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 8 vom 07.12.2006
Abgabe am 14.12.2006 vor der Vorlesung um 8:30 in H3

Aufgabe 1 Betrachte die BLOSUM62 Scorematrix. Welche Nicht-Diagonaleinträge sind positiv? Wie lassen diese sich möglicherweise durch chemische / physikalische Eigenschaften der betreffenden Aminosäuren erklären?

Aufgabe 2 Die BLAST Webseite (`http://130.14.29.110/BLAST/index.shtml`) ist sehr nützlich. Welche BLAST-Varianten gibt es (bisher) und wofür sind sie jeweils da? Wie kann man beim nucleotide-nucleotide-BLAST die Scores für matches und mismatches verändern?

Aufgabe 3 Wegen des größeren Umfangs zählt diese Aufgabe für zwei Aufgaben: Für zwei Strings s, t ist die "Anzahl der gemeinsamen q-grams auf Diagonale d" definiert als

$$c(d) := \#\left\{\, i \;\mid\; s_{i\ldots(i+q-1)} = t_{(i+d)\ldots(i+d+q-1)} \right\}.$$

Dabei läuft i in Abhängigkeit von d jeweils über alle Werte, für die die genannten Substrings noch vollständig in s und in t enthalten sind. Der FASTA-Score von s und t ist $C(s,t) := \max_d c(d)$; dabei läuft d über alle sinnvollen positiven und negativen Werte. In der Vorlesung wurde Pseudo-Code zur Berechnung von $C(s,t)$ angegeben. Setze diesen Code in ein lauffähiges Programm um. Teste das Programm anhand der Datei `GPCRs.fasta` mit fünf Proteinen, die auf der Vorlesungswebseite vorliegt: Berechne für $q \in \{1,2,3\}$ jeweils alle paarweisen FASTA-Scores.