Assignment 01

Algorithms for Sequence Analysis

Sven Rahmann and Jens Zentgraf

23.04.2024, due 29.04.2024 (23:59)

01.1: Naive Pattern Search Algorithm (4 Theory)

Consider

- an alphabet of size 4 $\Sigma = \{A, C, G, T\}$
- with the probabilities $(p_A = 0.26, p_C = 0.23, p_G = 0.24, p_T = 0.27)$,
- a random pattern
- of length $m = 6$.

a) What is the expected number of comparisons against a text window for the naive algorithm?

b) What is the number for $m \to \infty$?

01.2: Canonical Codes (4 Theory)

Reminder

The canonical code is the **minimum** of the encoding of the DNA k-mer and its reverse complement.

Calculate the canonical codes for the given k-mers.

a) TCGAT

b) TAGCTA

Is the given value a canonical code for $k = 5$?
If yes, what is the corresponding 5-mer?

c) 696

d) 975

01.3: Shift-And Algorithm (4 Theory)

- $\Sigma := \{\texttt{A}, \texttt{B}, \texttt{C}\}$,
- $P :=$ BBBABBABAB,
- $T :=$ BBABCBBABBBABBABABAB.

1. Calculate the masks for each character.
2. Execute the Shift-And algorithm, and provide the the bit vector D after each step.
3. Mark the bit vector if we have a hit

01.4: Horspool Implementation (Programming 4P)

The Horspool algorithm was discussed in the lecture. If we have a small alphabet $\Sigma = \{A, B\}$ and a long sequence and pattern, it can be useful two compute the shift table based on more than one character.

Example

$p = BAAAAB$

```
shifts
length 1: 1       2               3
          A B     AA AB BA BB     AAA AAB ABA ABB BAA BAB BBA BBB
          1 5      1  5  4  5       1   5   4   5   3   5   4   5
```

You can use and adapt the code provided on the lecture slides.

- Implement the support of shift pattern length $l = \{1, 2, \ldots\}$.
    - Add l as a parameter to the `preprocessing` and `search` functions.
    - Adapt the functions to support l.

For the text

$T = ABAABABABABBABABAABBBABACABBBABABBABABAABABCBABC$

and the pattern $P = ABABBABABA$:

- Print the matching positions.
- Count how often is the comparison between pattern and text done for $l \in \{1, 2, 3\}$
- How often is a shift of length n done for $l \in \{1, 2, 3\}$