Übungen zum Sequenzanalyse-Praktikum

Universität Bielefeld, WS 2020/21
Dr. Roland Wittler · M. Sc. Tizian Schulz
http://gi.cebitec.uni-bielefeld.de/teaching/2020winter/sequaprak
praktikum-seqan@CeBiTec.Uni-Bielefeld.DE

Übungsblatt 7 vom 15./16.12.2020
Abgabe bis Sonntag bzw. Montag, 24:00 Uhr.

Aufgabe 1 (Blast-Statistik)
Benutze für die folgende Aufgabe wieder die Sequenz aus der Datei *unknown_gene.fas*. Sie befindet sich im Verzeichnis `/vol/seqan/Praktikum/DB_BLAST`. Verwende weiterhin für diese Aufgabe *blastx* auf dem NCBI-Server mit Standardeinstellung, wenn es in den Teilaufgaben nicht anders gefordert ist.

1. Beschreibe in eigenen Worten, was ein E-Value ist. Verwende für alle nachfolgenden Berechnungen mit BLAST einen E-Value-Threshold von 10.
2. Vergleiche die letzten 42 Basen aus der Sequenz gegen die *SwissProt*-Datenbank. Welches Problem tritt auf und warum?
3. Vergleiche nun die ersten 210 Basen der Sequenz gegen die *SwissProt*-Datenbank. Was für E-Values bekommst du? Was sagen diese aus? Sind die Treffer signifikant?
4. Finde heraus, was der Unterschied zwischen *Max score* und *Total score* ist. Handelt es sich um einen normalisierten Score? Gib deine Quelle mit an.
5. Vergleiche jetzt die vollständige Sequenz gegen die *landmark*-Datenbank (Model Organismen) und die Datenbank „Non-redundant protein sequences“. Lasse dir dabei bis zu 5000 Treffer anzeigen. Betrachte die Vergleiche gegen die Sequenz mit der Accession Number *NP_001315428.1* bei der *landmark*-Suche und bei der „Non-redundant protein sequences“-Suche näher. Erkläre, warum sich die E-Values unterscheiden, obwohl *Max Score*, *Query cover* und *Ident* identisch sind. Ziehe bei deiner Erklärung die Formel zur Berechnung des E-Values in *Blast* mit ein, du findest sie auf der Seite `http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html`, Formel (3).
6. Bei deiner letzten Suche sind E-Values von 0.0 aufgetreten. Bedeutet dies, dass ein zufälliger Treffer unmöglich ist?

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, nenne die Formeln, die du benutzt und erläutere sie kurz. Die Theorie dazu hast du im Vortrag kennen gelernt, du kannst dir aber alles im Skript auf Seite 154 ff. noch einmal nachlesen.
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, n und l.

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 bestimmtes 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 und die Ergebnisse in geeigneter Form im Protokoll präsentieren.

5. Wie hoch ist die Wahrscheinlichkeit, dass eine zufällige Sequenz der Sequenz `TATTGAT` gleicht?

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

   - $q = 10$, $m = 50$, $n = 50$, $l = 10$
   - $q = 10$, $m = 50$, $n = 10000$, $l = 10$
   - $q = 10$, $m = 1000000$, $n = 1000000$, $l = 200$

   Erkläre, wie die Längen der Sequenzen, die Länge der zu findenden Treffer und die resultierenden Wahrscheinlichkeiten zusammenhängen.