Übungsblatt 5

k-mer Identifikation

Deadline: 8.6.2022 um 20:00 MEZ

Bioinformatik für Biochemiestudierende

Dr. Florian Klimm

Sommersemester 2022

Aufgabe 5.1 Identische Zufallssequenzen (10 %)

Wie lautet die Wahrscheinlichkeit, dass zwei Zufallssequenzen der Länge N mit einem Alphabet aus A Buchstaben identisch sind?

Aufgabe 5.2 Die Wahrscheinlichkeit von Zufallssequenzen (20 %)

Bestimmen Sie die Wahrscheinlichkeit $\Pr(N = 100, A = 2, \text{Pattern} = \text{”01”}, t = 1)$.

Aufgabe 5.3 *k*-mere bestimmen (70 %)

Programmieren Sie ein R Programm das die häufigsten k-mere in einem Text ermittelt und ausgibt.

Wenden Sie dieses Programm auf den folgenden String an

- ACGTTGCATGTCGCATGATGCATGAGAGCT

und bestimmen Sie die häufigsten k-mere für $k \in 1, 2, 3, 4, 5$.

Aufgabe 5.1 Identische Zufallssequenzen (10 %)

Wie lautet die Wahrscheinlichkeit, dass zwei Zufallssequenzen der Länge N mit einem Alphabet aus A Buchstaben identisch sind?

Sequenz 1 : N_1, A_1

Sequenz 2 : N_2, A_2

$N = N_1 = N_2$ & $A_1 = A_2 = A$

- an jeder Position tritt jeder Buchstabe mit $\frac{1}{A}$ Wahrscheinlichkeit auf

↳ N-Positionen
↳ jede Position: $\frac{1}{A}$

Wahrscheinlichkeit $= N \cdot \frac{1}{A}$

- wenn: $A_1 \neq A_2$

↳ Wahrscheinlichkeit: $\frac{1}{A^N}$

Aufgabe 5.2 Die Wahrscheinlichkeit von Zufallssequenzen (20 %)

Bestimmen Sie die Wahrscheinlichkeit $\Pr(N = 100, A = 2, \text{Pattern} = \text{"01"}, t = 1)$.

$$\Pr(N, A, \text{Pattern}, t) = \frac{\binom{N - t(k-1)}{t}}{A^{t \cdot k}}$$

$$\Pr(100, 2, \text{"01"}, 1) = \frac{\binom{100 - 1 \cdot (2-1)}{1}}{2^{1 \cdot 2}} = \frac{\binom{99}{1}}{2^2} = \frac{99}{4} = 24{,}75\,\%$$

Mit 27,75% Wahrscheinlichkeit ist das 2-mer in Sequenz zu finden.

$$\Pr(N, A, K, t) \approx \frac{\binom{N - t \cdot (k-1)}{t}}{A^{(t-1) \cdot K}}$$

$$\Pr(100, 2, 2, 1) \approx \frac{\binom{100 - 1 \cdot (2-1)}{1}}{2^{(1-1) \cdot 2}} \approx 0\,\%$$

Mit 0% Wahrscheinlichkeit kommt kein 2-mer häufiger als t=1 mal vor.