1. Übung zur Vorlesung “Sequenzanalyse”

Sebastian Böcker und Markus Fleischauer

Aufgabe 1 (5 Punkte)

Zeigen Sie, dass die Edit-Distanz zweier Strings u, v unverändert bleibt, wenn man diese Strings invertiert, d.h.

$$EditDistanz(u, v) \quad = \quad EditDistanz(u^{-1}, v^{-1})$$

wobei w^{-1} der invertierte String des Strings w ist, d.h. wenn $w = w_1 w_2 \dots w_k$ ist, dann $w^{-1} = w_k w_{k-1} \dots w_1$.

Aufgabe 2 (8 Punkte)

1. Berechnen Sie die q-gram Distanz der Sequenzen $u = \texttt{TACTTTCTAGCTTA}$ und $v = \texttt{ACTAGCTTTCTTAC}$:
   (a) für $q = 3$,
   (b) für $q = 5$.
2. Begründen Sie anhand dieses Beispiels, dass die q-gram Distanz keine Metrik ist.
3. Für die q-gram Distanz ist es wichtig ein geeignetes q zu benutzen. Zeigen Sie, dass $q = \frac{log(n) - log(c)}{log(|\Sigma|)}$ eine gute Wahl ist, um zu erreichen, dass für zufällige Strings der Länge n jedes q-gram im Mittel c-mal auftritt. (Dabei seien n und $|\Sigma|$ hinreichend groß.)

Aufgabe 3 (6 Punkte)

Betrachten wir den genetischen Code (für Wirbeltiere).

1. Gegeben seien zwei codierende DNA-Sequenzen der Länge $n = 3m$; ihre Hamming-Distanz betrage $h \in [0, n]$. Wie groß kann die Hamming-Distanz k der zugehörigen Proteinsequenzen der Länge m minimal und maximal sein? Geben Sie einfache Sequenzbeispiele für die Extremfälle an.
2. Jetzt seien zwei Proteinsequenzen der Länge m mit Hamming-Distanz k gegeben. Wie groß kann die Hamming-Distanz h von zugehörigen codierenden DNA-Sequenzen der Länge $n = 3m$ minimal und maximal sein? Geben Sie wiederum Beispiele an.

Aufgabe 4 (6 Punkte)

Sei $d_{LCS}(x, y) = |x| + |y| - 2 \cdot LCS(x, y)$ eine Distanz zwischen $x, y \in \Sigma^*$, wobei $LCS(x, y)$ (longest common subsequence distance) die Länge des längsten gemeinsamen Teilsequenz von x, y ist. Zeigen Sie, dass $d_{LCS}(x, y)$ eine Metrik ist.