HW2: Graph & Sequence Inference

CS 182 Spring 2024

Released: Thursday, February 15, 2024

Due: Thursday, February 22, 2024

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. 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 (14 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.

P2: Maximal-Scoring Subsequences (11 points)

Please read at least the abstract and introduction of the following paper for some context on the importance of MSS algorithms in computational biology!

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^{th}*-best subsequence* is that which maximizes $S_{i,j}$ among all subsequences of x which are disjoint from all $(k-1)^{th}$-, $\ldots$, 1^{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 dynamic programming solution that runs in linear time for determining the MSS. Given a sequence of scores $x = (x_1, x_2, ..., x_n)$, your algorithm should output the indices of the MSS, as well as the maximum score.
   hint: Think about generalizing the Smith-Waterman algorithm for local alignment to the 1-dimensional case. You should build a 1-D dynamic programming array. You can either describe your algorithm in words or use pseudocode. You can assume that there is only one MSS.

P3: Expected Number of Forks (15 points)

In the paper that introduces the Idury-Waterman Algorithm for genome assembly, Idury and Waterman prove the expected number of vertices, singletons, and forks in an experimentally constructed de Bruijn graph using Poisson statistics. In this problem, we will walk through the intuition for one of these proofs.

In particular, we will seek to explain the intuitions behind the following equation:

$$\mathbb{E}(|F|) = 2L' \sum_{i=0}^{\infty} e^{-c} \frac{c^i}{i!} \sum_{j=2}^{i} \binom{i}{j} (1-R)^j R^{i-j} (1-r)^j [1-(1-r)^j]$$

a) We can start by thinking about the outer term: $2L'$. What exactly does L' represent? And why might we need to multiply by 2 when specifically thinking about the number of forks?

b) From here, we can think about why we need to multiply the rest of the equation by $2L'$. What kind of value is $\sum_{i=0}^{\infty} e^{-c} \frac{c^i}{i!} \sum_{j=2}^{i} \binom{i}{j} (1-R)^j R^{i-j} (1-r)^j [1-(1-r)^j]$?

c) Next, we'll think about the large term itself. Why do we sum from $i = 0$ to ∞? What does the term $e^{-c} \frac{c^i}{i!}$ represent?

d) From here, we'll consider the nested sum, $\sum_{j=2}^{i} \binom{i}{j} (1-R)^j R^{i-j} (1-r)^j [1-(1-r)^j]$. What conditional probability does $\sum_{j=2}^{i} \binom{i}{j} (1-R)^j R^{i-j} (1-r)^j [1-(1-r)^j]$ represent? Write a mathematical expression for the final question, using the additional random variable Z, where $Z = 1$ if a $(k-1)$-tuple is associated with a fork and 0 otherwise.

e) Additionally, what does $\binom{i}{j} (1-R)^j R^{i-j}$ represent? How does $(1-r)^j [1-(1-r)^j]$ relate to our decision to multiply the whole expression by 2 in part a)?

P4: 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?