Übungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, WS 2024
Prof. Dr. Jens Stoye · Leonard Bohnenkämper

https://gi.cebitec.uni-bielefeld.de/teaching/2024winter/sa

Übungsblatt 3 vom 31.10.2024
Abgabe am 07.11.2024 bis 10:00 Uhr (morgens)

Aufgabe 1 (Rank und Unrank) (5 Punkte)
Gegeben sind das Alphabet $\Sigma = \{\texttt{A},\texttt{B}\}$ mit $r_\Sigma(\texttt{A}) = 0$, $r_\Sigma(\texttt{B}) = 1$ und die Wortlänge $q = 4$. Verwende die absteigende Variante der Codierung von $q-1$ nach 0.

1. Berechne den Rang des Wortes `BABA`. Gib alle Zwischenschritte an.
2. Berechne den Rang des Wortes `ABAB` ohne vollständige Neuberechnung, sondern durch ein Update in konstanter Zeit (aus dem Rang des Wortes `BABA`). Gib alle Zwischenschritte an.
3. Welches Wort hat den Rang 6?

Aufgabe 2 (Worte mit gleichem q-Gramm-Profil) (3 Punkte)
Gegeben sei der String $x =$ `ACTGACT`; finde alle Strings, die von x unterschiedlich sind, aber das gleiche 4-Gramm-Profil haben.

Aufgabe 3 (Laufzeit der Berechnung der q-Gramm-Distanz) (5 Punkte)
Begründe, warum die Berechnung der q-Gramm-Distanz von zwei Strings für festes q und $|\Sigma|$ in linearer Zeit berechnet werden kann. Gehe dafür die Schritte durch, die notwendig sind, um die Distanz zu berechnen. Nimm an, dass die q-Gramm Profile als Arrays gespeichert werden.

Wie sieht die Laufzeit aus, wenn q nicht mehr fest, sondern ein Eingabeparameter ist? Wie muss man q für Stringlänge n skalieren, damit der Algorithmus linear bleibt, also für $q = f(n)$, in welcher $\mathcal{O}$-Klasse muss f liegen?

Aufgabe 4 (q-Gramm- und Maximal-Matches-Distanzen als Filter) (7 Punkte)
Gegeben seien die Sequenzen: $\begin{cases} x &= \texttt{CAGAGTGCGG} \\ \\ y_1 &= \texttt{CAAATTCCGG} \\ y_2 &= \texttt{CAGGAGTGG} \\ y_3 &= \texttt{CAGTGAGCCG} \end{cases}$

Wir wollen entscheiden, ob die Sequenzen $y_1, \ldots, y_3$ eine *Edit*-Distanz von max. 2 zur Sequenz x haben können, ohne alle *Edit*-Distanzen zu berechnen.

1. Berechne die 2-Gramm-Profile der Worte x und $y_1, \ldots, y_3$. Filtere die Sequenzen $y_1, \ldots, y_3$ mit Hilfe der 2-Gramm-Distanz. Welche Sequenzen können ausgeschlossen werden?
2. Filtere die übrigen Sequenzen mit Hilfe der Maximal-Matches-Distanz. Welche Sequenzen bleiben als Kandidaten übrig?
3. Nenne einen weiteren Filter, den man auch noch verwenden könnte.