AlgSeq - Übungen (3)

(1) Wahrscheinlichkeiten von Häufigkeiten
Bestimmen Sie die häufigsten *k*-mer Motive (mit revers-komplementären Sequenzen), mit k ∈ [7;12] für gängige DNA-bindende Proteine, im *oriC* von *V. cholerae*, und berechnen Sie jeweils die (approximierte) Wahrscheinlichkeit der beobachteten Vorkommnisse.

Datei: `oric_Vibrio_cholerae.txt`

(2) Pattern Matching
Implementieren Sie einen Algorithmus in Python, der das Pattern matching Problem löst, und lokalisieren Sie damit alle häufigsten Nonamere des *oriC* von *V. cholerae* im bakteriellen Genom (10^6 Basen -- achten Sie auf eine Laufzeit-effiziente Implementierung).

Datei: `genom_Vibrio_cholerae.fasta`
(FASTA Format!)

(3) Klumpen Finden
Skizzieren Sie einen Algorithmus, der das Klumpen Finden Problem effizient löst und geben Sie eine Abschätzung der Laufzeit-Komplexität an. Implementieren Sie Ihren Algorithmus in Python und testen Sie diesen auf der Genomsequenz von *V. cholerae*, mit dem/den häufigsten Nonamer(en) im *oriC* dieses Bakteriums.

(4) T. petrophila

Wiederholen Sie Aufgabe (3) – (5) für die entsprechenden *oriC*- und Genom-Sequenzen von *T. petrophila*, einem anderen Bakterium.

Dateien:

`oric_Thermotoga_petrophila.txt`
`genom_Thermotoga_petrophila.fasta`

(5) G-C Ungleichgewichts-Diagramme

Benutzen Sie die genomischen Sequenzen von *V. cholerae*, T. petrophila und auch von *E. coli*, um den einen Bereich für die Lokalisierung des *oriC* anzugeben.

Datei: `genom_Escherichia_coli.fasta`

(6) OriC Lokalisierung

Benutzen Sie nun alle im Seminar besprochenen Möglichkeiten der *oriC* Lokalisierung, und geben Sie die Koordinaten des besten oriC-Kandidatenbereiches für jede der 3 Spezies (*V. cholerae*, T. petrophila und auch von *E. coli*) an.