

Präsenzübungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, WS 2024

Prof. Dr. Jens Stoye · Leonard Bohnenkämper

<https://gi.cebitec.uni-bielefeld.de/teaching/2024winter/sa>

Präsenzübungsblatt 1, vom 28.10.2024/29.10.2024

Aufgabe 1 (Anzahl von Subsequenzen und Substrings)

Versuche, dir die folgenden Formeln jeweils anschaulich klar zu machen und fertige kleine Skizzen an.

1. Wie viele Subsequenzen der festen Länge k hat ein Text der Länge n maximal?
2. Wie viele Subsequenzen mit den Längen $1 \leq k \leq n$ hat ein Text der Länge n insgesamt maximal?
3. Wie viele Substrings der festen Länge k hat ein Text der Länge n maximal?
4. Wie viele Substrings mit den Längen $1 \leq k \leq n$ hat ein Text der Länge n insgesamt maximal?
5. Finde ein Beispiel, für das die obigen Schranken nicht erreicht werden.

Aufgabe 2 (GC-Gehalt)

Das Genom von *Saccharomyces cerevisiae* besitzt einen GC-Gehalt von etwa 40% und eine Genomgröße von ca. 12 Mb.

1. Wie oft würde man hier die Schnittstelle des Restriktionsenzym *EcoRI* mit der Erkennungssequenz GAATTC erwarten?
2. Wie hoch ist die Wahrscheinlichkeit, dass ein zufälliger Abschnitt der Länge 6, der ein A am Anfang und am Ende, aber sonst keiner anderen Position enthält, auftritt? (Also: $Ax_1x_2x_3x_4A$, mit $(x_i)_{i=1}^4 \in \{C, G, T\}$)
3. Wie viele Möglichkeiten existieren, um zwei A-Basen in $x_1x_2x_3x_4$ zu insertieren?
4. Wie hoch ist die Wahrscheinlichkeit genau zwei As in einem zufälligen Abschnitt der Länge 6 unseres Hefegenoms zu finden?

Aufgabe 3 (O-Notation)

Sortiere die O -Klassen der Funktionen $a(n)$ bis $e(n)$ anhand ihrer Untermengenrelation $\subset$, z.B. $O(g(n)) \subset O(h(n))$.

- $a(n) = n^2 \cdot \log_{42}(n) + 2$
- $b(n) = 4^n + n^2$
- $c(n) = 4^{4n}$
- $d(n) = 3^{n+2}$
- $e(n) = \sqrt{n^{5.4}}$