

## Final Assignment

Write a program that implements a 2-state HMM for detecting G+C-rich regions in the [Vibrio cholerae IEC224 chromosome II sequence](#). Conceptually, state 1 in the HMM would correspond to the more frequent 'A+T rich' state, whereas state 2 would correspond to the less frequent G+C rich state. Specifically:

- a. The starting parameter values should be as follows:
  - i. Transition probabilities  $a_{11} = .999$ ,  $a_{12} = .001$ ,  $a_{21} = .01$ ,  $a_{22} = .99$ .
  - 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.

Your output should provide:

- c. the name and first line of the .fna file
  - d. the information described above (i.e. numbers of states and segments, and new probability values), for each of the 10 iterations. Give probabilities to 4 decimal places only.
  - e. 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.
- Note: If for any reason your code was not working completely by the deadline, you can submit the pseudocode and/or description of your approach for a partial grade.