Homework 5

CS 181, Fall 2022

Out: Nov. 28
Due: Dec. 5, 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 Markov chain visits a state, one symbol (A, T, C, or G) is generated. In state 1, symbol A is generated with probability 0.2, symbol T with probability 0.1, symbol C with probability 0.2, and symbol G with probability 0.5. In state 2, symbol A is generated with probability 0.1, symbol T with probability 0.3, symbol C with probability 0.2, and symbol G with probability 0.4. The Markov chain jumps from state 1 to state 2 with probability 0.2, and from state 2 to state 1 with probability 0.3. The initial probability distribution is 0.5 for state 1 and 0.5 for state 2.

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

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

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

d. Use the forward-backward algorithm along with your answers from parts b and c to determine $P(q_3 = 2|O = TTCGA)$, the probability that the HMM above was in state 2 at time $t = 3$ while generating the sequence $TTCGA$. (Time $t = 3$ represents the third observation point, $O_3 = C$.) Show your work.

Problem 2: Multiple Alignment and Homology with HMMs

While sequencing the genomes of rare microbes found in the coral reef exhibit, Poseidon notices a couple stretches of DNA that look very similar to a set of homologous genes present in other microbes. Curious as to whether these sequences are actually evolutionarily related to the known homologous genes, Poseidon 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 Poseidon'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 Poseidon 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

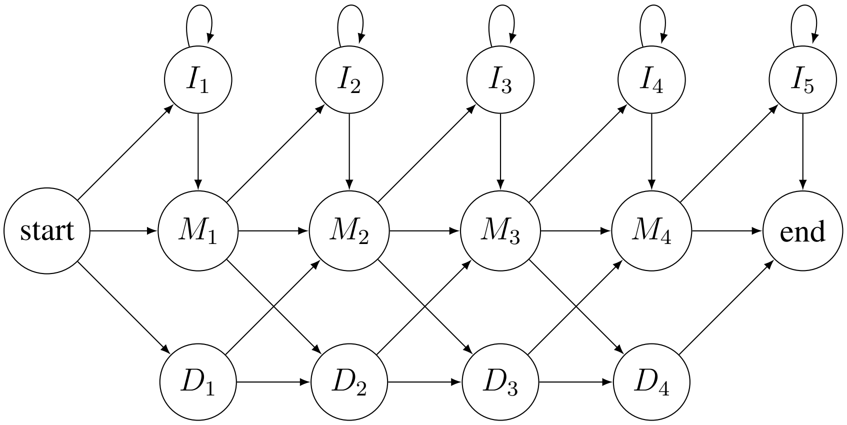
After determining that the consensus gene sequence is $TCTG$, you next construct an HMM with three states for every position in the consensus sequence:

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:

1. Each state M_i transitions to state I_{i+1} with probability 0.1 (reflecting the probability of a new insertion), state D_{i+1} with probability 0.1 (reflecting the probability of a new deletion), and state M_{i+1} with probability 0.8 (reflecting the probability of a match or mismatch). Similarly, M_4 transitions to I_5 with probability 0.1 and `end` with probability 0.9.
2. Likewise, the initial state is M_1 with probability 0.8, I_1 with probability 0.1, and D_1 with probability 0.1.
3. Each state I_i stays in state I_i with probability 0.5 (reflecting the probability of a longer insertion) and transitions to state M_i with probability 0.5 (reflecting the probability of not extending the insertion). Similarly, I_5 stays in I_5 with probability 0.5 and transitions to `end` with probability 0.5.
4. Each state D_i transitions to state D_{i+1} with probability 0.5 (reflecting the probability of a longer deletion) and transitions to state M_{i+1} with probability 0.5 (reflecting the probability of not extending the deletion). D_4 transitions to `end` with probability 1.
5. Each state M_i emits letters with probabilities proportional to how often those letters appear at position i in the homologous sequences above.
6. Each state I_i emits $A, T, C,$ and G with equal probability.
7. Each state D_i always emits the character $-$, representing the alignment of a letter in the consensus sequence with a gap in a homologous sequence.

Tasks:

a. For each state M_i, use the table of homologous sequences to determine the emission probabilities of each letter $A, T, C,$ and G.

b. Suppose you are given an alignment between the consensus sequence and a new sequence. Describe a reasonable heuristic you could use 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. Poseidon's first new sequence $TAACTG$, given the alignment:

```
T--CTG
TAACTG
```

iv. Poseidon'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 Poseidon's 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. Give at least one reason why you might prefer to share your genomic data with each type of entity over the other.

Problem 4: Computational Biology and Identity

Part 1: Genetics and Race

The biological basis of race has been an extremely fraught debate, especially because biological conceptions of race have been historically manipulated to justify scientific racism, eugenicist ideology, and perpetuate stereotypes against marginalized groups.

The completion of the Human Genome Project and the advent of relatively new computational biology algorithms that allow us to sequence DNA, find patterns, create relational models, etc. have allowed scientists to study human genomes in unprecedented detail, igniting conversations about whether race is a helpful concept in biology research.

a. Given your initial opinions, knowledge, and/or beliefs, to what degree do you think there is a biological basis for race? (2-3 sentences)

b. In 2004, Jorde and Wooding wrote an article titled *Genetic variation, classification and 'race'*. Read the questions below first. Then, review the sections titled "*Variation at the individual level*" and "*Genetic variation, race and medicine*" to support your responses.

i. Explain the difference between race and genetic ancestry. (Hint: Read the NHGRI definitions of race and genetic ancestry (1-2 sentences)

ii. Can someone's genetic information be used to infer their genetic ancestry? Why or why not? (1-3 sentences)

iii. Can someone's genetic information be used to infer their race? Why or why not? *(Hint: The section under figure 3 and the "Conclusions" section might be helpful)* (1-3 sentences)

iv. The article argues that "an individual's population affiliation [can] often be a faulty indicator of the presence or absence of an allele related to diagnosis or drug response." What reasoning do the authors use to justify this argument? Please discuss an example of a specific gene in your answer. (3-4 sentences).

Part 2: Diversity in Computational Biology

c. Science relies on the contributions of people from diverse backgrounds to bring in new perspectives and modalities of thinking. As Prof. Istrail mentioned in class, exceptional scientists, such as Margaret Dayhoff, have brought immense contributions to the field of computational biology.

   Research a computational biology scientist from a historically underrepresented background (e.g. gender, sexuality, race, etc.). Report the following in 3-4 sentences:

   i. Their name

   ii. One interesting fact

   iii. Their academic contribution to the field (e.g. landmark paper, professorship, award etc.)

Bonus: Make a Beautiful PowerPoint!

d. Expand on your answer in part C of this problem by delving more into your scientist and creating a 3-6 slide PowerPoint! You can get up to 2 points of extra credit for your creation for fulfilling the following requirements, as well for the presentation's overall quality and aesthetics.

   i. Intro Slide: Include basic info (e.g. name, date of birth, nationality, and notables firsts)

   ii. Content Slides (2-4 slides, but feel free to add more): Summarize at least 2 of the scientist's major academic contributions to the field

   iii. Conclusion Slide: Synthesize the scientist‘s lasting impacts on computational biology