

# Übungen zur Vorlesung Grundlagen der Sequenzanalyse

Universität Bielefeld, WS 2006/07

Dr. Sven Rahmann · Dipl.-Bioinf. Katharina Jahn

<http://gi.cebitec.uni-bielefeld.de/teaching/2006winter/sequenzanalyse/>

**Blatt 9 vom 14.12.2006**

**Abgabe am 21.12.2006 vor der Vorlesung um 8:30 in H3**

**Aufgabe 1** Zeichne den Dotplot für  $x = \text{ATGGGTTA}$  und  $y = \text{ATTAGGGG}$ . Zeichne auch den mit Wortlänge  $q = 2$  gefilterten Dotplot.

**Aufgabe 2** Erstelle den  $q$ -gram-Index für  $y = \text{ATGGGTTA\$CCGTTATC}$  für die übliche ranking-Funktion  $A \mapsto 0, C \mapsto 1, G \mapsto 2, T \mapsto 3$  und  $q = 2$ . (Beachte, dass  $q$ -grams, die das Trennzeichen  $\$$  enthalten, nicht indiziert werden!)

**Aufgabe 3 (zählt als 2 Aufgaben)** Schreibe ein Programm, das (mindestens) 10 000 zufällige Paare von Proteinsequenzen der Länge 500 erstellt und für jedes Paar mit dem Smith-Waterman Algorithmus (Scorematrix: BLOSUM 62, gap-open:  $11 + 1 = 12$ , gap-exend: 1) den Alignment-Score berechnet (das optimale Alignment selbst ist nicht von Interesse). Die Ausgabe des Programms besteht aus den (mindestens) 10 000 Scorewerten. Zeichne ein Histogramm dieser Scores. Zeichne auch ein Histogramm mit logarithmischer y-Achse.