Algorithmen auf Sequenzen

SoSe 2022

Projekt 3

Paarweise Sequenzalignments

Abgabetermin: 22.06.2022

Ansprechpersonen:

- Georg Grünert: georg.gruenert@mni.thm.de
- Sujaya Shrestha: sujaya.shrestha@mni.thm.de
- Cornelia Meckbach: cornelia.meckbach@mni.thm.de

Abgabe: Beschreiben Sie Ihr Vorgehen und Ihre wichtigsten Erkenntnisse in einem Dokument. Fügen Sie dem Dokument auch Ihren kommentierten Quellcode hinzu, sowie wichtige Ausgaben und speichern Sie es als PDF. Vergessen Sie nicht Ihren Namen dazuzuschreiben. Speichern Sie den Quellcode inkl. ReadMe-file (für Ausführungsdetails), sowie das oben beschriebene PDF-Dokument in einem Ordner und geben Sie diesen entweder als .zip oder als .tar.gz ab.

Wichtig: Die Abgabe zu dem Projekt erfolgt in einem „Mini-Bericht“ in dem Sie logisch strukturiert und losgelöst von der Reihenfolge der Aufgaben Ihr Vorgehen und Ihre Ergebnisse präsentieren.

Projekt 3: Paarweise Sequenzalignments

Aufgabe 1: Implementation

Implementieren Sie die beiden unten aufgeführten Algorithmen. Beachten Sie: Ihr Skript sollte zwei Sequenzen (im FASTA-Format), eine Substitutionsmatrix, und eine Gap-Penalty übergeben bekommen und den Score des besten Alignments, sowie das beste Alignment selbst zurückgeben.

a) Needleman-Wunsch-Algorithmus
b) Smith-Waterman-Algorithmus

Aufgabe 2: Testen

Laden Sie sich die Sequenzen des Onkogens KRAS des Menschen (NM_004985.3) und der Maus (NM_021284.7) herunter und speichern Sie sie zusammen in einer FASTA-Datei.

a) Was ist ein Onkogen?
b) Erstellen Sie eine Substitutionsmatrix für Nukleotide, in der Sie Matches mit 2 belohnen und Missmatches mit 0 bewerten.
c) Alignen Sie die beiden Sequenzen mit
   - dem Needleman-Wunsch-Algorithmus
   - dem Smith-Waterman-Algorithmus

   Verwenden sie dabei eine Gap-Penalty von -1.
d) Berechnen Sie für beide Alignments die Sequenzidentität, d.h. wieviel Prozent der Alignmentlänge ist identisch (Matches)?
e) Vergleichen Sie die beiden Alignments hinsichtlich finalem Score, Sequenzidentität und ihrer generellen Struktur.

Aufgabe 3: Signifikanz von Ähnlichkeit

Oft möchte man wissen, ob die Ähnlichkeiten zwischen zwei Sequenzen wirklich einen biologischen Hintergrund hat, oder aber einfach durch Zufall zustande gekommen sein könnte. Um das herauszufinden, werden die Sequenzen mehrfach (~1000 mal) miteinander alignt, dabei bleibt aber nur eine Sequenz in ihrer eigentlichen Form bestehen, die andere wird permutiert, d.h. die Nukleotide dieser Sequenz tauschen zufällig ihre Positionen.

Gegeben sind Ihnen zwei Dateien: „*PTCH2_TRA2.fasta*“ und „*Eyeless_Tra.fasta*". „*PTCH2_TRA2.fasta*“ beinhaltet die Aminosäuresequenzen des menschlichen PTCH2 Genes (NP_003729) und die Aminosäuresequenz von TRA2 von *C. elegans* (NC_003280). „*Eyeless_Tra.fasta*" beinhaltet TRA2 und das ey (eyeless) Protein von *D. melanogaster*.

a) Was sind die Funktionen von TRA2 in *C. elegans* und eyeless in *D. melanogaster*? (Zusatz: Welche Informationen finden Sie über PTCH2 in h.sapiens?)
b) Besorgen Sie sich die BLOSUM62 Matrix und lesen Sie sie in Python ein.
c) Alignen Sie die Sequenzen aus „*PTCH2_TRA2.fasta*“ mit dem Needleman-Wunsch-Algorithmus. Verwenden Sie die BLOSUM62 Matrix und eine Gap-Penalty von -5.
d) Erstellen Sie ~1000 der oben beschriebenen „Zufalls-Alignments“ mit dem Needleman-Wunsch-Algorithmus und speichern Sie deren Scores. (Wenn Ihr PC sehr langsam ist, verwenden sie nur 500 (oder 100) Alignments).
e) Erstellen Sie ein Histogramm, in dem Sie Anzahl der (Zufalls-) Alignments (y-Achse) gegen den Score (x-Achse) auftragen. Kennzeichnen Sie dabei den Score des originalen Alignments.
f) Ermitteln Sie die Wahrscheinlichkeit, dass ein Alignment mit einem Score ≥ dem Score des Alignments mit den originalen Sequenzen (Aufgabe 2.c) gefunden wird.
g) Wiederholen Sie c)-f) für die Sequenzen in „*Eyeless_Tra.fasta*“.
h) Welche Rückschlüsse ziehen Sie bezüglich der Ähnlichkeit der jeweiligen Sequenzen? Begründen Sie Ihre Annahmen.

Aufgabe 4: Einfluss von Alignmentparametern

Die behandelten Algorithmen erzeugen Alignments die optimal sind, bezogen auf die Zielfunktion. Verwenden Sie den Smith-Waterman-Algorithmus um die beiden RNA-Sequenzen in der Datei rRNASeq.fasta zu alignen.

a) Testen Sie unterschiedliche Werte als Gap-Penalties, Match- und Missmatch-Scores.
b) Welche Auffälligkeiten finden Sie im Hinblick auf die Länge und die Sequenzidentität der resultierenden Alignments? Welchen Einfluss haben die einzelnen Parameter?