Prof. Dr. Knut Reinert
Rene Rahn
Kathrin Trappe
Kathleen Gallo
Thomas Krannich

Institut für Informatik
AG Algorithmische Bioinformatik

Algorithmen und Datenstrukturen in der Bioinformatik
Siebtes Übungsblatt WS 12/13

Abgabe Freitag, 07.12., 15:00 Uhr

Name: Übungsgruppe: A☐ B☐ C☐

Matrikelnummer:

Niveau I

Aufgabe 1: BLAST

Gegeben ist ein Pattern $P =$ `HILAM` sowie eine Datenbank $D =$ `ILHIVAVHIIM`

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?

Aufgabe 2: Aho-Corasick Automaten und Pseudocodes

Mit Hilfe des *Aho-Corasick Automaten* wird in BLAST das multiple Stringmatching-Problem gelöst. Der Automat basiert auf einem *Trie*, der durch sogenannte *Supply-Links* erweitert wird. Der Supply-Link eines Knotens v zeigt auf das längste Suffix des Textes

a) Machen Sie sich mit den Pseudocodes zur Erstellung des Tries (Kapitel 5.7, Seite 5004) und zur Erweiterung des Tries durch Supply-Links (Kapitel 5.8, Seite 5007) vertraut. Erläutern Sie, was die Symbole δ und S bedeuten.

b) Bei der Suche kann es vorkommen, dass wenn ein Suchwort p gefunden wird, gleichzeitig auch ein anderes Suchwort s enthalten ist. In diesem Fall ist s ein Suffix von p. Um nun beim Finden von p auch einen Treffer für s auszugeben, muss sichergestellt werden, dass s in der Menge der finalen Suchwörter von p enthalten ist, also dass $s \in F(p)$. Wie wird das im Algorithmus umgesetzt?

Aufgabe 3: Bioinformatik Praxis

Sucht man eine Sequenz und es gibt bereits eine Blast-Applikation dafür, übergibt man dieser eine *queue*, eine Datenbank oder ihren Verweis sowie Parameter. Eine solche Onlne-Applikation wird von NCBI bereit gestellt. Gehen Sie auf diese Webseite, wählen Sie protein blast aus, lesen Sie sich alle Teilaufgaben zu 3. im Voraus durch und gehen Sie wie folgt vor:

a) Editieren Sie folgende Parameter: Choose Search Set/Database auf UniProtKB/Swiss-Prot(swissprot), Programm Selection auf blastp (sollte default sein), öffnen Sie Algorithm parameters und setten Sie General Parameters/Max target sequences auf 10, /Expect threshold auf 14 und /Word size auf 3 (sollte default sein) und /Scoring parameters/Matrix auf BLOSSUM80.

b) Übergeben Sie ganz oben in dem Kasten 'Enter Query Sequence' die random-Sequenz *PQLIIHCFDMYQVIH* und starten Sie die BLAST-Suche mit dem Button *BLAST*.

c) Notieren Sie, wie lange BLAST für die Suche der queue in der DB UniProt/Swissprot benötigt. Notieren Sie zudem die Top3 Hits der Suche (Name, Gene ID, query coverage und Score[bits]).

Aufgabe 4: Review

Es gibt diese Woche keine Aufgabe des Niveau 2. Bereiten Sie sich auf das Review am Mi., den 5.12. vor. Es können alle Themen der Vorlesung bis (inkl.) FastA drankommen.