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 5 vom 08./09.12.2020
Abgabe bis Sonntag bzw. Montag, 24:00 Uhr.

Aufgabe 1 (Sequenzlängenhistogramm)
Schreibe ein Programm, das eine Datei im multiple FASTA-Format einliest und entweder ein Histogramm der Sequenzlängen erstellt oder eine Ausgabe hat, aus der mit einem anderen Programm ein Histogramm erstellt werden kann, z. B. mit *gnuplot*, *R* oder einem Tabellenkalkulationsprogramm. Es sollen nur Sequenzenlängen beachtet werden, die kürzer oder gleich 1000 bp sind. Wähle eine Balkenbreite von 100 bp und achte auf eine sinnvolle Achsenbeschriftung.
Lade dir nun die kodierenden Sequenzen des Eintrags mit der *Accession.Version*-Nummer *KR230391.1* im FASTA-Format herunter und wende dein Programm darauf an. Bilde das erhaltene Histogramm bitte in deinem Protokoll ab.

Aufgabe 2 (UNIX-Kommandos)
Lade dir den Genbank-Eintrag mit der *Accession.Version*-Nummer: *KR230391.1* herunter. Verwende UNIX-Kommandos, um die folgenden Aufgaben zu bearbeiten. Gib bei der Beantwortung deiner Fragen auch die Kommandos an, die du verwendet hast.

1. Wie viele Gene werden in dem Eintrag gefunden?
2. Wie viele tRNAs kommen in komplementärer Richtung vor?
3. In dem Features-Bereich kommen verschiedene Attribute, wie z. B. *organism* oder *gene* vor. Zähle alle dieser unterschiedlichen Attribute auf. (Tipp: Die Zeilen mit den Attributen sind die einzigen, in denen ein Gleichheitszeichen vorkommt.)

Aufgabe 3 (BLAST)
Benutze für diese Aufgabe die BLAST-Version auf der NCBI-Website (`https://blast.ncbi.nlm.nih.gov/Blast.cgi`) und die Sequenz aus der Datei `/vol/seqan/Praktikum/DB_BLAST/unknown_gene.fas`.

1. Nutze den Algorithmus *blastn*, um ähnliche Sequenzen zu finden. Welches Gen wurde hier sequenziert und woran erkennst du das?
2. Ist die komplette mRNA des eingegebenen Gens vorhanden? Wie kannst du das herausfinden?
3. Wiederhole die Suche und stelle dabei aber den Parameter „Word Size“ auf 15. Vergleiche das Ergebnis mit dem vorherigen und erkläre, wie der Unterschied zu Stande kommt.
4. Suche jetzt mit *blastx* in der *SwissProt*-Datenbank. Vergleiche die Anzahl an gefundenen Sequenzen mit deinen Treffern in der Nukleotid-Datenbank. Diskutiere dein Ergebnis. (Lass an dieser Stelle die Word Size außer Acht.)
5. Warum kann es Sinn ergeben, eine Nukleotidsequenz erst in eine Aminosäuresequenz zu übersetzen, bevor man sie mit anderen Sequenzen vergleicht?