Übungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, SS 2020
Prof. Dr. Jens Stoye · Dr. Marília D. V. Braga · Leonie R. Brockmann · Rebecca K. Pfeil

https://gi.cebitec.uni-bielefeld.de/teaching/2020summer/sa

Übungsblatt 3 vom 7.5.2020
Abgabe am 14.5.2020 bis 12:00 Uhr (mittags)

Aufgabe 1 (Rank und Unrank) (3 Punkte)
Gegeben sind das Alphabet $\Sigma = \{\texttt{A},\texttt{C},\texttt{G},\texttt{T}\}$ mit $r_\Sigma(\texttt{A}) = 0$, $r_\Sigma(\texttt{C}) = 1$, $r_\Sigma(\texttt{G}) = 2$, $r_\Sigma(\texttt{T}) = 3$ und die Wortlänge $q = 5$. Verwende die absteigende Variante der Codierung von $q - 1$ nach 0.

1. Berechne den Rang des Wortes `AGACT`. Gib alle Zwischenschritte an.
2. Berechne den Rang des Wortes `GACTT` ohne vollständige Neuberechnung, sondern durch ein Update in konstanter Zeit (aus dem Rang des Wortes `AGACT`). Gib alle Zwischenschritte an.
3. Welches Wort hat den Rang 559?

Aufgabe 2 (Worte mit gleichem q-Gramm-Profil) (7 Punkte)
Gegeben sei der String $x =$ `CATGCATATGCA`; finde alle Strings, die von x unterschiedlich sind, aber das gleiche q-Gramm-Profil haben; und zwar:

1. für $q = 5$
2. für $q = 4$
3. für $q = 3$

Aufgabe 3 (Maximal-Matches-Distanz) (3 Punkte)
Gegeben seien die Sequenzen $x =$ `montreal` und $y =$ `motremotreaal`. Berechne $\delta(x||y)$ und $\delta(y||x)$. Gib jeweils die links-nach-rechts- und rechts-nach-links-Partitionen an.

Welche Beobachtung machst du? Was kannst du über die Maximal-Matches-Distanz aussagen?

Aufgabe 4 (q-Gramm- und Maximal-Matches-Distanzen als Filter) (7 Punkte)
Gegeben seien die Sequenzen:

$x =$ `AATCGAGGTAC`

$y_1 =$ `AAGATCGGACC`
$y_2 =$ `AATCGCGGTAC`
$y_3 =$ `AGGTACAATCG`
$y_4 =$ `CTGAACGTCTG`

Wir wollen entscheiden, ob die Sequenzen $y_1, \ldots, y_4$ eine Edit-Distanz von max. 1 zur Sequenz x haben können, ohne alle Edit-Distanzen zu berechnen.

1. Berechne die 2-Gramm-Profile des Wortes x und $y_1, \ldots, y_4$. Filtere die Sequenzen $y_1, \ldots, y_4$ 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.