Übungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, SS 2021
Prof. Dr. Jens Stoye · Dr. Marília D. V. Braga

https://gi.cebitec.uni-bielefeld.de/teaching/2021summer/sa

Übungsblatt 2 vom 29.4.2021
Abgabe am 6.5.2021 bis 12:00 Uhr (mittags)

Aufgabe 1 (Berechnung der Edit-Distanz) (5 Punkte)
Gegeben seien die Strings $x =$ `PLANET` und $y =$ `PARENTS`. Benutze zur Berechnung der Edit-Distanz der beiden Strings eine Edit-Matrix und gib die Distanz an. Benutze außerdem eine weitere Matrix, in der die optimalen Edit-Operationen gespeichert werden, um alle optimalen Edit-Sequenzen bestimmen zu können. Schreibe eine optimale Edit-Sequenz explizit auf.

Aufgabe 2 (Edit-Distanzen) (6 Punkte)
Die Rekurrenz zur Berechnung der Standard-Edit-Distanz mit Einheitskosten lautet
für $1 \leq i \leq |x|, 1 \leq j \leq |y|$:

$$D(i,j) = \min \begin{cases} D(i-1,\, j-1) + \mathbb{1}_{\{\{x[i] \neq y[i]\}\}} \\ D(i-1,\, j) + 1 \\ D(i,\, j-1) + 1 \end{cases}$$

Die Rekursionsbasis ist gegeben durch:

$$D(0,\, j) = j \text{ für } 0 \leq j \leq |y| \text{ und } D(i,\, 0) = i \text{ für } 0 \leq i \leq |x|$$

1. Wie sehen die Rekurrenz und die Basisfälle für die *Hamming-Distanz* aus?
2. Wie sehen die Rekurrenz und die Basisfälle für die *LCS-Distanz* aus?
3. Ist es möglich, die *Edit+Flip-Distanz* rekursiv zu berechnen? Wenn nein, warum nicht? Wenn ja, gib eine Formel und eine Rekursionsbasis an. Welche Schwierigkeiten könnten zudem hierbei auftreten?

Aufgabe 3 (Rank und Unrank) (4 Punkte)
Gegeben sind das Alphabet $\Sigma = \{$`A`,`B`,`C`,`D`,`E`$\}$ mit $r_\Sigma($`A`$) = 0$, $r_\Sigma($`B`$) = 1$, $r_\Sigma($`C`$) = 2$, $r_\Sigma($`D`$) = 3$, $r_\Sigma($`E`$) = 4$ und die Wortlänge $q = 6$. Verwende die absteigende Variante der Codierung von $q - 1$ nach 0.

1. Berechne den Rang des Wortes `BADECA`. Gib alle Zwischenschritte an.
2. Berechne den Rang des Wortes `ADECAB` ohne vollständige Neuberechnung, sondern durch ein Update in konstanter Zeit (aus dem Rang des Wortes `BADECA`). Gib alle Zwischenschritte an.
3. Welches Wort hat den Rang 2402?

Aufgabe 4 (Worte mit gleichem q-Gramm-Profil) (5 Punkte)
Gegeben sei der String $x =$ `ATCATCGAT`; finde alle Strings, die von x unterschiedlich sind, aber das gleiche q-Gramm-Profil haben; und zwar:

1. für $q = 4$
2. für $q = 3$
3. für $q = 2$