

Übungen zur Vorlesung Sequenzanalyse II

Universität Bielefeld, SS 2010
Prof. Dr. Jens Stoye · Dr. Inke Herms

<http://wiki.techfak.uni-bielefeld.de/gi/Teaching/2010summer/SequenzAnalyse>

Blatt 1 vom 16.04.2010

Abgabe in einer Woche vor Beginn der Vorlesung.
Bitte gib auch den Namen deines Tutors an.

Aufgabe 1 (Σ -Baum)

(4 Punkte)

Gegeben sei die Menge der Worte $W = \{cars, sad, salad, arc, card, at, scar, cat\}$.

- Zeichne den kleinsten Σ -Baum und den kleinsten kompakten Σ^+ -Baum, welche alle Worte aus W darstellen.
- Gib jeweils die Menge der Worte $x \in \Sigma^*$ an, für die $node(x)$ definiert ist.
- Welche Menge $words(T)$ von Worten wird durch die Bäume dargestellt?

Aufgabe 2 (Suffixbaum)

(4 Punkte)

- Zeichne den Suffixbaum von $s = \text{GTAACAGTACA}\$$. Sortiere die Blätter lexikographisch, wobei $\$ < A < C < G < T$, und beschrifte jedes Blatt mit dem Startindex des zugehörigen Suffixes in s .
- Gib eine Familie von Strings s_n^{\max} der Länge $|s_n^{\max}| = n$ an, so dass der Suffixbaum von $s_n^{\max}\$$ maximal viele innere Knoten hat. Weiterhin gib eine Familie von Strings s_n^{\min} der Länge $|s_n^{\min}| = n$ an, so dass der Suffixbaum von $s_n^{\min}\$$ minimal viele innere Knoten hat. Das Alphabet darf dabei beliebig gewählt sein und von n abhängen.

Aufgabe 3 (Effiziente Speicherung eines Suffixbaums)

(3 Punkte)

Wir können einen Suffixbaum effizient speichern, indem für jede Kante nur ein Index gespeichert wird (anstelle von Start- **und** Endposition des entsprechenden Substrings). Erläutere die entscheidende Idee. Lies für diese Aufgabe Abschnitt 8.4 im Skript.

Aufgabe 4 (Verallgemeinerter Suffixbaum)

(4 Punkte)

- Überlege dir am Beispiel von $s = \text{ABACA}$ und $t = \text{CBBAC}$, wie man den verallgemeinerten (generalisierten) Suffixbaum von $s\#t\$$ (mit $\# < \$ < A < B < C$) nutzen kann, um einen längsten gemeinsamen Substring von s und t (und seine Vorkommen) effizient zu finden.
- Nutze deine Überlegungen, um das längste palindromische Teilwort in $x = \text{BANANAS}$ zu finden.