May 31, 2021

Assignment 7
Algorithms for Sequence Analysis, Summer 2021

Algorithmic Bioinformatics · Prof. Dr. Sven Rahmann

Hand in date: Monday, June 07, before 20:00

Exercise 0: ±1RMQ (8 Programming)
Remember that the programming task from assignment sheet 6 is also due.

Exercise 1: Rank queries on a C64 (4 Theory)
The Commodore 64 (C64) was an 8-bit home computer. We want to build a rank data structure for the following bit sequence s, using a block size of $B = 8$ (because that is the register size), and a superblock size of $S = B^2/4 = 16$:

```
          1         2         3         4         5         6
i: 01234567890123456789012345678901234567890123456789012345678901234567890123456789
s: 10110110111011011100011011101111111000011011111110001010111001010111111
```

Write down all tables for the succinct rank data structure. How many bits do you need in comparison to the 70 bits for s?

Exercise 2: Wavelet tree (4 Theory)
Let $\Sigma := \{\texttt{a},\texttt{b},\texttt{c},\texttt{d},\texttt{e},\texttt{f},\texttt{g},\texttt{h}\}$. Compute the wavelet tree of

$$s = \texttt{dbaggdhcffcbedgfbhbfddged}.$$

Illustrate how to use binary rank queries on the wavelet tree to find

(a) $s[15]$,

(b) $\mathrm{rank}_{\texttt{g}}(12)$ in s.

Exercise 3: Fibonacci strings (4 Theory)
The well-known (integer) Fibonacci sequence is defined by the recurrence

$$F_0 := 1, \qquad F_1 := 1, \qquad F_n := F_{n-2} + F_{n-1} \text{ for } n \geq 2\,.$$

Similarly, we define the sequence of *Fibonacci strings* by

$$f_0 := \texttt{a}, \qquad f_1 := \texttt{b}, \qquad f_n := f_{n-2}f_{n-1} \text{ for } n \geq 2\,.$$

It follows that the length of f_n is $|f_n| = F_n$. We have $f_2 = \texttt{ab}$, $f_3 = \texttt{bab}$, $f_4 = \texttt{abbab}$, ... Note that the concatenation order ($f_{n-2}f_{n-1}$ vs. $f_{n-1}f_{n-2}$) matters. Fibonacci strings are interesting because they contain long repeats and compress well.

(a) Describe an algorithm that, given n and $k < F_n$, computes $f_n[k]$ efficiently. Your solution should work for large n and k, e.g., what is $f_{1000}[999\,999\,999]$?

(b) By experimentation, conjecture and proof, determine the maximum lcp value for $f_n\$$ as a function of n.

Exercise 4: Lempel-Ziv factorizations (4 Theory)
For the Fibonacci string f_6 (length 13), find

(a) the LZ77 factorization as defined in the lecture, using the suffix tree of $f_6\$$,

(b) the LZ78 factorization as defined in the lecture, by constructing the factor trie.

Show the suffix tree and factor trie, respectively.