

# Übungen zur Vorlesung Sequenzanalyse I

Universität Bielefeld, Wintersemester 2010/2011  
Dipl.-Inform Peter Husemann · Dr. Roland Wittler

<http://wiki.techfak.uni-bielefeld.de/gi/Teaching/2010winter/SequenzAnalyse>

**Blatt 7 vom 26.11.2010**

**Abgabe in einer Woche vor Beginn der Vorlesung.**

## Aufgabe 1 Filter

(3 Punkte)

1. Beschreibe kurz und in eigenen Worten, was mit *Filtern* im Zusammenhang mit Sequenzvergleichen gemeint ist?
2. Welche zwei wichtigen Eigenschaften muss eine Distanzfunktion haben, damit sie als Filter für die Edit-Distanz taugt?
3. Worin unterscheidet sich eine (Genauigkeits-)Heuristik von einem Filter?

## Aufgabe 2 Modifizierte Maximal-Matches Distanz

(7 Punkte)

Die Maximal-Matches „Distanz“  $\delta$  ist *nicht symmetrisch*. Sie kann allerdings so modifiziert werden, dass sie den Eigenschaften einer Metrik genügt:  $d_{||}(x, y) := \log_2(\delta(x||y) + 1) + \log_2(\delta(y||x) + 1)$ .

Gegeben seien die Sequenzen  $x = \text{ANANAS}$ ,  $y = \text{ANNASHASE}$  und  $z = \text{NANA}$ .

1. Berechne die Distanzen  $d_{||}(x, y)$ ,  $d_{||}(x, z)$  und  $d_{||}(y, z)$ . (Gib die optimalen Partitionen explizit an und unterstreiche die entsprechenden  $b_i$ )
2. Zeige (für das Beispiel der gegebenen Sequenzen für  $x$ ,  $y$  und  $z$ ), dass alle Eigenschaften einer Metrik für  $d_{||}(\cdot, \cdot)$  gelten.
3. Berechne (z. B. mit deinem Programm von Zettel 4, oder per Hand) die Edit-Distanz mit Einheitskosten für alle Paare der Sequenzen  $x$ ,  $y$  und  $z$ .  
Lässt sich anhand der berechneten Distanzen auf einen Zusammenhang zwischen der modifizierten Maximal-Matches Distanz  $d_{||}(\cdot, \cdot)$  und der Edit-Distanz mit Einheitskosten  $d(\cdot, \cdot)$  schließen?

## Aufgabe 3 Äquivalente Score-Funktion

(3 Punkte)

1. Erkläre was es bedeutet, wenn eine Score-Funktion *äquivalent* zu einer Kosten-Funktion ist.
2. Gib zu der folgenden Kosten-Funktion eine äquivalente Score-Funktion an, bei der  $\text{score}(\mathcal{C}) = 1$  sein soll:

$$\begin{aligned}\text{cost}(\mathcal{C}) &= 0 \\ \text{cost}(\mathcal{S}_{a,c}) &= 2 \\ \text{cost}(\mathcal{I}_c) &= \text{cost}(\mathcal{D}_c) = 1.5\end{aligned}$$

Gib auch eine Begründung oder Herleitung für deine Wahl an.