

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 6 vom 30.11.2012

Abgabe in einer Woche vor Beginn der Vorlesung.

Aufgabe 1 (Statistik)

(2 Punkte)

Mache Dich mit dem *Infinite Monkey Theorem* vertraut, um die folgende Aufgabe zu lösen:

Ein Affe tippt den ganzen Tag auf einer Schreibmaschine, die nur die 26 Buchstaben des deutschen Alphabets in Grossbuchstaben darstellt. Die Wahrscheinlichkeit, dass der Affe einen Buchstaben trifft, sei für jeden Buchstaben gleich.

Wie oft muss der Affe tippen, bis die Wahrscheinlichkeit, dass er deinen Vornamen tippt, größer oder gleich 0,9 ist?

Aufgabe 2 (q-gram Statistik)

(4 Punkte)

Seien X und Y zwei Zufallssequenzen nach dem i.i.d.-Modell. Betrachte nun die zwei q-grams

$X[i \dots i + q - 1]$ und $Y[j \dots j + q - 1]$. Gib die Formeln für die folgenden Wahrscheinlichkeiten an und erläutere jeweils deine Herleitung.

1. Wie hoch ist die Wahrscheinlichkeit, dass die q-grams **genau ein** Mismatch enthalten?
 $\mathbb{P}(d_H(X[i \dots i + q - 1], Y[j \dots j + q - 1]) = 1) = \dots$
2. Wie hoch ist die Wahrscheinlichkeit, dass die q-grams **höchstens ein** Mismatch enthalten?
 $\mathbb{P}(d_H(X[i \dots i + q - 1], Y[j \dots j + q - 1]) \leq 1) = \dots$
3. Wie hoch ist die Wahrscheinlichkeit, dass die q-grams **höchstens zwei** Mismatches enthalten?
 $\mathbb{P}(d_H(X[i \dots i + q - 1], Y[j \dots j + q - 1]) \leq 2) = \dots$

Aufgabe 3 (§ – Der Wächter)

(1 Punkt)

Bei der Erstellung eines Suffixbaums wird am Ende des Strings ein eindeutiges Zeichen, z.B. $\$ \in \Sigma$, angehängt. Warum?

Aufgabe 4 (Suffixbäume)

(8 Punkte)

1. Beschreibe zwei Anwendungen aus der bioinformatischen Praxis, für die man einen Suffixbaum verwenden kann.
2. Gegeben ist der String $s\$ = \text{abcacbbaca\$}$, wobei $\$ < a < b < c$.
 - (a) Zeichne den Suffixbaum für s , sortiere dabei die Blätter lexikographisch.
 - (b) Beschrifte die Blätter mit dem Start-Index des zugehörigen Suffixes in s . Die Indizierung beginnt bei 1.
 - (c) Beschrifte jeden Knoten mit der Anzahl der unter ihm liegenden Blätter.
3. Analysiere die *worst-case*- und *average-case*-Laufzeit des WOTD-Algorithmus in eigenen Worten.
4. Zeige, dass der Speicherverbrauch eines Suffixbaums linear bezüglich der Eingabe ist.

Aufgabe 5 (Verallgemeinerter Suffixbaum)

(4 Punkte)

Gegeben seien die Strings $s = \text{GAAG}$ und $t = \text{AAGG}$.

1. Zeichne den generalisierten Suffixbaum von s und t (mit $\# < \$ < A < G$).
2. Überlege dir, wie man den längsten gemeinsamen Substring von s und t finden kann und gib seine Vorkommen an.
3. Formuliere allgemein, wie man längste gemeinsame Substrings zweier Strings im generalisierten Suffixbaum finden kann.
4. Mache dir Gedanken darüber, wie man das längste palindromische Teilwort in einem Wort mit einem generalisierten Suffixbaum finden kann. Verwende dazu das Beispiel $x = \text{BANANAS}$. Beschreibe deine Idee. Es ist nicht nötig, den Baum explizit zu zeichnen.