Übungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, SS 2023
Prof. Dr. Jens Stoye · Tizian Schulz

https://gi.cebitec.uni-bielefeld.de/teaching/2023summer/sa

Übungsblatt 3 vom 20.4.2023
Abgabe am 27.4.2023 bis 12:00 Uhr (mittags)

Aufgabe 1 (Rank und Unrank) (5 Punkte)
Gegeben sind das Alphabet $\Sigma = \{\texttt{A},\texttt{B},\texttt{C}\}$ mit $r_\Sigma(\texttt{A}) = 0$, $r_\Sigma(\texttt{B}) = 1$, $r_\Sigma(\texttt{C}) = 2$ und die Wortlänge $q = 5$. Verwende die absteigende Variante der Codierung von $q-1$ nach 0.

1. Berechne den Rang des Wortes ACBAC. Gib alle Zwischenschritte an.
2. Berechne den Rang des Wortes CBACA ohne vollständige Neuberechnung, sondern durch ein Update in konstanter Zeit (aus dem Rang des Wortes ACBAC). Gib alle Zwischenschritte an.
3. Welches Wort hat den Rang 158?

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

Aufgabe 3 (Laufzeit der Berechnung der q-Gramm-Distanz) (3 Punkte)
Begründe, warum die Berechnung der q-Gramm-Distanz von zwei Strings in linearer Zeit berechnet werden kann. Gehe dafür die Schritte durch, die notwendig sind, um die Distanz zu berechnen.

Aufgabe 4 (Maximal-Matches-Distanz) (5 Punkte)
1. Gegeben seien die Sequenzen $x =$ 220320202 und $y =$ 0203030030.
   (a) Berechne die folgenden Partitionen: $P_{\text{LR}}(x,y)$, $P_{\text{RL}}(x,y)$, $P_{\text{LR}}(y,x)$, $P_{\text{RL}}(y,x)$.
   (b) Wie ist die Maximal-Matches-Distanz $\delta(x||y)$ von x bezüglich y definiert? Gib $\delta(x||y)$ und $\delta(y||x)$ an.
2. Zeige an einem von dir ausgedachten Beispiel, dass die Maximal-Matches-Distanz keine Metrik ist.

Aufgabe 5 (q-Gramm- und Maximal-Matches-Distanzen als Filter) (7 Punkte)
Gegeben seien die Sequenzen:
$$\begin{cases} x = \texttt{ATGCCACAGTT} \\ \\ y_1 = \texttt{ACCATTGCAGT} \\ y_2 = \texttt{CAGTTATGCCT} \\ y_3 = \texttt{CGAGCATTCGA} \end{cases}$$

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.