Übungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, SS 2023
Prof. Dr. Jens Stoye · Tizian Schulz

https://gi.cebitec.uni-bielefeld.de/teaching/2023summer/sa

Übungsblatt 12 vom 22.6.2023
Abgabe am 29.6.2023 bis 12:00 Uhr

Aufgabe 1 (Divide-and-Conquer-Alignment) (5 Punkte)
Gegeben sind die Sequenzen $s_1 = GAG$, $s_2 = TG$ und $s_3 = TA$. Benutze für deine Berechnungen Einheitskosten.

1. Was ist der Unterschied zwischen einem optimalen und einem C-optimalen Schnitt?
2. Erstelle die Zusatzkostenmatrizen für die drei Sequenzen.
3. Gib alle C-optimalen Schnitte an.
4. Nenne alle möglichen optimalen multiplen Alignments, die dir die C-optimalen Schnitte angeben. Gib auch ihre Kosten an.
5. Wieso verkleinert sich der Suchraum von $\mathcal{O}(n^k 2^k)$ beim Sum-of-Pairs-optimalen multiplen Alignment auf $\mathcal{O}(n^{k-1})$ beim Divide-and-Conquer-Alignment?

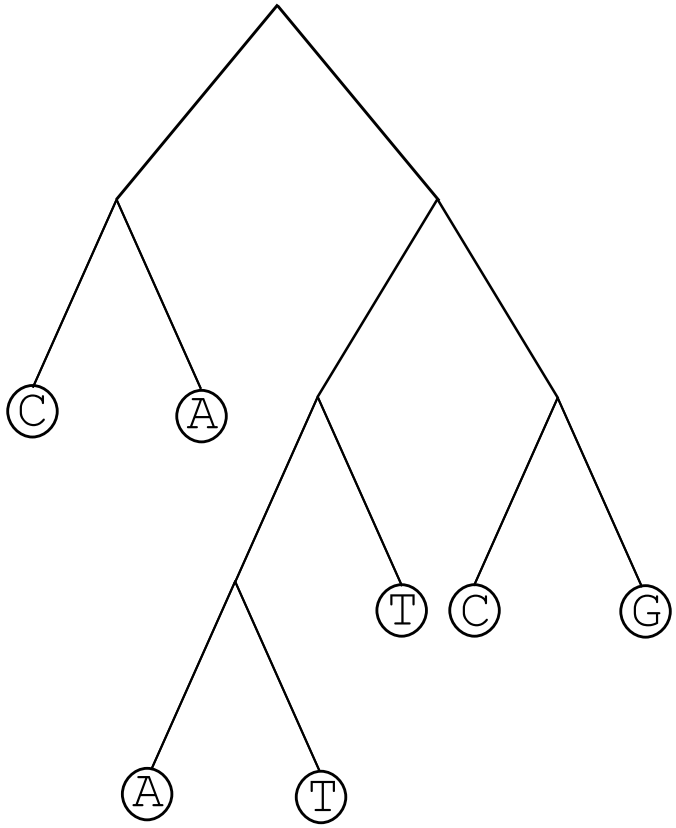
Aufgabe 2 (Baumalignment) (10 Punkte)

1. Ein optimales Baumalignment für die Sequenzen $s_1, \ldots, s_k$ entspricht im Sequenzraum (V, E) einem minimalen Steinerbaum für die Teilmenge $\{s_1, \ldots, s_k\} \subseteq V$.
   (a) Wie sind dabei V und E definiert?
   (b) Wie sind die Kantengewichte gewählt?
2. Für die innere Minimierung der Formel für das Baumalignment kann der Fitch-Algorithmus verwendet werden. Finde eine optimale Beschriftung der inneren Knoten des Baums auf der Rückseite unter Einheitskosten. Trage auch Zwischenergebnisse in den Baum ein.

Aufgabe 3 (Genomalignment mit MUMs) (10 Punkte)

Gegeben seien die Genome: $\begin{cases} \mathcal{G} = \texttt{CACGGGTATAAGTCTT} \\ \mathcal{H} = \texttt{ACGAATAAAGGTGCTT} \end{cases}$

1. Verwende den verallgemeinerten Suffixbaum $\mathcal{G}\#\mathcal{H}\$$, um alle MUMs der Mindestlänge 3 von $\mathcal{G}$ und $\mathcal{H}$ zu finden.
2. Berechne ein globales Alignment von $\mathcal{G}$ und $\mathcal{H}$, indem du die MUMs verkettest (Gewicht=Länge) und die Lücken durch paarweise optimale Alignments füllst.


C
A
T
C
G
A
T