Übungen zur Vorlesung Sequenzanalyse

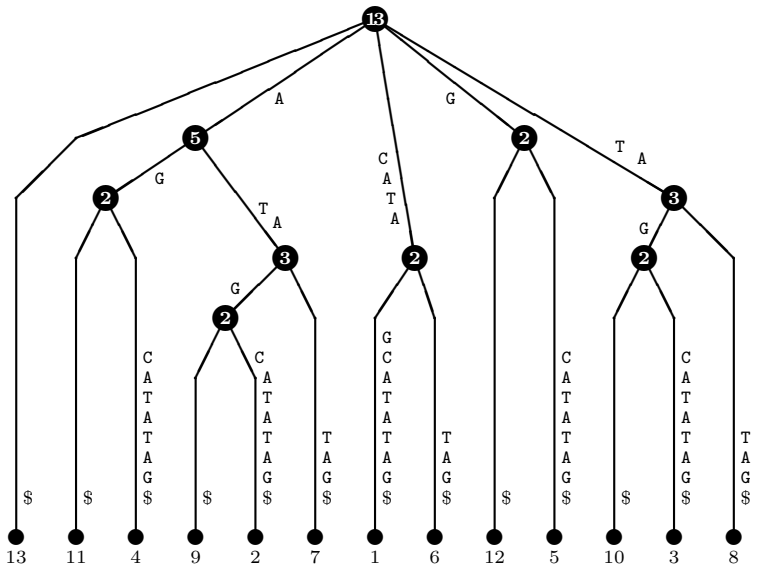
Universität Bielefeld, SS 2020
Prof. Dr. Jens Stoye · Dr. Marília D. V. Braga · Leonie R. Brockmann · Rebecca K. Pfeil

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

Übungsblatt 9 vom 18.6.2020
Abgabe am 25.6.2020 bis 12:00 Uhr (mittags)

Aufgabe 1 (Suffixbaum und lcp-Array) (2 Punkte)

Gegeben seien die Sequenz $s =$ `CATAGCATATAG` und der Suffixbaum von s\$ mit lexikographisch sortierten Kanten:



Berechne mit Hilfe dieses Baums das `lcp`-Array von s\$, ohne `rank` oder `pos` explizit auszurechnen.

Aufgabe 2 (Manber-Myers-Algorithmus) (9 Punkte)

1. Gegeben sei die Sequenz $s =$ `TCATAGTCGCATCATATAG`.
   (a) Erstelle das Suffix-Array `pos`(s\$) unter Verwendung des Manber-Myers-Algorithmus. Gib das Zwischenergebnis nach jeder Phase des Algorithmus' an.
   (b) Berechne die Arrays `rank`(s\$) und `lcp`(s\$).
   (c) Verwende Binärsuche im Suffix-Array `pos`(s\$), um alle Vorkommen des Musters $p =$ `ATA` in s zu finden. Gib alle Zwischenscritte an.
2. Wie viele Phasen benötigt (im *worst case*) der Manber-Myers-Algorithmus für einen String der Länge n? Gib eine Beispielsequenz an, bei der diese Anzahl Phasen erreicht wird.

Aufgabe 3 (Burrows-Wheeler-Transformation) (9 Punkte)

1. Gegeben sei der String $t =$ `TTAGGCTTAGCTTGGC`.
   (a) Berechne die Burrows-Wheeler-Transformierte $s = \text{BWT}(t\$)$. Gib alle Zwischenschritte an.
   (b) Demonstriere den String-Matching-Algorithmus unter Verwendung der Sequenz $s = \text{BWT}(t\$)$ beispielhaft an der Suche des Musters $p =$ `AGCTT` in t.
   (c) Schreibe $r = \text{RLE}(s)$ als komprimierten String mit Hilfe von *run-length encoding* auf. Fasse dabei nur Buchstaben zusammen, die mindestens dreimal hintereinander vorkommen.
2. Gegeben sei der String $r =$ `CT5GA$A4TCC`.
   (a) Dekomprimiere r in $s = \text{RLE}^{-1}(r)$.
   (b) Rekonstruiere den String $t\$ = \text{BWT}^{-1}(s)$. Gib alle Zwischenschritte an.