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

Wintersemester 2022/23
Wiederholungsklausur
20. April 2023

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)

Verwende den Algorithmus von Carrillo und Lipman zur Berechnung eines Sequenzen-Alignments zwischen zwei Sequenzen $s = TAGA$ und $t = CAA$. Hierzu sind für das Distanzmaß die **Gap-Kosten** von 3 und **Mismatch-Kosten** von 2 zu verwenden. Die **globale obere Schranke** für die Distanz von s und t ist mit 6 vorgegeben.

Hinweis: In der Vorlesung wurde dies für 3 oder mehr Sequenzen besprochen, natürlich funktioniert das Verfahren auch mit nur 2 Sequenzen.

Gib die kombinierte **Prefix-/Suffix-Matrix** $P + S$ und dessen Herleitung an und **markiere alle Zellen**, die in den **Heap** aufgenommen wurden. Gib dabei ebenfalls die Berechnung der verwendeten **obere Schranke** im Relevanz-Test für das Sequenzpaar (s, t) an.

Lösungsskizze (nicht ausreichend für die volle Punktzahl)

P		C	A	A
	0	3	6	9
T	3	2	5	8
A	6	5	2	5
G	9	8	5	4
A	12	11	8	5

S		C	A	A
	5	6	9	12
T	4	3	6	9
A	5	2	3	6
G	6	3	0	3
A	9	6	3	0

$P + S$		C	A	A
	5	*9*	15	21
T	*7*	*5*	*11*	17
A	11	*7*	*5*	*11*
G	15	11	*5*	*7*
A	21	17	*11*	*5*

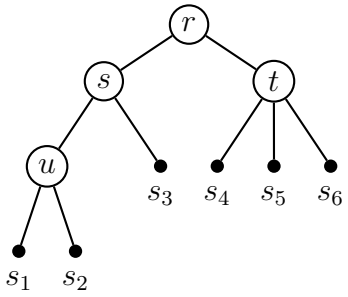
Die in den Heap aufgenommen Elemente sind rot bzw. kursiv dargestellt.

Die im Relevanz-Test verwendete obere Schranke für s und t lautet:

$$C_{s,t} := C - \sum_{(i,j) \neq (s,t)} d(s_i, s_j) = C - 0 = 6 - 0 = 6.$$

Vorname: _____ Name: _____ Matrikelnummer: _____

Aufgabe 2 (8 Punkte)



d	s_1	s_2	s_3	s_4	s_5	s_6
s_1	0	1	3	2	2	3
s_2		0	2	2	3	3
s_3			0	3	3	3
s_4				0	1	1
s_5					0	2
s_6						0

Berechne für den rechts angegebenen Baum und der zugehörigen Distanzmatrix **ein** optimales geliftetes phylogenetisches mehrfaches Sequenzen-Alignment gemäß der dynamischen Programmierung. Hierzu müssen nur die gelifteten Sequenzen für jeden inneren Knoten angegeben werden (nicht das mehrfache Sequenzen-Alignment selbst). Es müssen nur legale Paare berücksichtigt werden.

Lösungsskizze (nicht ausreichend für die volle Punktzahl)

$$
\begin{aligned}
D[u,s_1] &= (d(s_1,s_1)+0)+(d(s_1,s_2)+0) = 0+1 = 1\\
D[u,s_2] &= (d(s_2,s_1)+0)+(d(s_2,s_2)+0) = 1+0 = 1\\
D[t,s_4] &= (d(s_4,s_4)+0)+(d(s_4,s_5)+0)+(d(s_4,s_6)+0) = 0+1+1 = 2\\
D[t,s_5] &= (d(s_5,s_4)+0)+(d(s_5,s_5)+0)+(d(s_5,s_6)+0) = 1+0+2 = 3\\
D[t,s_6] &= (d(s_6,s_4)+0)+(d(s_6,s_5)+0)+(d(s_6,s_6)+0) = 1+2+0 = 3\\
D[s,s_1] &= (d(s_1,s_1)+D[u,s_1])+(d(s_1,s_3)+0) = 1+3 = 4\\
D[s,s_2] &= (d(s_2,s_2)+D[u,s_2])+(d(s_2,s_3)+0) = 1+2 = 3\\
D[s,s_3] &= \min\{d(s_3,s_1)+D[u,s_1], d(s_3,s_2)+D[u,s_2]\}+(d(s_3,s_3)+0) = 3+0 = 3\\
D[r,s_1] &= (d(s_1,s_1)+D[s,s_1])\\
&\quad + \min\{d(s_1,s_4)+D[t,s_4], d(s_1,s_5)+D[t,s_5], d(s_1,s_6)+D[t,s_6]\} = 4+4 = 8\\
D[r,s_2] &= (d(s_2,s_2)+D[s,s_2])\\
&\quad + \min\{d(s_2,s_4)+D[(t,s_4], d(s_2,s_5)+D[t,s_5], d(s_2,s_6)+D[t,s_6]\} = 3+4 = 7\\
D[r,s_3] &= (d(s_3,s_3)+D[s,s_3])\\
&\quad + \min\{d(s_3,s_4)+D[t,s_4], d(s_3,s_5)+D[t,s_5], d(s_3,s_6)+D[t,s_6]\} = 3+5 = 8\\
D[r,s_4] &= \min\{d(s_4,s_1)+D[s,s_1], d(s_4,s_2)+D[s,s_2], d(s_4,s_3)+D[s,s_3]\}\\
&\quad +(d(s_4,s_4)+D[t,s_4]) = 5+2 = 7\\
D[r,s_5] &= \min\{d(s_5,s_1)+D[s,s_1], d(s_5,s_2)+D[s,s_2], d(s_5,s_3)+D[s,s_]\}\\
&\quad +(d(s_5,s_5)+D[t,s_5]) = 6+3 = 9\\
D[r,s_6] &= \min\{d(s_6,s_1)+D[s,s_1], d(s_6,s_2)+D[s,s_2], d(s_6,s_3)+D[s,s_3]\}\\
&\quad +(d(s_6,s_6)+D[t,s_6]) = 6+3 = 9
\end{aligned}
$$

Damit sind die Lösungen:

$$(r,s,t,u) \in \{(s_2,s_2,s_4,s_2),(s_4,s_2,s_4,s_2)\}.$$

Vorname: _____ Name: _____ Matrikelnummer: _____

Aufgabe 3 (8 Punkte)

Sei X die Zufallsvariable, die zählt, wie oft man eine Münze werfen muss, bis zum ersten Mal Kopf erscheint. Für $k \in \mathbb{N}$ gilt $\text{Ws}[X = k] = (1-p)^{k-1}p$, wenn p die Wahrscheinlichkeit ist, dass bei einem Wurf Kopf erscheint.

Angenommen, bei einem Versuch wird 6-mal geworfen bis zum ersten Mal Kopf erscheint. Überprüfe mit Hilfe des Likelihood-Ratio-Tests die Null-Hypothese ($p = \frac{1}{2}$) gegen die Alternativ-Hypothese ($p = \frac{1}{4}$) für das Signifikanz-Niveau $\alpha = 0.02$.

Hilfe: Für $F(r,p) = \sum_{i=r}^{\infty}(1-p)^{i-1}p = (1-p)^{r-1}$ gilt:

r	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$F(r,1/4)$	1.00	0.75	0.56	0.42	0.32	0.24	0.18	0.13	0.10	0.08	0.06	0.04	0.03	0.02	0.02	0.01
$F(r,1/2)$	1.00	0.50	0.25	0.12	0.06	0.03	0.02	0.01	0.004							
$F(r,3/4)$	1.00	0.25	0.06	0.02	0.004											

Lösungsskizze (nicht ausreichend für die volle Punktzahl)

Für die Likelihood-Ratio gilt:

$$\Lambda(k) = \frac{\left(\frac{1}{2}\right)^{k-1}\frac{1}{2}}{\left(\frac{3}{4}\right)^{k-1}\frac{1}{4}} = \frac{2^{k-1}\cdot 2}{3^{k-1}} = 3\cdot\left(\frac{2}{3}\right)^{k}.$$

Für das Signifikanz-Niveau α müssen wir jetzt λ so wählen, dass

$$\text{Ws}\left[\Lambda(X) \leq \lambda\right] = \alpha = 0.02.$$

Es gilt genau dann $\Lambda(k) \leq \lambda$, wenn (beachte hierbei, dass $\log(2/3) < 0$)

$$k \geq \frac{\log(\lambda/3)}{\log(2/3)} =: x.$$

Wir müssen also λ (bzw. x) so bestimmen, dass $\text{Ws}\left[k \geq x\right] = \alpha = 0.02$ mit $x = \frac{\log(\lambda/3)}{\log(2/3)}$. Dies gilt nach der Tabelle für $x \approx 7$ (für die Wahrscheinlichkeit wird die Verteilung der Nullhypothese zugrunde gelegt, also $F(r,1/2)$). Damit gilt $\lambda \approx 3\cdot\left(\frac{2}{3}\right)^7$.

Nun berechnen wir $\Lambda(6)$:

$$\Lambda(6) = 3\cdot\left(\frac{2}{3}\right)^6 > 3\cdot\left(\frac{2}{3}\right)^7 \approx \lambda.$$

Also können wir die Nullhypothese nicht verwerfen.

Vorname: _____ Name: _____ Matrikelnummer: _____

Aufgabe 4 (8 Punkte)

Sei $Z_n \in [0:1]$ für $n \in \mathbb{N}_0$ eine Zufallsvariable, die den Ausgang des n-ten Wurfs einer Münze beschreibt (wobei beide Ausgänge gleichwahrscheinlich sind).

Betrachte die Markov-Kette $X_0 := Z_0$ und für $n > 0$

$$X_n := Z_n + 2 \cdot Z_{n-1}.$$

a) Bestimme das zu $(X_n)_{n\in\mathbb{N}}$ gehörige Markov-Modell (Q, P, π).

b) Zeige, dass das Markov-Modell aus a) ergodisch ist.

c) Gib die stationäre Verteilung des Markov-Modells aus a) an.

Lösungsskizze (nicht ausreichend für die volle Punktzahl)

a) Es gilt $Z_{n-1} = (Z_{n-1} + 2 \cdot Z_{n-2}) \bmod 2$. Somit folgt

$$\begin{aligned} X_n &= Z_n + 2 \cdot Z_{n-1} \\ &= Z_n + 2 \cdot ((Z_{n-1} + 2 \cdot Z_{n-2}) \bmod 2) \\ &= Z_n + 2 \cdot (X_{n-1} \bmod 2) \end{aligned}$$

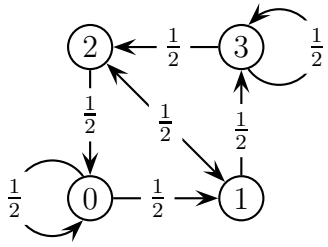
und $(X_n)_{n\in\mathbb{N}}$ ist eine Markov-Kette.

Wir wählen $Q = \{q_0, q_1, q_2, q_3\}$, wobei i in q_i den Wertebereich $\{0, 1, 2, 3\}$ von X_n durchläuft. Dann ist $\pi = (0.5, 0.5, 0.0, 0.0)$ und

$$P = \begin{pmatrix} 0.5 & 0.5 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.5 \end{pmatrix}.$$

b) Eine durch ein Markov-Modell induzierte Markov-Kette ist ergodisch, wenn sie irreduzibel und aperiodisch ist

Der Graph der Übergangswahrscheinlichkeiten P ist stark zusammenhängend, was man an der folgenden Abbilding leicht sieht. Also kann jeder Zustand von jedem anderen Zustand mit positiver Wahrscheinlichkeit in endlich vielen Schritten erreicht werden, somit ist die Markov-Kette irreduzibel.



Die Periode d_q eines Zustands $q \in Q$ ist definiert als

$$d_q := \text{ggT}\left\{ k \in \mathbb{N} : \begin{array}{lll} \exists (q_0, \ldots, q_k) \in Q^{k+1} & \wedge & q_0 = q_k = q \\ \wedge \quad \forall i \in [0:k-1] p_{q_i, q_{i+1}} > 0 & & \end{array} \right\}.$$

Offensichtlich ist $d_0 = d_3 = 1$, da 1 in die ggT-Berechnung einfließt. Für d_1 gehen die Pfade (q_1, q_2, q_1) und (q_1, q_3, q_2, q_1) ein. Da $\text{ggT}(2,3) = 1$ ist $d_1 = 1$. Für d_2 gehen die Pfade (q_2, q_1, q_2) und (q_2, q_0, q_1, q_2) ein. Da $\text{ggT}(2,3) = 1$ ist $d_1 = 1$. Somit sind alle Zustände aperiodisch und damit auch die Markov-Kette.

c) Die stationäre Verteilung ist (wie man leicht nachrechnet) $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$:

$$\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) \cdot \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix} = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right).$$

Vorname: _____ Name: _____ Matrikelnummer: _____

Aufgabe 5 (8 Punkte)

Betrachte das unten angegebene MAX2SAT-Problem.

a) Zeige, dass MAX2SAT $\in \mathcal{NPO}$.

b) Konstruiere eine polynomielle 2-Approximation für MAX2SAT.

Beachte: Das zu MAX2SAT gehörige Entscheidungsproblem ist $\mathcal{NP}$-hart.

Hinweise: Versuche einen Greedy-Algorithmus über die Variablen, die in F vorkommen, zu entwerfen. Korrektheitsbeweise und Laufzeitanalyse nicht vergessen.

MAX2SAT

Eingabe: Ein Boolesche Formel $F = \bigwedge_{i=1}^{k} C_i$ in konjunktiver Normalform, wobei jede Klausel aus genau 2 verschiedenen Literalen besteht.
Lösung: Ein Belegung $B : V(F) \to \mathbb{B}$
Optimum: Maximiere $|\{i \in [1:k] : B(C_i) = 1\}$.

Lösungsskizze (nicht ausreichend für die volle Punktzahl)

a) Zuerst muss in polynomieller Zeit entscheidbar sein, ob die Eingabe eine Boolesche Formel in 2-konjunktiver Normalform beschreibt. Für die Repräsentation von Booleschen Formeln wählen wir eine Zeichenreihe mit geeigneten Trennzeichen, z.B. $(,), \wedge, \vee, \neg$, die Variablen werden dabei als Zeichenreihe aus $\{0,1\}^*$ dargestellt (binäre Kodierung des Index von x_i). Dann kann leicht überprüft werde, ob die Eingabe eine Boolesche Formel in 2-konjunktiver Normalform ist. Neben der Syntax muss nur noch geprüft werden, dass genau zwei verschiedene Literale pro Klausel vorkommen. Eine Boolesche Formel mit k Klauseln mit je zwei Literalen besitzt dann offensichtlich eine Eingabegröße von $s(k,n) = \Omega(k \log(n))$ (wobei n der maximale Index einer Booleschen Variable in der Eingabe ist). Mit dieser Realisierungen von Booleschen Formeln ist das in Zeit $O(k \log(n)) = O(s(k,n))$ möglich.

   Weiter muss gezeigt werden, dass eine Lösung polynomiell in der Eingabegröße beschränkt ist. Für jede Lösung B genügt eine Angabe der Belegung der maximal $2k$ vorkommenden Variablen, somit ist $\|B\| = O(k \log(n)) = O(s(k,n))$.

   Weiterhin muss das Maß einer Lösung in polynomieller Zeit berechenbar sein. Hier ist das Maß einer Lösung $B : V(F) \to \mathbb{B}$ die Anzahl erfüllter Klauseln, Dies ist durch Einsetzen in Zeit $O(\log(n))$ pro Klausel möglich. Insgesamt ist der Zeitbedarf also $O(k \log(n)) = O(s(k,n)$, das Maß also in polynomieller Zeit berechenbar.

b) O.B.d.A. nehmen wir an, dass Klauseln der Art $(x \vee \bar{x})$ für eine Variable $x \in V(F)$ nicht vorkommen. Diese sind ja unabhängig von der Belegung immer erfüllt und verbessern nur die Approximationsgüte.

   Sei $x \in V(F)$ eine beliebige Variable und komme diese in genau k' Klauseln vor. Wir belegen nun x so, dass mindestens die Hälfte dieser k' Klauseln erfüllt sind. Nach unserer Annahme gilt für jede solcher Klausel C, dass entweder $C|_{x=0}$ oder $C|_{x=1}$ erfüllt ist (da x in C vorkommt). Wir eliminieren jetzt alle diese k' Klauseln und bekommen so eine Boolesche Formel F', mit der wir dann genauso verfahren.

Offensichtlich wird eine Belegung konstruiert, so dass mindestens die Hälfte der Klauseln erfüllt wird. Da es im besten Falle eine Belegung gibt, die alle Klauseln erfüllt, haben wir mindestens eine 2-Approximation.

Die Laufzeit ist sicherlich für jede Variable in Zeit $O(s(k,n))$ möglich, also insgesamt in Zeit $O(k \cdot s(k,n) = O((s(k,n)^2)$