

Übungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, SS 2021

Prof. Dr. Jens Stoye · Dr. Marília D. V. Braga

<https://gi.cebitec.uni-bielefeld.de/teaching/2021summer/sa>

Übungsblatt 6 vom 27.5.2021

Abgabe am 3.6.2021 bis 12:00 Uhr (mittags)

Aufgabe 1 (Suboptimale Alignments)

(6 Punkte)

1. Beschreibe in eigenen Worten, was man unter *überlappenden* lokalen Alignments versteht.
2. Warum ist man bei der Bestimmung von suboptimalen Alignments in erster Linie an *nichtüberlappenden* Alignments interessiert? Welche Probleme möchte man vermeiden?
3. Gegeben seien die Sequenzen $x = \text{ACAGCTTA}$ und $y = \text{ATTCATGTG}$ sowie die folgende Score-Funktion: Match = 3, Mismatch = -2, Indel = -2.
 - (a) Berechne die Alignment-Matrix für lokale Alignments mit dem Smith-Waterman-Algorithmus und gib ein optimales Alignment an.
 - (b) Berechne die Matrix für nichtüberlappende, suboptimale lokale Alignments nach *Waterman-Eggert* und gib das erste suboptimale Alignment an.
 - (c) Aktualisiere die Matrix und gib das nächste suboptimale Alignment an.

Aufgabe 2 (Approximative Textsuche)

(4 Punkte)

Gegeben seien das Muster $s = \text{TCAG}$ und der Text $t = \text{AACGTCAGTCGAGTG}$.

1. Verwende die *Cutoff*-Variante von Sellers' Algorithmus mit Einheitskosten und berechne die Edit-Matrix D von s und t mit bis zu $k = 1$ Fehlern. Markiere in jeder Spalte den *last essential index*.
2. Markiere in D alle Positionen, an denen ein Vorkommen von s mit bis zu $k = 1$ Fehlern in t endet. Gib für jede dieser Endpositionen ein korrespondierendes Alignment an.
3. Warum ist es sinnvoll, in der Praxis nicht *alle* gefundenen Endpositionen auszugeben? (Stichwort: *Runs*.) Welche Endpositionen würde man im obigen Beispiel nicht ausgeben?

Aufgabe 3 (Forward-Backward Technik)

(4 Punkte)

Gegeben seien die Sequenzen $x = \text{ATGCAATC}$ und $y = \text{CTCAGAT}$. Verwende Einheitskosten für globale Alignments.

1. Berechne die Edit-Matrix D und die Edit-Matrix D^{rev} der reversen Strings.
2. Berechne die Gesamtkostenmatrix T und die Zusatzkostenmatrix C .
3. Gib ein optimales globales Alignment von s und t an und markiere es in D , D^{rev} , T und C .

Aufgabe 4 (Paarweises Alignment mit linearem Speicherbedarf)

(6 Punkte)

Zeige anhand der Sequenzen aus Aufgabe 3, wie das optimale Alignment berechnet werden kann, während nur linearer Speicherplatz benötigt wird. Simuliere dazu die Schritte des in der Vorlesung besprochenen Divide-and-Conquer Verfahrens auf Kopien der Edit-Matrix. (Es kann teilweise auf die in Aufgabe 4 berechneten Werte zurückgegriffen werden.)

1. Markiere in einer neuen, leeren Matrix der entsprechenden Größe, für welche Bereiche D und D^{rev} berechnet werden würden bzw. berechne die Werte wenn nötig neu und trage sie ein.
2. Trage die Werte von T für die mittlere Zeile $m' = \lceil m/2 \rceil$ ein. Markiere ein Minimum in T und die entsprechende Spalte, an der die Matrix geteilt wird.
3. Führe rekursiv die oberen Schritte für die beiden Teilmatrizen aus, bis jeweils nur noch eine Zeile übrig ist.

Du kannst diese Seite für deine Lösung verwenden:

Aufgabe 3: Einheitskosten

1.

<i>D</i>	€	<i>C</i>	<i>T</i>	<i>C</i>	<i>A</i>	<i>G</i>	<i>A</i>	<i>T</i>
€								
<i>A</i>								
<i>T</i>								
<i>G</i>								
<i>C</i>								
<i>A</i>								
<i>A</i>								
<i>T</i>								
<i>C</i>								

<i>D</i> ^r	<i>C</i>	<i>T</i>	<i>C</i>	<i>A</i>	<i>G</i>	<i>A</i>	<i>T</i>	€
<i>A</i>								
<i>T</i>								
<i>G</i>								
<i>C</i>								
<i>A</i>								
<i>A</i>								
<i>T</i>								
<i>C</i>								
€								

2.

<i>T</i>	€	<i>C</i>	<i>T</i>	<i>C</i>	<i>A</i>	<i>G</i>	<i>A</i>	<i>T</i>
€								
<i>A</i>								
<i>T</i>								
<i>G</i>								
<i>C</i>								
<i>A</i>								
<i>A</i>								
<i>T</i>								
<i>C</i>								

<i>C</i>	€	<i>C</i>	<i>T</i>	<i>C</i>	<i>A</i>	<i>G</i>	<i>A</i>	<i>T</i>
€								
<i>A</i>								
<i>T</i>								
<i>G</i>								
<i>C</i>								
<i>A</i>								
<i>A</i>								
<i>T</i>								
<i>C</i>								