Programming Homework 1
Graded Quiz • 30 min

Due Jun 21, 1:59 AM CDT

✓ **Congratulations! You passed!**

TO PASS 70% or higher

Keep Learning

Retake the assignment in **7h 52m**

GRADE
100%

Programming Homework 1

LATEST SUBMISSION GRADE

100%

1. How many times does `AGGT` or its reverse complement (`ACCT`) occur in the lambda virus genome? E.g. if `AGGT` occurs 10 times and `ACCT` occurs 12 times, you should report 22.

   1 / 1 point

   ✓ **Correct**

2. How many times does `TTAA` or its reverse complement occur in the lambda virus genome?

   1 / 1 point

   Hint: `TTAA` and its reverse complement are equal, so remember not to double count.

   ✓ **Correct**

3. What is the offset of the leftmost occurrence of `ACTAAGT` or its reverse complement in the Lambda virus genome? E.g. if the leftmost occurrence of `ACTAAGT` is at offset 40 (0-based) and the leftmost occurrence of its reverse complement `ACTTAGT` is at offset 29, then report 29.

   1 / 1 point

   ✓ **Correct**

4. What is the offset of the leftmost occurrence of `AGTCGA` or its reverse complement in the Lambda virus genome?

   1 / 1 point

   ✓ **Correct**

5. As we will discuss, sometimes we would like to find *approximate* matches for *P* in *T*. That is, we want to find occurrences with one or more differences.

   1 / 1 point

   For Questions 5 and 6, make a new version of the `naive` function called `naive_2mm` that allows up to 2 mismatches per occurrence. Unlike for the previous questions, **do not consider the reverse complement here**. We're looking for approximate matches for *P* itself, not its reverse complement.

   For example, `ACTTTA` occurs twice in `ACTTACTTGATAAAGT`, once at offset 0 with 2 mismatches, and once at offset 4 with 1 mismatch. So `naive_2mm('ACTTTA', 'ACTTACTTGATAAAGT')` should return the list `[0, 4]`.

   Hint: See [this notebook]() for a few examples you can use to test your `naive_2mm` function.

   How many times does `TTCAAGCC` occur in the Lambda virus genome when allowing up to 2 mismatches?

   ✓ **Correct**

6. What is the offset of the leftmost occurrence of `AGGAGGTT` in the Lambda virus genome when allowing up to 2 mismatches?

   1 / 1 point

   ✓ **Correct**

7. Finally, download and parse the provided FASTQ file containing real DNA sequencing reads derived from a human:

   1 / 1 point

   https://d28rh4a8wq0iu5.cloudfront.net/ads1/data/ERR037900_1.first1000.fastq

   Note that the file has *many* reads in it and you should examine all of them together when answering this question. The reads are taken from this study:

   Ajay, S. S., Parker, S. C., Abaan, H. O., Fajardo, K. V. F., & Margulies, E. H. (2011). Accurate

   and comprehensive sequencing of personal genomes. *Genome research*, 21(9), 1498-1505.

This dataset has something wrong with it; one of the sequencing cycles is poor quality.

Report which sequencing cycle has the problem. Remember that a sequencing cycle corresponds to a particular offset in *all* the reads. For example, if the leftmost read position seems to have a problem consistently across reads, report **0**. If the fourth position from the left has the problem, report **3**. Do whatever analysis you think is needed to identify the bad cycle. It might help to review the "Analyzing reads by position" video.

✓ **Correct**