Übungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, WS 2013/2014
Prof. Dr. Jens Stoye · Linda Sundermann

http://wiki.techfak.uni-bielefeld.de/gi/Teaching/2013winter/SequenzAnalyse

Übungsblatt 9 vom 19.12.2013
Abgabe Dienstag, den 07.01.14, vor Beginn der Vorlesung

Aufgabe 1 (Links-Rechts-Partition) (4 Punkte)

Die Links-Rechts-Partition $P_{lr}(s,t)$ eines Strings s bezüglich eines Strings t (siehe Abschnitt 3.8 im Skript) kann mit Hilfe eines Suffixbaumes effizient berechnet werden.

1. Überlege dir einen Algorithmus, der die Links-Rechts-Partition $P_{lr}(s,t)$ in linearer Zeit berechnet. Gib die einzelnen Schritte deines Algorithmus explizit und verständlich an.
2. Verwende diesen Algorithmus, um $P_{lr}(s,t)$ für $s =$ `BAALUBALUU` und $t =$ `LUBBALUBUUAZ` zu berechnen.

Aufgabe 2 (MUMs) (4 Punkte)

1. Wofür können *Maximal Unique Matches* verwendet werden?
2. Gib einen Algorithmus an, der die MUMs der Länge l oder größer findet. In welcher Komplexitätsklasse liegt der Algorithmus und warum?

Aufgabe 3 (Maximale Repeats) (5 Punkte)

1. Finde alle maximalen Repeats in $s =$ `AGTGCAATGTGCAT` unter Verwendung des im Skript geschilderten Algorithmus (Abschnitt 6.6.4). Beschreibe dein Vorgehen beim Annotieren des Suffixbaumes.
2. **Satz:** In jedem String der Länge n gibt es höchstens n maximale Repeats.

   Argumentiere unter Berücksichtigung des Suffixbaumes, warum die Aussage korrekt ist. Bedenke: Es stimmt nicht, dass an jeder Position nur ein maximaler Repeat beginnen oder enden kann.

Aufgabe 4 (Suffixarray) (3 Punkte)

Gegeben ist der String $s =$ `ATGTCACTGTACA`. Gib das Suffixarray `pos`, sein Inverses `rank` und das `lcp`-Array von $s\$$ an. Beachte $\$ < A < C < G < T$ und Indizierung beginnend mit 0.

Bitte wenden.

***** Weihnachtsbonuszettel *****

Aufgabe 5 (Kürzester eindeutiger Teilstring) (4 Punkte)

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

Aufgabe 6 (Verallgemeinerter Suffixbaum) (5 Punkte)

Gegeben seien die Strings s=`TAAT` und t=`GATA`.

1. Zeiche den generalisierten Suffixbaum von s und t (mit $\# < \$ < A < G$).
2. Finde den längsten gemeinsamen Substring von s und t und gib seine Vorkommen an.
3. Formuliere allgemein, wie man längste gemeinsame Substrings zweier Strings im generalisierten Suffixbaum finden kann.
4. Nutze deine Überlegungen, um das längste palindromische Teilwort in $x =$ `OTTOTO` zu finden.

Aufgabe 7 (Suffixarray Interpretation) (2 Punkte)

Inwiefern ist das Suffixarray `pos` ein q-gram Index für alle $q \geq 1$ gleichzeitig?

Aufgabe 8 (Manber-Myers Algorithmus) (4 Punkte)

Betrachte das Beispiel im Skript auf den Seiten 74 und 75.

1. Wie viele Phasen braucht man für einen String der Länge 24?
2. Führe den Manber-Myer Algorithmus schrittweise für den String `GCGTGGCTCTCGC` aus.

Wir wünschen euch schöne Weihnachten und einen guten Rutsch ins neue Jahr! :)