Übungen zur Algorithmischen Bioinformatik I

Blatt 11

Xiheng He

Juli 2021

1. Aufgabe: Z-Boxes and Z-Algorithm (10 Punkte)

Fundamental preprocessing of a string computes so called **Z-boxes** and yields probably the simplest linear-time exact matching algorithm.

Definition (Z-Box): Given a string S and a position $i > 1$ in S. $Z_i(S)$ is the **length** of the longest substring in S that starts at i and matches a prefix of S.

$$Z_i(S) = max_k\{(i[S)[1:k] = k]S\} = max_k\{S[i:i+k] = S[1:k]\}$$

Given the following binary string $B = 010100101011$

(a) For all i, compute the $Z_i(B)$ values.

i	1	2	3	4	5	6	7	8	9	10	11
z-box	-	010	-	0	01010	-	0101	-	01	-	-
z-value	0	3	0	1	5	0	4	0	2	0	0

(b) Draw all Z-boxes (i.e.$Z_i(B)) > 0$ for string B.

Hinweis: Boxen mit blauer Farbe sind Z-boxen.

- $Z_2 = 3$: 0 1 0 1 0 0 1 0 1 0 1 1
- $Z_4 = 1$: 0 1 0 1 0 0 1 0 1 0 1 1
- $Z_5 = 5$: 0 1 0 1 0 0 1 0 1 0 1 1
- $Z_7 = 4$: 0 1 0 1 0 0 1 0 1 0 1
- $Z_9 = 2$: 0 1 0 1 0 0 1 0 1 0 1 1

(c) What is $Z_1(S)$?

$$Z_1(S) = \begin{cases} k+1 & \text{falls } S[0:k] = S[1:1+k] \\ 0 & \text{sonst} \end{cases}$$

(d) What is the complexity of a naive algorithm to compute all Z-Boxes of a string S of length n? Explain!

Für einen naiven Algorithmus liegt die Komplexität der Laufzeit in $O(m^2)$ wobei m die Länge der Strings S ist. Beide While-Schleife und For-Schleife können maximal $m-1$ mal durchlaufen. Somit liegt die Komplexität der Laufzeit in $O(m^2)$ für einen naiven Algorithmus (siehe Pseudocode).

```
Z-Box_naive (int[] S, int m)
begin
    for (i = 1; i < m; i++) do
        Z_i := 0, j := 0 ;                                   /* start from 0 */
        while ((i + j < m) && (S[j] = S[i + j])) do
            Z_i := Z_i + j;
            j := j + 1;
end
```

(e) Let $Z_i(S) > 1$ what is the end k of the Z-Box at position i, i.e. $Z[i, \ldots, k]$.

Leider kann man hier nicht wirklich nachvollziehen, wofür k steht. In **Definition (Z-Box)** ist $Z_i(S) = S[i : i + k] = S[1 : k]$, aber hier ist $Z_i(S) = Z[i, \ldots, k]$. Exakter Wert kann nicht ausgegeben werden aber wenn $Z_i(S) > 1$ gilt:

$$Z_i(S) > 1 \Longrightarrow |Z[i, \ldots, k]| > 1 \Longrightarrow k > i \wedge k < |S|$$

Führen Sie per Hand den Z-Box Algorithmus für $B = 010100101011$ durch. Geben Sie in einer Tabelle für jedes k die Werte von $k', q, |\beta|, l_k, r_k$ und Z_k in dieser Reihenfolge an. Falls einige dieser Werte für einen Schritt irrelevant sind, kennzeichnen Sie dies.

Achtung: $k' \neq k - l + 1, k' = k - l$

irrelevant Variablen: -

k	1	2	3	4	5	6	7	8	9	10	11
Z_k	0	3	0	1	5	0	4	0	2	0	0
l_k	0	2	2	4	5	5	7	7	9	9	9
r_k	0	4	4	4	9	9	10	10	10	10	10
k'	-	-	1	0	-	1	2	1	2	1	2
$\|\beta\|$	-	-	2	1	-	4	4	3	2	1	0
q	-	-	-	-	-	-	11	-	11	-	11