

# Übungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, WS 2013/2014

Prof. Dr. Jens Stoye · Linda Sundermann

<http://wiki.techfak.uni-bielefeld.de/gi/Teaching/2013winter/SequenzAnalyse>

## Übungsblatt 7 vom 03.12.2013

Abgabe in einer Woche vor Beginn der Vorlesung.

### Aufgabe 1 (Dot-Plot)

(3 Punkte)

Gegeben seien die zwei Sequenzen  $s_1 = \text{ATGCTTAGC}$  und  $s_2 = \text{GTCTGGAGCTAGTT}$ . Vergleiche die beiden Sequenzen mit einem Dot Plot. Markiere anschließend die Bereiche, die einen Match der Länge 3 oder größer haben. Warum setzt man im Bereich der Sequenzanalyse oft solche Längfilter ein?

### Aufgabe 2 (Der $q$ -gram Index)

(3 Punkte)

Gegeben sei der String  $s_1 = \text{GATCTGATCTATC}$ . Erstelle einen  $q$ -gram Index für  $s_1$  mit  $q = 2$ . Schreibe den Index in einer Tabelle der Form

$q$ -gram	Positionen	Anzahl Vorkommen
-----------	------------	------------------

auf. Es reicht, nur die  $q$ -grams aufzuschreiben, die auch im String vorkommen. Sortiere alle  $q$ -grams lexikographisch (also  $A < C < G < T$ ).

### Aufgabe 3 (Implementierung des $q$ -gram Index)

(5 Punkte)

Schreibe ein kleines Programm, das den  $q$ -gram Index eines Strings berechnet. Verwende bei der Berechnung die geschickte Variante, die im Skript auf Seite 51 unten beschrieben ist. Als Eingabe soll das Programm einen String und ein bestimmtes  $q$  entgegen nehmen und den  $q$ -gram Index ähnlich der Tabelle in Aufgabe 2 auf diesem Zettel ausgeben. Verwende eine Programmiersprache, die mit deinem Tutor abgesprochen ist und sende ihm dein Programm per Email zu.

### Aufgabe 4 (BLAST)

(5 Punkte)

Gegeben sei die Query  $x = \text{PAPALAPAP}$  und  $q = 4$ . Der Score für einen Match sei  $+3$  und für einen Mismatch  $-1$ . Betrachte nun die Nachbarschaft von  $x$ :

- 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  $N_9(x)$ .
- Bestimme die folgenden Listen  $P_k(z)$  (Beachte die verschiedenen Werte für  $k!$ ):
  - $P_4(\text{AAAA}), P_8(\text{AAAA})$
  - $P_8(\text{PAPA}), P_9(\text{PAPA})$
  - $P_8(\text{LALA}), P_9(\text{LALA})$