Übungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, SS 2020
Prof. Dr. Jens Stoye · Dr. Marília D. V. Braga · Leonie R. Brockmann · Rebecca K. Pfeil

https://gi.cebitec.uni-bielefeld.de/teaching/2020summer/sa

Übungsblatt 8 vom 11.6.2020
Abgabe am 18.6.2020 bis 12:00 Uhr (mittags)

Aufgabe 1 (Links-Rechts-Partition) (4 Punkte)
Die Links-Rechts-Partition $P_{lr}(s,t)$ einer Sequenz s bezüglich einer Sequenz t (siehe Abschnitt 3.8 im Skript) kann mit Hilfe eines Suffixbaums effizient berechnet werden.

1. Überlege dir einen Algorithmus, der die Links-Rechts-Partition $P_{lr}(s,t)$ in linearer Zeit berechnet. (Hinweis: Verwende den Suffixbaum von t\$.) Gib die einzelnen Schritte deines Algorithmus' explizit und verständlich an.
2. Verwende diesen Algorithmus, um $P_{lr}(s,t)$ für $s =$ `ABCCDBAABA` und $t =$ `BAABC` zu berechnen. Gib alle Zwischenschritte an.

Aufgabe 2 (Suffixbaum-Anwendungen) (8 Punkte)
Gegeben sei die Sequenz $s =$ `CATAGCATATAG`.

1. Berechne den Suffixbaum von s\$. Sortiere dabei die von einem Knoten ausgehenden Kanten lexikographisch (mit \$ < `A` < `C` < `G` < `T`).
2. Kürzeste eindeutige Substrings:
   (a) Finde alle eindeutigen Substrings in s mit Hilfe des Suffixbaums von s\$. Welche sind die kürzesten eindeutigen Substrings?
   (b) Beschreibe eine mögliche Anwendung für kürzeste eindeutige Substrings.
3. Maximale Repeats:
   (a) Finde alle maximalen Repeats in s mit Hilfe des Suffixbaums von s\$. Gib alle Zwischenschritte des verwendeten Algorithmus' an.
   (b) **Satz:** In jedem String der Länge n gibt es höchstens n Teilworte, die maximale Repeats sind.
   Argumentiere unter Berücksichtigung des Suffixbaums, warum diese Aussage korrekt ist.

Aufgabe 3 (Anwendungen des generalisierten Suffixbaums) (8 Punkte)
Gegeben seien die Sequenzen s=`TTCATAGTTC` und t=`CATCATATAG`.

1. Berechne den generalisierten Suffixbaum T von s und t. Sortiere dabei die von einem Knoten ausgehenden Kanten lexikographisch (mit # < \$ < `A` < `C` < `G` < `T`).
2. Maximale gemeinsame Substrings (MEMs) mit Mindestlänge ℓ zweier Sequenzen s und t sind maximale Repeats in s#t\$, von denen je ein Vorkommen in s und ein Vorkommen in t liegt, und die mindestens ℓ Zeichen lang sind.
   Verwende den in Aufgabenteil 1 erstellten verallgemeinerten Suffixbaum T, um alle MEMs mit Mindestlänge $\ell = 2$ von s und t zu finden. Gehe dafür wie folgt vor:
   (a) Finde Kandidaten: Markiere jeden inneren Knoten v von T, für den gilt:
      - Die String-Tiefe von v muss mindestens $\ell = 2$ betragen.
      - Der Teilbaum unter v enthält mindestens ein Blatt, das ein in s beginnendes Suffix repräsentiert und mindestens ein Blatt, das ein in t beginnendes Suffix repräsentiert.
   (b) Gib für jeden dieser Kandidaten v alle Vorkommen (Start- und Endpositionen) von string(v) in s und in t an.
   (c) Welche Paare dieser Vorkommen sind maximal und repräsentieren daher MEMs von s und t?
3. Welche MEMs sind auch MUMs von s und t mit Mindestlänge $\ell = 2$?
4. Modifiziere den Algorithmus von Aufgabenteil 2, um direkt MUMs zu finden, ohne den Umweg über MEMs. Welche inneren Knoten sind Kandidaten für MUMs der Mindestlänge 2?