Algorithmen auf Sequenzen

SoSe 2022

Projekt 2

Maximale Repeats -Lösungen

Abgabetermin: 25.05.2022

Ansprechpersonen:

- Georg Gruenert: 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 2: Sekundärstrukturen von RNAs

MiRNAs sind nicht codierende, funktionelle RNAs, die komplexe-räumliche Strukturen ausbilden können. Ein Beispiel für Sekundärstrukturelemente von RNAs ist in der unteren Grafik gegeben. Ein wichtiges Sekundärstrukturelement ist der sogenannte Stem (-*loop*), in dem das RNA-Molekül eine Doppelhelix-ähnliche Struktur mit sich selbst formt. Sequenzabschnitte, die einen Stem bilden, sind revers komplementär zueinander, d.h. der reverse komplementäre Sequenzabschnitt ist identisch zu dem eigentlichen Abschnitt.

Aufgabe 1: RNAs 

a. Nennen Sie stichpunkthaft einige biologische Funktionen von miRNAs.
b. Gegeben ist eine miRNA mit folgender Sequenz:
   `>NR_029848.1 Homo sapiens microRNA 361 (MIR361), microRNA`
   `GGAGCUUAUCAGAAUCUCCAGGGGUACUUUAUAAUUUCAAAAAGUCCCCCAGGUGUGAUUCUGAUUUGCUUC`
   Die Sequenz finden Sie auch in der Datei *miRNA.fasta*. Lesen Sie die Sequenz ein.

c. Erstellen Sie eine Funktion, die das reverse Komplement für eine RNA-Sequenz berechnet.

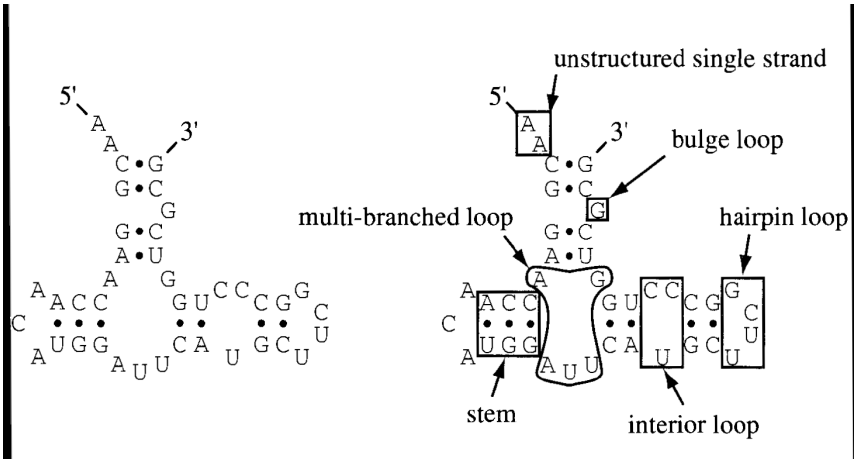

Abbildung 1 RNA-Sekundärstrukturelemente. (Quelle: Biological Sequence Analysis. R. Durbin, Cambridge University Press)

Aufgabe 2: Maximale Repeats

Ein maximales Paar (*maximal pair*) ist in der theoretischen Informatik ein paar einander entsprechender Substrings eines Strings, repräsentiert durch deren Indizes. Maximal dabei bedeutet, dass die Substrings nicht nach außen hin verlängert werden können, um noch ein größeres Match zu erreichen. Ein maximaler Repeat ist wiederrum die eigentliche Sequenz eines maximalen Paares.

Anmerkung: Um *Stem*-Strukturen zu detektieren, sucht man nach einander entsprechenden Bereichen in der eigentlichen RNA-Sequenz und deren reversen Komplement. Da man so formal zwei Sequenzen betrachtet, trifft die obige Definition von maximalen Paaren/Repeats nicht ganz zu, wir übernehmen trotzdem die Bezeichnungen und algorithmisch macht es keinen großen Unterschied.

a) Erstellen Sie einen naiven Algorithmus, der alle maximalen Paare zwischen einer RNA - Sequenz und deren reversen Komplement findet. Geben Sie den Algorithmus in Pseudocode an.
b) Schätzen Sie die obere asymptotisch Laufzeit Ihres Algorithmus an. Begründen Sie Ihre Abschätzung.
c) Implementieren Sie Ihren Algorithmus in Python und ermitteln Sie für die in Aufgabe 1 gegebene Sequenz die maximalen Paare inklusive Repeats, deren Länge >3 ist. (Rechnen Sie die Indizes der reversen Sequenz auf die der ursprünglichen Sequenz zurück).

Aufgabe 3: RNA Sekundärstrukturen

Um die Sekundärstruktur von RNAs genauer zu ermitteln, gibt es ergänzende Methoden, die aus der Information von möglichen Basenpaarungen, die finale Sekundärstruktur berechnen, deren freie Energie minimal ist. Ein solches Programm ist zum Beispiel RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi).

a) Berechnen Sie die Sekundärstruktur der gegebenen RNA mit RNAfold.
b) Selektieren Sie aus Ihren Repeats aus Aufgabe 2c) Stems, deren paarende Sequenzabschnitte nicht überlappen und zueinander einen Mindestabstand von 3 bp in der eigentlichen miRNA-Sequenz haben.

AUGCUUAAAGGGUUUCCAA

✓ Abstand beträgt hier genau 3 bps & keine Überlappung

✘ Abstandsbetrag zwischen Start und Stop >3 bp, aber Überlappung (Beispiel: ATATATATATAT)

c) Betrachten Sie die RNAfold- Struktur für die „minimum free energy prediction“. Vergleichen Sie Ihre detektierten Stems von RNAfold mit Ihren maximalen Pairs. (Hinweis: Das Ct Format oder VIEW IN FORNA kann Ihnen vielleicht helfen)
d) Theoretisch gesehen: Können alle Ihre detektieren Stems gleichzeitig in einer möglichen Sekundärstruktur vorkommen?

Aufgabe 4: Längere Sequenzen

Die Identifikation von nicht-codierenden, funktionalen RNAs im Genom ist eine große Herausforderung. RNA-Gene sind nicht zwangsläufig so aufgebaut wie Proteinkodierende Gene oder aber, die RNAs selbst sind in proteincodierenden Genen integriert. Eine Möglichkeit der Detektion von nicht-kodierenden RNAs ist die Suche nach Stem-Strukturen.

a) Messen Sie die Zeit, die Ihr Algorithmus benötigt, um Stems in der gegebenen miRNA zu detektieren.
b) Berechnen Sie die Zeit, die Ihr Algorithmus benötigt, um Stems in der menschlichen 12S rRNA zu detektieren (*rRNA_12S_human.fasta*). Verwenden Sie Ihre Laufzeitabschätzung aus Aufgabe 2b). (Hinweis: Falls Sie doch einen recht effizienten Algorithmus erstellt haben, probieren Sie es auch gerne aus und vergleichen Sie die tatsächliche Laufzeit mit Ihren Berechnungen).
c) Angenommen, Sie würden das gesamte menschliche Genom mit Ihrem Algorithmus nach maximalen Paaren absuchen, wie lange würde das ungefähr dauern?