1. Give a clear statement of the problem.

   We are provided with 10 simulated DNA sequences of 300 bases in length that were simulated using an unknown two state HMM. Given the two states G-C (l) and A-T (k), we need to find realistic estimates for the Transition Matrix and Emission Matrix of the data.

2. Give a brief description of the method of your procedure. This should be a few sentences describing the problem, solution, and results. It should be similar to an abstract in a scientific paper.

   Finding realistic estimates for the Transition Matrix and Emission Matrix for the data requires us to implement the forward-backward algorithm first. We then use the resulting forward and backward states along with the forward probability in order to re-estimate the Transition matrix and Emission Matrix. These values are evaluated for each of the ten sequences and then the log likelihood is computed. Until the current log likelihood and previous log likelihood converge to less than 10^-4, the procedure reiterates.

3. Present the procedure in detail, including any relevant mathematics. If the solution involves programming, give an overview in words of how the program operates. The presentation should contain sufficient detail so that a reader can follow each step without difficulty. *

   Note: The following procedure recursively calls itself until the previous log likelihood and current log likelihood difference converges on a number less than 10^-4. The functions within the iterate procedure that are called are pseudo coded below this procedure.

```
procedure iterate(records, states, A, E, prev_log, count):

    initialize A, E, log_likelihood, p_fwd_total to 0's
    for each sequence in the records:
        forward, backward, p_fwd = forward-backward(sequence, states, A, E)

        A_temp, E_temp = baumWelch(sequence, states, forward, backward, A, E)

        #complete summations
        log_likelihood += log(p_fwd)
        p_fwd_total += p_fwd
        add A_temp to A
        add E_temp to E

        divide A and E by p_fwd_total

        normalize the A and E matrices
        #check that convergence is met
        if (log_likelihood - prev_log) < 10^-4 :
            return A, E, log_likelihood

        #break iterations if taking too long to converge
        else if count > 150:
            return A, E, log_likelihood

        else:
            return iterate(records, states, A, E, prev_log, count)

procedure fwd-bwd (sequence, states, A, E):

    #forward
    init forward list
    for i, x_i in enumerate(sequence):
        for st in state:
            base case is prev_f_sum = 0.5
            else prev_f_sum = sum(f_prev[k] * A[k][st] for k in states)
            f_current[st] = E[st][x_i] * prev_f_sum
        add f_current to forward
    p_fwd is sum of f_currents for each state

    #backward
    init backward list
    for x, x_i in enumerate(sequence):
        for st in states:
            base case is b_current = 1
            else b_current = sum(A[st][l] * E[l][x_i] * b_prev for l in states)
        insert b_current at front of backward
    p_bwd = sum(0.5 * E[l][sequenece[0]] * b_current[l] for l in states)

    assert b_fwd == p_fwd

    return foward, backward, p_fwd

procedure baumWelch(sequence, states, forward, backward, A, E):

    #emission calculation
    for i in L:
        for st in states:
            sum the new emission for the current state and nucleotide sequence
    #transition calculation
    for i in L:
        for k in states:
            for l in states:
                sum new transition with (forward[i][k] * A[k][l] * backward[i+1][l] * E[l][sequence[i+1]])

    return two new matrices
```

4. If the solution involves programming code, output must be clearly labeled with each item having a descriptive title. Programming code must be documented with comments describing each major step in the program and all relevant variables. Source code should be submitted in a separate file. *

   Source code with comments is attached. The output clearly displays the initial matrices, log likelihood and the new transition matrices that are found after convergence.

5. Graphs must be titled and have each axis labeled clearly

   No graphs in this assignment.