

Übungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, WS 2025/26

Prof. Dr. Jens Stoye

Leonard Bohnenkämper

Tutorien: Lennart Finke, Sofie Jans

<https://gi.cebitec.uni-bielefeld.de/teaching/2025winter/sa1>

Übungsblatt 5 vom 27.11.2025

Abgabe bis 03.12.2025 bis 9:30 Uhr per Mail an den Tutor/die Tutorin

Aufgabe 1 (Dot-Plot)

(8 Punkte)

In dieser Aufgabe implementieren wir einen einfachen Dotplot. Vervollständige in Übungsrepository¹ das python-Skript `dotplot.py`. Du kannst deinen Zwischenstand immer mit der Beispieldatei `toy.fasta` überprüfen.

1. Vervollständige die Funktion `dot_coordinates`. Für diesen Teil der Aufgabe kannst du `filtersize` ignorieren und stattdessen alle Zeichen-Matches ausgeben.
2. Implementiere nun einen gefilterten Dotplot in `dot_coordinates`. Dazu nutze nun die Variable `filtersize`. (*Tipp: Am einfachsten ist es, zunächst zu betrachten, ob ein entsprechender Substring der Länge k an Positionen i, j gleich ist und dann für jedes paar $(i + k', j + k')$ für $0 \leq k' < k$ einen Punkt hinzuzufügen.*)
3. Natürlich können bei DNA-Sequenzen Ähnlichkeiten sowohl auf dem Vorwärts- oder Rückwärtsstrang liegen. Daher ergibt es Sinn, auch das reverse Komplement einer der Sequenzen zu betrachten. Da dieses bereits in der ersten Woche implementiert wurde, können wir es nun nutzen, um die Koordinaten der Punkte für das Reverskomplement zu erstellen (Variable `filtered_r` in `dotplot`; die Musterlösung ist bereits im Repository eingebaut). Um diese Koordinaten im gleichen System wie die des Vorwärtsstrangs darzustellen, müssen wir das Koordinatensystem übersetzen. Fülle daher die Funktion `reverse_t_coordinates` aus.
4. Plote nun die Sequenzen der Datei `example.fasta` mit einem Filter von 5, 15 und 30 und füge die entstehenden Bilder in deine Abgabe ein. Solltest du den Filter nicht implementiert haben, nutze hier stattdessen die Datei `toy.fasta` (mit Filtergröße 1 - also keinem Filter).
 - (a) Beschreibe die Struktur, die du für Filtergröße 15 siehst. Wie könnte sie entstanden sein?
 - (b) Was sind die Vor- und Nachteile eines größeren oder kleineren Filters?

Bitte wenden!

¹<https://gitlab.ub.uni-bielefeld.de/lbohnenkemper/sqa-ex/-/tree/main/w06>

Aufgabe 2 (Blast-Statistik)

(8 Punkte)

Benutze für die folgende Aufgabe die Sequenz aus der Datei *unknown_gene.fna*. Sie befindet sich im Übungsrepository². Verwende weiterhin für diese Aufgabe *blastx* auf dem NCBI-Server (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>). Nutze die Voreinstellungen, außer für die folgenden Parameter: Setze den E-Value-Threshold auf 10 und die maximalen Zielsequenzen auf 5000. Achte darauf, die Datenbank in jeder Teilaufgabe wie gefordert anzupassen.

1. Beschreibe in eigenen Worten, was ein E-Value ist.
2. Vergleiche die letzten 30 Basen aus der Sequenz gegen die SwissProt-Datenbank. Welches Problem tritt auf und warum?
3. Vergleiche nun die ersten 100 Basen der Sequenz gegen die SwissProt-Datenbank. Was für E-Values bekommst du? Was sagen diese aus? Sind die Treffer signifikant?
4. Vergleiche jetzt die vollständige Sequenz gegen die landmark-Datenbank (Modellorganismen) 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_001375381.1 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. Beziehe bei deiner Erklärung die Formel zur Berechnung des E-Values in *Blast* mit ein. Du findest sie auf der Seite <https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html>, Formel (3).
5. Bei deiner letzten Suche sind E-Values von 0.0 aufgetreten. Bedeutet dies, dass ein zufälliger Treffer unmöglich ist?

Aufgabe 3 (Sum-of-Pairs Score)

(4 Punkte)

Berechne den Sum-Of-Pairs Score des folgenden Alignments. Gib deinen Rechenweg mit an. Nutze dazu das folgende Score-Schema: Matchscore 2, Mismatchscore -1 sowie Gapscore -2 .

$$\begin{pmatrix} T & C & T & A & G \\ T & - & T & - & C \\ - & C & C & - & - \\ - & C & C & A & G \end{pmatrix}$$

²https://gitlab.ub.uni-bielefeld.de/lbohnkaemper/sqa-ex/-/blob/main/w06/unknown_gene.fna