step again, and iterate. The is the famous Baum-Welch algorithm.)

c. Does posterior decoding always give a valid state sequence (i.e. a state sequence permissible by the initial and transition matrix of the model)? Why?

3. Implement a forward-backward algorithm and the viterbi algorithm, and compute the likelihood of a sequence I will put on the web, and the estimates of the hidden states using viterbi and posterior decoding. Compute the posterior probabilities of the state-sequences for both decoding. Be careful about the data underflow issue. Rescale the forward and backward probability if necessary. (Bonus, upgrade your algorithm to a generalized HMM that use a uniform duration distribution of 1, …, K.) Discussions among students are allowed, but you should not copy each other's code.

```
Here is the HMM model. It has three states: bk, intron, exon.

initial_prob=[0.9,0.05,0.05]',


The transition probabilities between these states are as follows
(p(j|i) corresponds to the (i,j) entry in the matrix)):

transmat=[0.90,0.10,0.00;
      0.00,0.90,0.10;
      0.01,0.09,0.90]


And here is the emission matrix (p(character k|i) corresponds to the
(i,k) entry in the matrix):

obsmat=[0.25,0.25,0.25,0.25;
      0.35,0.35,0.15,0.15;
      0.15,0.15,0.35,0.35]

The observation sequence is posted on the web. It is sequence of 200
letters long.
```

4. What is the difference between the MEME and the AlignACE algorithm for motif detection? Are they really different in terms of the underlying model? Discuss how to justify your answer.

5. Read the MEME paper and implement the MEME algorithm for motif detection.