Algorithmic Bioinformatics
Prof. Dr. Sven Rahmann

UNIVERSITÄT DES SAARLANDES

ZBI ZENTRUM FÜR BIOINFORMATIK

Assignment 1 - Algorithms for Sequence Analysis, Summer 2021

Exercise 1: Naïve Pattern Matching (4 Theory)
Consider an alphabet of size 4 (like DNA) and a random (uniform i.i.d.) pattern of length $m = 6$. What is the expected number of comparisons against a random text window for the naïve algorithm? What is the number for $m \to \infty$?

Exercise 2: DFAs & NFAs (4 Theory)

(a) Design an NFA over $\Sigma = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$ that accepts all strings that end in a *non-empty* suffix whose sum of the number of `a`s and `c`s is divisible by 4 or 6, or that is equal to `bb`. Use as few states as possible. (Read carefully!)

(b) Write down this NFA formally.

(c) Use the subset construction to design an equivalent DFA. How many states do you need? Note: You should specify the idea; you do not need to show all details of the transition function.

Exercise 3: Knuth-Morris-Pratt (4 Theory)
The Knuth-Morris-Pratt algorithm finds exact matches of a pattern in a text in optimal worst-case linear time by using a function called lps. The lps function is computed for each pattern $P[1 \ldots m]$ once and will be used to find P in any input text. For each $q \in \{1, \ldots, m\}$, $\mathrm{lps}(q)$, is the length of the longest prefix of P that is a suffix of $P[1 \ldots q]$.

(a) Study the code slide from the lecture that shows the construction of the lps function.

(b) Argue why the function needs $O(m)$ time to build lps.

(c) Show the correctness of the construction. Use the invariant given on the slide for your argument.

Exercise 4: Shift-Or Algorithm (4 Theory)
The Shift-And algorithm uses the following update step for a bit vector `D` of active states:

```
D = ((D << 1) | 1) & masks[c]
```

Therefore, we need to carry out one *shift* operation, one *OR* operation, one *AND* operation, and one table lookup.

(a) Explain why the OR-operation (`| 1`) is necessary.

(b) We can speed up the algorithm by inverting the bit logic, such that 0 means active, 1 means inactive. Of course, also all bit masks will have to be inverted, etc. Work out the details of this variant of the algorithm, including initialization. How many and which operations are needed for one update step?

(c) The resulting algorithm is called Shift-Or algorithm. Why?

Exercise 5: Shift-Or Algorithm (4 Programming)
Implement the Shift-And algorithm and the Shift-Or algorithm from the previous problem in an efficient language (C, Java, compiled Python, etc.) with a small command line interface (CLI). A Python example is provided.
Your program has to be callable by `./task_01_5` and must at least support the following:

```
./task_01_5 -P ABC -t text.txt -a and
./task_01_5 -p patterns.txt -t text.txt -a or
./task_01_5 -P ABC -T ABCDEF
```

Here, `-P` indicates an immediate pattern, `-p` a file name containing one pattern per line (output will also be one line per pattern), `-T` an immediate text, `-t` a file with a text. The algorithm is specified with `-a and` (which is the default) or `-a or`.
The output should be a comma-separated list of all positions in the text where a pattern occurs. Positions should be 0-based, i.e., counting starts at 0. Sample input and output files are provided.

Remarks

- 50% of points in each category theory and programming (over all exercises and not each assignment sheet separately) are necessary to take the exam.
- You are allowed to work in groups of <u>two</u> and only one of the group members need to submit.
- Submission is via GitHub Classroom (as demonstrated in the lecture).
- Source code must be sufficiently commented and documented to be understandable.
- When using a compiled language, compilation instructions and tools must be provided (e.g., a Makefile).
- In addition to source code, the output must be provided.
- Also, a file AUTHORS with your name(s) must be provided.
- Copying between groups will result in zero points for all involved groups!

Hand in date: Monday, April 26, before 20:00