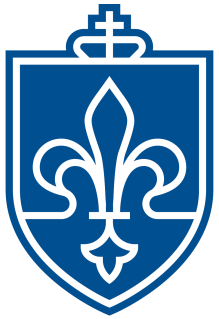
Lecture 15: Midterm Review

BCB 5300 Algorithms in Computational Biology

Fall 2019

Tae-Hyuk (Ted) Ahn

Department of Computer Science
Program of Bioinformatics and Computational Biology
Saint Louis University



SAINT LOUIS UNIVERSITY™

EST. 1818

Outline

- Midterm prep and sample problems

Complexity

Q: Find the complexity of the following set loops, where n is given as input. Express your answer using the Θ(·) notation.

```
i <-- n;
while(i > 1) {
    j = i;
    while (j < n) {
        k <-- 0;
        while (k < n) {
            k = k + 2;
        }
        j <-- j * 2;
    }
    i <-- i / 2;
}
```

Complexity

Q: The statements below show some features of “Big-Oh” notation for the functions f = f(n) and g = g(n). Determine whether each statement is TRUE or FALSE and correct the formula in the latter case.

(1) Rule of sums: $O(f + g) = O(f) + O(g)$

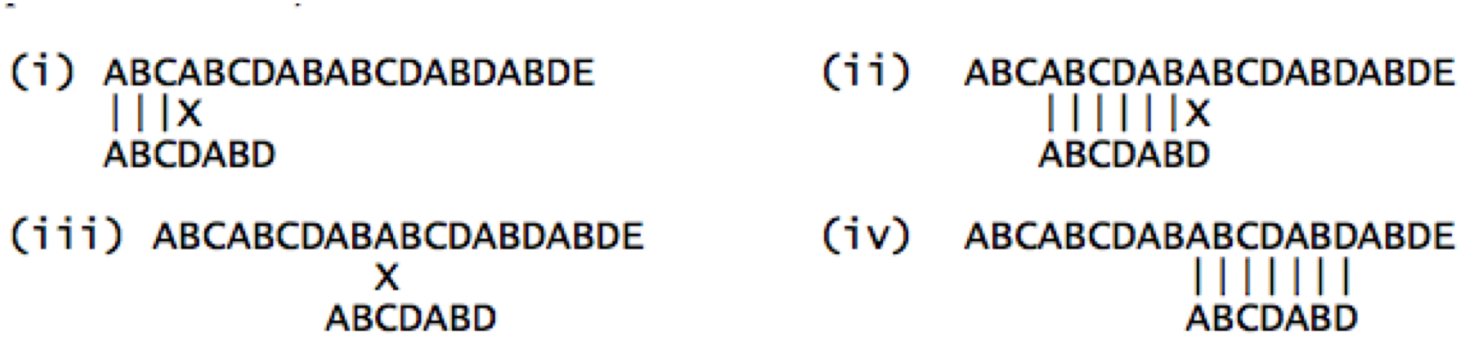
(2) Rule of products: $O(f \cdot g) = O(f) \cdot O(g)$

(3) $5n + 8n^2 + 100n^3 = O(n^4)$

(4) $2n + 7n^3 + n^4 = O(n^3 \log n)$

KMP Algorithm

Q: The steps outlined below represent an execution of the KMP algorithm (text on top, pattern on bottom)



In step (i), the pattern is matched until a mismatch occurs at the 4th character. Since no suffix of the matched portion matches a prefix of the whole pattern (no prefix and suffix of ABC match each other), the pattern is shifted beyond the aligned region and the matching continues from the beginning of the pattern. In (ii) a mismatch is found at position 7 in the pattern and the pattern is shifted (iii) so that the common prefix and suffix of the matched regions are aligned (the first AB in the pattern is placed where the second AB matched at step (ii)). The third position in the pattern is compared to the text and results in a mismatch. The pattern is then shifted past this position and a match is found.

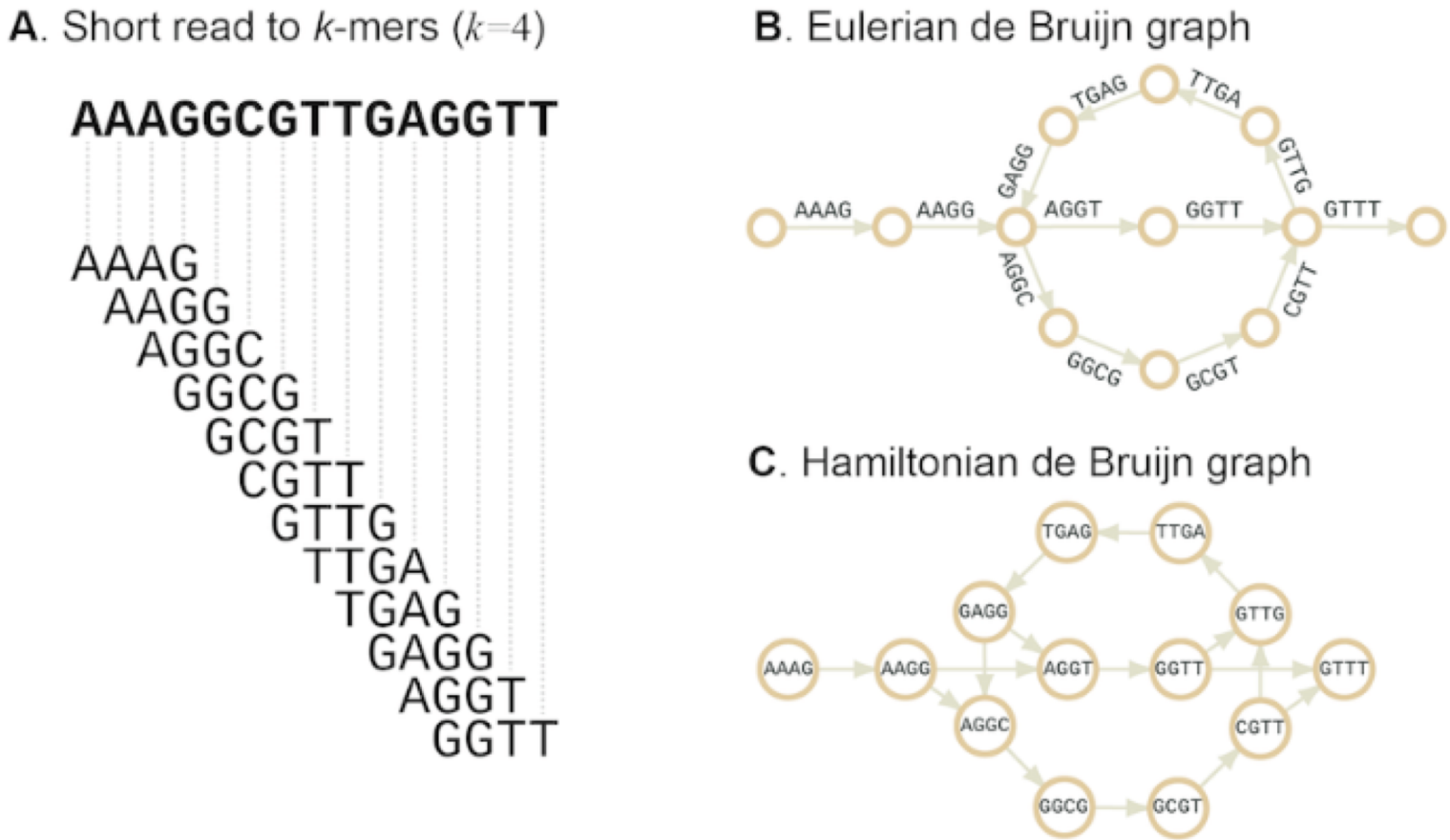
Using this execution as an example, what is the time complexity of KMP algorithm?

Motif Search

Motif finding

- What is the motif finding problem (biologically and computationally)?
- Why motif finding problem is challenging?
- Performance of Brute Force and Greedy Motif Search algorithms?

Genome Assembly



(A) In the de Bruijn graph approach, short reads are split into short k-mers before the de Bruijn graphs are built. (B) In the Hamiltonian approach, the k-mers (or sequences) are the nodes, whereas they are the edges in the Eulerian approach. The k-mers are connected to neighbors by overlapping prefix and suffix (k-1)-mers.

- Find the Eulerian path
- Is the Eulerian path unique?
- Reconstruct the sequence from this path.
- The Hamiltonian and Eulerian approaches to sequence assembly (complexity and tradeoffs)

Dynamic Programming

The Coin Row Problem:

Suppose you have a row of coins with values that are positive integers $c_1, \cdots, c_n$. These values might not be distinct. Your task is to pick up coins have as much total value as possible, subject to the constraint that you don't ever pick up two coins that lie beside each other.

- How would you solve this using dynamic programming?
- Make your pseudocode for it.
- Solve the problem for coins with values c_1 to c_6 as follows: (7, 9, 10, 9, 3, 5, 2).
- Can you think about Greedy algorithm?