Technische Universität München
Fakultät für Informatik
Lehrstuhl für Effiziente Algorithmen
Dr. Volker Heun

Sommersemester 2002
Semestralklausur
16.07.02

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. 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.

Hinweise zum Multiple-Choice-Teil

- Der rechte Rand ist frei zu lassen. Die Rückseiten und die letzten Seiten dürfen als Schmierpapier verwendet werden.
- Falls ein Kästchen versehentlich angekreuzt wurde, so ist ist dieses vollständig auszufüllen und unmittelbar rechts bzw. links daneben eine neues Kästchen zu malen: ■ ☐. Es gilt dann als nicht angekreuzt.
- Falls dieses Kästchen dennoch wieder angekreuzt werden soll, so ist das rechts bzw. links daneben gemalte Kästchen anzukreuzen: ■ ☒.
- Bei Antworten, die einen Zahlenwert verlangen, ist der konkrete Zahlenwert oder eine kurze Formel einzutragen, also z.B. 24 bzw. 4!.

Viel Glück und Erfolg!

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

Vorzeitig abgegeben um ....

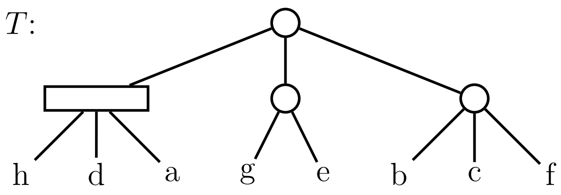
	Seite 2	Seite 3	Seite 4	A1	A2	A3	A4	A5	$\sum$
Erstk.									
Zweitk.									

Meiner Meinung nach war die Klausur . . . leicht ☐ ☐ ☐ ☐ ☐ schwer

Vorname: ______________ Name: ______________ Matrikelnummer: ______________

Aufgabe M1 (2 Punkte)

Sei im Folgenden T der unten angegebene echte PQ-Baum über $\Sigma = \{a, b, c, d, e, f, g, h\}$.



Kreuze zu jeder Aussage an, ob diese richtig [R] oder falsch [F] ist.

- $egadhcfb \in \text{consistent}(T)$. .... ☑ R | F ☐
- Es gilt $\text{consistent}(T) \cap \Pi(\Sigma, \{e, f, g, h\}) = \emptyset$. .... ☐ R | F ☑
- Wie viele verschiedene zu T äquivalente PQ-Bäume gibt es, deren Frontier mit f beginnen? .... $\boxed{\mathbf{2^4 = 16}}$

Aufgabe M2 (1 Punkt)

Kreuze zu jeder Aussage an, ob diese richtig [R] oder falsch [F] ist.

- Jeder echte PQ-Baum über Σ besitzt maximal $|\Sigma|$ P-Knoten. .... ☑ R | F ☐
- Jeder echte PQ-Baum besitzt mehr P-Knoten als Q-Knoten. .... ☐ R | F ☑

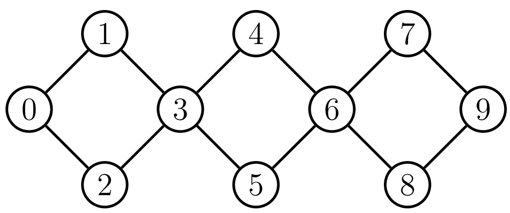
Aufgabe M3 (1 Punkt)

Kreuze zu jeder Aussage an, ob diese richtig [R] oder falsch [F] ist.

- Jeder Baum ist ein Intervall-Graph. .... ☐ R | F ☑
- Jeder Intervall-Graph ist ein Einheits-Intervall-Graph. .... ☐ R | F ☑

Aufgabe M4 (1 Punkt)

Betrachte folgende Eingabe (V, M, F) für das allgemeine Intervall Sandwich Problem. Hierbei ist $V = [0 : 9]$, der Graph (V, M) ist unten angegeben und F wird in den Teilaufgaben spezifiziert.



Kreuze zu jeder Aussage an, ob diese richtig [R] oder falsch [F] ist.

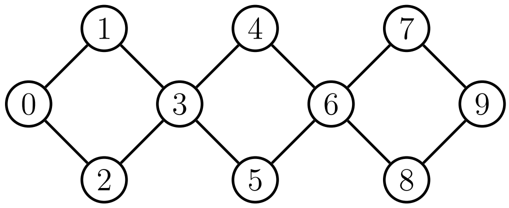
- Für die Menge $F = \emptyset$ ist die Menge $\{0, 1, 8, 9\}$ ein zulässiger Kern für das allgemeine Intervall Sandwich Problem. .... ☑ R | F ☐
- Für die Menge $F = \{\{0, 3\}, \{1, 2\}, \{3, 6\}, \{4, 5\}\}$ ist die Menge $\{6, 7, 8, 9\}$ ein zulässiger Kern für das allgemeine Intervall Sandwich Problem. .... ☑ R | F ☐

1 ☐
2 ☐

Vorname: ______________ Name: ______________ Matrikelnummer: ____________

Aufgabe M5 (1 Punkt)

Betrachte die folgende Eingabe (V, M, F) für das Bounded Degree Interval Sandwich Problem, wobei der Ausgabegraph maximalen Grad 4 besitzen soll. Hierbei ist $V = [0:9]$, $F = \emptyset$ und der Graph (V, M) ist unten angegeben.



Kreuze zu jeder Aussage an, ob diese richtig [R] oder falsch [F] ist.

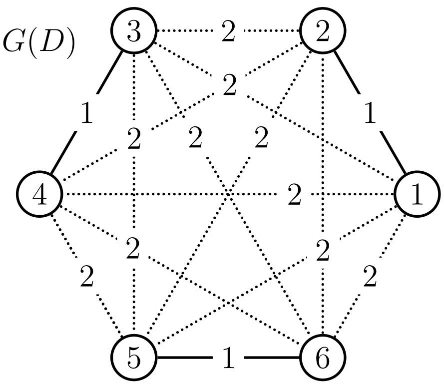
- Die Menge $V = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ ist ein zulässiger Kern für das Bounded Degree Interval Sandwich Problem mit maximalem Grad 4. .... ☑ R | F ☐
- Die Menge $\{3, 4, 6\}$ ist ein zulässiger Kern für das Bounded Degree Interval Sandwich Problem mit maximalem Grad 4. .... ☐ R | F ☑

Aufgabe M6 (1 Punkt)

Wir betrachten die folgende symmetrische Distanzmatrix D, die unten angegeben ist und rechts daneben als gewichteter Graph $G(D)$ illustriert ist.

D	1	2	3	4	5	6
1	0	1	2	2	2	2
2		0	2	2	2	2
3			0	1	2	2
4				0	2	2
5					0	1
6						0

$G(D)$



Kreuze zu jeder Aussage an, ob diese richtig [R] oder falsch [F] ist.

- Es existiert ein kompakter additiver Baum für D. .... ☐ R | F ☑
- Es existiert ein ultrametrischer Baum für D. .... ☑ R | F ☐

1

2

Vorname: ______________ Name: ______________ Matrikelnummer: ____________

Aufgabe M7 (1 Punkt)

Im Folgenden sind zwei binäre 4×5-Charaktermatrizen M_1 und M_2 angegeben.

M_1	a	b	c	d	e
1	0	1	1	1	1
2	0	0	0	1	0
3	0	1	1	1	0
4	1	1	0	0	0

M_2	a	b	c	d	e
1	0	1	1	1	1
2	0	0	0	0	0
3	0	1	1	1	0
4	1	1	0	0	0

Kreuze zu jeder Aussage an, ob diese richtig [R] oder falsch [F] ist.

- M_1 besitzt eine perfekte Phylogenie. .... ☐ R | F ☑
- M_2 besitzt eine perfekte Phylogenie. .... ☑ R | F ☐

Aufgabe M8 (2 Punkte)

Im Folgenden ist eine binäre 4×5-Charaktermatrizen M_3 angegeben.

M_3	a	b	c	d	e
1	0	1	0	0	0
2	1	0	1	0	1
3	0	0	1	1	1
4	0	0	1	0	0

Konstruiere mit Hilfe des in der Vorlesung angegebenen Algorithmus einen phylogenetische Baum für M_3:

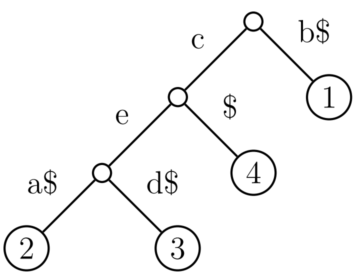
Lösungsvorschlag

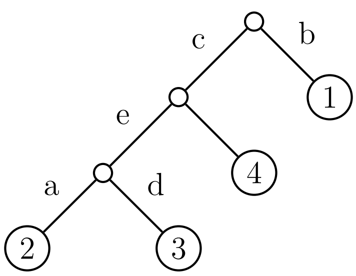
M_3 besitzt eine perfekte Phylogenie. Wir sortieren zuerst die Spalten von M_3 absteigend und ermitteln daraus die zugeordneten Zeichenreihen:

M_3'	b	c	e	a	d
1	1	0	0	0	0
2	0	1	1	1	0
3	0	1	1	0	1
4	0	1	0	0	0

s_1 : b\$
s_2 : cea\$
s_3 : ced\$
s_4 : c\$

Aus diesen Strings können wir folgenden Trie bauen der auch die perfekte Phylogenie der Matrix M_3 ist (hierbei sind rechts die Hilfs-Markierung \$ bereits wieder entfernt):



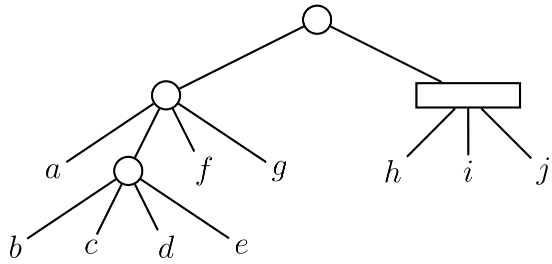


1
2

Vorname: ______________ Name: ______________ Matrikelnummer: ____________

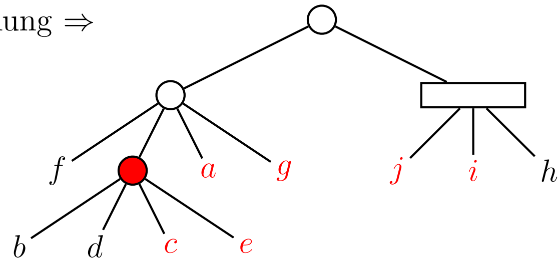
Aufgabe 1 (6 Punkte)

Betrachte den rechts abgebildeten PQ-Baum. Modifiziere den PQ-Baum gemäß der in der Vorlesung angegebenen Schablonen, so dass zusätzlich auch die Restriktion $\{a, c, e, g, i, j\}$ erfüllt wird. Dabei sind alle Zwischenschritte anzugeben und zu dokumentieren.

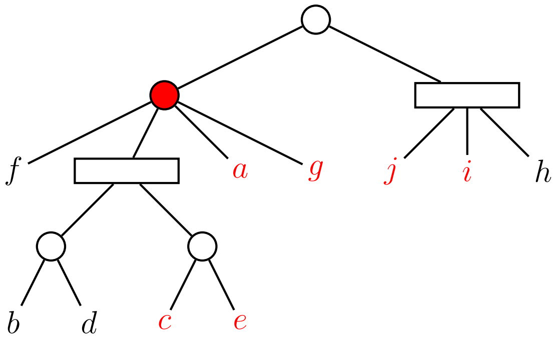


Lösungsvorschlag

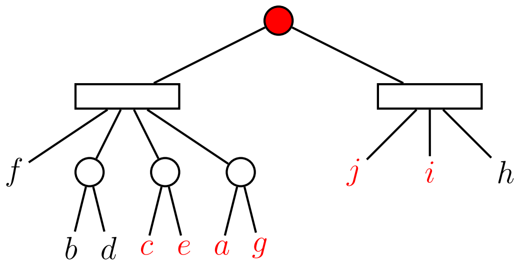
Äquivalenzumformung $\Rightarrow$



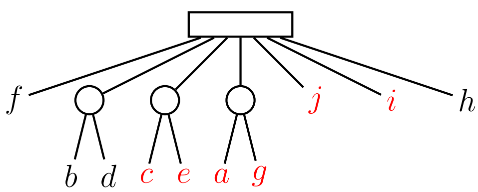
Mit P_3 $\Rightarrow$



Mit P_5 $\Rightarrow$



Mit P_6 $\Rightarrow$



Vorname: ______________ Name: ______________ Matrikelnummer: ____________

Aufgabe 2 (6 Punkte)

Die Distanzmatrizen D_1 und D_2 seien durch die folgenden Tabellen definiert.

D_1	1	2	3	4	5	6	7	8
1	0	5	8	8	1	5	7	7
2		0	8	8	5	2	7	7
3			0	6	7	7	3	6
4				0	7	7	6	4
5					0	5	7	7
6						0	7	7
7							0	6
8								0

D_2	1	2	3	4	5	6	7	8
1	0	4	4	7	7	1	7	7
2		0	3	7	7	4	7	7
3			0	7	7	4	7	7
4				0	6	7	6	2
5					0	7	5	6
6						0	7	7
7							0	6
8								0

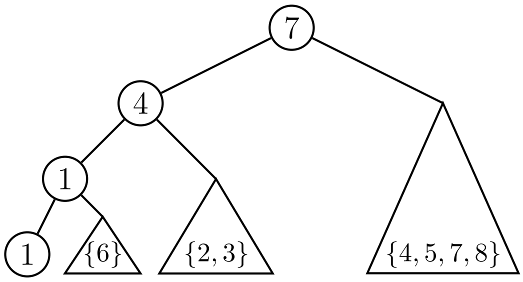
Entscheide, welche der Matrizen D_1 und D_2 ultrametrisch ist oder nicht. Im Falle, dass eine Matrix ultrametrisch ist, ist die explizite Konstruktion des zugehörigen ultrametrischen Baumes mit dem in der Vorlesung vorgestellten Algorithmus ausführlich mit allen Schritten anzugeben. Im Falle, dass eine Matrix nicht ultrametrisch ist, ist dies zu begründen.

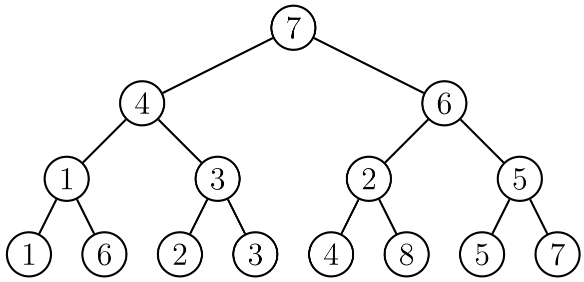
Lösungsvorschlag

D_1 ist nicht ultrametrisch, da in der Matrix 8 verschiedene positive Werte vorkommen und diese nicht an die maximal $7 = 8 - 1$ inneren Knoten des ultrametrischen Baumes verteilt werden können (jeder Matrixeintrag muss als Knotenmarkierung auftreten).

D_1 ist nicht ultrametrisch, da die Matrixeinträge $D_1(1,3)$, $D_1(1,5)$ und $D_1(3,5)$ nicht die ultrametrische Dreiecksungleichung erfüllen: $8 = D_(1,3) \not\leq \max\{D_1(1,5), D_1(5,3)\} = 7$.

D_2 ist ultrametrisch, was wir konstruktiv zeigen werden. Wir stellen zuerst fest, dass $\{d_{11}, \ldots, d_{18}\} = \{1, 4, 7\}$ ist. Wir partitionieren daher $[2 : 8]$ in $\{\{6\}, \{2,3\}, \{4,5,7,8\}\}$ und konstruieren rekursiv für jede Indexmenge einen ultrametrischen Baum. Für $\{6\}$ und $\{2,3\}$ ist dies trivial. Für $\{4,5,7,8\}$ bestimmen wir $\{d_{45}, d_{47}, d_{48}\} = \{2, 6\}$. Die zugehörige Partition der Indexmenge $\{5,7,8\}$ liefert $\{\{5,7\},\{8\}\}$. Damit erhalten wir den folgenden ultrametrischen Baum:





Vorname: ______________ Name: ______________ Matrikelnummer: ______________

Aufgabe 3 (6 Punkte)

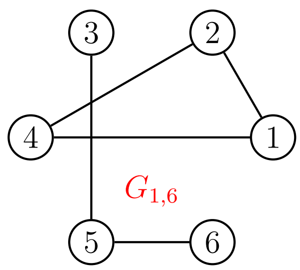
Gegeben seien die beiden folgenden 9×9-Matrizen D_ℓ und D_h. Entscheide, ob es eine ultrametrische Matrix $D \in [D_\ell, D_h]$ gibt oder nicht. Konstruiere dazu entweder einen ultrametrischen Baum für M nach einer der in der Vorlesung angegebenen Algorithmen, oder gib eine Begründung, dass dies nicht möglich ist. Alle Zwischenschritte müssen dabei mit angegeben werden.

D_ℓ	1	2	3	4	5	6
1	0	4	4	1	6	8
2		0	4	3	6	6
3			0	7	2	6
4				0	6	6
5					0	5
6						0

D_h	1	2	3	4	5	6
1	0	6	9	3	9	9
2		0	9	5	8	9
3			0	8	5	9
4				0	8	8
5					0	7
6						0

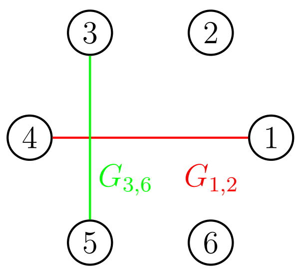
Lösungsvorschlag

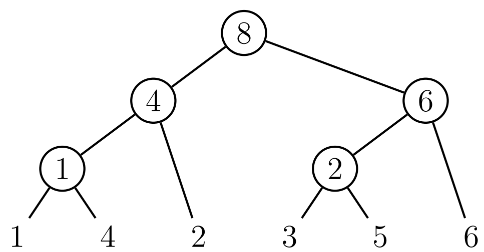
1. Variante Wir stellen zuerst fest, dass $8 = D_\ell(1,6)$ der maximale Eintrag von D_ℓ ist und konstruieren $G_{1,6}$. Für die Zusammenhangskomponenten konstruieren wir dann rekursiv die ultrametrischen Bäume.



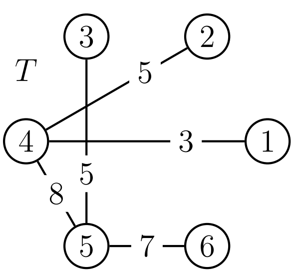
D_ℓ	1	2	4	3	5	6
1	0	4	1			
2		0	3			
4			0			
3				0	2	6
5					0	5
6						0

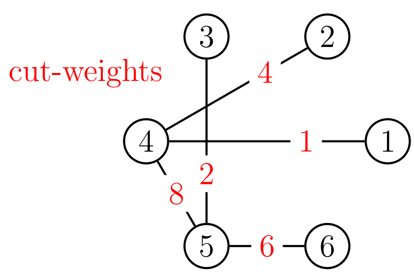
D_h	1	2	4	3	5	6
1	0	6	3			
2		0	5			
4			0			
3				0	5	9
5					0	7
6						0





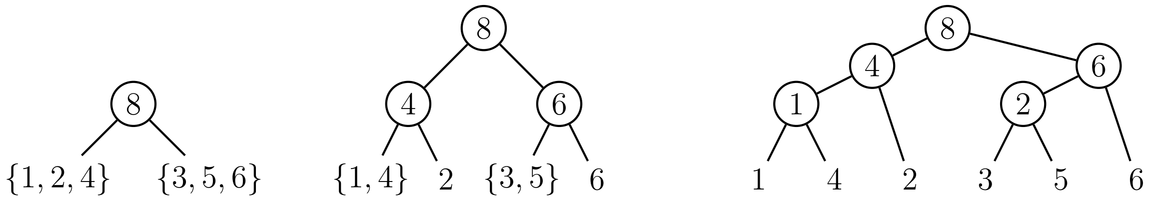
2. Variante Wir bestimmen zuerst den minimalen Spannbaum T von D_h und ermitteln die cut-weights der Kanten $E(T)$.





Vorname: ______________ Name: ______________ Matrikelnummer: ____________

Nun entfernen wir die Kanten des minimalen Spannbaumes T nach fallenden cut-weights und konstruieren für die entstehenden Bipartitionen sukzessive einen ultrametrischen Baum der der Sandwich-Bedingung genügt.



Vorname: ______________ Name: ______________ Matrikelnummer: ____________

Aufgabe 4 (6 Punkte)

Sei M eine binäre $n \times m$-Charaktermatrix, die eine perfekte Phylogenie besitzt und deren Zeilen paarweise verschieden sind (d.h. es gibt keine zwei Objekte, die genau dieselben Charaktere besitzt). Zeige, dass die $n \times n$-Matrix D_M additiv ist, wobei

$$D_M(i,j) = |\{c \in [1:m] \mid M(c,i) \neq M(j,c)\}|.$$

Lösungsvorschlag

1. Variante Sei T der zu M gehörige phylogenetischen Baum (der nach Voraussetzung existieren muss). Wir bauen jetzt diesen Baum zu einem additiven Baum um, indem wir an die Kanten des Baumes Kantengewichte einführen. An jeder Kante ersetzen wir die Eigenschaften durch ihre Anzahl. Seien v und w jetzt zwei Knoten des Baumes, die mit i bzw. j markiert sind (die Knotenmarkierung des phylogenetischen Baumes behalten wir bei). Auf dem eindeutigen Pfad von v zu w im phylogenetischen Baum ist die Menge der Kantenmarkierungen dieses Pfades im phylogenetischen Baum gerade gleich $\{c \in [1:m] \mid M(c,i) \neq M(c,j)\}$. Da jede Eigenschaft c im phylogenetischen Baum nur an genau einer Kante vorkommen kann, ist das Gewicht desselben Pfades im additive Bäume gerade gleich der Kardinalität der obigen Menge, also gleich $D_M(i,j) = |\{c \in [1:m] \mid M(c,i) \neq M(c,j)\}|$. Somit ist der umgebaute Baum T ein additiver Baum für D_M und D_M somit additiv.

2. Variante Sei M eine $n \times m$-Matrix mit einer perfekten Phylogenie. Wir weisen jetzt nach, dass D_M additiv ist.

Nichtnegativität: Da $D_M(i,j)$ als die Kardinalität einer Menge definiert ist, muss sie immer nichtnegativ sein.

Definitheit: $D_M(i,i) = |\{c \in [1:m] \mid M(i,c) \neq M(i,c)\}| = 0$. Da wir ausgeschlossen hatten, dass M zwei gleiche Zeilen hat, gilt $D_M(i,j) > 0$ für beliebige Zeilen $i \neq j$.

Symmetrie: Es gilt

$$\begin{aligned} D_M(i,j) &= |\{c \in [1:m] \mid M(i,c) \neq M(j,c)\}| \\ &= |\{c \in [1:m] \mid M(j,c) \neq M(i,c)\}| = D_M(j,i). \end{aligned}$$

4-Punkte-Bedingung: Betrachten wir vier beliebige Zeilen i, j, k und l. Es ist zu zeigen, dass aus

$$D_M(i,k) + D_M(j,l) \leq \max\{D_M(i,l) + D_M(j,k), D_M(i,j) + D_M(k,l)\}$$

die Gleichheit $D_M(i,l) + D_M(j,k) = D_M(i,j) + D_M(k,l)$ folgt.

Für eine beliebige Spalte c betrachten wir die Veränderung von $D_M(i,k)+D_M(j,l)$, $D_M(i,l)+D_M(j,k)$ und $D_M(i,j)+D_M(k,l)$:

Nr.	$M(i,c)$	$M(j,c)$	$M(k,c)$	$M(l,c)$	$D_M(i,k)+$ $D_M(j,l)$	$D_M(i,l)+$ $D_M(j,k)$	$D_M(i,j)+$ $D_M(k,l)$
1	0	0	0	0	+0	+0	+0
2	1	1	1	1	+0	+0	+0
3	1	0	0	0	+1	+1	+1
4	0	1	0	0	+1	+1	+1
5	0	0	1	0	+1	+1	+1
6	0	0	0	1	+1	+1	+1
7	1	1	1	0	+1	+1	+1
8	1	1	0	1	+1	+1	+1
9	1	0	1	1	+1	+1	+1
10	0	1	1	1	+1	+1	+1
11	1	0	1	0	+0	+2	+2
12	0	1	0	1	+0	+2	+2
13	1	0	0	1	+2	+0	+2
14	0	1	1	0	+2	+0	+2
15	1	1	0	0	+2	+2	+0
16	0	0	1	1	+2	+2	+0

Man sieht, dass die drei Terme nur dann verschieden erhöht werden, wenn in jeweils zwei Zeilen eine 0 und in den anderen zwei eine 1 steht. Aufgrund der Eigenschaft perfekter Phylogenien gilt für zwei Spalten A und B, entweder $A \cap B = \emptyset$ oder $A \subset B$ bzw. $B \subset A$. Daher schließen sich die Zeilenpaare $(11,12)$, $(13,14)$ und $(15,16)$ für verschiedene Eigenschaften c_1, c_2 gegenseitig aus, d.h. es können nur Ereignisse aus einem dieser Paare stattfinden.

Somit sind entweder alle drei Terme $D_M(i,k)+D_M(j,l)$, $D_M(i,l)+D_M(j,k)$ und $D_M(i,j)+D_M(k,l)$ gleich groß, oder zwei der Terme haben sind durch eine oder mehrere Eigenschaften jeweils stärker gewachsen. Da aber jeweils einer der Terme bei der Gleichung (iv) auch rechts steht, ist diese immer erfüllt.

Vorname: ______________ Name: ______________ Matrikelnummer: ____________

Aufgabe 5 (6 Punkte)

Sei M eine binäre $n \times m$-Charaktermatrix.

a) Zeige, dass M die *Consecutive Ones Property* für die Spalten besitzt, wenn M einen perfekten phylogenetischen Baum besitzt.

b) Konstruiere mit der Aussage aus Teil a) einen effizienten Algorithmus zur Konstruktion phylogenetischer Bäume für binäre Charaktermatrizen.

Erinnerung: Eine Matrix besitzt die *Consecutive Ones Property* für die Spalten, wenn es eine Permutation der Zeilen gibt, so dass in jeder Spalte der Matrix die 1en aufeinander folgend auftreten.

Lösungsvorschlag

a) Sei M eine Matrix, die einen perfekten phylogenetischen Baum besitzt. Sei dann $F := (S_{\pi(1)}, S_{\pi(2)}, \ldots, S_{\pi(n)})$ die Reihenfolge in welcher die Spezies bei einer DFS im phylogenetischen Baum angetroffen werden. Da in F Spezies, welche in einem Teilbaum liegen, hintereinander vorkommen, gilt für eine beliebige Eigenschaft c, dass alle Spezies, die die Eigenschaft c haben, in einem gemeinsamen Teilbaum liegen und daher in F hintereinander vorkommen. Daher zeigt die Permutation π, dass M die „Consecutive Ones"-Eigenschaft für Spalten besitzt.

b) Wir können die Spalten der Matrix M als Restriktionen für Permutationen auffassen, indem wir zur i-ten Spalte die Restriktionsmenge $S_i = \{k \in [1 : n] \mid M(k, i) = 1\}$ bilden. Für diese Restriktionsmengen bauen wir den entsprechenden PQ-Baum (der nach (a) für eine Matrix mit einer perfekten Phylogenie auch existiert). In diesem Baum gilt, dass alle Spezies, die eine gemeinsame Eigenschaft haben auch in einem gemeinsamen Unterbaum vorkommen. Da für zwei beliebige Restriktionsmengen S_i und S_j gilt, dass $S_i \cap S_j = \emptyset$, kommen im PQ-Baum keine Q-Knoten vor (eine Restriktion geht niemals über einige verschiedenene Teilbäume hinweg, entweder über alle oder einen!). Es gilt ferner, dass zu jeder Restriktion genau ein P-Knoten zugeordnet werden kann: Betrachten wir das Schema zum Einbau einer Restriktion S_i in einen bestehenden PQ-Baum, dann gilt induktiv für einen beliebigen Knoten p, dass alle Blätter unterhalb dieses Knotens zu einer Restriktion S_j gehören. Da entweder $S_j \subset S_i$ oder $S_i \subset S_j$, gibt es nur folgende Fälle:

(a) Alle Nachfahren von p sind markiert und es ist nichts zu tun.

(b) Alle Elemente aus S_i hängen unterhalb von p.

   i. Falls $S_i \supset S_k$ so gibt es Kinder $K = \{k_1, k_2, \ldots, k_r\}$, so dass $S_l \subset S_i$ falls $l \in K$ und $S_l \cap S_i = \emptyset$ sonst. Somit ensteht ein neuer P-Knoten, der die Kinder K zusammenfasst und für die Restriktion S_i steht.

   ii. Andernfalls gibt es ein Kind mit $S_i \subset S_k$ und es kommt keine Regel an p zum Einsatz.

Fall ii) kann nur für ein Kind auftreten, da für je zwei Kinder $S_{k_1} \cap S_{k_2} = \emptyset$ gilt.

Daher können wir den zu den P-Knoten führenden Kanten jeweils die entsprechende Eigenschaft i zuordnen und erhalten den perfekten phylogenetischen Baum in Zeit $O(nm)$.