Übungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, WS 2012/2013
Prof. Dr. Jens Stoye · Nina Luhmann · Linda Sundermann

http://wiki.techfak.uni-bielefeld.de/gi/Teaching/2012winter/SequenzAnalyse

Übungsblatt 7 vom 14.12.2012
Abgabe in einer Woche vor Beginn der Vorlesung.

Aufgabe 1 (Maximale Repeats) (5 Punkte)

1. Finde alle maximalen Repeats in $s =$ `TATGTACCGTATAC` unter Verwendung des im Skript geschilderten Algorithmus (Abschnitt 9.7.4, auf der Homepage zur Übung befindet sich eine überarbeitete Version). Beschreibe dein Vorgehen beim Annotieren des Suffixbaumes.

2. **Satz:** In jedem String der Länge n gibt es höchstens n maximale Repeats.

   Argumentiere unter Berücksichtigung des Suffixbaumes, warum die Aussage korrekt ist. Bedenke: Es stimmt nicht, dass an jeder Position nur ein maximales Repeat beginnen oder enden kann.

Aufgabe 2 (Suffixarray) (5 Punkte)
Gegeben ist der String $s =$ `TTGATCGATTCGCGA`. Beachte $\$ < A < C < G < T$ und Indizierung beginnend mit 0.

1. Gib das Suffixarray *pos* für s an.

2. Implementiere zwei Funktionen, die das *rank*- und *lcp*-Array in linearer Zeit berechnen, wenn das Array *pos* gegeben ist. Die Funktionen sollen so in ein Programm eingebettet sein, dass der Benutzer nur das *pos*-Array übergeben muss. Verwende als Programmiersprache entweder Java, C++ oder Perl. Wenn du eine andere Sprache benutzen möchtest, frage vorher deinen Tutor. Gib dein Programm auch in elektronischer Form an deinen Tutor ab.

3. Wie lauten das *rank*- und *lcp*-Array für s?

Aufgabe 3 (Suffixarray Interpretation) (2 Punkte)
Inwiefern ist das Suffixarray `pos` ein q-gram Index für alle $q \geq 1$ gleichzeitig?

Aufgabe 4 (Manber-Myers Algorithmus) (4 Punkte)
Betrachte das Beispiel 10.4 zur Konstruktion von Suffixarrays im Skript auf den Seiten 106 und 107.

1. Wie viele Phasen braucht man für einen String der Länge 24? Warum?

2. Führe den Manber-Myer Algorithmus schrittweise für den String `GCATAAATAAA` aus.