Übungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, WS 2015/2016
Prof. Dr. Jens Stoye · M.Sc. Linda Sundermann

http://wiki.techfak.uni-bielefeld.de/gi/Teaching/2015winter/SequenzAnalyse

Übungsblatt 5 vom 01.12.2015
Abgabe in einer Woche vor Beginn der Vorlesung.

Aufgabe 1 (Paarweises Alignment mit linearem Speicherbedarf) (9 Punkte)

1. Fertige eine Skizze an, die zeigt, wie mit Hilfe des Hirschberg Algorithmus ein paarweises Alignment in linearem Platz berechnet werden kann.
2. Berechne nun für die Strings $x =$`ACATG` und $y =$`TACCG` ein Alignment mit linearem Speicherbedarf, wobei $|x| = m = 5$, INDEL $= -1$, MATCH $= 2$ und MISMATCH $= -1$. Führe dabei folgenden Schritte aus:
   (a) Berechne zuerst die *forward* Edit-Matrix von x und y.
   (b) Berechnen nun die *backward* Edit-Matrix von x und y.
   (c) Durch welches Feld in der Zeile mit dem Index $z = \lfloor \frac{m}{2} \rfloor$ läuft das optimale Alignment? (Die Indizierung beginne bei 0.) Begünde deine Antwort.
   (d) Teile nun die Strings, so wie in Aufgabenteil 2c ermittelt, auf und berechne für beide Teile erneut, durch welches Feld in der Zeile $z' = \lfloor \frac{m'}{2} \rfloor$ und $z'' = \lfloor \frac{m''}{2} \rfloor$ das optimale Alignment läuft, wobei m' und m'' die Längen des Präfixes und Suffixes von x sind. Berechne diesmal die *forward* und die *backward* Edit-Matrix nicht komplett, sondern nur jeweils bis zur Zeile z' und z''.
   (e) Teile das Präfix und das Suffix von x noch einmal auf, so wie in Aufgabenteil 2d berechnet. Da die Teilstrings nun sehr kurz sind, berechne die Edit-Matrix und das Teilalignment auf normale Weise.
   (f) Füge nun das komplette paarweise Alignment aus den Alignments der kleinen Edit-Matrizen zusammen.

Aufgabe 2 (Der *q*-gram Index) (3 Punkte)

Gegeben sei der String $s_1 =$ `TACTTGCATGCTAT`. Erstelle einen q-gram Index für s_1 mit $q = 2$. Schreibe den Index in einer Tabelle folgender Form auf:

q-gram	Positionen	Anzahl Vorkommen

Es reicht, nur die q-grams aufzuschreiben, die auch im String vorkommen. Sortiere alle q-grams lexikographisch (also $A < C < G < T$).

Aufgabe 3 (Suffixbäume) (4 Punkte)

1. Beschreibe zwei Anwendungen aus der bioinformatischen Praxis, für die man einen Suffixbaum verwenden kann.
2. Gegeben ist der String $s\$ =$ `TATTAATAAA$`, wobei $\$ < A < T$.
   (a) Zeichne den Suffixbaum für s, sortiere dabei alle ausgehenden Kanten lexikographisch.
   (b) Beschrifte die Blätter mit dem Start-Index des zugehörigen Suffixes in s. Die Indizierung beginnt bei 1.
   (c) Beschrifte jeden Knoten mit der Anzahl der unter ihm liegenden Blätter.

Bitte wenden!

Aufgabe 4 (Kürzester eindeutiger Substring) (4 Punkte)

1. Gib eine Definition des Problems des kürzesten eindeutigen Substrings in eigenen Worten an.
2. Beschreibe eine mögliche Anwendung des Problems.
3. Gib einen Linearzeit-Algorithmus an, mit dem man einen kürzesten eindeutigen Substring eines Strings s finden kann, wenn der Suffixbaum von $s\$$ gegeben ist.