

5. Übung zur Vorlesung "Sequenzanalyse"

Sebastian Böcker, Kai Dührkop, Markus Fleischauer

Abgabe: 09.06.2017

Aufgabe 1 (3 Punkte)

Gegeben sei eine Alignmentdatenbank über dem Alphabet $\{A, C, G\}$ mit insgesamt 100000 **gapfrei** alignierten Positionen und 1% Mismatches. A kommt 70000 mal und C 40000 mal vor. Angenommen die Anzahl der Substitutionen von A und G ist $m(A, G) = 500$.

Berechnen Sie den log-odds-Score $\sigma^{(1)}(A, G)$.

Aufgabe 2 (5 Punkte)

Angenommen, unsere Scores sind ganze positive Zahlen $0, 1, 2, \dots$. Durch Betrachten eines Scorehistogramms stellen wir fest, dass $\mathbb{P}(\text{Score} = s) = 0.02 \cdot (0.98)^s$ gilt.

1. Zeigen Sie, dass dies in der Tat eine Verteilung definiert, d.h., es gilt $\sum_{s=0}^{\infty} \mathbb{P}(\text{Score} = s) = 1$.
2. Bestimmen Sie den minimalen Wert T , so dass $\mathbb{P}(\text{Score} \geq T) \leq 0.01$.

Aufgabe 3 (5 Punkte)

Wir betrachten zwei unabhängige zufällige Sequenzen der Länge n über einem Alphabet der Größe σ . Jeder Buchstabe, darunter A , ist mit gleicher Wahrscheinlichkeit an jeder Position anzutreffen, unabhängig von den anderen Positionen (iid Modell). Wie groß ist die Wahrscheinlichkeit, dass die erste Sequenz genauso viele A s enthält wie die zweite? (Es genügt die Lösung als Summe anzugeben).

Aufgabe 4 (7 Punkte)

Literatursuche: Die Karlin-Altschul Theorie setzt *unendlichlange* Sequenz voraus. Deshalb ist im BLAST eine Korrektur für die Sequenzlänge implementiert. Wie funktioniert diese Korrektur?