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

Wintersemester 2016/17
Semestralklausur
16. Februar 2017

Algorithmische Bioinformatik II

Vorname	Name	Matrikelnummer

Reihe	Platz	Unterschrift

Hiermit stimme ich einer Veröffentlichung meines Klausurergebnisses dieser Semestralklausur unter Verwendung meiner Matrikelnummer im Internet zu. *Ja* ☐ *Nein* ☐

_____
(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 $s_1 = \text{TACGG}$, $s_2 = \text{ACTGG}$, $s_3 = \text{ATGG}$ und $s_4 = \text{ATGGC}$. 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	2	3
s_2	2	0	1	2
s_3	2	1	0	1
s_4	3	2	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

Wir wählen s_3 als Zentrum, da $\sum_{i=1}^{4} d(s_i, s_c)$ für $c = 3$ mit dem Wert 4 minimal wird.

Für die paarweisen optimalen Alignments gilt:

$$(\bar{s}_1, \bar{s}_2) = \binom{TAC-GG}{-ACTGG} \qquad (\bar{s}_1, \bar{s}_3) = \binom{TACGG}{-ATGG} \qquad (\bar{s}_1, \bar{s}_4) = \binom{TACGG-}{-ATGCC}$$

$$(\bar{s}_2, \bar{s}_3) = \binom{ACTGG}{A-TGG} \qquad (\bar{s}_2, \bar{s}_4) = \binom{ACTGG-}{A-TGGC}$$

$$(\bar{s}_3, \bar{s}_4) = \binom{ATGG-}{ATGGC}$$

Wir bauen jetzt das MSA auf. Zuerst werden die paarweisen Alignments von $(\overline{s}_3, \overline{s}_1)$ mit $(\overline{s}_3, \overline{s}_2)$ gemischt:

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

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

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

Vorname: _____ Name: _____ Matrikelnummer: _____

Aufgabe 2 (8 Punkte)

Betrachte die Sequenzen $s_1 = \text{CAT}$, $s_2 = \text{ATA}$ und $s_3 = \text{TCA}$. Berechne die C-optimalen Schnittpositionen mit Respekt zu $c_1 = 1$ 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

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

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

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

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

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

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

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

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

C		T	C	A
	0	1	2	4
A	0	0	1	2
T	2	0	0	2
A	4	2	1	0

c=1		T	C	A
	2	2	2	6
A	3	2	2	5
T	5	2	**1**	5
A	9	6	4	5

Somit ist $(1, 2, 2)$ C-optimaler Schnitt und damit ergeben sich folgende mehrfache Alignments:

–	C	A	T
A	T	A	–
T	C	A	–

Vorname: _____ Name: _____ Matrikelnummer: _____

Aufgabe 3 (8 Punkte)

Bestimme für die folgenden Blöcke von Sequenzen die zugehörigen Häufigkeiten $H(a, b)$ für die BLOSUM50-Matrix.

$$
\begin{array}{ll}
s_1^{(1)} = \text{ACCCA} & s_1^{(2)} = \text{CBBCACBAC} \\
s_2^{(1)} = \text{ABACA} & s_2^{(2)} = \text{CCBCABCAA} \\
s_3^{(1)} = \text{CBABA} & s_3^{(2)} = \text{BCBBABBAB} \\
s_4^{(1)} = \text{ACABB} & s_4^{(2)} = \text{CBACABBBA}
\end{array}
$$

Lösungsskizze

Die Partitionierung nach mindestens 50%-Sequenzähnlichkeit ergibt:

$$
\begin{array}{lrcl}
\text{Block 1:} & [1:4] & = & [1:3] \cup \{4\} \\
\text{Block 2:} & [1:4] & = & [1:4]
\end{array}
$$

Dabei sind im zweiten Block u.a. folgende Ähnlichkeiten von mindestens 55%: $s_1^{(2)}$ mit $s_2^{(2)}$, $s_2^{(2)}$ mit $s_3^{(3)}$ und $s_1^{(2)}$ mit $s_4^{(2)}$.

Somit ist nur Block 1 auszuwerten:

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

Vorname: _____ Name: _____ Matrikelnummer: _____

Aufgabe 4 (8 Punkte)

Sei $Z_n \in \{0,1\}$ eine Zufallsvariable, die den Ausgang des n-ten Wurfs einer Münze beschreibt (wobei die beiden Ausgänge gleichwahrscheinlich sind).

Betrachte $X_0 := 2 \cdot Z_0$ und für $n > 0$

$$X_n := 3 \cdot Z_n + 2 \cdot Z_{n-1}.$$

a) Begründe, dass $(X_n)_{n\in\mathbb{N}}$ eine Markov-Kette erster Ordnung ist.

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

c) Gib für das Markov-Modell aus b) die stationäre Verteilung an.

d) Begründe, dass das Markov-Modell aus b) ergodisch ist

Lösungsskizze

a) Es gilt für $n \geq 1$, dass $Z_{n-1} = (3 \cdot Z_{n-1} + 2 \cdot Z_{n-2}) \bmod 2$ und somit:

$$\begin{aligned} X_n &= 3 \cdot Z_n + 2 \cdot Z_{n-1} \\ &= 3 \cdot Z_n + 2 \cdot ((3 \cdot Z_{n-1} + 2 \cdot Z_{n-2}) \bmod 2) \\ &= 3 \cdot Z_n + 2 \cdot (X_{n-1} \bmod 2) \end{aligned}$$

b) Wir wählen $Q = \{q_0, q_2, q_3, q_5\}$, wobei i in q_i den Wertebereich $\{0,2,3,5\}$ von X_n durchläuft. Dann ist $\pi = (0.5, 0.5, 0.0, 0.0)$ und

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

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

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

Der Graph der Übergangswahrscheinlichkeiten P ist zusammenhängend, 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 $d_i := d_{q_i}$ eines Zustands $q_i \in Q$ ist definiert als

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

Offensichtlich ist $d_0 = d_5 = 1$, da 1 in die ggT-Berechnung einfließt. Für d_2 gehen die Pfade (q_2, q_3, q_2) und (q_2, q_3, q_5, q_2) ein. Da $\mathrm{ggT}(2,3) = 1$ ist $d_2 = 1$. Für d_3 gehen die Pfade (q_3, q_2, q_3) und (q_3, q_2, q_0, q_3) ein. Da $\mathrm{ggT}(2,3) = 1$ ist $d_3 = 1$. Somit sind alle Zustände aperiodisch und damit auch die Markov-Kette.

Vorname: _____ Name: _____ Matrikelnummer: _____

Aufgabe 5 (8 Punkte)

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

MinEdgeCover (MinEC)

Eingabe: Ein ungerichteter Graph $G = (V, E)$ mit $E = \{e_1, \ldots, e_m\}$.
Lösung: Eine Teilmenge $E' \subseteq E$ der Kanten, so dass jeder Knoten mindestens ein Endpunkt einer Kante in E' ist, d.h. $\forall v \in V : \exists e \in E' : e \cap \{v\} \neq \emptyset$.
Optimum: Minimiere $|E'|$.

MinSAT

Eingabe: Eine Boolesche Formel F über $V(F) = X$.
Lösung: Eine erfüllende Belegung $B : X \to \mathbb{B}$, d.h. $B(F) = 1$.
Optimum: Minimiere $\mu(B) = |\{x \in X : B(x) = 1\}|$.

Achtung: Bei MinSAT ist **nicht** die Anzahl erfüllter Klauseln zu minimieren, sondern die Anzahl der auf wahr gesetzten Variablen.

Lösungsskizze

Somit muss also $f(G, \epsilon)$ eine Boolesche Formel F sein. Dazu ordnen wir jeder Kante $e_i \in E = \{e_1, \ldots, e_m\}$ eine Variable $x_i \in X = \{x_1, \ldots, x_m\}$ zu und für jeden Knoten $v \in V$ erzeugen wir eine Klausel $\{e_{i_1}, \ldots, e_{i_{\ell_i}}\}$, wobei $\{e_{i_1}, \ldots, e_{i_{\ell_i}}\}$ alle Kanten sind, die v_i als einen Endpunkt enthalten (d.h. $e_i \cap v_{i_j} \neq \emptyset$ für alle $i_j \in [1 : \ell_i]$. F ist dann die Konjunktion dieser Klauseln.

g ordnet einer erfüllenden Belegung $B : X \to \mathbb{B}$ eine Menge E' wie folgt zu:

$$E' := \{e_i \in E : B(x_i) = 1\}.$$

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

f, g und α sind offensichtlich in polynomieller Zeit berechenbar.

Offensichtlich gilt, dass für eine erfüllende Belegung B für F jede Klausel $\{e_{i_1}, \ldots, e_{i_{\ell_i}}\}$ erfüllt ist. Dies impliziert dann natürlich einen Edge-Cover für G vermöge g, da dann jede Klausel durch eine erfüllende Variable abgedeckt ist, d.h. jeder Knoten ist mindestens Endpunkt einer Kante in E'. Umgekehrt kann man jedem Edge-Cover eine erfüllende Belegung für F zuordnen.

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 $\Gamma_\mu(x, g(x, y', \epsilon)) \leq 1 + \epsilon$.

Sei $B \in S'(f(G, \epsilon))$ eine erfüllende Belegung für $f(G, \epsilon) = F$ und sei B^* eine erfüllende Belegung für F, die eine minimale Anzahl von Variablen auf `true` setzt. Gelte hierfür

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

Also gilt dann nach Definition von g

$$\frac{|E'|}{|E^*|} = \frac{\mu(B)}{\mu(B^*)} \leq 1 + \epsilon.$$