Übungen zur Vorlesung
Algorithmen auf Sequenzen

TU Dortmund, SS 2008
Prof. Dr. Sven Rahmann

Blatt 9 vom 04.07.2008
Abgabe am Fr 11.06.2008 in der Vorlesung

Aufgabe 1 Gib den PAA an, mit dem sich die Verteilung der Zufallsvariablen X, die die Anzahl der Treffer der Stringmenge {`GAGA, CTGA, AGAC, GA`} in einem DNA-String der Länge n angibt, berechnen lässt. Sollte man bei $n = 100\,000$ den Linearzeit-Algorithmus oder den Verdoppelungs-Algorithmus nehmen, wenn man an maximal 10 000 Treffern interessiert ist?

Aufgabe 2 Sei $T = T_1 \dots T_n$ ein Text, T^i das Suffix, das an Position i beginnt, und π eine beliebige Permutation von $\{1, \dots, n\}$. Definiere $lcp_\pi(1) := 0$ und $lcp_\pi(r) :=$ Länge des längsten gemeinsamen Präfixes von $T^{\pi(r-1)}$ und $T^{\pi(r)}$ für $r > 1$. Betrachte die Summe $S_\pi := \sum_{r=1}^{n} lcp_\pi(r)$. Zeige: S_π ist genau dann maximal, wenn π das Suffixarray von T ist.

Aufgabe 3 Zeichne den Suffixbaum von `banana`$\$_1$`ananas`$\$_2$. Wie sehen Suffixarray und lcp-Array aus? Hinweis: Es sei stets $\$_i < \$_{i+1}$.

Aufgabe 4 Beschreibe eine einfache Prozedur, die einen Stack verwendet, um aus *pos* und *lcp* die Liste der d-Intervalle d-$[r, r']$ (innere Knoten des Suffixbaumes) bottom-up zu erzeugen.