

Übungen zur Algorithmischen Bioinformatik I

Blatt 6

Abgabetermin: Montag, 12.6.2017, 10 s.t.

1. Aufgabe:

Betrachten Sie den unten angegebenen Boyer-Moore-Algorithmus, der für die Bestimmung der Shifts bei einem Mismatch nur die Extended-Bad-Character-Rule berücksichtigt. Geben Sie eine Konstante $c > 0$ sowie eine unendliche Familie $\mathcal{F} = \{(s, t) : s, t \in \Sigma^*\}$ an, die für jedes Paar $(m, n) \in \mathbb{N}^2$ ein Paar $(s, t) \in \mathcal{F}$ mit $|s| \geq m$, $|t| \geq n$ und $|s| \leq |t|$ enthält und bei dem diese Variante für eine erfolglose Suche mindestens $c \cdot (|s| \cdot |t|)$ Zeichenvergleiche ausführt.

BC-Boyer-Moore (char $t[]$, int n , char $s[]$, int m)

```
int  $i := 0$ ,  $j := m - 1$ ;  
while ( $i \leq n - m$ ) do  
    while ( $t[i + j] = s[j]$ ) do  
        if ( $j = 0$ ) then return TRUE;  
         $j--$ ;  
         $i := i + j - \max\{k : (k < j \wedge s[k] = t[i + j]) \vee (k = -1)\}$ ;  
         $j := m - 1$ ;  
return FALSE;
```

2. Aufgabe:

Betrachten Sie den Boyer-Moore-Algorithmus, der für die Bestimmung des Shifts bei einem Mismatch nur die Weak-Good-Suffix-Rule berücksichtigt:

$$S'[j] = \min \left\{ \begin{array}{l} \sigma : (s_{j+1} \cdots s_{m-1} \in \mathcal{R}(s_{j+1-\sigma} \cdots s_{m-1}) \wedge \sigma \leq j) \vee \\ \sigma : (s_0 \cdots s_{m-1-\sigma} \in \mathcal{R}(s_0 \cdots s_{m-1}) \wedge \sigma > j) \vee (\sigma = m) \end{array} \right\}$$

Geben Sie eine Konstante $c > 0$ sowie eine unendliche Familie $\mathcal{F} = \{(s, t) : s, t \in \Sigma^*\}$ an, die für jedes Paar $(m, n) \in \mathbb{N}^2$ ein Paar $(s, t) \in \mathcal{F}$ mit $|s| \geq m$, $|t| \geq n$ und $|s| \leq |t|$ enthält und bei dem diese Variante für eine erfolglose Suche mindestens $c \cdot (|s| \cdot |t|)$ Zeichenvergleiche ausführt.

3. Aufgabe (Bonus-Aufgabe):

Geben Sie einen Algorithmus in Pseudo-Code an, der die Tabelle ebc (definiert auf Folie 99 zur Textsuche) in Zeit $O(|\Sigma| \cdot m)$ berechnet.

4. Aufgabe (Bonus-Aufgabe):

Gegeben seien $s \in \Sigma^m$ und $t \in \Sigma^n$ mit $m \leq n$. Zeigen Sie, wie man Hilfe der Berechnung von Z-Werten für die Zeichenreihe $s\$t$ mit $\$ \notin \Sigma$ in Zeit $O(n + m)$ feststellen kann, wo überall s in t als Teilwort auftritt.