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

Wintersemester 2016/17
Wiederholungsklausur
03. Mai 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 ...........

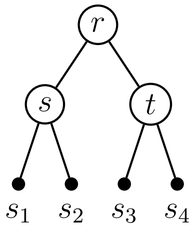
Vorzeitig abgegeben um ...........................

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

Vorname: ____________ Name: ____________ Matrikelnummer: ____________

## Aufgabe 1 (8 Punkte)

Berechne für den rechts angegebenen vollständigen Baum ein optimales **uniform** geliftetes Alignment mittels der dynamischen Programmierung.




| $d$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ |
|---|---|---|---|---|
| $s_1$ | 0 | 3 | 1 | 3 |
| $s_2$ | | 0 | 4 | 5 |
| $s_3$ | | | 0 | 2 |
| $s_4$ | | | | 0 |

**Lösungsskizze**

$$\begin{aligned}
D[s, s_1] &= (d(s_1, s_1){+}0){+}(d(s_1, s_2){+}0) = 0{+}3 = 3 \\
D[s, s_2] &= (d(s_2, s_1){+}0){+}(d(s_2, s_2){+}0) = 3{+}0 = 3 \\
D[t, s_3] &= (d(s_3, s_3){+}0){+}(d(s_3, s_4){+}0) = 0{+}2 = 2 \\
D[t, s_4] &= (d(s_4, s_3){+}0){+}(d(s_4, s_4){+}0) = 2{+}0 = 2 \\
D[r, s_1] &= (d(s_1, s_1){+}D[s, s_1]){+}(d(s_1, s_3){+}D[t, s_3]) = (0{+}3) + (1{+}2) = 6 \\
D[r, s_2] &= (d(s_2, s_2){+}D[s, s_2]){+}(d(s_2, s_4){+}D[t, s_4]) = (0{+}3) + (5{+}2) = 10 \\
D[r, s_3] &= (d(s_3, s_1){+}D[s, s_1]){+}(d(s_3, s_3){+}D[t, s_3]) = (1{+}3) + (0{+}2) = 6 \\
D[r, s_4] &= (d(s_4, s_2){+}D[s, s_2]){+}(d(s_4, s_4){+}D[t, s_4]) = (5{+}3) + (0{+}2) = 10
\end{aligned}$$

Damit sind die Lösungen:

$$\begin{aligned}
r &= s_1 \mid s_3 \\
s &= s_1 \\
t &= s_3
\end{aligned}$$

Vorname: ______________ Name: ______________ Matrikelnummer: ____________

## Aufgabe 2 (8 Punkte)

Verwende den Algorithmus von Carrillo und Lipman zur Berechnung eines Sequenzen-Alignments zwischen zwei Sequenzen $s = ATTG$ und $t = TG$. 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 7 vorgegeben.

*Hinweis:* In der Übung wurde dies für 3 oder mehr Sequenzen implementiert, 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

| $P$ | | T | G |
|---|---|---|---|
| | 0 | 3 | 6 |
| A | 3 | 2 | 5 |
| T | 6 | 3 | 4 |
| T | 9 | 6 | 5 |
| G | 12 | 9 | 6 |

| $S$ | | T | G |
|---|---|---|---|
| | 6 | 9 | 12 |
| A | 3 | 6 | 9 |
| T | 0 | 3 | 6 |
| T | 3 | 0 | 3 |
| G | 6 | 3 | 0 |

| $P+S$ | | T | G |
|---|---|---|---|
| | *6* | *12* | 18 |
| A | *6* | *8* | 14 |
| T | *6* | *6* | *10* |
| T | *12* | *6* | *8* |
| G | 18 | *12* | *6* |

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

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

$$C_{s,t} := C - \sum_{(i,j)\neq(s,t)} d(s_i, s_j) = C - 0 = 7 - 0 = 7.$$

Vorname: ____________ Name: ____________ Matrikelnummer: ____________

## Aufgabe 3 (8 Punkte)

Betrachte die folgende Matrix mit den Werten $n_{a,b}$ für die PAM-Matrizen.

| $n_{a,b}$ | $A$ | $B$ | $C$ |
|---|---|---|---|
| $A$ | 0 | 100 | 200 |
| $B$ | 100 | 0 | 200 |
| $C$ | 200 | 200 | 0 |

a) Gib die allgemeine Formel für 1-PAM-Matrizen für die Werte $w(\cdot,\cdot)$ an:

b) Berechne die Werte $w(\cdot,\cdot)$ der zugehörigen 1-PAM.

c) Was bedeutet PAM? Was ist der prinzipielle Unterschied zwischen einer 2-PAM und einer Substitutionsmatrix basierend auf einer Mutationsrate von 2%?

*Hinweis:* Gib auch die Zwischenergebnisse an. Am Ende müssen die Logarithmen nicht ausgerechnet werden, auftretende Brüche aber weitestgehend gekürzt werden.

**Lösungsskizze**

a) Es gilt für $a \neq b \in \Sigma$.

$$w(a,a) = \log\left(\frac{p_{a,a}}{p_a}\right) \qquad \text{und} \qquad w(a,b) = \log\left(\frac{p_{a,b}}{p_b}\right)$$

mit $p_a = \frac{1}{2n}\sum_{b\in\Sigma} n_{a,b}$. und $p_{a,b} = \frac{n_{a,b}}{200\cdot n\cdot p_a}$ sowie $p_{a,a} = 1 - \sum_{\substack{b\in\Sigma\\ b\neq a}} p_{a,b}$.

b) Es gilt zunächst

$$\begin{aligned}
p_A &= \frac{100+200}{1000} = \frac{3}{10}\\
p_B &= \frac{100+200}{1000} = \frac{3}{10}\\
p_C &= \frac{200+200}{1000} = \frac{4}{10}\\
p_{A,B} &= \frac{100}{200\cdot 500\cdot \frac{3}{10}} = \frac{1}{300}\\
p_{A,C} &= \frac{200}{200\cdot 500\cdot \frac{3}{10}} = \frac{1}{150}\\
p_{B,A} &= \frac{100}{200\cdot 500\cdot \frac{3}{10}} = \frac{1}{300}\\
p_{B,C} &= \frac{200}{200\cdot 500\cdot \frac{3}{10}} = \frac{1}{150}\\
p_{C,A} &= \frac{200}{200\cdot 500\cdot \frac{4}{10}} = \frac{1}{200}\\
p_{C,B} &= \frac{200}{200\cdot 500\cdot \frac{4}{10}} = \frac{1}{200}
\end{aligned}$$

$$\begin{aligned}
p_{A,A} &= 1 - \frac{1}{300} - \frac{1}{150} = \frac{147}{150} = \frac{49}{50}\\
p_{B,B} &= 1 - \frac{1}{300} - \frac{1}{150} = \frac{147}{150} = \frac{49}{50}\\
p_{C,C} &= 1 - \frac{1}{200} - \frac{1}{200} = \frac{198}{200} = \frac{99}{100}
\end{aligned}$$

$$\begin{aligned}
w(A,B) &= \log\left(\frac{\frac{1}{300}}{\frac{3}{10}}\right) = \log\left(\frac{1}{90}\right)\\
w(A,C) &= \log\left(\frac{\frac{1}{150}}{\frac{4}{10}}\right) = \log\left(\frac{1}{60}\right)\\
w(B,A) &= \log\left(\frac{\frac{1}{300}}{\frac{3}{10}}\right) = \log\left(\frac{1}{90}\right)\\
w(B,C) &= \log\left(\frac{\frac{1}{150}}{\frac{4}{10}}\right) = \log\left(\frac{1}{60}\right)\\
w(C,A) &= \log\left(\frac{\frac{1}{200}}{\frac{3}{10}}\right) = \log\left(\frac{1}{60}\right)\\
w(C,B) &= \log\left(\frac{\frac{1}{200}}{\frac{3}{10}}\right) = \log\left(\frac{1}{60}\right)\\
w(A,A) &= \log\left(\frac{1-\frac{3}{300}}{\frac{3}{10}}\right) = \log\left(\frac{297}{90}\right) = \log\left(\frac{33}{10}\right)\\
w(B,B) &= \log\left(\frac{1-\frac{3}{300}}{\frac{3}{10}}\right) = \log\left(\frac{297}{90}\right) = \log\left(\frac{33}{10}\right)\\
w(C,C) &= \log\left(\frac{1-\frac{2}{200}}{\frac{4}{10}}\right) = \log\left(\frac{198}{80}\right) = \log\left(\frac{99}{40}\right)
\end{aligned}$$

c) PAM steht für *Percent Accepted Mutation* oder auch *Point Accepted Mutation*

Bei einer Substitutionsmatrix mit einer Mutationsrate von 2% wird zugrunde gelegt, dass eine Mutation einer Aminosäure mit einer Wahrscheinlichkeit von genau 2% auftritt. Bei einer 2-PAM werden hingene in einem Schritt zwei Mutationen mit Wahrscheinlichkeit von 1% zugrunde gelegt, so dass die zugrunde gelegte Mutationsrate kleiner als 2% ist, da hierbei auch das Mutationspaar ($A \rightarrow B, B \rightarrow A$) nach außen neutral aussieht.

Vorname: ______________ Name: ______________ Matrikelnummer: ____________

## Aufgabe 4 (8 Punkte)

Wir betrachten vier verschiedene Typen von Münzen $A_i$ für $i \in \{1,2,3,4\}$. Eine Münze vom Typ $A_i$ zeigt mit Wahrscheinlichkeit $\frac{i}{10}$, Kopf und mit der entsprechenden Gegenwahrscheinlichkeit Zahl. Wir wählen zufällig eine Münze (aus den vier Typen) und erhalten anschließend beim viermaligen Werfen dieser Münze genau zweimal Kopf.

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

b) Bestimme die Likelihood-Funktion für dieses Ergebnis bezüglich des Parameterraums $\Theta = \{A_1, A_2, A_3, A_4\}$ und gib dann den Maximum-Likelihood-Schätzer für den Typ der gewählten Münze aus dem Parameterraum $\Theta = \{A_1, A_2, A_3, A_4\}$ an.

c) Bestimme die Posteriori-Wahrscheinlichkeit für dieses Ergebnis bezüglich des Parameterraums $\Theta = \{A_1, A_2, A_3, A_4\}$ und gib dann den Maximum-A-Posteriori-Schätzer für den Typ der gewählten Münze aus dem Parameterraum $\Theta = \{A_1, A_2, A_3, A_4\}$ an, wobei die Priori-Wahrscheinlichkeit, dass wir einen Würfel vom Typ $A_i$ zufällig ziehen, bei $\frac{5-i}{10}$ liegt.

**Lösungsskizze**

Sei $X_\theta \in \{0,1,2,3,4\}$ eine Zufallsvariable, die gemäß $\theta \in \Theta = \{A_1, A_2, A_3, A_4\}$ verteilt ist und die Anzahl von 4 Würfen angibt, in denen Kopf erschienen ist.

a) Der Maximum-Likelihood-Schätzer ist $\operatorname{argmax}\{\mathrm{Ws}[x \mid M(\theta)] : \theta \in \Theta\}$.

Der Maximum-A-Posteriori-Schätzer ist $\operatorname{argmax}\{f(\theta \mid x) : \theta \in \Theta\}$, wobei $f(\theta \mid x)$ eine vorgegebene Dichtefunktion für den Parameterraum ist, wenn die Daten gegeben sind (Posterior).

b) Es gilt:

$$
\begin{aligned}
\mathrm{Ws}[X_\theta = 2 \mid \theta = A_1] &= \binom{4}{2}\left(\frac{1}{10}\right)^2 \cdot \left(\frac{9}{10}\right)^2 = 6 \cdot \frac{81}{10^4} \\
\mathrm{Ws}[X_\theta = 2 \mid \theta = A_2] &= \binom{4}{2}\left(\frac{2}{10}\right)^2 \cdot \left(\frac{8}{10}\right)^2 = 6 \cdot \frac{256}{10^4} \\
\mathrm{Ws}[X_\theta = 2 \mid \theta = A_3] &= \binom{4}{2}\left(\frac{3}{10}\right)^2 \cdot \left(\frac{7}{10}\right)^2 = 6 \cdot \frac{441}{10^4} \\
\mathrm{Ws}[X_\theta = 2 \mid \theta = A_4] &= \binom{4}{2}\left(\frac{4}{10}\right)^2 \cdot \left(\frac{6}{10}\right)^2 = 6 \cdot \frac{576}{10^4}
\end{aligned}
$$

Also ist $\operatorname{argmax}\{\mathrm{Ws}[X_\theta = 1 \mid \theta] : \theta \in \Theta\} = A_4$. Somit ist $A_4$ der ML-Schätzer.

c) Sei $Y \in \Theta = \{A_1, A_2, A_3, A_4\}$ eine Zufallsvariable. Es gilt nach der Regel von Bayes:

$$
\mathrm{Ws}[Y = \theta \mid X = 2] = \frac{\mathrm{Ws}[Y = \theta] \cdot \mathrm{Ws}[X = 2 \mid Y = \theta]}{\mathrm{Ws}[X = 2]}.
$$

Weiterhin gilt für $\theta \in \{A_1, A_2, A_3, A_4\}$:

$$\begin{aligned}
\mathrm{Ws}[Y = A_1] \cdot \mathrm{Ws}[X = 2 \mid Y = A_1] &= \frac{4}{10} \cdot 6 \cdot \frac{81}{10^4} = 6 \cdot \frac{324}{10^5} \\
\mathrm{Ws}[Y = A_2] \cdot \mathrm{Ws}[X = 2 \mid Y = A_2] &= \frac{3}{10} \cdot 6 \cdot \frac{256}{10^4} = 6 \cdot \frac{768}{10^5} \\
\mathrm{Ws}[Y = A_3] \cdot \mathrm{Ws}[X = 2 \mid Y = A_3] &= \frac{2}{10} \cdot 6 \cdot \frac{441}{10^4} = 6 \cdot \frac{882}{10^5} \\
\mathrm{Ws}[Y = A_4] \cdot \mathrm{Ws}[X = 2 \mid Y = A_4] &= \frac{1}{10} \cdot 6 \cdot \frac{576}{10^4} = 6 \cdot \frac{576}{10^5}
\end{aligned}$$

Somit gilt:

$$\operatorname{argmax}\{\mathrm{Ws}[Y{=}\theta \mid X{=}1] : \theta{\in}\Theta\} = \operatorname{argmax}\{\mathrm{Ws}[Y{=}\theta]\cdot \mathrm{Ws}[X{=}1 \mid Y{=}\theta] : \theta{\in}\Theta\}.$$

Also ist $\operatorname{argmax}\{\mathrm{Ws}[Y = \theta \mid X = 1] : \theta \in \Theta\} = A_3$, und somit ist $A_3$ der MAP-Schätzer.

Vorname: ______________ Name: ______________ Matrikelnummer: ____________

## Aufgabe 5 (8 Punkte)

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

MinTSP (Undirected Traveling Salesperson)

**Eingabe:** Ein vollständiger ungerichteter Graph $G = (V, E)$ und Kantengewichten $w$.
**Lösung:** Ein ungerichteter Hamiltonscher Kreis, d.h. eine Permutation $(v_{\pi(1)}, \ldots, v_{\pi(n)})$ der Knotenmenge $V$ mit $\{v_{\pi(i)}, v_{\pi((i \bmod n)+1)}\} \in E$ und $n = |V|$.
**Optimum:** Minimiere $\sum_{i=1}^{n} w(v_{\pi(i)}, v_{\pi((i \bmod n)+1)})$ mit $n = |V|$.

MinDirectedTSP (Directed Traveling Salesperson)

**Eingabe:** Ein vollständiger gerichteter Graph $G = (V, E)$ und Kantengewichten $w$.
**Lösung:** Ein gerichteter Hamiltonscher Kreis, d.h. eine Permutation $(v_{\pi(1)}, \ldots, v_{\pi(n)})$ der Knotenmenge $V$ mit $(v_{\pi(i)}, v_{\pi((i \bmod n)+1)}) \in E$ und $n = |V|$.
**Optimum:** Minimiere $\sum_{i=1}^{n} w(v_{\pi(i)}, v_{\pi((i \bmod n)+1)})$ mit $n = |V|$.

**Lösungsskizze**

Wir werden im Folgenden die gerichteten Kanten durch ungerichtete Kanten ersetzen, die aber auch alle in derselben Richtung durchlaufen werden müssen. Zur Erinnerung: Eine PTAS-Reduktion ist ein Tripel $(f, g, \alpha)$.

Hier ist $f(G, \epsilon)$ ist ein ungerichteter Graph, der aus $G$ durch Ersetzen jedes Knotens $v$ und jeder gerichteten Kante $(v, x)$ wie folgt entsteht (und von $\epsilon$ unabhängig ist): Jeder Knoten $v$ wird durch drei Knoten $v', v^*, v''$ und die ungerichteten Kanten $\{v', v^*\}$ und $\{v^*, v''\}$ ersetzt. Jede gerichtete Kante $(v, x)$ wird durch die ungerichtete Kante $\{v'', x'\}$ ersetzt. Für die Gewichte gilt $w(v', v^*) = w(v^*, v'') = 0$ und $w(v'', x') = w(v, x)$.

Wir halten fest, dass jeder Hamiltonsche Kreis in $f(G, \epsilon)$, der $v'$ (für ein $v \in V$) besucht, anschließend erst $v^*$ und dann $v''$ besuchen muss (andernfalls kann $v^*$ nie besucht werden).

$g$ ist quasi die identische Abbildung mit $g(G, c', \epsilon) = c$ (ist also eigentlich unabhängig von $G$ und $\epsilon$). Genau genommen werden im Hamiltonschen Kreis $c'$ die Teilpfade $(v', v^*, v'')$ durch $v$ ersetzt (aus dem dieses Tripel $(v', v^*, v'')$ generiert wurde), so das man $c$ erhält.

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

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

Da nach Konstruktion der Kantengewichte mittel $f$ und bei der Rekonstruktion der Hamiltonsche Kreise mittels $g$ die Gewichte der entsprechenden Hamiltonsche Kreise gleich bleiben, also $w(c') = w(g(G, c', \epsilon))$ gilt, bleiben also auch die Güten der zugehörigen Lösungen gleich, $\mu'(f(G), c') = \mu(G, g(c'))$. Somit bleibt auch die Approximationsgüte erhalten.

Wie man leicht nachweist, sind $f$, $g$ und $\alpha$ in polynomieller Zeit berechenbar.