# Übungen zum Sequenzanalyse-Praktikum

Universität Bielefeld, SoSe 2014
Prof. Dr. Jens Stoye · M.Sc. Nina Luhmann · M.Sc. Linda Sundermann
http://wiki.techfak.uni-bielefeld.de/gi/Teaching/2014summer/SequaPrak
praktikum-seqan@CeBiTec.Uni-Bielefeld.DE

**Übungsblatt 7 vom 26.05.2014**
**Abgabe am Donnerstag, den 29.05.2014**

### Aufgabe 1 (Blast-Statistik)

In dieser Aufgabe sollt du mit drei Sequenzen arbeiten, die du in der Datei $Seq.fas$ auf der Vorlesungsseite findest. Nutze NCBI Blast, um herauszufinden, mit welchen Wahrscheinlichkeiten die Sequenzen auftreten. Wenn nichts anderes gesagt wird, verwende *nucleotide blast*, die Nukleotid-Datenbank (*Nucleotide collection(nr/nt)*) und Standardeinstellungen.

1. Beschreibe in eigenen Worten in ein bis zwei Sätzen, was ein E-Value ist.
2. Wie viele Sequenzen befinden sich in der Nukleotid-Datenbank, wie viele in der *Drosophila*-Datenbank? (*Drosophila melanogaster* ist die gemeine Fruchtfliege. Du findest eine Datenbank mit Sequenzen aus *Drosophila*, wenn du auf der NCBI Blast Startseite bei *BLAST Assembled RefSeq Genomes* das Stichwort *Drosophila melanogaster* auswählst.)
3. Suche die erste Sequenz. Welches Problem tritt auf. Warum gibt es dieses Problem?
4. Suche jetzt die zweite Sequenz in der *Drosophila*-Datenbank.
   Bekommst du Ergebnisse? Wie lautet der E-Value für deinen ersten Treffer? Was bedeutet dieser Wert?
5. Suche jetzt die dritte Sequenz. Benutze einmal die Nukleotid-Datenbank und einmal die *Drosophila*-Datenbank. Was für Treffer erhältst du? Wie sehen die E-Values aus? Wie die Identitäten und Scores? Darfst du die Ergebnisse einfach vergleichen?

### Aufgabe 2 (Statistiken von q-Gram-Matches)

In dieser Aufgabe sollst du ein Programm schreiben, das per Kommandozeile aufgerufen werden kann und verschiedene $q$-Gram-Statistiken berechnet. Zum Aufruf gehört die Übergabe von Parametern in folgender Reihenfolge: *Sequenz* $q$ $m$ $n$ $l$. Dabei sei *Sequenz* eine DNA-Sequenz, $q$ die Länge eines $q$-Grams, $m$ die Länge einer ersten und $n$ die Länge einer zweiten Sequenz. Die Mindestlänge eines Treffers sei $l$. Als Ausgabe soll das Programm die Ergebnisse zu allen Teilaufgaben liefern, die du im folgenden programmierst.
Wenn du dein Protokoll erstellst, stelle bitte kurz die Formeln vor, die du zur Berechnung benutzt. Die Theorie dazu hast du im Vortrag kennen gelernt, du kannst dir alles im Skript auf Seite 152 und 153 noch einmal durchlesen.
Für die Bearbeitung der Aufgaben nehmen wir an, dass alle DNA-Basen in $|\Sigma| = \{A, C, G, T\}$ mit der selben Wahrscheinlichkeit $\frac{1}{4}$ auftreten.

1. Schreibe eine Funktion, die die Wahrscheinlichkeit ausgibt, dass deine übergebene DNA-Sequenz genau so auftritt.

Nun brauchst du die Parameter $q$, $m$ und $n$.

2. Berechne den E-Value, also die erwartete Anzahl an exakten $q$-Gram-Matches, für zwei Sequenzen der Länge $m$ und $n$ und für ein $q$.
3. Was ist nun die Wahrscheinlichkeit, mindestens ein exaktes $q$-Gram zu finden?
4. Schreibe zum Schluss noch eine Funktion, die die ungefähre Wahrscheinlichkeit berechnet, einen Treffer der Mindestlänge $l$ zu finden.

Nun sollst du deine Funktionen noch kurz testen.

5. Was ist die Wahrscheinlichkeit, die der Sequenz $TATCGT$ zu Grunde liegt?

6. Berechne den E-Value, die Wahrscheinlichkeit mindestens ein exaktes $q$-Gram zu finden und die Wahrscheinlichkeit, einen Treffer der Länge $l$ zu finden für folgende Parameter:

   - $q = 20$, $m = 100$, $n = 100$, $l = 20$
   - $q = 20$, $m = 100$, $n = 1000000$, $l = 20$
   - $q = 20$, $m = 1000000$, $n = 1000000$, $l = 200$

   Wie lauten diese Werte und was sagen sie dir? Was kannst du über die Beziehung der Längen der Sequenzen, der zu findenden Treffer und die resultierenden Wahrscheinlichkeiten sagen?