Homework 5

CS 181, Fall 2024

Out: Nov. 19
Due: Dec. 6, 11:59 PM

Please upload your solutions on Gradescope. You can use LaTeX or a word document to write up your answers, but we prefer you use LaTeX. You may scan hand-written work or images for parts of solutions **only if** they are extremely clean and legible. Please ensure that your name does not appear anywhere in your handin.

Problem 1: Hidden Markov Models

A HMM has been constructed to generate a sequence, O, of symbols consisting of 2 states $S = \{1, 2\}$. Each time the model visits a state, one symbol (A, C, T, or G) is generated. The emission probability distribution for state 1, or b_1, is:

Emission	A	C	T	G
Probability	0.1	0.1	0.2	0.6

On the other hand, b_2 is:

Emission	A	C	T	G
Probability	0.3	0.2	0.4	0.1

The state transition matrix A is:

State	1	2
1	0.7	0.3
2	0.4	0.6

Finally, the initial probability distribution, π, is:

State	1	2
Probability	0.5	0.5

a) Calculate the most likely sequence of states using the Viterbi algorithm for $O = TCG$. Show your work.

b) Use the forward algorithm to determine the probability that the sequence TCG was generated by the HMM above. Show your work.

c) Briefly compare and contrast the Viterbi algorithm and the forward algorithm. In particular, explain the Viterbi and forward algorithms in terms of which sequences of states they consider, and which components are factored into their final calculation. Hint: think about the difference in the recurrence step of each algorithm.

d) Use the backward algorithm to determine the probability that the sequence TCG was generated by the HMM above. How does this probability compare to your answer from part b? Show your work.

e) Use the forward-backward algorithm along with your answers from parts b and c to determine $P(q_2 = 2|O = TCG)$, the probability that the HMM above was in state 2 at time $t = 2$ while generating the sequence TCG. Time $t = 2$ represents the second observation point, $O_2 = C$. Show your work.

Problem 2: Multiple Alignment and Homology with HMMs

Sorin recently adopted a rare Zebu cow he lovingly named Lumpy. While sequencing Lumpy's genome, Sorin notices a couple stretches of DNA that look very similar to a set of known homologous genes present in other bovines. Curious as to whether these sequences are actually evolutionarily related to the known homologous genes, Sorin turns to you, a distinguished computational biologist in CS181!

Remembering that you have recently learned how HMMs can probabilistically model sequence data, you decide to use an HMM to try to answer Hannah's question. To begin, you first use the alignment algorithms you learned at the beginning of CS181 to perform multiple alignment on the set of homologous genes she has provided:

1	T	-	C	T	G
2	T	-	A	T	G
3	T	-	T	T	G
4	T	-	C	T	A
5	T	C	T	A	G
6	T	-	G	T	G
7	T	-	G	-	G
8	T	-	C	C	G
Consensus	T	-	C	T	G

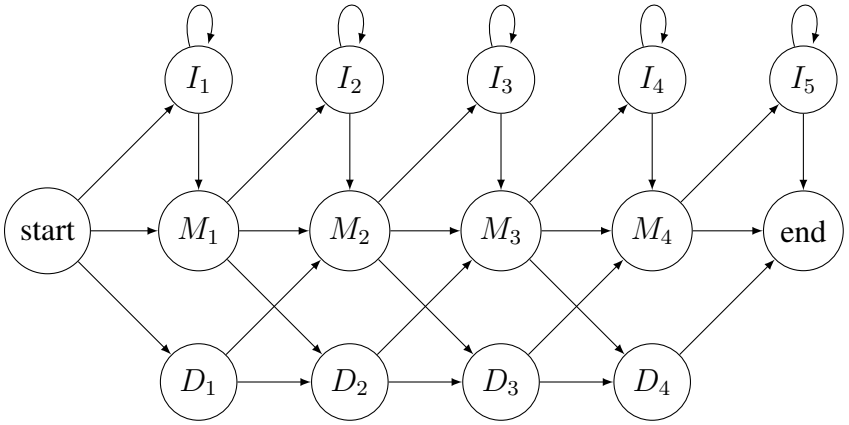
After determining that the consensus gene sequence is $TCTG$, you next construct an HMM with three states for every position in the consensus sequence:

1. M_i, a state representing matches and mismatches
2. I_i, a state representing insertions
3. D_i, a state representing deletions

At each position in the consensus sequence, you determine that 3 types of alignments from a homologous sequence to the consensus sequence are possible. Each of these alignments can be represented by a sequence of transitions in the HMM:

1. A match or mismatch, represented by transitioning directly to M_i
2. An insertion of some number of nucleotides in the homologous sequence followed by a match or mismatch, represented by transitioning to I_i, staying in I_i for some time, and then transitioning to M_i
3. A deletion in the homologous sequence, represented by transitioning to D_i

After one of these transition sequences is performed, the alignment will proceed to the next position in the consensus sequence. The diagram below depicts all possible state transitions in this HMM. Notice the addition of one extra insertion state at the end to allow for insertions after the final character in the consensus sequence.



Finally, you define the following transition and emission probabilities for your HMM to represent the likelihood that the consensus sequence is mutated in various ways:

Initial State	M_1	I_1	D_1
Probability	0.8	0.1	0.1

Transitions	M_{i+1}	D_{i+1}	I_{i+1}
M_i	0.8	0.1	0.1
D_i	0.5	0.5	-
I_i	0.5	-	0.5

Emissions	A	C	T	G	-
M_i	?	?	?	?	-
D_i	-	-	-	-	1
I_i	0.25	0.25	0.25	0.25	-

Final Transitions	I_5	end
M_4	0.1	0.9
D_4	-	1
I_5	0.5	0.5

A transition from M_i to M_{i+1} indicates a match or mismatch; M_i to D_{i+1} indicates a new deletion region; M_i to I_{i+1} indicates a new insertion. The transition probabilities from the "Final Transitions" matrix reflect the cases that there are additional insertions or that the sequence ends.

You will calculate the emission probabilities for each state M_i below.

Tasks:

a. For each state M_i, determine the emission probabilities of each letter $A, T, C,$ and G. Refer to the table of homologous sequences to calculate the proportion of each letter found at each index.

b. Suppose you are given an alignment between the consensus sequence and a new sequence. Describe a reasonable heuristic you could use in lieu of the canonical algorithms to estimate the most likely sequence of states from this HMM that generated the new sequence.

c. Using your heuristic from part (b), estimate the most likely sequence of states from this HMM that generated each of the following sequences:

   i. The consensus sequence $TCTG$, given the alignment:
   `TCTG`
   `TCTG`

   ii. The homologous sequence TCG, given the alignment:
   `TCTG`
   `TC-G`

   iii. Hannah's first new sequence $TAACTG$, given the alignment:
   `T--CTG`
   `TAACTG`

   iv. Hannah's second new sequence $TAAG$, given the alignment:
   `TCTG`
   `TAAG`

d. Determine the probability that your HMM will generate each sequence in part (c) while also following your proposed most likely sequence of states. Show your work.

e. While computing probabilities in part (d), you may have noticed that the probabilities of observing longer sequences of letters will necessarily tend to be smaller than the probabilities of observing shorter sequences of letters. As a result, when comparing sequence data of different lengths, we typically normalize our resulting probabilities by dividing by the probability of observing each emitted sequence due to random chance. Given that your HMM can emit 5 different characters (`A`, `T`, `C`, `G`, `-`), normalize each of your probabilities from part (d) in this way.

f. Do you think that the new sequences are likely to be evolutionarily related to the homologous sequences given? Explain your reasoning.

Bonus (very much extra, don't waste too much time on this): A multivariate HMM H is given by a tuple (S, V, M, B, π) where S is again a set of states, M is again a state transition matrix, and π is again an initial state distribution. As the name suggests, however, V is no longer a single random variable with its set of outcomes, but rather a set of random variables, each with its own outcome set. That is, rather than emitting a single observation, each state now emits a tuple of observations. As you might imagine, this means that B must now give a joint distribution over all the emission variables for each state. Given the high-dimensional nature of most modern data, you can see how multivariate HMMs might be considered

more common and more natural than the univariate HMMs you've learned about in class. And with some thought, the algorithms we've learned generalize nicely to the multivariate case.

Let's extend this model a little bit. Now, suppose that we have a set of $M_{\bar{x}}$ where x is an observation. That is, we have a transition matrix for each observation, so that the next transition depends on the current emission. Clearly, this is no longer an HMM, but all our algorithms carry over nicely, and this type of model is much more expressive. We'll call it a hidden conditional Markov Model (HCMM).

Your task is to formulate a multivariate homology HCMM which takes in *aligned* homologous sequences like those given in part c). That means you should give (S, V, M, B, π), but don't forget to make V a set of symbol sets, B a set of joint distributions, and M a set of transition matrices indexed by emissions!

Problem 3: HMMs and Ancestry Deconvolution

Many ancestry testing companies like 23andMe use algorithms involving HMMs to trace your ancestry. (If you're interested, you can read more about 23andMe's ancestry deconvolution algorithm here.)

a. Suppose you are given a DNA sequence and a set of ancestries from which this DNA sequence could be descended. How could you use an HMM to determine what fraction of the DNA sequence is descended from each ancestry? In your response, be sure to answer the following questions:

   i. What are the hidden states of the HMM?

   ii. What are the emissions of the HMM?

   iii. What biological phenomenon is represented by transitions between hidden states?

   iv. What algorithm(s) from this class would you use to determine the fraction of the DNA sequence descended from each ancestry?

After you get your ancestry results, ancestry testing companies often continue to store your genomic data for future use. Read this article about how stored genomic data can be used and this article about how to protect your genomic data.

b. Give three examples from the articles of how your genomic data could be useful.

c. Compare and contrast how medical institutions and non-healthcare private companies can use genomic data.

Problem 4: HMMs and CpG Islands

The following excerpt defines CpG islands and their biological significance.

> "In the human genome wherever the dinucleotide CG occurs (frequently written CpG to distinguish it from the C-G base pair across the two strands) the C nucleotide (cytosine) is typically chemically modified by methylation. There is a relatively high chance of this methyl-C mutating into a T, with the consequence that in general CpG dinucleotides are rarer in the genome than would be expected from the independent probabilities of C and G. For biologically important reasons the methylation process is suppressed in short stretches of the genome, such

> as around the promoters or 'start' regions of many genes. In these regions we see many more CpG dinucleotides than elsewhere, and in fact more C and G nucleotides in general. Such regions are called CpG islands." (Durbin, Eddy, Krogh, Mitchison, 1998)

We want to model this biological phenomena as a HMM. The HMM should help us determine if (i) given a short sequence, this sequence comes from a CpG island or not and (ii) given a long sequence, how to find all of the CpG island occurrences, if any.

a. What are the hidden states of the HMM?

b. What are the emissions of the HMM?

c. Construct an appropriate transition probability matrix with all of the values populated. Justify your chosen values. (Consider if certain values should be greater than or less than others.)

d. To improve the accuracy of our HMM, we need to ensure that the emission and transition probabilities are reasonable and well-calibrated. How can we enhance the model's accuracy?