Übungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, WS 2014/2015
Dr. Roland Wittler · Nina Luhmann · Linda Sundermann

http://wiki.techfak.uni-bielefeld.de/gi/Teaching/2014winter/SequenzAnalyse

Übungsblatt 6 vom 18.11.2014
Abgabe in einer Woche vor Beginn der Vorlesung.

Aufgabe 1 (Dot Plot) (3 Punkte)
Zeichne einen *Dot Plot* für $x =$ `FEUERREGEN` und $y =$ `FEGEFEUER`. Markiere die Treffer, die nach einer Filterung mit $q = 2$ bzw. $q = 4$ noch sichtbar sind. Warum setzt man im Bereich der Sequenzanalyse oft solche Längenfilter ein?

Aufgabe 2 (Implementierung des q-gram Index) (5 Punkte)
Implementiere ein Programm, das den q-gram Index eines Strings x berechnet. Verwende bei der Berechnung die effizientere Variante, die im Skript auf Seite 57 unten beschrieben ist. Als Eingabe soll ein String x und ein bestimmtes q übergeben werden. Der q-gram Index soll in tabellenform ausgegeben werden. Verwende eine Programmiersprache, die mit deinem Tutor abgesprochen ist und sende ihm deinen Quellcode per Email zu.

Aufgabe 3 (k-Nachbarschaften) (4 Punkte)
Gegeben sei die Query $x =$ `TCTCATCTC` und $q = 4$. Der Score für einen Match sei $+3$ und für einen Mismatch -1. Betrachte nun die Nachbarschaft von x:

1. Gib für $i = 3$ und $i = 5$ jeweils zwei Tupel (z, i) an, die sich in der Nachbarschaft $N_8(x)$ befinden, aber nicht in $N_9(x)$.
2. Die k-Nachbarschaft kann ähnlich zu einem q-gram Index repräsentiert werden: Für jedes $z \in \Sigma^q$ gib die Listen $P_k(z)$ von Positionen i an, so dass $(z, i) \in N_k(x)$.

   Bestimme die folgenden Listen $P_k(z)$ (Beachte die verschiedenen Werte für k!):

   (a) $P_4($`CCCC`$)$, $P_8($`CCCC`$)$
   (b) $P_8($`TCTC`$)$, $P_9($`TCTC`$)$
   (c) $P_8($`ACAC`$)$, $P_9($`ACAC`$)$

Aufgabe 4 (Σ-Baum) (4 Punkte)
Gegeben sei die Menge der Worte $W = \{star, sad, salad, art, card, at, scar, cars, cat\}$.

1. Zeichne den kleinsten Σ-Baum und den kleinsten kompakten Σ^+-Baum, welche alle Worte aus W darstellen.
2. Gib jeweils die Menge der Worte $x \in \Sigma^*$ an, für die $node(x)$ definiert ist.
3. Welche Menge words(T) von Worten wird durch die Bäume dargestellt?