Übungen zur Vorlesung Sequenzanalyse I

Universität Bielefeld, WS 2011/2012
Dr. Alexander Sczyrba · Nina Luhmann · Linda Sundermann

http://wiki.techfak.uni-bielefeld.de/gi/Teaching/2011winter/SequenzAnalyse

Übungsblatt 5 vom 02.12.2011
Abgabe in einer Woche vor Beginn der Vorlesung.

Aufgabe 1 (q-Gram-Distanz) (5 Punkte)

1. Gegeben sind $x = \texttt{TCTTTCTTTCTTCTTCTTCCC}$ und $y = \texttt{TTTCCCTTCTTCTTTCTTTCTTC}$ und es sei $q = 4$. Bestimme die q-Gram-Profile beider Sequenzen (Schreibe die q-Gramme dabei in lexikographischer Ordnung auf) und berechne die q-Gram-Distanz $d_q(x, y)$.
2. Finde für $v = \texttt{CCCTTCCTCCTTTC}$ und $q = 4$ eine Sequenz w, für die gilt $d_q(v, w) = 0$ und $w \neq v$. (Es gibt eine elegante Lösung, bei der du ohne zielloses Ausprobieren zum Ziel kommst.)
3. Ist die q-Gram-Distanz eine Metrik? Begründe deine Antwort.

Aufgabe 2 (Rank und Unrank) (4 Punkte)
Zur schnellen Berechnung eines q-Gram Profiles kann jedes q-Gram mittels einer Rankingfunktion auf eine natürliche Zahl abgebildet werden. Ein Beispiel für eine solche Funktion ist

$$r(x) = \sum_{i=1}^{q} r_\Sigma\big(x[i]\big) \cdot |\Sigma|^{i-1}$$

welche einer Sequenz $x \in \Sigma^q$ ihren Rang $r(x) \in \mathbb{N}_0^+$ zuweist. Dabei wird die Funktion $r_\Sigma(\cdot)$ benutzt, die das Alphabet Σ auf die Zahlen $\{0, \ldots, |\Sigma| - 1\}$ abbildet.
Gegeben sei nun das Alphabet $\Sigma = \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\}$ ($r_\Sigma(\texttt{A}) = 0$, $r_\Sigma(\texttt{C}) = 1$, ...) und die Wortlänge $q = 4$.

1. Berechne den Rang des Wortes $x = \texttt{ATCG}$. (Gib dabei den Rechenweg an.)
2. Berechne, ohne vollständige Neuberechnung, sondern durch ein Update in konstanter Zeit, ausgehend von dem Rang von x den Rang des Wortes $y = x[2], x[3], x[4], \texttt{G} = \textit{TCG}\texttt{G}$. (Gib ebenfalls den Rechenweg an.)
3. Welche Sequenz $z \in \Sigma^q$ hat den Rang 110? (Gib den Rechenweg an.)

Aufgabe 3 (Distanzen im Vergleich: q-Gram- vs. Edit-Distanz) (3 Punkte)
Erkläre in eigenen Worten, wieso

$$d_q(x, y)/(2q) \leq d(x, y) \tag{1}$$

gilt, wobei x und y zwei Strings sind und $d_q(x, y)$ die q-Gram-Distanz, sowie $d(x, y)$ die Edit-Distanz zwischen ihnen beschreibt.