Klausur zur Vorlesung Sequenzanalyse

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

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

Klausur vom 20.07.2023, 10:15–11:45 Uhr in H6

Matrikelnummer:

☐ benotet ☐ unbenotet

Hinweise: Bitte **erst alle Aufgaben einmal durchlesen** und mit den für dich einfachsten anfangen. Es sind **keine Hilfsmittel** (wie z. B. Aufzeichnungen, Taschenrechner, Handys, Uhren mit Kommunikationsfunktion etc.) zugelassen. Weitere leere Blätter sind bei der Klausuraufsicht erhältlich; **bitte keine eigenen Blätter verwenden**.

Aufgabe	**1**	**2**	**3**	**4**	**Summe**
Punkte	/25	/25	/25	/25	/100

Aufgabe 1 Suffixbäume und Suffixarrays

1. Gegeben sei die Sequenz $s = \texttt{GACACGCGTC}$. (15 Punkte)

   (a) Führe den WOTD-Algorithmus schrittweise zur Erstellung des Suffixbaums für die Sequenz $s\$$ aus. Sortiere dabei die Kanten lexikographisch (wobei $\$ < \texttt{A} < \texttt{C} < \texttt{G} < \texttt{T}$).

   (b) Beschrifte die Blätter mit dem Start-Index des jeweils zugehörigen Suffixes in $s\$$. Die Indizierung soll bei 0 beginnen.

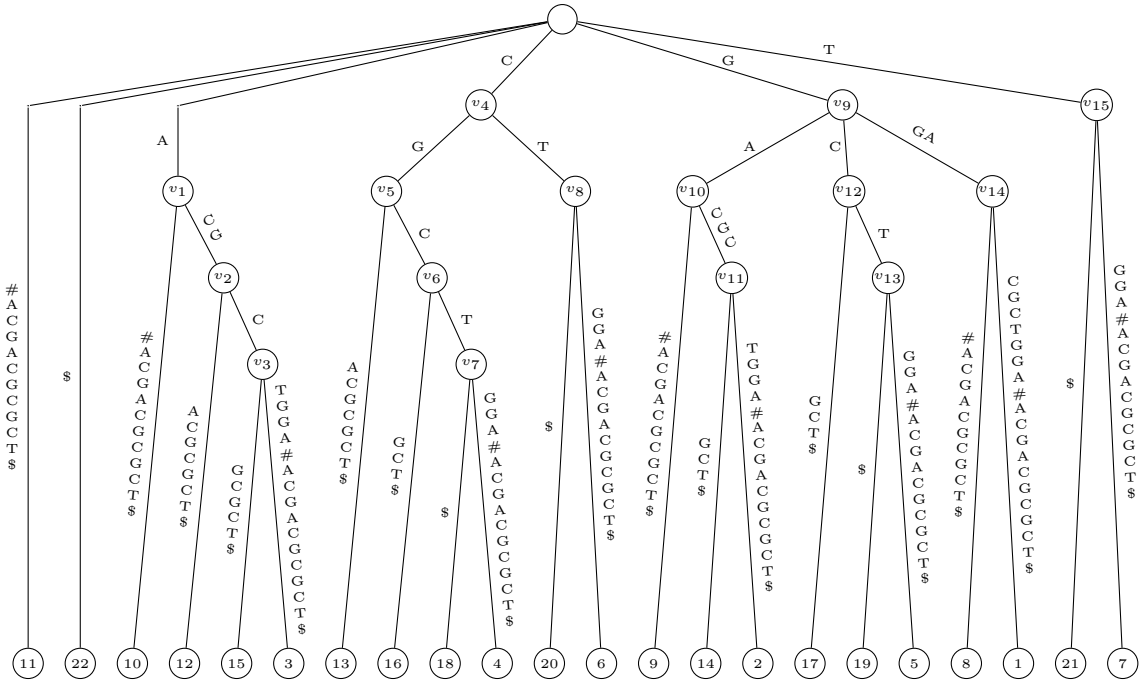
   (c) Gib für die obige Sequenz die Suffix-Arrays $\texttt{pos}(s\$)$ und $\texttt{lcp}(s\$)$ an.

r	0	1	2	3	4	5	6	7	8	9	10
pos[r]											

r	0	1	2	3	4	5	6	7	8	9	10	11
lcp[r]												

2. Ein *maximal exact match* (MEM) ist ein Teilwort, das in zwei Sequenzen vorkommt und nicht gleichzeitig in beiden Sequenzen nach links oder rechts erweitert werden kann. (10 Punkte)

Unten ist der generalisierte Suffixbaum T der Sequenzen $s =$ `GGACGCTGGA` and $t =$ `ACGACGCGCT` gegeben. Erkläre am Beispiel von T, wie MEMs einer bestimmten Mindestlänge ℓ unter Verwendung eines generalisierten Suffixbaumes gefunden werden können und gib alle MEMs von s und t für $\ell = 2$ an.



Aufgabe 2 Paarweises Sequenzalignment

1. Gegeben sei das Alignment A und die Scorematrix M: (5 Punkte)

$$A = \begin{pmatrix} - & - & - & \mathtt{C} & \mathtt{G} & \mathtt{C} & \mathtt{C} & \mathtt{T} & \mathtt{G} & \mathtt{A} \\ \mathtt{G} & \mathtt{C} & \mathtt{T} & \mathtt{C} & \mathtt{T} & \mathtt{C} & - & \mathtt{T} & \mathtt{G} & \mathtt{T} \end{pmatrix}$$

M	A	C	G	T
A	1	−2	−2	−1
C	−2	2	−1	−2
G	−2	−1	2	−2
T	−1	−2	−2	1

Berechne die Kosten bzw. den Score von A unter Verwendung von:

(a) Einheitskosten

(b) M und affinen Gapscores ($d = -2$, $e = -1$)

2. Nach Berechnung des semiglobalen Alignments von $P = \mathtt{ABBBB}$ und $T = \mathtt{ABABAABBBAAAABB}$ unter Verwendung von Einheitskosten ergibt sich die unten angegebene Matrix. (10 Punkte)

(a) Ergänze die Werte in den fehlenden Spalten.

(b) Markiere in jeder Spalte, welche Zellen unter Verwendung der *Cutoff*-Variante von Sellers' Algorithmus für $k = 2$ nicht berechnet werden würden und umkreise den *last essential index* jeder Spalte.

(c) Gib die Koordinaten und die minimalen Kosten aller *runs of local minima* an.

	T	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
P		ϵ	A	B	A	B	A	A	B	B	B	A	A	A	A	B	B
0	ϵ	0	0	0		0				0		0	0				0
1	A	1	0	1		1				1		0	0				1
2	B	2	1	0		0				1		1	1				1
3	B	3	2	1		1				0		2	2				0
4	B	4	3	2		1				1		1	2				1
5	B	5	4	3		2				2		1	2				2

3. Gegeben seien die Sequenzen $x =$ `TCATCGAC` und $y =$ `CGATAGCGAT`. (10 Punkte)

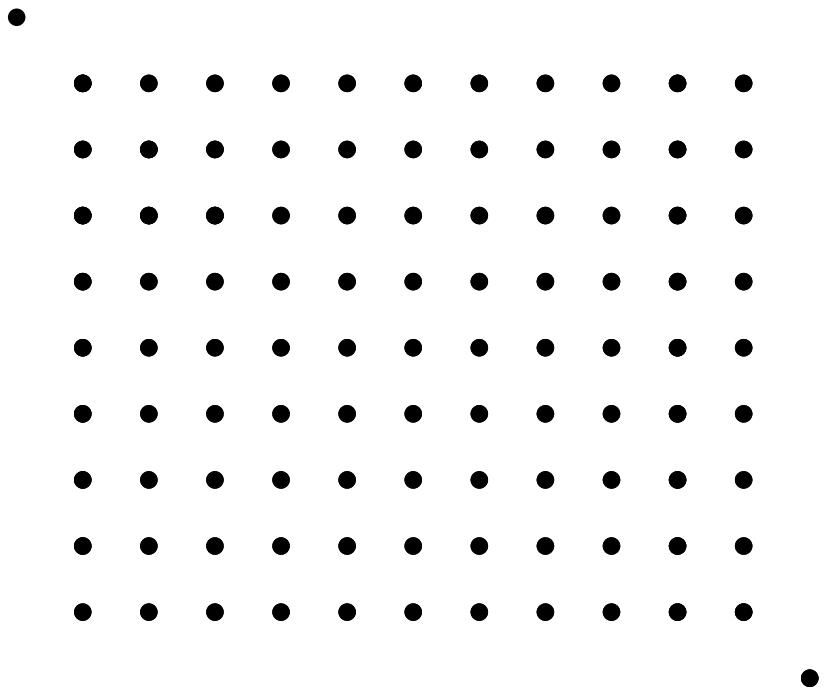
   (a) Berechne das optimale, lokale Sequenzalignment von x und y. Benutze dafür die unten angegebene Matrix und die Scorefunktion: MATCH = 2; MISMATCH = -1; INDEL = -1.

   (b) Markiere den Teil der Matrix, der neu berechnet werden muss, um suboptimale Alignments nach Waterman und Eggert zu finden.

   (c) Du findest unten den zum Alignment von x und y gehörenden Alignmentgraphen ohne Kanten. Zeichne dort alle Kanten ein, die dein optimales, lokales Alignments verwendet. Markiere diejenigen dieser Kanten, die bei der Berechnung des ersten suboptimalen Alignments *nicht* mehr verwendet werden dürfen.

	ϵ	C	G	A	T	A	G	C	G	A	T
ϵ											
T											
C											
A											
T											
C											
G											
A											
C											

Optimales Alignment:



Optimale paarweise Sequenzalignments der Sequenzen $s_1, \ldots, s_5$ aus Aufgabe 3:

$$A_{s_1,s_2} = \Big(\quad \Big)$$

$$A_{s_1,s_3} = \begin{pmatrix} \mathtt{C} & \mathtt{T} & \mathtt{T} & \mathtt{A} & \mathtt{G} & \mathtt{A} & \mathtt{G} & \mathtt{-} & \mathtt{T} \\ - & - & - & - & \mathtt{G} & \mathtt{G} & \mathtt{G} & \mathtt{C} & \mathtt{T} \end{pmatrix}$$

$$A_{s_1,s_4} = \begin{pmatrix} \mathtt{C} & \mathtt{T} & \mathtt{T} & \mathtt{A} & \mathtt{G} & \mathtt{A} & \mathtt{G} & \mathtt{T} \\ \mathtt{A} & \mathtt{C} & \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{G} & \mathtt{T} \end{pmatrix}$$

$$A_{s_1,s_5} = \begin{pmatrix} \mathtt{C} & \mathtt{T} & \mathtt{T} & \mathtt{A} & \mathtt{G} & \mathtt{A} & \mathtt{G} & \mathtt{T} \\ - & - & \mathtt{A} & \mathtt{A} & \mathtt{G} & \mathtt{G} & \mathtt{G} & \mathtt{C} \end{pmatrix}$$

$$A_{s_2,s_3} = \begin{pmatrix} \mathtt{A} & \mathtt{C} & \mathtt{C} & \mathtt{C} & \mathtt{G} & \mathtt{T} & \mathtt{G} \\ \mathtt{G} & \mathtt{G} & \mathtt{G} & \mathtt{C} & - & \mathtt{T} & - \end{pmatrix}$$

$$A_{s_2,s_4} = \begin{pmatrix} \mathtt{A} & \mathtt{C} & \mathtt{C} & \mathtt{C} & - & - & \mathtt{G} & \mathtt{T} & \mathtt{G} \\ \mathtt{A} & \mathtt{C} & \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{G} & \mathtt{T} & - \end{pmatrix}$$

$$A_{s_2,s_5} = \begin{pmatrix} \mathtt{A} & \mathtt{C} & \mathtt{C} & \mathtt{C} & \mathtt{G} & \mathtt{T} & \mathtt{G} & - \\ \mathtt{A} & \mathtt{A} & - & - & \mathtt{G} & \mathtt{G} & \mathtt{G} & \mathtt{C} \end{pmatrix}$$

$$A_{s_3,s_4} = \begin{pmatrix} - & - & - & \mathtt{G} & \mathtt{G} & \mathtt{G} & \mathtt{C} & \mathtt{T} \\ \mathtt{A} & \mathtt{C} & \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{G} & \mathtt{T} \end{pmatrix}$$

$$A_{s_3,s_5} = \begin{pmatrix} - & - & \mathtt{G} & \mathtt{G} & \mathtt{G} & \mathtt{C} & \mathtt{T} \\ \mathtt{A} & \mathtt{A} & \mathtt{G} & \mathtt{G} & \mathtt{G} & \mathtt{C} & - \end{pmatrix}$$

$$A_{s_4,s_5} = \begin{pmatrix} \mathtt{A} & \mathtt{C} & \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{G} & \mathtt{T} \\ - & - & \mathtt{A} & \mathtt{A} & \mathtt{G} & \mathtt{G} & \mathtt{G} & \mathtt{C} \end{pmatrix}$$

Aufgabe 3 Multiples Sequenzalignment

1. Für die Sequenzen $s_1, \ldots, s_5$ soll das Center-Star-Alignment A_c unter Verwendung von Einheitskosten berechnet werden. Dafür sind bereits die optimalen paarweisen Alignments für die meisten Sequenzkombinationen auf der Rückseite des vorherigen Blattes angegeben. (15 Punkte)

   (a) Berechne das fehlende optimale paarweise Alignment der Sequenzen s_1 und s_2 und trage es neben den anderen Alignments ein.

	ϵ	A	C	C	C	G	T	G
ϵ								
C								
T								
T								
A								
G								
A								
G								
T								

   (b) Berechne die Gesamtdistanzen d_p für $1 \leq p \leq 5$ und gib die Center-Sequenz s_c an.

   $d_1 =$

   $d_2 =$

   $d_3 =$

   $d_4 =$

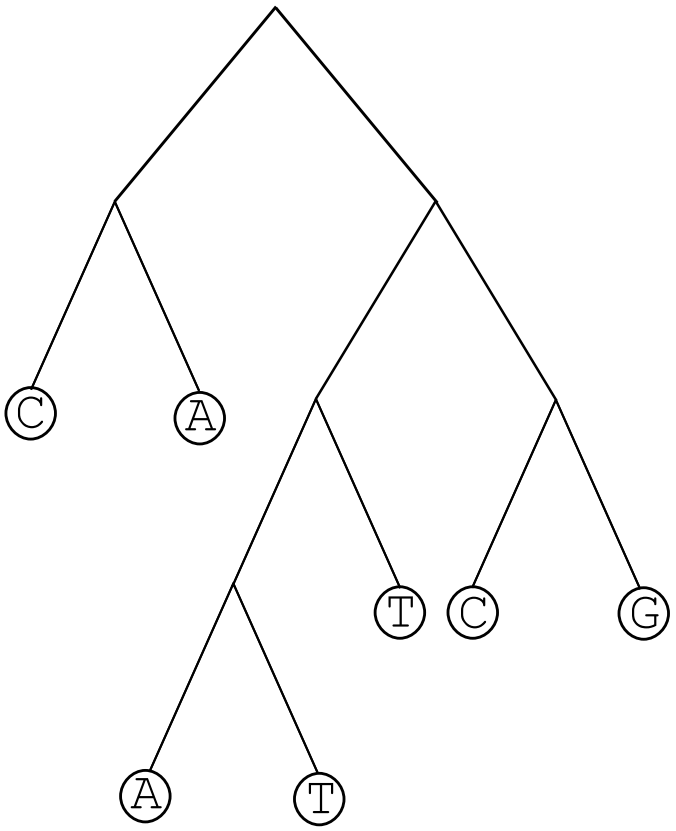
   $d_5 =$

   $s_c =$

   (c) Konstruiere A_c und gib es an.

   (d) Das Alignment A_c ist eine 2-Approximation. Was sagt das über seine Sum-of-Pairs-Kosten $D_{SP}(A_c)$ aus?

2. Gegeben sei der unten dargestellte Baum T. Führe den Fitch-Algorithmus für T aus, indem du für jeden internen Knoten die Menge an potentiellen Beschriftungen angibst und eine mögliche optimale Beschriftung für jeden Knoten markierst. (5 Punkte)



3. Was ist die Idee des Baumalignments? Diskutiere Vor- und Nachteile gegenüber dem Sum-of-Pairs-Alignment. (5 Punkte)

Aufgabe 4 *q*-Gramm-Distanz
Gegeben seien die Sequenzen $x = \texttt{ABBAAABBABB}$ und $y = \texttt{BBBABABAB}$ über dem Alphabet $\Sigma = \{\texttt{A}, \texttt{B}\}$ und $q = 4$.

1. Berechne die q-Gramm-Profile von x und y. Ordne dabei die q-Gramme gemäß ihres Ranges (wobei $\texttt{A} < \texttt{B}$). (10 Punkte)

rank	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$p_4(x)$																
$p_4(y)$																

2. Berechne die q-Gramm-Distanz von x und y. (3 Punkte)

3. Zeichne den de-Bruijn-Teilgraphen $B(x, q)$ und gib alle Sequenzen mit einem q-Gramm-Profil identisch zu dem von x an. (7 Punkte)

4. Nenne die Eigenschaft einer Metrik, die die q-Gramm-Distanz nicht erfüllt und gib ein Beispiel an, das dies belegt. (5 Punkte)