Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Volker Heun

Wintersemester 2022/23
Semestralklausur
23. Februar 2022

Algorithmische Bioinformatik II

Vorname	Name	Matrikelnummer

Reihe	Platz	Unterschrift

Allgemeine Hinweise zur Semestralklausur

- Vor der Prüfung ist diese Seite mit Vornamen, Namen, Matrikelnummer, Reihe und Platz **leserlich mit Druckbuchstaben** zu versehen und zu unterschreiben.
- Bitte **nicht** in *roter* oder *grüner* Farbe bzw. **nicht** mit Bleistift schreiben.
- Der Studentenausweis und ein amtlicher Lichtbildausweis sind bereit zu halten.
- Die reine Bearbeitungszeit beträgt **120 Minuten**.
- Es sind insgesamt 40 Punkte zu erreichen, zum Bestehen sind 17 Punkte nötig.

Viel Erfolg!

Hörsaal verlassen von .... bis .... von .... bis ....

Vorzeitig abgegeben um ....

	Hz	A1	A2	A3	A4	A5	$\sum$
Erstkorrektur							
Nachkorrektur							
Zweitprüfer							

Vorname: ______________ Name: ______________ Matrikelnummer: ____________

Aufgabe 1 (8 Punkte)

Gegeben seien s_1 = GTCTG, s_2 = TGTCT, s_3 = AGTT und s_4 = AGTCT. Konstruiere für diese Sequenzen ein mehrfaches Alignment mit Hilfe der Center-Star-Methode.

d	s_1	s_2	s_3	s_4
s_1	0	2	3	2
s_2	2	0	2	1
s_3	3	2	0	1
s_4	2	1	1	0

Hierbei gilt $w(a,b) = 1$ und $w(a,a) = 0$ für alle $a \neq b \in \overline{\Sigma}$.

Lücken sollen dabei wiederverwendet werden.

Lösungsskizze (nicht ausreichend für die volle Punktzahl)

Wir wählen s_4 als Zentrum, da $\sum_{i=1}^{4} d(s_i, s_c)$ für $c = 4$ mit dem Wert 4 minimal wird.

Für die paarweisen optimalen Alignments gilt:

$$(\bar{s}_1, \bar{s}_2) = \binom{-GTCTG}{TGTCT-} \qquad (\bar{s}_1, \bar{s}_3) = \binom{-GTCTG}{AGT-T-} \qquad (\bar{s}_1, \bar{s}_4) = \binom{-GTCTG}{AGTCT-}$$

$$(\bar{s}_2, \bar{s}_3) = \binom{TGTCT}{AGT-T} \qquad (\bar{s}_2, \bar{s}_4) = \binom{TGTCT}{AGTCT}$$

$$(\bar{s}_3, \bar{s}_4) = \binom{AGT-T}{AGTCT}$$

Wir bauen jetzt das MSA auf. Zuerst werden die paarweisen Alignments von $(\overline{s}_4, \overline{s}_1)$ mit $(\overline{s}_4, \overline{s}_2)$ gemischt:

$$\begin{array}{ccccccc} s_4 & A & G & T & C & T & - \\ s_1 & - & G & T & C & T & G \\ s_2 & T & G & T & C & T & - \end{array}$$

Nun wird noch s_4 mittels $(\overline{s}_4, \overline{s}_3)$ hinzugemischt:

$$\begin{array}{ccccccc} s_4 & A & G & T & C & T & - \\ s_1 & - & G & T & C & T & G \\ s_2 & T & G & T & C & T & - \\ s_3 & A & G & T & - & T & - \end{array}$$

Vorname: ______________ Name: ______________ Matrikelnummer: ____________

Aufgabe 2 (8 Punkte)

Betrachte die Sequenzen $s_1 = \text{TATA}$, $s_2 = \text{TAG}$ und $s_3 = \text{CTA}$. Berechne die C-optimalen Schnittpositionen mit Respekt zu $c_1 = 2$ und die daraus resultierenden mehrfachen Alignments gemäß des Divide-and-Conquer-Alignment-Algorithmus, wobei nach der ersten Rekursion bereits jeweils ein optimales Alignment für die jeweiligen Präfixe bzw. Suffixe berechnet wird.

Für die Kostenfunktion des SP-Distanzmaßes gelte $w(a, a) = 0$ und $w(a, b) = 1$ für alle $a \neq b \in \overline{\Sigma}$.

Lösungsskizze (nicht ausreichend für die volle Punktzahl)

P		T		A		G	
	0		1		2		3
T							
	1		0		1		2
A							
	2		1		0		1
T							
	3		2		1		1
A							
	4		3		2		2

S		T		A		G	
	2		3		4		4
T							
	2		2		3		3
A							
	1		2		2		2
T							
	2		1		1		1
A							
	3		2		1		0

C		T		A		G	
	0		2		4		5
T							
	1		0		2		3
A							
	1		**1**		**0**		**1**
T							
	3		1		0		0
A							
	5		3		1		0

P		C		T		A	
	0		1		2		3
T							
	1		1		1		2
A							
	2		2		2		1
T							
	3		3		2		2
A							
	4		4		3		2

S		C		T		A	
	2		2		3		4
T							
	1		1		2		3
A							
	1		0		1		2
T							
	2		1		0		1
A							
	3		2		1		0

C		C		T		A	
	0		1		3		5
T							
	0		0		1		3
A							
	1		**0**		**1**		**1**
T							
	3		2		0		1
A							
	5		4		2		0

P		C		T		A	
	0		1		2		3
T							
	1		1		1		2
A							
	2		2		2		1
G							
	3		2		3		2

S		C		T		A	
	2		1		2		3
T							
	3		2		1		2
A							
	3		2		1		1
G							
	3		3		1		0

C		C		T		A	
	0		0		2		4
T							
	2		1		0		2
A							
	3		2		1		0
G							
	4		3		2		0

c=2		C		T		A	
	2		**1**		4		6
T							
	4		2		2		4
A							
	4		2		2		**1**
G							
	6		4		4		2

Somit sind $(2, 0, 1)$ und $(2, 2, 3)$ jeweils C-optimaler Schnitt und damit ergeben sich folgende mehrfache Alignments:

$$\begin{array}{cc|ccc} \text{T} & \text{A} & \text{T} & \text{A} & - \\ - & - & \text{T} & \text{A} & \text{G} \\ - & \text{C} & \text{T} & \text{A} & - \end{array} \qquad \begin{array}{ccc|cc} - & \text{T} & \text{A} & \text{T} & \text{A} \\ - & \text{T} & \text{A} & \text{G} & - \\ \text{C} & \text{T} & \text{A} & - & - \end{array}$$

Vorname: ________________ Name: ________________ Matrikelnummer: ______________

Aufgabe 3 (8 Punkte)

Bestimme für die folgenden Blöcke von Sequenzen die zugehörigen Häufigkeiten $H(a,b)$ für die BLOSUM50-Matrix ($a,b \in \{A,B,C\}$).

$$\begin{array}{ll} s_1^{(1)} = \mathrm{BACCA} & s_1^{(2)} = \mathrm{CBBCACB} \\ s_2^{(1)} = \mathrm{BCCCA} & s_2^{(2)} = \mathrm{CCBCABC} \\ s_3^{(1)} = \mathrm{BAABB} & s_3^{(2)} = \mathrm{BCBBABB} \\ s_4^{(1)} = \mathrm{CCACB} & s_4^{(2)} = \mathrm{BBABCBB} \\ s_5^{(1)} = \mathrm{CAACB} & \end{array}$$

Lösungsskizze (nicht ausreichend für die volle Punktzahl)

Die Partitionierung nach mindestens 50%-Sequenzähnlichkeit ergibt:

$$\begin{array}{lcl} \text{Block 1:} \quad [1:4] & = & \{1,2\} \cup [3:5] \\ \text{Block 2:} \quad [1:4] & = & [1:4] \end{array}$$

Dabei sind im ersten Block u.a. folgende Ähnlichkeiten von mindestens 50%: $s_1^{(1)}$ mit $s_2^{(1)}$ sowie $s_3^{(1)}$ mit $s_4^{(1)}$ und $s_4^{(1)}$ mit $s_5^{(1)}$, aber nicht $s_1^{(1)}$ und $s_2^{(1)}$ zu einer übrigen Sequenz im ersten Block.

Im zweiten Block sind u.a. folgende Ähnlichkeiten von mindestens 50%: $s_1^{(2)}$ mit $s_2^{(2)}$, $s_2^{(2)}$ mit $s_3^{(2)}$ und $s_3^{(2)}$ mit $s^4(2)$. Also gelangen alle Sequenzen in einen Cluster.

Somit ist nur Block 1 auszuwerten:

$$\begin{aligned} H(A,A) &= \frac{0+4+0+0+0}{2\cdot 3} = \frac{4}{6} = \frac{2}{3}, \\ H(A,B) &= \frac{0+0+0+0+6}{2\cdot 3} = \frac{6}{6} = 1, \\ H(A,C) &= \frac{0+3+6+0+0}{2\cdot 3} = \frac{9}{6} = \frac{3}{2}, \\ H(B,B) &= \frac{4+0+0+0+0}{2\cdot 3} = \frac{4}{6} = \frac{2}{3}, \\ H(B,C) &= \frac{4+0+0+2+0}{2\cdot 3} = \frac{6}{6} = 1, \\ H(C,C) &= \frac{0+2+0+8+0}{2\cdot 3} = \frac{10}{6} = \frac{5}{3}. \end{aligned}$$

Aufgabe 4 (8 Punkte)

Wir betrachten eine Supermarktkasse, an die sich Personen anstellen. Sei X die Zufallsvariable für die Anzahl Personen, die sich an dieser Kasse pro Stunde anstellen. Die Zufallsvariablen X ist dabei Poisson-verteilt zum Parameter $\lambda > 0$:

$$\mathrm{Ws}[X = N \mid \lambda] = \frac{\lambda^N}{N!}e^{-\lambda}.$$

a) Gib die allgemeinen Formeln sowohl für den Maximum-Likelihood-Schätzer als auch den Maximum-A-Posteriori-Schätzer an.

b) Angenommen an der Kasse haben sich in einer Stunde N Personen angestellt. Bestimme die Likelihood-Funktion für dieses Ergebnis und gib dann den Maximum-Likelihood-Schätzer für λ an.

c) Angenommen an der Kasse haben sich in einer Stunde N Personen angestellt. Bestimme die Posteriori-Wahrscheinlichkeit für dieses Ergebnis bezüglich des Parameterraums $\lambda > 0$, wobei der Prior $f(\lambda) = e^{-\lambda}$ ist und gib dann den Maximum-A-Posteriori-Schätzer für λ an.

Lösungsskizze (nicht ausreichend für die volle Punktzahl)

a) Der Maximum-Likelihood-Schätzer ist gegeben durch $\operatorname{argmax}\{L(\lambda) : \lambda > 0\}$, wobei $L(\lambda) = \mathrm{Ws}[X = N \mid \lambda]$ ist.

Der Maximum-A-Posteriori-Schätzer ist $\operatorname{argmax}\{\mathrm{Ws}[\lambda \mid X = N] : \lambda > 0\}$, wobei nun $\mathrm{Ws}[\lambda \mid X = N]$ die Posteriori-Wahrscheinlichkeit ist mit

$$\mathrm{Ws}[\lambda \mid X = N] = \frac{f(\lambda) \cdot \mathrm{Ws}[X = N \mid \lambda]}{\mathrm{Ws}[X = N]}.$$

b) Es gilt:

$$L(\lambda) = \mathrm{Ws}[X = N \mid \lambda] = \frac{\lambda^N}{N!}e^{-\lambda}.$$

Um das Extremum zu bestimmen, leiten wird die Log-Likelihoodfunktion nach λ ab:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\lambda}\ln(L(\lambda)) &= \frac{\mathrm{d}}{\mathrm{d}\lambda}\ln\left(\frac{\lambda^N}{N!}e^{-\lambda}\right) \\
&= \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\ln\left(\lambda^N\right) - \ln(N!) + \ln\left(e^{-\lambda}\right)\right) \\
&= \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(N \cdot \ln(\lambda) - \ln(N!) - \lambda\right) \\
&= \frac{\mathrm{d}}{\mathrm{d}\lambda}N \cdot \ln(\lambda) - \frac{\mathrm{d}}{\mathrm{d}\lambda}\ln(N!) - \frac{\mathrm{d}}{\mathrm{d}\lambda}\lambda \\
&= \frac{N}{\lambda} - 1.
\end{aligned}$$

Für das Maximum muss $\frac{N}{\lambda} - 1 = 0$ sein, also $\frac{N}{\lambda} = 1$ und somit $\lambda = N$. Um zu prüfen, ob es ein Maximum ist, bilden wir die zweite Ableitung:

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\ln(L(\lambda)) &= \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{N}{\lambda} - 1\right)\\ &= -\frac{N}{\lambda^2}\\ &< 0.\end{aligned}$$

Somit ist $\lambda = N$ der Maximum-Likelihood-Schätzer.

c) Es gilt:

$$\mathrm{Ws}[\lambda \mid X = N] = \frac{f(\lambda)\cdot \mathrm{Ws}[X = N \mid \lambda]}{\mathrm{Ws}[X = N]} = \frac{e^{-\lambda}\cdot \frac{\lambda^N}{N!}\cdot e^{-\lambda}}{\mathrm{Ws}[X = N]} = \frac{\frac{\lambda^N}{N!}\cdot e^{-2\lambda}}{\mathrm{Ws}[X = N]}.$$

Um das Extremum zu bestimmen, logarithmieren wir die Gleichung und leiten nach λ ab:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\lambda}\ln(\mathrm{Ws}[\lambda \mid X = N]) &= \frac{\mathrm{d}}{\mathrm{d}\lambda}\ln\left(\frac{\lambda^N}{N!}\cdot e^{-2\lambda}\cdot\frac{1}{\mathrm{Ws}[X = N]}\right)\\ &= \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\ln\left(\lambda^N\right) - \ln(N!) + \ln\left(e^{-2\lambda}\right) - \ln(\mathrm{Ws}[X = N])\right)\\ &= \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(N\cdot\ln(\lambda) - \ln(N!) - 2\lambda - \ln(\mathrm{Ws}[X = N])\right)\\ &= \frac{\mathrm{d}}{\mathrm{d}\lambda}N\cdot\ln(\lambda) - \frac{\mathrm{d}}{\mathrm{d}\lambda}\ln(N!) - \frac{\mathrm{d}}{\mathrm{d}\lambda}2\lambda - \frac{\mathrm{d}}{\mathrm{d}\lambda}\ln(\mathrm{Ws}[X = N])\\ &= \frac{N}{\lambda} - 2.\end{aligned}$$

Für das Maximum muss $\frac{N}{\lambda} - 2 = 0$ sein, also $\frac{N}{\lambda} = 2$ und somit $\lambda = \frac{N}{2}$. Um zu prüfen, ob es ein Maximum ist, bilden wir die zweite Ableitung:

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\ln(\mathrm{Ws}[\lambda \mid X = N]) &= \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{N}{\lambda} - 2\right)\\ &= -\frac{N}{\lambda^2}\\ &< 0.\end{aligned}$$

Somit ist $\lambda = \frac{N}{2}$ der MAP-Schätzer.

Vorname: ______________ Name: ______________ Matrikelnummer: ______________

Aufgabe 5 (8 Punkte)

Zeige, dass MinBP $\leq_{\text{PTAS}}$ MinEBP. Gib dazu explizit eine PTAS-Reduktion (f, g, α) an und weise die erforderlichen Eigenschaften einer PTAS-Reduktion nach.

MinBinPacking (MinBP)

Eingabe: Eine Folge $(s_1, \ldots, s_n) \in \mathbb{N}^n$ und $B \in \mathbb{N}$.
Lösung: Eine Partition $P = (P_1, \ldots, P_m)$ von $[1 : n]$, so dass $\sum_{i \in P_j} s_i \leq B$ für alle $j \in [1 : m]$ gilt.
Optimum: Minimiere m.

MinEvenBinPacking (MinEBP)

Eingabe: Eine Folge $(s_1, \ldots, s_n) \in \mathbb{N}^n$ und $B \in \mathbb{N}$, wobei s_i für alle $i \in [1 : n]$ eine gerade Zahl ist.
Lösung: Eine Partition $P = (P_1, \ldots, P_m)$ von $[1 : n]$, so dass $\sum_{i \in P_j} s_i \leq B$ für alle $j \in [1 : m]$ gilt.
Optimum: Minimiere m.

Lösungsskizze (nicht ausreichend für die volle Punktzahl)

Da in der Eingabe für MinEBP nur gerade Zahlen s_i erlaubt sind, verdoppeln wir einfach diese Gewichte und auch die Schranke B. Sei also $X = ((s_1, \ldots, s_n), B)$ eine Eingabe für MinBP, dann konstruieren wir eine Eingabe $X' = ((s'_1, \ldots, s'_n), B')$ für MinEBP wie folgt: Für $i \in [1 : n]$ ist $s'_i = 2 \cdot s_i$ und $B' = 2 \cdot B$. Somit ist also $f(X, \epsilon) = X'$

g ordnet einer Lösung von MinEBP, also einer Partition $P = (P_1, \ldots, P_m)$, dieselbe Partition für MinBP zu, d.h. $g(P, X, \epsilon) = P$.

α ist die Identität: $\alpha(\epsilon) = \epsilon$.

f, g und α sind in polynomieller Zeit berechenbar. Für f muss nur jeder Eingabewert mit 2 multipliziert werden (was in der Binärdarstellung durch Anhängen einer 0 möglich ist). Da g und α jeweils die Identität ist, kann die Eingabe jeweils einfach übernommen werden. Somit müssen für diese drei Funktionen im Wesentlichen nur die Eingabewerte kopiert werden, was in linearer Zeit möglich ist.

Nach Konstruktion, also der Verdopplung der Gewichte $s'_i = 2 \cdot s_i$ der einzelnen zu packenden Objekte und der Schranke für die einzelnen Kisten, ist klar, dass für eine Partition $P = (P_1, \ldots, P_m)$ von $[1 : n]$ für die Eingabe $X' = ((s'_1, \ldots, s'_n), B')$ mit $\sum_{i \in P_j} s'_i \leq B'$ dann auch gilt, dass

$$2 \cdot \sum_{i \in P_j} s_i = \sum_{i \in P_j} 2 \cdot s_i = \sum_{i \in P_j} s'_i \leq B' = 2 \cdot B$$

und somit dann natürlich auch $\sum_{i \in P_j} s_i \leq B$ gilt. Damit ist P auch eine zulässige Lösung von $X = ((s_1, \ldots, s_n), B)$. Offensichtlich ist eine Lösung P von X' genau dann optimal, wenn P für X optimal ist

Es bleibt zu zeigen: Für alle $X \in I$, $\epsilon \in \mathbb{Q}^*_+$ und $y' \in S'(f(X, \epsilon))$ (mit $\Gamma_\mu(\bar{X}, \bar{Y}) := \frac{\mu(\bar{X}, \bar{Y})}{\mu^*(\bar{X})}$) gilt: Ist $\Gamma_{\mu'}(f(X, \epsilon), Y') \leq 1 + \alpha(\epsilon)$, dann ist $\Gamma_\mu(X, g(X, Y', \epsilon)) \leq 1 + \epsilon$.

Vorname: ______________ Name: ______________ Matrikelnummer: ____________

Sei $P \in S'(f(X,\epsilon))$ eine Partition von $[1:n]$ für $f(X,\epsilon) = X'$ und sei P^* eine Partition für X' zu einer optimalen Lösung mit m^* Elementen. Gelte hierfür

$$\frac{\mu(P)}{\mu(P^*)} = \Gamma_{\mu'}(f(X,\epsilon), P) \leq 1 + \alpha(\epsilon) = 1 + \epsilon.$$

Also gilt dann nach Definition von g auch $\frac{\mu(g(P,X,\epsilon))}{\mu(P^*)} = \frac{\mu(P)}{\mu(P^*)} \leq 1 + \epsilon$ für MinEBP.