

Übungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, SS 2018

Dr. Daniel Dörr

<https://gi.cebitec.uni-bielefeld.de/teaching/2018summer/sa>

Übungsblatt 11 vom 28.06.2018

Abgabe bis spätestens Donnerstag, 05.07.2018 23:59 GMT+2 via Email oder Postfach (U10-151)

Aufgabe 1 (Center-Star-Approximation (Wdh. Aufgabe 10.2))

(5 Punkte)

Gegeben sind die Sequenzen $s_1 = \text{GACCC}$, $s_2 = \text{GACT}$ und $s_3 = \text{GCTC}$. Benutze für deine folgenden Berechnungen Einheitskosten.

1. Berechne die *Center-Sequenz* s_c .
2. Erstelle das multiple Alignment A_c und gib seine Sum-of-Pairs-Kosten an.
3. Was kannst du über die Optimalität des gefundenen Alignments sagen?
4. Beschreibe in eigenen Worten die Laufzeit- und Speicherkomplexität der Center-Star-Approximation. Unterscheide dabei zuerst die einzelnen Phasen und erkläre dann das Gesamtergebnis.

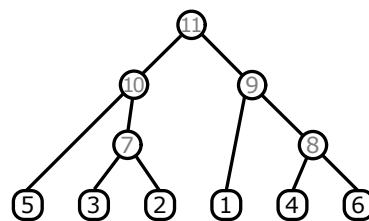
Aufgabe 2 (Baumalignment)

(4 Punkte)

1. Um die Kosten für ein Baumalignment exakt zu berechnen, kann der Sankoff-Algorithmus verwendet werden. Erkläre in 3–5 Sätzen, wie dieser Algorithmus funktioniert. Erläutere auch, welche Rolle der Fitch-Algorithmus dabei spielt.
2. Gegeben seien folgendes Alignment der Blattsequenzen s_1, \dots, s_k

$$A = \begin{matrix} s_1 & \begin{pmatrix} A & T & T & C \end{pmatrix} \\ s_2 & \begin{pmatrix} - & - & T & G \end{pmatrix} \\ s_3 & \begin{pmatrix} A & C & T & C \end{pmatrix} \\ s_4 & \begin{pmatrix} - & T & T & G \end{pmatrix} \\ s_5 & \begin{pmatrix} A & G & A & T \end{pmatrix} \\ s_6 & \begin{pmatrix} - & T & T & - \end{pmatrix} \end{matrix}$$

und phylogenetischer Baum T :



Berechne für Alignment A und phylogenetischen Baum T die kostenminimierende Beschriftung der inneren Knoten mit Hilfe des Fitch-Algorithmus. Gib dabei für die inneren Knoten jeweils die mit ihnen assoziierten Informationen der Bottom-Up- und der Top-Down-Phase an. Gebe als Endergebnis das komplette Alignment inklusive der Sequenzen s_7, \dots, s_{11} der inneren Knoten an.

Aufgabe 3 (Forward-Backward Technik)

(5 Punkte)

Gegeben seien die beiden Strings $x = \text{GCAGC}$ und $y = \text{CGCT}$. Die Indizierung beginnt bei 1.

1. Berechne die Matrizen D , $D^{rev} = D^{-1}$ und C unter dem Einheitskostenmodell.
2. Lies alle optimalen Alignments aus C ab und gib diese an.
3. Welche minimalen zusätzlichen Kosten hat ein Alignment unter der Einschränkung, dass $x[1]$ und $y[2]$ aligniert sind?