Übungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, SS 2019
Dr. Daniel Dörr · Michel T. Henrichs

https://gi.cebitec.uni-bielefeld.de/teaching/2019summer/sa

Übungsblatt 10 vom 10.6.2019
Abgabe am 17.6.2019 bis 12:00 Uhr (mittags)

Aufgabe 1 (Manber-Myers Algorithmus) (4 Punkte)

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

Aufgabe 2 (Suffixbaum und lcp-Array) (3 Punkte)
Gegeben den Suffixbaum zu einem String $s\$$ mit lexikographisch sortierten Kanten. Wie kann man von diesem Baum das `lcp`-Array von $s\$$ ablesen, ohne `rank` oder `pos` explizit ausrechnen zu müssen? Gib so das `lcp`-Array zum Baum in Abbildung 7.2 auf Seite 65 des Skripts an.

Aufgabe 3 (Mustersuche im Suffixarray) (3 Punkte)
Gegeben sei ein Text der Länge n sowie das zugehörige Suffixarray `pos`. Auch ohne die Arrays `rank` und `lcp` zu berechnen, können alle Vorkommen eines Musters der Länge m im Text bestimmt werden. Beschreibe ein Vorgehen, dass das in einer besseren Laufzeit als $\mathcal{O}(mn)$ schafft und analysiere dessen Laufzeit.

Aufgabe 4 (Burrows-Wheeler Transformation) (4 Punkte)
Gegeben sei der String $t =$ `TGATGCATCATGCAT$`.

1. Berechne die Burrows-Wheeler Transformierte $bwt(t)$ für t.
2. Schreibe $rle(bwt(t))$ als komprimierten String mit Hilfe von run-length encoding auf. Fasse dabei nur Buchstaben zusammen, die mindestens dreimal hintereinander vorkommen.
3. Demonstriere den String Matching Algorithmus unter Verwendung der $bwt(t)$ beispielhaft an der Suche des Musters $p =$ `ATGC`.