Lab 4

1. Write a program that implements a 2-state HMM for detecting G+C rich regions in the Vibrio cholerae IEC224 chromosome II sequence (Genbank file). Conceptually, state 1 will correspond to the more frequent 'A+T rich' state, whereas state 2 will correspond to the less frequent G+C-rich state. Specifically:
    a. The starting parameter values should be as follows:
        i. Transition probabilities a_ij are a_11 = .999, a_12 = .001, a_21 = .01, a_22 = .99.
        ii. Initiation probabilities for each state (i.e. the transition probabilities from the 'begin' state into state 1 or 2) should be .996 for state 1, and .004 for state 2; these should be held fixed throughout the Viterbi training
        iii. Emission probabilities (which should also be held fixed) are
            1. e_A = e_T = .291, e_G = e_C = .209 for state 1;
            2. e_A = e_T = .169, e_G = e_C = .331 for state 2.
    b. Use Viterbi training to find improved parameter estimates for the transition probabilities, holding the emission and initiation probabilities fixed at the above values. Run the training for 10 iterations, where for each iteration you:
        i. Use dynamic programming to find the highest probability underlying state sequence.
        ii. Using this state sequence, compute
            1. The number of states of each of the two types (1 and 2), and the number of segments of each type (where a segment consists of a contiguous series of states of the same type, that is preceded and followed by states of the opposite type or the beginning or end of the sequence).
            2. New transition probabilities to be used in the next iteration.
2. Your output should provide
    a. the name and first line of the .fna file
    b. the information described above (in 2. -- i.e. numbers of states and segments, and new probability values), for each of the 10 iterations. Give probabilities to 4 decimal places only.
    c. the list of G+C-rich segments (corresponding to the segments having state 2 as the underlying state) after the final (10th) round of Viterbi training, sorted by genomic position.
    d. your description of the first 5 of the segments found above, as found by looking up the Genbank annotations.
3. Using the same HMM and dataset as in step 1, write a program that implements EM (Baum-Welch) training instead. Use the same starting parameter values, but in contrast to step 1, you should not hold any parameter values fixed -- allow all of them to change with each iteration. Compute the log-likelihood (to the base 2) of the sequence at each iteration, and run the program until the increase in log-likelihood between successive

iterations becomes less than .1. You should check that the loglikelihood increases with each iteration -- if it doesn't, something is wrong with your program.

4. Your output should provide
    a. the name and first line of the .fna file
    b. the number of iterations until convergence
    c. the final log-likelihood
    d. the final emission and transition probabilities -- please output these in scientific notation, to four significant digits (i.e., 9.000e-1)