Übungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, WS 2024
Prof. Dr. Jens Stoye · Leonard Bohnenkämper
https://gi.cebitec.uni-bielefeld.de/teaching/2024winter/sa

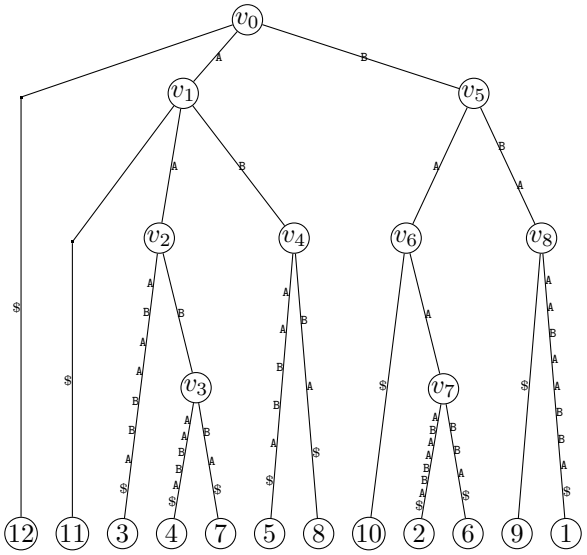
Übungsblatt 9 vom 12.12.2024
Abgabe am 19.12.2024 bis 10:00 Uhr (morgens)

Aufgabe 1 (Maximale Repeats) (6 Punkte)

Betrachte für diese Aufgabe den Suffixbaum für $s\$ =$ `123456789.12` / `BBAAABAABBA$`.

1. Finde alle maximalen Repeats mit Mindestlänge 2 in s mit Hilfe des Suffixbaums (s. unten) von $s\$$. Gib alle Zwischenschritte des verwendeten Algorithmus' an.
2. **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 2 (Anwendungen des verallgemeinerten Suffixbaums) (8 Punkte)

Gegeben seien die Sequenzen $s =$ `12345678` / `TCATCGGA` und $t =$ `.1234567` / `CATGGCAT`.

1. Berechne den verallgemeinerten Suffixbaum von s und t, d.h. den Suffixbaum T von $s\#t\$$. 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.

(Bitte wenden!)

(b) Gib für jeden dieser Kandidaten v alle Vorkommen (Start- und Endpositionen) von $\text{STRING}(v)$ 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 von T sind Kandidaten für MUMs der Mindestlänge 2?

Aufgabe 3 (Inverted Repeats) (6 Punkte)

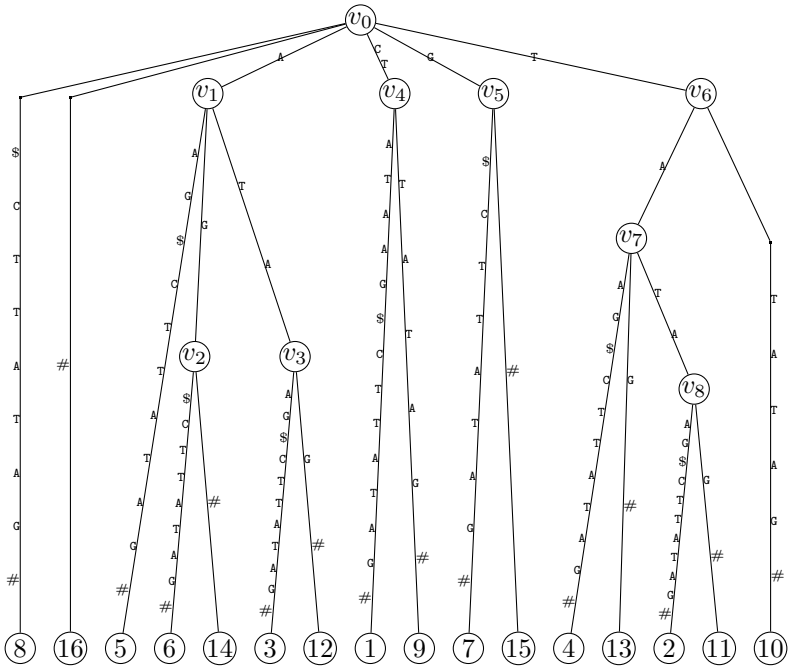
Inverted Repeats[1] sind Sequenzmuster, die in verschiedenen biologischen Bereichen relevant sind. Zum Beispiel spielen sie eine Rolle bei der Sekundärstruktur von RNA, Transposons und sind assoziiert mit großen strukturellen Variationen. Formal gesehen ist ein Inverted Repeat eines Strings s ein Paar von Sequenzabschnitten, von dem eines das reverse Komplement eines anderen ist, also $s = s_0 t s_1 u s_2$, wobei u das reverse Komplement von v ist (und umgekehrt) und s_0, s_1, s_2 beliebige strings sind.

Als Beispiel ist in $s = \begin{matrix}\texttt{123456789.1}\\ \texttt{ATGACACACAT}\end{matrix}$ $((1,3),(9,11))$ (also `ATG` und `CAT`) ein inverted Repeat. Hier sind wir an maximalen inverted Repeats interessiert, also solchen, die sich nicht nach links oder rechts erweitern lassen.

1. Entwirf einen Algorithmus, um maximale inverted Repeats zu finden.
(*Hinweis: Du kannst einen Algorithmus aus der Vorlesung modifizieren.*)
(*Hinweis 2: Wenn du nicht weiter weißt, kannst du auch versuchen, die nächste Aufgabe ohne den Algorithmus zu lösen.*)

2. Nutze deinen Algorithmus, um in `CTATAAG` maximale inverted Repeats der Länge 2 und höher zu finden. Einen dazu möglicherweise hilfreichen Suffixbaum findest du unten.

3. Welches Problem stellst du fest? Wie könnte man es beheben?



[1] `https://en.wikipedia.org/wiki/Inverted_repeat`