Prof. Dr. Knut Reinert
René Rahn
Jakob Schulze
Kathleen Gallo
Denise Thiel

Institut für Informatik
AG Algorithmische Bioinformatik

Freie Universität Berlin

Algorithmen und Datenstrukturen in der Bioinformatik siebtes Übungsblatt WS 14/15

Abgabe Donnerstag 12:00

Niveau I

Aufgabe 1: Blast

Gegeben ist ein Pattern P = HILAM sowie eine Datenbank D = ILHIVAMVHIIM

a) Finden Sie alle k-mere, die zu einem beliebigen k-mer aus P einen Score von mindestens 13 aufweisen. Die Fenstergrösse ist auf k = 3 festgelegt. Benutzen Sie zum Scoring die BLOSUM62-Matrix.

b) Erstellen Sie aus der Liste der k-mere einen Aho-Corasick-Automaten. Markieren Sie dabei die finalen Knoten und zeichnen Sie alle nicht-trivialen Supply-Links ein, also jene die nicht zur Wurzel zurückgehen.

c) Suchen Sie die seeds, also alle Vorkommen eines der Pattern in der Datenbank D mit dem Aho-Corasick Algorithmus. Als Ergebnis genügt es die Stellen zu markieren.

d) Erweitern sie die gefundenen seeds nach rechts oder links, sofern dies ohne gaps möglich ist. Welcher Bereich liefert das beste lokale Alignment ohne gaps?

e) Wenn man Blast von NCBI mit dem gegebenen Pattern, der Datenbank und den Parametern anwenden will, werden keine Treffer gefunden. Warum? (Tipp: E-Wert)

Aufgabe 2: q-gram-Lemma

Beweisen Sie das q-gram-Lemma: Zwei Strings der Länge w, die sich in höchstens k Stellen unterscheiden, haben mindestens w + 1 ? (k + 1)q gemeinsame q-gramme.

Aufgabe 3: Speicherplatz

Berechnen Sie den Speicherverbrauch für ein q-Gramm Index über DNA in C++. Nehmen Sie q=4 an und ein 32-bit Betriebssystem. Damit Sie eine Intuition für den Speicherverbrauch bekommen, geben Sie den verbrauchten Speicher in KB (Kilobyte) an.

Niveau II

a) Wie wirken sich die Parameter w (Seedlänge), T (similarity threshold) und C auf die Laufzeit einer Blast-Suche aus? Wie verändern sie die Anzahl und Qualität der Funde?

b) Gegeben ist eine Datenbank mit 10.000 zufälligen Nukleotid-Sequenzen der Länge 1.000. Die Nukleotide sind gleichverteilt und unabhängig voneinander. Sie haben eine queue von 100 Nukleotiden. Wie oft erwarten Sie im Schnitt die queue in der DB?