ASE Übung 1 - Informationstheorie

Codierung

Sollen DNA Sequenzen in Textform gespeichert werden, so wird pro Sequenzelement (Nukleotid) ein ASCII (American Standard Code for Information Interchange) Zeichen von der Größe 1 Byte = 8 Bits benötigt. Berechnen Sie, wie viele Bits pro Symbol optimalerweise benötigt werden, wenn nur die 4 Sequenzsymbole A, C, G, und T angenommen werden (d.h. keine mehrdeutigen Symbole).

Das menschliche Genom besteht aus etwa 3×10^9 Basenpaaren. Berechnen Sie den benötigten Speicherplatz, wenn das Genom als ASCII- bzw. mittels optimaler Bit-Codierung gespeichert wird, und die entsprechende Platzersparnis der letzteren. Beachten Sie dabei die Präfixe für Vielfache von Bytes: 1 Kilobyte = 10^3 Bytes, 1 Megabyte = 10^6 Bytes, 1 Gigabyte = 10^9 Bytes.

Berechnen Sie ebenfalls die Anzahl der benötigten Bits für eine optimale Codierung von Proteinsequenzen und die generelle Platzersparnis im Vergleich zu ASCII-Text.

Entropie

Berechnen Sie die Entropien der folgenden drei DNA Sequenzabschnitte. Nutzen Sie hierbei die Formel der *Shannon-Entropie*:

$$S = -\sum_i p(x_i) \log_2 p(x_i)$$

Hierbei steht $p(x_i)$ für die relative Häufigkeit eines Nukleotids innerhalb seiner Sequenz. Sollte es Ihnen nicht möglich sein, den Logarithmus zur Basis 2 direkt zu berechnen, nutzen Sie die Logarithmus-Basisumrechnung:

$$\log_b x = \frac{\log_a x}{\log_a b}$$

Sequenz 1: `GATGGTGTAGCTCAGCGGTTAGAGCGGTTGACTGTTAATC`

Sequenz 2: `CACGGGCCGTGGGTCCGTGGCCACCCTCCCGCGCCGGAGG`

Sequenz 3: `GGGCCGCCCGCCGGCGCCCGCGGCCCCGCGGGCGCCCGCG`

Wie verhält sich der Entropiewert hinsichtlich der Häufigkeitsverteilungen der Nukleotide? Wann würde die Entropie maximal (2) bzw. minimal (0) sein?

Fragen

1. Mit welcher einfachen Formel kann die für eine optimale Codierung benötigte Anzahl von Bits eines Alphabets mit A Zeichen ermittelt werden?
2. Gegeben sei die Codierung A=00, C=01, G=10, T=11. Mit welcher einfachen Bitoperation kann die jeweils komplementäre Base bestimmt werden?
3. Im Genom beobachten Sie Abschnitte mit sehr hohen Entropiewerten (nahe 2) zwischen längeren Abschnitten mit geringerer Entropie. Was könnte dies bedeuten?