Homework 2

CS 181, Fall 2022

Out: Sep. 30
Due: Oct. 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: Motifs

Up to this point, we have mostly represented DNA strands as single strings of nucleotides, such as *AATCGAGG*. However, DNA molecules actually consist of two strands of complementary nucleotides, as shown below:

```
AATCGAGG
TTAGCTCC
```

Furthermore, each strand of DNA in a DNA molecule has a polarity that determines the direction in which it can bind to other DNA strands. We label the two ends of a DNA strand with 5' and 3' to denote the direction in which the DNA strand is written. When two complementary DNA strands bind to each other to form a DNA molecule, the two strands must have opposite polarities:

```
5′ AATCGAGG 3′
3′ TTAGCTCC 5′
```

By convention, we typically assume that DNA strands are written in the 5' to 3' direction unless their polarities are explicitly given. However, when working with complete DNA molecules, it is often important to specify the direction of each strand, as the polarities of the DNA strands can affect biological behaviors. In this problem, we will explore how motif identification is affected by the polarities of DNA strands.

In class you will soon be learning about motifs. Wikipedia defines a motif as "a nucleotide or amino-acid sequence pattern that is widespread and has, or is conjectured to have, biological significance". Combinatorial pattern matching algorithms are often used to find motifs, but in this homework we'll see how alignment algorithms can be used for the same task.

For this problem, we'll be exploring a very important motif, the restriction site. Restriction sites are palindromic sequences (sequences which are identical to their reverse complements, e.g. *CCCGGG*) that serve as recognition sites for restriction enzymes. Restriction enzymes are proteins that look for these restriction sites and cut them like a pair of molecular scissors, breaking the DNA into fragments. Because the restriction sites are palindromic, the enzymes bind onto both sides of the DNA and cleave the strands at a specific position on the site. For example, the restriction enzyme *TaqI* recognizes the palindromic restriction site *TCGA*, and cleaves it between the *T* and *C* position. Thus, we will say that *TaqI* has a recognition site of *T/CGA*, to show the restriction site and where the site is cleaved. Here is an example of how a DNA strand might be cleaved using *TaqI*:

```
   5′ AATCGAGG 3′
   3′ TTAGCTCC 5′
          ↓
5′ AAT   CGAGG 3′
3′ TTAGC   TCC 5′
```

Importantly, *TaqI* can only cleave the restriction sequence *TCGA* if the sequence appears in the DNA molecule from 5' to 3'. Consequently, if the polarities of the strands in the above example were switched, *TaqI* would no longer be able to cleave the DNA molecule:

```
3′ AATCGAGG 5′
5′ TTAGCTCC 3′
```

a. For the following sequence, list all the DNA fragments that would be produced if it was exposed to the restriction enzyme *Hin*dIII, which has the recognition site *A/AGCTT*. How would reversing the polarity of the following sequence change the DNA fragments produced after exposure to *Hin*dIII?

```
5′ TTCGAAGTCTTGGGCAAGCTTAGGCTAAGCTTCGAATCCAACGTCGTTTCGAAG 3′
```

Sometimes *dot plots* are used to better visualize how two sequences are aligned. Dot plots make use of a table that is similar to an alignment table. We use the characters of two different strings to label the rows and columns of the table. Each cell is then filled with a dot if its row and column correspond to a matching character, and left empty otherwise. Here is an example:

	A	T	C	G
G				●
G				●
C			●	
G				●

b. Create a dot plot for the long sequence above and the recognition site of *Hin*dIII. Because we are very kind, we have provided you with the table so you can just print it and fill it out. Alternatively, you can find the LaTeX source in the support files.

	T	T	C	G	A	A	G	T	C	T	T	G	G	G	C	A	A	G	C	T	T	A	G	G	C	T	A	A	G	C	T	T	C	G	A	A	T	C	C	A	A	C	G	T	C	G	T	T	T	C	G	A	A	G
A																																																						
A																																																						
G																																																						
c																																																						
T																																																						
T																																																						

c. Describe the relationship between the dot plot and the DNA fragments you previously identified. Specifically, explain how the dot plot can be used to identify restriction sites for both possible polarities of the DNA sequence given.

Note: Restriction enzymes are the basis for the polymerase chain reaction (PCR), which is used to amplify DNA samples and earned its inventor Kary Mullis the Nobel Prize in 1993. It is no exaggeration to say that most of modern DNA biotechnology would not be possible without PCR. Indeed, there is a long tradition in biology of using tools and mechanisms devised by nature to suit the needs of science and society. A more recent example of this is the CRISPR/Cas system, which is a prokaryotic immune mechanism that is now being used for precise genome editing.

Problem 2: Free End Gap Alignment

Now that we have an intuition for what motifs might look like in a sequence, let's design an alignment algorithm to find them for us. Global alignment won't work for the motif problem because it will align our short motif to a large region of the longer string with many gaps. Since we expect our motif to match on a small area, we might guess that local alignment is appropriate. But local alignment will also have trouble because we want to align the *entirety* of our motif, and local alignment may pick out a subsequence of the motif. To sidestep these problems, we introduce a new kind of alignment: *Free End Gap Alignment*. We want to align the entire motif to a subsequence of the longer sequence, with as few gaps in the aligned section as possible. To this end, we will discount any gaps that occur on the ends of a sequence. For example, the optimal **global** alignment of `GTAGGCTTAAGGTTA` and `TAGATA` is

```
GTAGGCTTAAGGTTA
-TAG----A---T-A
```

with a score of -3 (assuming our usual scoring scheme of +1 for match, -1 for gap and mismatch). An optimal Free End Gap Alignment, however, would be

```
GTAGGCTTAAGGTTA
-TAGA--TA------
```

with a higher score of 2. Indeed, this latter alignment better captures the essence of what it means to find a motif in a longer sequence. We formalize the Free End Gap Alignment problem as follows:

FREE END GAP ALIGNMENT

Input: Strings u, v, and a scoring function δ that takes in two characters as arguments.

Output: An alignment of u, v, for which the score is maximal (as defined by δ) and for which end gaps are free (i.e. not penalized) in both sequences.

a. Create and fill out the free end gap alignment dynamic programming table, with scores and back pointers, to align the above sequences `GTAGGCTTAAGGTTA` and `TAGATA`. For scoring, use the match bonus +1, mismatch penalty -1, and indel penalty -1. When there is a tie between a mismatch and an indel alignment, use a back pointer that indicates a mismatch.

   Hint: Think about how initialization and backtracking will change from global alignment in order to not penalize end gaps.

b. Design a dynamic programming solution to the Free End Gap Alignment problem. Your solution must include

   - a new recurrence relation, if necessary
   - a description of how your table is initialized
   - how to backtrack through your table, including where to start and where to end

   If you wish, you may also include diagrams, pseudocode, mathematical expressions, and plain English text in your answer. However, please do not submit a block of code with no accompanying explanation.

   Now to see the algorithm at work. We've implemented Free End Gap Alignment for you already. Download the support files from the course website to get the code. The implementation takes in two arguments, a long sequence and a short sequence to align it against (the scoring scheme is the standard scheme mentioned prior). The code will then output *all* optimal alignments. Here's an example of how to run our implementation:

```
> sh FEGAlignment.sh GTAGGCTTAAGGTTA TAGATA
GTAGGCTTAAGGTTA
-TAGA--TA------
```

c. Use our implementation to count the number of DNA fragments that are created by exposing the following sequence (all three lines make up one contiguous sequence, written 5' to 3') to the restriction enzyme *Bam*HI, which has the recognition site *G/GATCC*. How would your answer change if the polarity of the following sequence was reversed?

```
TCCATTGATGCCACGGCGGATCCTGGAGAGCAGCAGCGACTTGCATACATCAGATCAGAGTAATACTAGC
ATGCGATAAGTCCCTAACTGACTATGGATCCTTCTAGAGTCAACTTCAGGACATATGGTCTCTGGATCCC
GTGGATCCTTCCTAGGAATCAGATTGGATCCTGGTTAACCATCAAACAGGTCTTGAGTCTAAAATTGTCG
```

Problem 3: Homology

Finding conserved patterns across different species is important for evolutionary biology. Consider the following sequences:

A : 5′ `AGCTTCGAAGTTATCTTGGACGGACTTG` 3′
B : 5′ `AGTTTCCCAGGATATCTTCGAACGACTG` 3′
C : 5′ `AGGCTTCCCATCCTCCTATAAAGGTAGG` 3′

We wish to find whether B or C is a homologous protein to A, that is, a protein that originates from some ancestral species. The optimal alignment for A and B is

```
A: AGCTTCGAAGT-TATCTTGGA-CGGACTTG
B: AGTTTCCCAGGATATCTTCGAACG-ACT-G
```

with a score of 12. The optimal alignment between A and C is

```
A: AG-CTTCGAAGT--TATCT-TGGACGGACTTG-
C: AGGCTTCCCA-TCCTC-CTATAAA-GG--TAGG
```

with a score of 1. After splicing the DNA into cells, you find that the DNA actually transcribes the following protein sequences (using the one-letter amino acid code). Notice how similar A and C are as proteins despite being much more dissimilar than A and B as DNA sequences:

```
A: SSYLGR
B: SDIFER
C: SSYKGR
```

a. Look up the terms *intron* and *exon* and give a biological reason for this observation. Why might it be a better idea to align by amino acid sequence rather than DNA? Identify the introns and exons in the DNA sequences above (there are multiple solutions, but find one which minimizes the number of introns and exons).

b. There are 20 different amino acids and only 4 different nucleotides (used in DNA). Give a probabilistic reason why aligning by amino acid sequence might be better than by DNA sequence.

Problem 4: Statistical Foundations of Sequence Alignment

The most popular type of substitution matrices are known as BLOSUM matrices. Blocks, or fixed regions in a given set of aligned sequences, are used to create these substitution matrices. We are studying sequence similarities among some of the most endangered marine creatures and will analyze BLOSUM matrices derived from their genomes.

Given an alphabet, $\sum = \{\alpha, \beta, \gamma, \delta\}$, the following block contains the sequences of a single, 3-residue protein from each of the nine most endangered marine species: Vaquita, Whale Sharks, Hawksbill sea turtle, Sea otter, Whales, River Dolphins, Florida manatee, Galapagos Penguin, and Hawaiian monk seal.

Vaquita	α	γ	δ
Whale Sharks	β	γ	α
Hawksbill sea turtle	γ	γ	γ
Sea otter	β	β	γ
Whales	β	γ	δ
River Dolphins	γ	α	δ
Florida manatee	α	β	β
Galapagos Penguin	γ	γ	δ
Hawaiian monk seal	β	γ	β

By determining the evolutionary information in the above block, we will develop an improved scoring scheme for sequence alignment.

a. What is the total number of residues we have in this block? How many ways can we pick a pair of letters in each column? How many possible ways of picking letters from the above table are there? Assume the paired letters must be in the same column.

b. Determine the observed frequencies for each alignment pair.

c. For each alignment pair, determine the expected frequency based on the above block, and calculate the log-likelihood score, f, that compares the observed and expected frequencies for each pair.

d. Explain the meaning of positive, negative, and zero log-likelihood scores. Given this interpretation, why would using log-likelihood scores as the scoring scheme be desirable for sequence alignment?

e. Interpret the final log-likelihood scores in this problem in terms of what they indicate about conservation of letters in this alphabet. Comment on how realistic this scoring scheme would be if we were to use it to align actual genome sequences.

f. **Bonus:** More formally, we can express the log-likelihood score of aligning two letters a and b as:

$$f_{ab} = 2 \log_2 \frac{p_{ab}}{q_{ab}}$$

where:

i. $p_{ab} = \mathbb{P}(a, b|M)$ is the joint probability of observing aligned letters a and b under the alignment model M

ii. $q_{ab} = \mathbb{P}(a, b|R) = \mathbb{P}(a|R)\mathbb{P}(b|R)$ is the joint probability of observing aligned letters a and b under a random model where letters appear independently.

Recall that local alignment requires the expected score of aligning two letters to be non-positive to ensure that alignments remain local. Therefore, if we want to use BLOSUM matrices for local alignment, we need to ensure that the expected score f_{ab} is not positive.

Prove that the expected score $\mathbb{E}[f_{ab}]$ for a randomly aligned pair of letters is not positive for any alignment model M.

Hint: For any concave function g, $\mathbb{E}[g(X)] \leq g(\mathbb{E}[X])$.

Problem 5: Reference Genomes

The first draft of the human genome was generated by The Human Genome Project and published in 2001 (Our very own Professor Istrail helped work on the human genome!). Since then, the human reference genome has been continuously revised and improved to reflect new scientific advances. While the first draft of the human genome was an amazing feat, let's dive a bit deeper into the decisions made during the initial construction.

What is a reference genome and what is their use? Here is an explanation given by the NHS:

> *"Every time a genome as large as ours is sequenced, the genetic material must be broken up into short overlapping fragments, often numbering in the millions. Once the genome has been sequenced, the readings of these individual fragments need to be put back together in order to be analysed by scientists looking for variation in regions of DNA that could have an impact on our health. In order to do this, scientists refer to what is known as a 'reference genome' - a template genome incorporating the most up to date information we have on human genomics."*

Part 1: Sample Collection Methodology

The construction of the human genome was a complex process that essentially involved carefully piecing together segments of DNA (or *contigs*) into a single, continuous genome. In order to do so, human DNA had to be collected and prepared for the genome assembly algorithm. Read the following excerpt from Venter et. al (2001) that describes how the DNA used to create the first version of the human genome was obtained:

> *"[T]he opportunity to donate DNA for this purpose was broadly advertised near the two laboratories engaged in library construction. Volunteers of diverse backgrounds were accepted on a first-come, first-taken basis. Samples were obtained after discussion with a genetic counsellor and written informed consent. The samples were made anonymous as follows: the sampling laboratory stripped all identifiers from the samples, applied random numeric labels, and transferred them to the processing laboratory, which then removed all labels and relabelled the samples. All records of the labelling were destroyed. The processing laboratory chose samples at random from which to prepare DNA and immortalized cell lines. Around 5-10 samples were collected for every one that was eventually used. Because no link was retained between donor and DNA sample, the identity of the donors for the libraries is not known, even by the donors themselves."*

Next, read the following excerpts from Sherman et.al (2018), a paper published seventeen years later. In this paper, authors constructed sequences missing from the reference genome using DNA samples collected from individuals of African descent:

- *"The lack of diversity in the reference genome poses many challenges when analyzing individuals whose genetic background does not match the reference...differences between populations are quite large; examination of 26 populations across five continents revealed that 86% of discovered variants were present in only one continental group."*
- *"Our analysis revealed 296,485,284 bp in 125,715 distinct contigs present in the populations of African descent, demonstrating that the African pan-genome contains ∼10% more DNA than the current human reference genome. Although the functional significance of nearly all of this sequence is unknown, 387 of the novel contigs fall within 315 distinct protein-coding genes, and the rest appear to be intergenic."*

Answer the following questions:

a. What were design choices made during the sample collection process that might have influenced the types or groups of people samples were collected from? What issues might have arisen from these choices? (3-4 sentences)

b. If you were the lead researcher in this project, describe how you might change your sample collection methodology to solve the issues you've identified above. Consider how you would recruit new patients and the types of people you would include in your study. (2-3 sentences)

Part 2: Ethics of Large-Scale DNA Sequencing

In 1996, the National Center for Human Genome Research (NCHGR) and the Department of Energy (DOE) released a joint document to provide guidance and "address ethical issues that must be considered in designing strategies for recruitment and protection of DNA donors for large-scale sequencing."

Read the following sections from the document:

- 2. Privacy and Confidentiality
- 3. Source/Recruitment of DNA Donors for Library

Answer the following questions:

c. Per the guidelines, the researchers on The Human Genome Project made the decision to strip all identifiers from the DNA samples in order to protect donor privacy, including demographic information. In your opinion, what are the ethical trade-offs of collecting vs. not collecting demographic data? (3-4 sentences)

d. Pick another guideline provided in the document. Make an analytical argument for why or why not you think it is an ethically sound guideline. Your reasoning should incorporate concepts from at least one of the perspectives from the Intro to Ethical Theories document. (3-4 sentences) (Hint: You may find the executive summary at the bottom of the NCHGR-DOE document helpful).