8. Übung zur Vorlesung “Sequenzanalyse”

Sebastian Böcker, Marcus Ludwig, Kai Dührkop, Fleming Kretschmer

Ausgabe: 3.1.2022
Abgabe: 9.1.2022

Aufgabe 1 (5 Punkte)

Gegeben seien vier Sequenzen $s_1 = TACA$, $s_2 = CTAC$, $s_3 = GTAG$, $s_4 = ATGC$. Berechnen Sie das multiple Sequenzalignment dieser vier Sequenzen anhand der Center-Star-Methode mit Einheitskosten.

Aufgabe 2 (7 Punkte)

Zeigen Sie, dass die Center-Star-Methode eine 2-Approximation der optimalen Lösung ist.
Hinweis: Wir nehmen an, dass die Distanzfunktion $d(x, y)$ zwischen zwei Buchstaben eine Metrik ist und $d(-, -) = 0$ gilt. Zeigen Sie zunächst, dass für das berechnete, approximierte Alignment M gilt

$$d(S_i, S_j) \leq D(S_i, S_c) + D(S_c, S_j) \text{ für } 1 \leq i, j \leq k, i \neq j$$

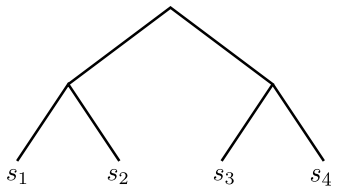
Benutzen Sie diese Aussage für den Beweis der 2-Approximation:

$$\frac{d(M_c)}{d(M^*)} \leq \frac{2(k-1)}{k} < 2$$

Dabei ist k die Anzahl der Sequenzen. c ist der Index der Center-Sequenz. M ist das durch die Center-Star-Methode approximierte multiple Alignment und M^* ist das optimale multiple Alignment. $d(S_i, S_j)$ ist die Distanz des paarweisen Alignments von S_i und S_j welches durch M induziert wird. $D(S_i, S_j)$ ist die Distanz des optimalen paarweisen Alignments der Sequenzen S_i und S_j.

Aufgabe 3 (5 Punkte)

Berechnen Sie das multiple Sequenzalignment der Sequenzen $s_1 = AGTCAT$, $s_2 = GTACT$, $s_3 = ATTATC$, und $s_4 = GGCCT$ mit Einheitskosten und anhand des unten gegeben Leitbaums. Alignieren Sie dabei das Alignment (s_1, s_2) mit (s_3, s_4) an der Wurzel des Baumes.



Erklären Sie, warum das Problem schwieriger wird, wenn anstatt Einheitskosten die affinen Gapkosten verwendet werden.

Aufgabe 4 (3 Punkte)

Gewichten Sie die Sequenzen s_1 bis s_9 anhand des unten stehenden Leitbaums. Die Kantengewichte seien dabei 1.