Overview

Various theoretical frameworks can enable substantial inference from DNA sequences, with applications ranging from genomics to biochemistry. In this HW, you will encounter several algorithms employed in different fields of sequence analysis.

This assignment is worth a total of 50 points.

Reading

• Why are de Bruijn graphs useful for genome assembly? (Compeau et al., 2011)
• Scalable Genome Assembly through Parallel de Bruijn Graph Construction for Multiple k-mers (Mahadik et al., 2019)
• Applications and statistics for multiple high-scoring segments in molecular sequences (Karlin and Altschul, 1993)

Handin

Submit your answers to the following problems as a PDF on Gradescope using your anonymous email address. You may include images of hand-drawn diagrams if necessary, but all written responses must be typed up. Do not include any identifying information on your handin.

P1: de Bruijn Graphs (25 points)

Answer the following questions about the target DNA sequence below:

ATTATTCTTG

1. What are all the 3-mers of this sequence (including repeats)?
2. What are all the distinct 2-mers necessary to create these 3-mers?
3. What are all the 4-mers of this sequence (including repeats)?
4. What are all the distinct 3-mers necessary to create these 4-mers?

Suppose that we carried out a perfect sequencing of our target sequence; i.e., we have complete coverage with no sequencing errors, and all our reads are of length $k$ with overlaps between adjacent reads of length $k - 1$. Note that your answers to P1.1 and P1.3 result from perfect sequencing with $k = 3$ and $k = 4$, respectively.
Now recall that de Bruijn graphs are directed, and that all edges should be labeled uniquely.

5. Construct a de Bruijn graph for \( k = 3 \), using your answers to P1.1 as edges and P1.2 as nodes.
6. Construct a de Bruijn graph for \( k = 4 \), using your answers to P1.3 as edges and P1.4 as nodes.
7. How many Eulerian paths does your graph from P1.5 contain? What DNA sequences can be inferred?
8. How many Eulerian paths does your graph from P1.6 contain? What DNA sequences can be inferred?
9. de Bruijn graphs can sometimes fail to yield the true original DNA sequence. Name at least two potential drawbacks to using de Bruijn graphs for sequence inference and illustrate with examples from your graphs if possible.

Now consider the following set of reads:

\[
\begin{align*}
&\text{ATC} & \text{CCT} & \text{CTG} & \text{GAT} & \text{GGA} & \text{TCC} & \text{TGG} \\
\end{align*}
\]

10. Construct a de Bruijn graph for \( k = 3 \), using the reads above.
11. Determine all possible target sequences which can be inferred from your graph from P1.10.

Suppose that upon reviewing the sequencing information, we discover two additional reads: \( \text{ATA} \) and \( \text{TAT} \).

12. Construct a de Bruijn graph for \( k = 3 \), using all of the reads above.
13. Determine all possible target sequences which can be inferred from your graph from P1.12.

**P2: Maximal-Scoring Subsequences (15 points)**

Consider a sequence \( x = (x_1, x_2, \ldots, x_n) \) of (not necessarily positive) real numbers, called “scores”. We define the score of an arbitrary subsequence \((x_i, x_{i+1}, \ldots, x_j)\) as

\[
S_{i,j} = \sum_{i \leq k \leq j} x_k
\]

where \( 1 \leq i \leq j \leq n \).

**Defn:** A **maximal-scoring subsequence** (MSS) is a contiguous subsequence of \( x \) which maximizes \( S_{i,j} \) over all \( 1 \leq i \leq j \leq n \). The \( k \text{th}-best subsequence \) is that which maximizes \( S_{i,j} \) among all subsequences of \( x \) which are disjoint from all \((k - 1)\text{th}, \ldots, 1\text{st}-\)best subsequences. Such subsequences may not contain zero-scoring prefixes or suffixes.

1. Why is it beneficial to only consider disjoint regions of \( x \) in looking for successive best subsequences?
2. Why is it beneficial to exclude zero-scoring prefixes/suffixes in looking for successive best subsequences?
3. Describe a naive algorithm to find the MSS for \( x = (x_1, \ldots, x_n) \). Give an overview of its methodology and runtime. You may assume there is only one MSS.

**Bonus (5 points):** Consider the following alternative definition for a maximal-scoring subsequence:

**Defn:** A subsequence \( s \) of \( x \) is an MSS of \( x \) if and only if the following two conditions hold:

- all proper subsequences of \( s \) have a lower score
- no proper supersequence of \( s \) in \( x \) satisfies the previous condition

Describe a linear-time algorithm to find the MSS for \( x = (x_1, \ldots, x_n) \). Give an overview of its methodology and justify its runtime. You may assume there is only one MSS.
P3: Reading Questions (10 points)

Read through the three papers linked above and answer the following questions:

1. Compeau et al. describe the historical origins of de Bruijn graph theory. Why are $k$-mers typically used to define edges rather than nodes in constructing the graph?

2. Compeau et al. also discuss numerous strategies to combat the drawbacks of de Bruijn graphs. How do modern-day sequencing technologies account for the fact that some reads may be missed?

3. Compeau et al. outline several powerful heuristics enabled by next-generation sequencing (NGS). How do paired reads assist in resolving one of the major problems of de Bruijn graph inference?

4. ScalaDBG, an algorithm recently developed by Mahadik et al., is designed to parallelize de Bruijn graph construction and inference for greater efficiency in genome assembly. Briefly describe the motivation and methodology behind the two “patching” algorithms detailed in the paper.

5. Karlin and Altschul outline several possible applications of the MSS problem in the realm of biological inference. Briefly describe at least three such applications and expand upon the methodology and results of at least one of them in greater detail.