Algorithms in Computational Biology – Homework Exercise 3

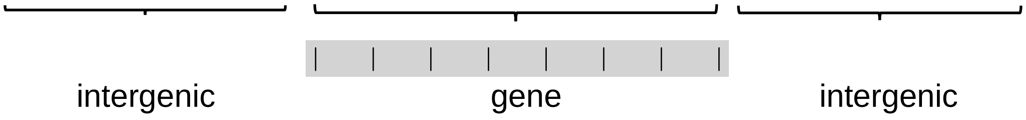
Publication date: *Thursday, December 10*
Due date: *Sunday, December 27 (9pm IST)*

Researchers discovered a new virus (*oy vei*) whose DNA has the following peculiar features:

- Protein coding genes have high composition of `C`'s and `G`'s (40% each) than `T`'s (20%), and no `A`'s.
- DNA sequences outside of protein coding genes (termed intergenic) have the opposite bias in composition: 20% `C`'s and `G`'s and 30% `A`'s and `T`'s.
- Protein coding genes are always flanked by an `A` (before the gene) and `T` (after the gene). The flanking bases are not part of the protein coding sequence.
- Similar to other organisms, protein coding genes in this virus consist of a series of codons of length 3. The length of a protein coding gene is geometrically distributed with average length of 5 codons (see note below). The length of an inter-genic segment (between terminating `T` and next starting `A` and) is also geometrically distributed with average length of 20 bases. Note that a gene is never empty, but two genes may be separated by a terminating `T` followed by a starting `A`.

The following is a typical sequence in the virus' DNA with its gene annotation given below (including the boundaries of the seven codons in the gene):

```
... ACTAATGGTAGATCCGCTTGCCGGTGCTTCTCGTTCAGAAATCA ...
```



Note: a random variable X is said to be geometrically distributed $X\sim Geom(p)$ if: $P(X=x) = (1-p)^{x-1}p$. The mean of such a variable is $E[X]=1/p$. Notice that the length of a consecutive sequence of annotation of the same state in an HMM is geometrically distributed with parameter p determined by the probability of transition out of that state.

a. What are the nine possible annotations of the following fragment of viral DNA sequence? Enumerate all possible annotations and explain your answer.

```
CCATCGCACTCCGATGTGGCCGGTGCTCACGTTGCCT
```

b. Describe a Hidden Markov model (HMM) that models the DNA sequence of the virus together with the appropriate annotations.

- The emission characters of your HMM should be A,C,G,T.
- Label the states by S1, S2, etc., and write a brief description for each state. Examples: "S1 - intergenic base" ; "S2 - first base of start codon".
- Clearly specify the transition and emission probability matrices.
- You may draw the finite state machine to help describe the structure of your HMM.

There is an HMM that models the virus DNA with six states. I couldn't find anything smaller, but maybe you can…

Submit your HMMs for review by e-mail (ilan.gronau@idc.ac.il) by Thursday, December 17, to make sure that you proceed with the assignment with the correct HMM.

c. Specify the nine different paths in your HMM that correspond to the annotations you specified in (a). Connect each path with an annotation you specified in (a). Paths should be written below the sequence as follows:

```
C C A T C G C A     . . .    C
1-1-1-1-1-1-1-2-    . . .   -1
```

(numbers indicate state ids)

d. Implement a program for the Viterbi algorithm on this HMM.

- Your program should receive a sequence over the alphabet `A,C,G,T` as input and produce the sequence of most likely annotations (hidden states).
- The transition and emission probabilities may be hard coded into the program, and you may also hard code the input DNA sequence, but make sure it is easy enough to change the input sequence when needed.
- The program should output the most probable sequence of states given the input sequence and the log-probability of the hidden state jointly with the data (log(P(*X,S*|HMM))).
- Make sure that your program performs calculations using log-probabilities (use **natural log (*ln*())** as shown in Lecture #6).
- You may assume that the first base in the sequence is intergenic.
- Submit your documented code in an appendix file.

e. Run your program on the sequence specified in (a) above. Which of the possible paths is the most likely, and what is its log-probability?

f. Implement a program that computes a posterior-decoding for this HMM.

- Your program should receive a sequence over the alphabet `A,C,G,T` as input and produce for each position *i* along the sequence the HMM state *s* with highest posterior probability for that position: P(S_i=*s*|X,HMM).
- The program should print for each position the state and its posterior **conditional probability** (values should sum to 1 across states), and also print the log likelihood for the entire sequence (log(P(*X*|HMM))).
- Do this by implementing procedures for computing the forward and backward matrices, and follow the coding guidelines specified in (d) above. Namely, make sure that your program performs calculations using log-probabilities.
- Make use of the technique presented in Lecture #6 to compute log-of-sum-of-exponents. Submit your documented code in an appendix file.

Debugging tip:

- Start by implementing a version of the forward and backward algorithms that use standard probability calculations (without using the log-of-sum-of-exponents trick).
- Validate your calculations by making sure that the same value of the likelihood is retrieved by taking the inner product of every column of the Forward matrix with its counterpart in the Backward matrix.
- Apply the log-of-sum-of-exponents trick to each algorithm and make sure that the obtained values are logs of the original ones.

g. Run your program from (f) on the sequence specified in (a) above. What is the most likely state for each position in the sequence? Compare your results to the ones obtained in (e).

h. After further study of this virus, the scientists discovered that in some occasions its protein coding genes do contain `A`'s. So, they changed the model to allow `A`'s in genes with frequency of 5%, and `T`'s with frequency 15%. Other aspects of the model remained unchanged. Change your programs from (d) and (f) to reflect this model change and re-run them on the sequence specified in (a). Report the results. Did anything change?

i. **Bonus (5pts.):** Provide an example of an observed DNA sequence for which the most likely annotation (output of Viterbi's algorithm) and the posterior decoding results in a different sequence of states. Specify the DNA sequence and show the two resulting annotation sequences. This question involves empirical experimentation. Think of what might cause discrepancy between the two types of paths and design data accordingly. Use the program you wrote for (f) above to experiment with different inputs until you find one that satisfies the requirements. For complete bonus credit, provide an example in which the posterior decoding annotation that has probability zero under the model (i.e., contains a zero-probability transition).

Submission Instructions:

- Submit your work on the course **Moodle website** by Sunday, Dec 27 @21:00.
- For items (a), (b), (c) submit your written/typed solution as usual.
- For the experimental results in items (e), (g), (h), (i) provide your program's output and an explanation of the results.
- Submit the code for items (d), (f), (h) as appendix to your solution.
- Type your solutions or write legibly and scan. **If you scan, make sure the scan came out fine.**
- **Submit your work in pairs!** One student should submit a solution file with both of your student IDs specified. The other student should submit a simple text file with your two student ids, to help us match back the grade to both of you.
- If you consult with other pairs on ideas, specify their names clearly on the first page and make sure that they **acknowledge your collaboration** as well.
- You have two weeks to complete the assignment. **Plan your time wisely**. Extensions due to special circumstances, will be granted only upon **request by e-mail at least 48 hours** before the deadline. No last minute extensions!
- Please post any questions that you have on the course Piazza website: https://piazza.com/idc.ac.il/fall2020/cs3571/.