# Übungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, SS 2022
Prof. Dr. Jens Stoye · Dr. Marília D. V. Braga

https://gi.cebitec.uni-bielefeld.de/teaching/2022summer/sa

**Übungsblatt 8 vom 26.5.2022**
**Abgabe am 2.6.2022 bis 12:00 Uhr (mittags)**

**Aufgabe 1 (Suffixbaum-Anwendungen)** (8 Punkte)

Gegeben sei die Sequenz $s = \texttt{BADACBADADAC}$.

1. Konstruiere den Suffixbaum von $s\$$. Sortiere dabei die von einem Knoten ausgehenden Kanten lexikographisch (mit $\$ < \texttt{A} < \texttt{B} < \texttt{C} < \texttt{D}$).

2. Maximale Repeats:

   (a) Finde alle maximalen Repeats in $s$ mit Hilfe des Suffixbaums von $s\$$. Gib alle Zwischenschritte des verwendeten Algorithmus' an.

   (b) **Satz:** In jedem String der Länge $n$ gibt es höchstens $n$ Teilworte, die maximale Repeats sind.
   Argumentiere unter Berücksichtigung des Suffixbaums, warum diese Aussage korrekt ist.

**Aufgabe 2 (Anwendungen des generalisierten Suffixbaums)** (12 Punkte)

Gegeben seien die Sequenzen $s = \texttt{QQONQNPQQO}$ und $t = \texttt{ONQONQNQNP}$.

1. Konstruiere den generalisierten Suffixbaum $T$ von $s$ und $t$. Sortiere dabei die von einem Knoten ausgehenden Kanten lexikographisch (mit $\# < \$ < \texttt{N} < \texttt{O} < \texttt{P} < \texttt{Q}$).

2. Maximale gemeinsame Substrings (MEMs) mit Mindestlänge $\ell$ zweier Sequenzen $s$ und $t$ sind maximale Repeats in $s\#t\$$, von denen je ein Vorkommen in $s$ und ein Vorkommen in $t$ liegt, und die mindestens $\ell$ Zeichen lang sind.

   Verwende den in Aufgabenteil 1 erstellten verallgemeinerten Suffixbaum $T$, um alle MEMs mit Mindestlänge $\ell = 2$ von $s$ und $t$ zu finden. Gehe dafür wie folgt vor:

   (a) Finde Kandidaten: Markiere jeden inneren Knoten $v$ von $T$, für den gilt:
   - Die String-Tiefe von $v$ muss mindestens $\ell = 2$ betragen.
   - Der Teilbaum unter $v$ enthält mindestens ein Blatt, das ein in $s$ beginnendes Suffix repräsentiert und mindestens ein Blatt, das ein in $t$ beginnendes Suffix repräsentiert.

   (b) Gib für jeden dieser Kandidaten $v$ alle Vorkommen (Start- und Endpositionen) von $\textsc{string}(v)$ in $s$ und in $t$ an.

   (c) Welche Paare dieser Vorkommen sind maximal und repräsentieren daher MEMs von $s$ und $t$?

3. Welche MEMs sind auch MUMs von $s$ und $t$ mit Mindestlänge $\ell = 2$?

4. Modifiziere den Algorithmus von Aufgabenteil 2, um direkt MUMs zu finden, ohne den Umweg über MEMs. Welche inneren Knoten von $T$ sind Kandidaten für MUMs der Mindestlänge 2?