Präsenzübungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, SS 2023
Prof. Dr. Jens Stoye · Tizian Schulz

https://gi.cebitec.uni-bielefeld.de/teaching/2023summer/sa

Präsenzübungsblatt 9, vom 6.6.2023

Aufgabe 1 (Suffixbaum-Suffixarray)

Gegeben ist der String $s = \texttt{CATATAGTAGCGTCATAGT}$.

1. Zeichne den Suffixbaum für $s\$$, sortiere dabei alle ausgehenden Kanten lexikographisch (wobei $\$ < \texttt{A} < \texttt{C} < \texttt{G} < \texttt{T}$).
2. Beschrifte die Blätter mit dem Start-Index des zugehörigen Suffixes in $s\$$. Die Indizierung beginnt bei 0.
3. Berechne mit Hilfe dieses Baums das `lcp`-Array und das Suffix-Array `pos` von $s\$$.
4. Verwende Binärsuche im Suffix-Array $\texttt{pos}(s\$)$, um alle Vorkommen der folgenden *Patterns* in s zu finden und gib ihre Startpositionen an:
   - $p_1 = \texttt{CAT}$
   - $p_2 = \texttt{TAG}$
   - $p_3 = \texttt{ATA}$

Aufgabe 2 (Exakte Textsuche in BW-transformierten Texten)
Gegeben sei der Text $t = \texttt{TACGCACGGCAAGCCCGTGC\$}$, seine Borrows-Wheeler-Transformation $bwt(t) = \texttt{CCTCAGGGGCAACTGCACC\$G}$ und sein Suffix Array

$$pos(t) = [20, 10, 1, 5, 11, 19, 9, 4, 13, 14, 2, 6, 15, 18, 8, 3, 12, 7, 16, 0, 17]\,,$$

wobei hier der Text t mit Startposition 0 indiziert wurde.

1. Konstruiere F, die lexikographische Sortierung von t.
2. Suche die folgenden 10 Muster in $bwt(t)$ und gib ihre Startpositionen mithilfe von $pos(t)$ an:
   - `GCA`
   - `CGTG`
   - `GCAC`
   - `TGC`
   - `CGCAC`
   - `GC`
   - `AAGC`
   - `CGCA`
   - `ACGC`
   - `CA`