

# Übungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, WS 2012/2013

Prof. Dr. Jens Stoye · Nina Luhmann · Linda Sundermann

<http://wiki.techfak.uni-bielefeld.de/gi/Teaching/2012winter/SequenzAnalyse>

## Übungsblatt 3 vom 02.11.2012

Abgabe in einer Woche vor Beginn der Vorlesung.

### Aufgabe 1 (Die q-gram-Distanz)

(4 Punkte)

Gegeben seien die beiden Strings  $x = ATTATTATTAT$  und  $y = TAATAATTAA$ , und sei  $q = 3$ .

1. Bestimme die q-gram-Profile für  $x$  und  $y$ .
2. Berechne die q-gram-Distanz zwischen  $x$  und  $y$ .

### Aufgabe 2 (Praktische Berechnung der q-gram-Distanz)

(5 Punkte)

Gegeben seien das Alphabet  $\Sigma = \{A, C, G, T\}$  der Länge  $\sigma = 4$ , die Buchstaben-Codierung  $r_\Sigma(A) = 0$ ,  $r_\Sigma(C) = 1$ ,  $r_\Sigma(G) = 2$  und  $r_\Sigma(T) = 3$  und die ranking-Funktion

$$r(z) = \sum_{i=1}^q r_\Sigma(z[i]) \cdot \sigma^{q-i}.$$

1. Was bedeutet es, wenn ein q-gram-Profil „dicht“ ist?
2. Berechne den Ranking-Wert des q-grams  $s = ATCG$ .
3. Berechne den Ranking-Wert des q-grams  $s' = TCGA$  mit Hilfe des Werts von  $s$ .
4. Gegeben ist der Ranking-Wert 16. Für welches q-gram der Länge 4 steht er?

### Aufgabe 3 (Der de-Bruijn-Graph)

(5 Punkte)

1. Was ist ein Euler-Pfad in einem Graphen  $G = (V, E)$ , wobei  $V$  die Knoten und  $E$  die Kanten des Graphen sind?
2. Warum ist ein de-Bruijn-Graph immer balanciert?
3. Zeichne den de-Bruijn-Graphen für den String  $s = ATCAGTAGTACAGT$  und  $q = 4$ .
4. Gibt es einen String  $t$ , für den gilt:  $d_q(s, t) = 0$  und  $s \neq t$ ? Wenn ja, schreibe eine Möglichkeit für  $t$  auf, wenn nicht; begründe deine Antwort.

### Aufgabe 4 (Die Maximal-Matches-Distanz)

(6 Punkte)

Gegeben seien die beiden Strings  $x = TATAGCCATTAG$  und  $y = GCTATCATA$ .

1. Berechne die Links-Rechts- und die Rechts-Links-Partition von  $x$  in Bezug zu  $y$  und gib jeweils ihren Wert an.
2. Berechne  $|P_{lr}(y, x)|$  und  $|P_{rl}(y, x)|$  und gib jeweils auch die Partitionen an.
3. Wie ist die Maximal-Matches-Distanz von  $x$  zu  $y$  definiert?
4. Zeige an einem von dir ausgedachten Beispiel, dass die Maximal-Matches-Distanz keine Metrik ist.

### Aufgabe 5 (Filter)

(3 Punkte)

1. Was ist mit *Filtern* im Zusammenhang mit Sequenzvergleichen gemeint? Warum verwendet man Filter?
2. Welche Distanzmodelle kann man beim Filtern der Edit-Distanz benutzen? Wie verhalten sich diese zur Edit-Distanz?
3. Worin unterscheidet sich eine Heuristik von einem Filter?