Final Exam
CS 181, Fall 2018

Out: Dec. 13, 5:00 PM
Due: Dec. 17, 11:59 PM

Instructions

- There are five problems in total.
- Submit a single PDF document (final.pdf) containing all of your solutions.
- You must type up your solutions. You may include scans of hand-drawn figures and illustrations.
- To hand in, log into a department linux machine. Put the file that you want the TAs to see into a directory titled 'final' as well as a blank README. Then navigate to that directory and type cs1810_handin final.
- This exam is strictly non-collaborative. You may not work with anyone else and you are not allowed to consult any sources other than the textbook, the lecture notes, your own notes, and past assignments from this semester.
- The course staff will only answer questions related to clarifications, explanations, or typos. Hints will not be given. If you ask a question on Piazza, please post privately.
- Do as much as you can to convey your solution clearly. Clean solutions will receive more credit than solutions that are difficult to understand.
1 HMMs (10 Points)

a) A HMM has been constructed to generate a sequence, \( x \), of symbols consisting of 2 states. 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 0.1, symbol T with probability 0.2, symbol C with probability 0.2, and symbol G with probability 0.5. In state 2, symbol A is generated with probability 0.3, symbol T with probability 0.2, symbol C with probability 0.3, and symbol G with probability 0.2. The Markov chain jumps from state 1 to state 2 with probability 0.3, and from state 2 to state 1 with probability 0.2. The initial probability distribution is 0.5 for state 1 and 0.5 for state 2.

Calculate the most likely sequence of states using the Viterbi algorithm for ATCAG. Show your work.

b) Use the forward algorithm to determine the probability that the sequence ATCAG was generated by an HMM according to the emission and transition probabilities listed in Problem 1. Show your work.

2 Phylogeny (25 points)

UPGMA is one of the simplest and most effective methods of hierarchical clustering. Recall from our unit on phylogeny that the distance between clusters \( i \) and \( j \) is:

\[
d_{ij} = \frac{1}{|C_i||C_j|} \sum_{q \in C_i, \ p \in C_j} d_{pq}
\]

The cardinality of \( C_x \), denoted \(|C_x|\) is defined as the average distance between each pair of elements in the cluster \( C_x \).

a) Say we join two clusters \( C_a \) and \( C_b \) into a new cluster, \( C_e \) such that \( C_e = C_a \cup C_b \). Show that for any arbitrary other cluster, \( C_f \), that :

\[
d_{ef} = \frac{d_{af}|C_a| + d_{bf}|C_b|}{|C_a| + |C_b|}
\]

\textbf{Hint:} \(|C_e| = |C_a| + |C_b|\)

b) One way UPGMA is different from other clustering algorithms is how heights of branches are determined. If we join clusters \( i \) and \( j \), the height of this cluster on the evolutionary tree is:

\[
h_{ij} = d_{ij}/2
\]

For any node on a UPGMA tree, show that a parent node’s height is always greater than it’s children’s.

\textbf{Hint:} Consider the steps of UPGMA and what components change at each iteration.
c) Another clustering algorithm you’re familiar with is the neighbor joining algorithm. One property of the neighbor joining algorithm is that the distances on branches (edges) are additive. For example if \( d_{ab} = 3 \), and \( d_{bc} = 4 \), we can conclude that \( d_{ac} = 7 \). While this makes it easy to find distances of branches between species, this isn’t the only measure we can use on these types of trees for additive edge length trees. Another metric is distance between leaves and it’s calculation is not as simple.

We use the following formula for distances between arbitrary leaves \( i \) and \( j \):

\[
D_{ij} = d_{ij} - (r_i + r_j)
\]

where \( d_{ij} \) is the total length of edges between \( i \) and \( j \) and \( r_i \) is:

\[
r_i = \frac{1}{|L| - 2} \sum_{k \in L} d_{ik}
\]

where \( L \) is the set of leaves, and \(|L|\) is the size of the set of leaves. Here we have four leaves \( S, J, L, \) and \( H \) named after your four wonderful TA’s and the following tree:

![Diagram of tree with distances](image)

Calculate distance between all pairs of leaves in the tree using the formulas above and show your work. What do you notice about the leaves with the smallest distance?

d) Name two reasons using this method for calculating distance between leaves might be better than using the sum of the edges between them.

### 3 Sequence Logos (25 points)

In class we presented the concept of phylogenetic footprinting, which allows us to identify transcription factor binding sites (non-coding regions of DNA that have varying levels of conservation between related species). An alternative way to represent consensus sequences makes use of **sequence logos**. Just like
phylogenetic footprinting, sequence logos are derived using information theoretic concepts and ultimately compute the information content of each position along a finite conserved region. Sequence logos have the added benefit of lending themselves well to neat, colorful graphical representations. The most commonly used method of sequence logo construction was proposed by Schneider et al. in 1990. The method we will focus on is an earlier construction, also by Schneider et al. (1986). This is what a sequence logo looks like:

The height of each letter is related (but not proportional) to its frequency at the indicated position.

Although we urge you peruse the papers above to take in the sublime beauty of sequence logo construction in its entirety, the following definitions are all that are required to complete this problem.

Suppose we have a multiple alignment of \( n \) sequences each of length \( m \). The uncertainty about which base pair occurs at position \( i \in \{1, \ldots, m\} \) can be measured by the entropy \( H_i \), defined by

\[
H_i = -\sum_{a \in \mathcal{X}} f_{a,i} \log_2 f_{a,i},
\]

where \( \mathcal{X} \) is the set of possible characters (for DNA, \( \mathcal{X} = \{A, C, G, T\} \)) and \( f_{a,i} \in [0, 1] \) is the frequency of character \( a \) at position \( i \). Note that for all \( i \),

\[
\sum_{a \in \mathcal{X}} f_{a,i} = 1.
\]

The **information content** (which is plotted on the vertical axis of the sequence logo above) of a position \( i \) is given by

\[
R_i = \log_2 |\mathcal{X}| - (H_i + e_n),
\]

where \( e_n \) is a small-sample correction factor defined by

\[
e_n = \frac{1}{\ln 2} \cdot \frac{|\mathcal{X}| - 1}{2n}.
\]

To construct the sequence logo, we set the height of character \( a \) at position \( i \) to be \( f_{a,i}R_i \).

Consider the following multiple alignment of 6 DNA sequences, each of length 7:

```
AACGGTTC
CGCTACG
```
GCCAAGT
TGCTAAA
GTATATC
TGCTAGA

a) Use the formulas above to compute the information content of each position. In other words, for
\( i = 1, 2, \ldots, 7 \), compute \( R_i \). Additionally, for \( i = 3 \) and \( i = 6 \), compute the height of each of the
four nucleotides A, T, C, and G at position \( i \).

b) Compare and comment on the relative values of \( R_3 \) and \( R_6 \).

c) Describe in words what the height of each letter in the graphic representation means. In particular,
think about how a character can have two different heights at positions \( i \) and \( j \) in even though its
relative frequency at \( i \) is the same as its relative frequency at \( j \). For example, the height of A at
position \( i = 2 \) is different from its height at position \( i = 3 \), even though \( f_A,2 = f_A,3 = 1/6 \). Why is
this seemingly unnatural property actually desirable in sequence logos?

4 A Hidden Markov Model (25 points)

Chromatin is a complex of DNA and protein molecules. It is responsible for the structure of chromosomes as seen under the microscope. Chromatin comes in two forms: euchromatin and heterochromatin. In this problem we will restrict ourselves to euchromatin. In euchromatin, certain segments of the DNA are wrapped around proteins called histones. The resulting structure resembles a series of beads (DNA-bound histones) on a string (intervening linker DNA).

Histone proteins play a role in many DNA-dependent processes, such as transcription, replication, repair, and recombination, so it is important to study them and attempt to infer where they might be located. We can consider the DNA in euchromatin to have two states: one in which it is bound to histones and one in which it is not.

One experimental technique that can be used to locate where proteins bind to DNA is called Chromatin Immunoprecipitation and Sequencing (ChIP-seq). A ChIP-seq experiment has the following steps: (1) proteins are chemically linked to the DNA to which they are bound; (2) the DNA is fragmented into small pieces; (3) fragments containing bound protein are selectively extracted; (4) the extracted fragments are sequenced. Note that the extraction process is imperfect and thus DNA sequence reads are produced from
both bound and unbound DNA sequences, albeit with an enrichment for bound sequences.

After aligning these reads to the source genome of length $n$, the output of a ChIP-seq experiment can be represented as a vector of read counts $r = (r_1, \ldots, r_n)$, where $r_i$ is the number of reads whose alignment to the genome contains position $i$. For example, if $r_{22390} = 5$, then this means that 5 reads cover location 22390 in the genome. Note that $0 \leq r_i < \infty$ for all $i \in \{1, \ldots, n\}$. The results of a ChIP-seq experiment can be used to locate histones, since we expect higher observed read counts to be correlated with the presence of a histone at a given position in the genome.

In this problem your task is to come up with a hidden Markov model for determining which genomic regions are bound by histones. Present your model by answering the following questions. The notation below is consistent with the notation in the lecture notes.

a) What is the sequence of observation symbols $\sigma = \sigma_1 \sigma_2 \ldots \sigma_T$? What is the meaning of each $\sigma_t$? What is the value of $T$?

b) How many hidden states should the model contain (what is the value of $N$)? What does each hidden state $s_t$ represent? Please be as specific as possible.

c) Using your answers and notation from parts (a) and (b), define the probabilities and/or probability distributions needed to completely specify the HMM.

d) Discuss what you expect the (relative) magnitudes of these probabilities to be.

e) Give an equation for the most likely euchromatin structure for a DNA sequence, given the data $\sigma$. What algorithm would you use to compute the structure itself (i.e. not the probability)?

5 Assembly (15 points)

The $k$-mer decomposition of a string $s$ is defined to be the collection of all length $k$ substrings of $s$ (including repeats). Symbolically, the $k$-mer decomposition of $s$ is the collection

$$(s_1s_2\ldots s_k, s_2s_3\ldots s_{k+1}, \ldots, s_{n-k+1}s_{n-k+2}\ldots s_n),$$

where $s_i$ denotes the $i$th character in $s$.

a) What is the 3-mer decomposition of CTTATTGTTC?

b) What is the 4-mer decomposition of CTTATTGTTC?

c) Recall that in genome assembly, the only information we have is the collection of short reads. Imagine trying to reconstruct the string CTTATTGTTC from its 3- and 4-mer decompositions. In the context of this particular string, what information does the 4-mer decomposition contain that the 3-mer decomposition does not?