Abteilung für Genominformatik **Universität Hamburg**
S. Kurtz, F. Meyer und S. Steinbiß **Zentrum für Bioinformatik**

Grundlagen der Sequenzanalyse
Wintersemester 2012/2013
Übungen zur Vorlesung: Ausgabe am 08.01.2013

Punkteverteilung: Aufgabe 11.1: 4 Punkte, Aufgabe 11.2: 3 Punkte, Aufgabe 11.3: 4 Punkte

Abgabe bis zum 14.1.2013.

Aufgabe 11.1 1. Implementieren Sie den Algorithmus zur Berechnung von $score_{\text{fasta}}$ für eine Datenbank-Sequenz u und eine Anfragesequenz w aus der Vorlesung. Dabei soll die Codierung der q-Worte in konstanter Zeit erfolgen (analog zur Codierung der q-Worte bei der Berechnung der q-Wort-Distanz).

2. Berechnen Sie $score_{\text{fasta}}$ für die Sequenzen

   (a) $u = agtgcacacatc$ und $w = atcacacttagc$ für $q \in \{1, 2, 3\}$.

   (b) $u = agcgatag$ und $w = agtgacag$ für $q \in \{2, 3\}$.

Aufgabe 11.2 Sei $\mathcal{A} = \{c, g\}$ eine Teilmenge des DNA-Alphabetes. Gegeben sei für alle $\alpha, \beta \in \mathcal{A}$ (d.h. ohne Indels) eine Scorefunktion

$$\sigma(\alpha \to \beta) = \begin{cases} 1 & \text{falls} \quad \alpha = \beta \\ 0 & \text{sonst} \end{cases}$$

1. Gegeben sei eine Anfragesequenz $w = ggccgc$. Konstruieren Sie die Umgebung $Env_k(w)$ von w gemäß des Blast-Ähnlichkeitsmodells für $q = 4$ und $k = 3$.

2. Charakterisieren Sie die folgenden Umgebungen des Blast-Ähnlichkeitsmodells in Ihren eigenen Worten (für beliebige Alphabete):

   - Die Umgebung für $k = 0$.
   - Die Umgebung für $k > q$.
   - Die Umgebung für $k = 1$.
   - Die Umgebung für $k = q$.

Hinweis: Bei dieser Aufgabe ist es nicht notwendig, ein Programm zu schreiben!

Aufgabe 11.3 Implementieren Sie die im Skript beschriebene Strategie zur Erweiterung von (ungapped) BLAST-hits zu MSPs. Dabei sollen die FASTA-Sequenzdateien (eine pro Anfrage- und Datenbanksequenz), der Parameter q, die Scores für Matches und Mismatches sowie der Parameter X_d über Kommandozeilenoptionen angegeben werden können. Lesen Sie die BLAST-hits aus einer Textdatei oder `stdin` in folgendem Format ein:

```
[3,15]
[259,876]
[33,128]
...
```

Dabei bezeichnet die erste Zahl die Startposition des Hits in der Datenbanksequenz und die zweite Zahl die Startposition des Hits in der Anfragesequenz. Hinweis: Es reicht, jeweils eine Sequenz pro Datei zu betrachten. Sollten mehrere Sequenzen in den FASTA-Dateien vorhanden sein, wird die erste verwendet.

Als Ergebnis geben Sie bitte die MSPs mit ihren Positionen und Sequenzen wie folgt aus:

```
(113-135) (309-331) length 23
TTGGAGGTTATGGAGCATACTAT
ACGCTTATTATGGAGCTCGACAT

(3091-3129) (35-73) length 39
GCGACGAGTTACTGGCCCTGATTTCTCCGCTTCTAATAC
CGGTCGTGACATTATCCCTGATTTTCTCACTACTATTAG

...
```

In STiNE finden Sie zum Testen Ihres Programms ein Archiv mit den Dateien `testdb.fasta` und `testquery.fasta` sowie Dateien mit Hits und Referenzausgaben für verschiedene Werte von X_d (Matchscore: 1, Mismatchscore: -2, $q = 8$).

Die Lösungen zu diesen Aufgaben werden am 15.01.2013 besprochen.