5. Übung zur Vorlesung “Sequenzanalyse”

Markus Fleischauer und Sebastian Böcker

Aufgabe 1 (3 Punkte)

q-Gramm Lemma: Gegeben sei ein Alignment der Länge n mit höchstens e Mismatches zwischen zwei Strings X, Y. Es gilt, dass X und Y mindestens $n + 1 - q(e + 1)$ gemeinsame q-Gramme enthalten.

Beweisen Sie das q-Gramm Lemma.

Aufgabe 2 (4 Punkte)

1. Erstellen Sie die Tabelle $first$ und pos für s =ATGGGTTACCGTTATC und $q = 2$. Der Index eines q-Grams ist definiert wie in Aufgabe 3.
2. Erstellen Sie die Tabelle $first$ und pos für s =ABBAAABBABABBBBA und $q = 3$. Dabei sei $\Sigma = \{A, B\}$, $r_\Sigma(A) = 1$ und $r_\Sigma(B) = 2$. Der Index $r(z)$ eines q-Grams ist definiert wie in Aufgabe 3.

Aufgabe 3 (6 Punkte)

Index eines q-Gramms: Sei die bijektive Funktion $r_\Sigma(A) = 1$, $r_\Sigma(C) = 2$, $r_\Sigma(G) = 3$, $r_\Sigma(T) = 4$ eine Rangfunktion über die Menge $\Sigma = \{A, C, G, T\}$. Der Index $r(z)$ eines q-Gramms $z = (z_1, \ldots, z_q) \in \Sigma^q$ wird als

$$r(z) = \left(\sum_{i=1}^{q} (r_\Sigma(z_i) - 1) \cdot |\Sigma|^{q-i} \right) + 1$$

definiert.

Geben Sie einen Algorithmus an, der die Tabelle $first$ und pos in Zeit $O(|s| + |\Sigma|^q)$ berechnen kann. Hinweis: Dazu muss der Index eines q-Gramms **beim Durchlaufen von s** in $O(1)$ (unabhängig von q) berechnet werden.