9. Übung zur Vorlesung
"Einführung in die Bioinformatik I, 1. Teil"

Wintersemester 2016/2017

Prof. Sebastian Böcker, Marcus Ludwig, Emanuel Barth, Maximilian Collatz

Ausgabe: 21. Dezember 2016,
Abgabe: 04. Januar 2017 zu Beginn der Übung

Suffixbäume

Aufgabe 1 (5 Punkte):

1. Bestimmen Sie für den String `abrakadabra` und `abaababab` den jeweiligen komprimierten Suffixbaum. Ist der komprimierte Suffixbaum nicht eindeutig, geben Sie bitte alle an.
   Hinweis: Es ist ausreichend, eine alternative Beschriftung in anderer Farbe anzubringen oder separat alternative Beschriftungen zu kennzeichnen.
2. Bestimmen Sie für die Strings `AGTT`, `TAC` und `GTAG` den komprimierten generalisierten[1] Suffixbaum.

Aufgabe 2 (5 Punkte): Wie kann man mit Hilfe eines Suffixbaumes in einem Text der Länge n den längsten Teilstring finden, der in diesem Text auch rückwärts vorkommt? In welcher asymptotischen Laufzeit ist dies möglich?

Aufgabe 3 (5 Punkte): Geben Sie einen möglichst schnellen Algorithmus an, der mit Hilfe eines Suffixbaumes für zwei Strings S_1, S_2 alle gemeinsamen Teilstrings berechnet, die länger als eine gegebene Länge l sind. Begründen Sie Korrektheit und Laufzeit des Algorithmus.

Aufgabe 4 (5 Punkte): Geben Sie einen Algorithmus an, der für k Strings $S_1, S_2, \ldots S_k$ (Gesamtlänge n) und alle $q \in \{1, 2, \ldots k\}$ in $O(kn)$ Zeit die Länge des längsten Teilstrings berechnet, der in mindestens q der eingegebenen Strings vorkommt. Begründen Sie Korrektheit und Laufzeit Ihres Algorithmus.

Bonusaufgabe (4 Punkte): Beweisen Sie mit Hilfe der Definition von Θ, dass $\max\{f(n), g(n)\} = \Theta(f(n) + g(n))$, mit $f(n) \geq 0$ und $g(n) \geq 0$.

[1] Der *generalisierte* Suffixbaum für k Strings $S_1, \ldots, S_k$ ist der Suffixbaum für $S_1\$_1S_2\$_2 \ldots S_k\$_k$, wobei $\$_1, \$_2, \ldots, \$_k$ paarweise verschiedene Zeichen sind, die nicht im verwendeten Alphabet vorkommen.