Biostatistik und Bioinformatik (WS 2010/11)

Klausur

15. Februar 2011

Name ..... **Matrikelnummer** .....

40 bestanden (>25?)

Punktzahl ..... **Note** ..... 19 nicht best. (<25?)

Biostatistik

1. In einem Mischbestand zweier *Senecio*-Arten wurden folgende Variablen gemessen bzw. protokolliert und für Pflanzen ohne (0; *S. ovatus*) bzw. mit Drüsenhaaren (1; *S. hercynicus*) die angegebenen Mittelwerte (MW) und Standardabweichungen (SD) berechnet:

Drüsenhaare	*n*	Köpfchenbreite (mm)	Hüllblattlänge (mm)	Röhrenblütenlänge (mm)
0	23	MW: 3,1; SD: 0,4	MW: 6,4; SD: 0,8	MW: 12,7; SD: 2,3
1	17	MW: 3,4; SD: 0,8	MW: 7,0; SD: 0,8	MW: 13,7; SD: 2,9

Prüfen Sie, ob die Aufsammlung in Bezug auf das Fehlen bzw. Vorhandensein von Drüsenhaaren heterogen ist oder ob von einer Gleichverteilung in der zugrunde liegenden Grundgesamtheit auszugehen ist! Geben Sie dabei (a) den zu verwendenden Test, (b) den Wert für die Teststatistik, (c) den kritischen Wert für die Teststatistik ($\alpha = 0{,}05$) und (d) ihre Testentscheidung an! (**4 Punkte**)

(a) Chi-Quadrat-Test

(b) $X^2 = \frac{(23-20)^2}{20} + \frac{(17-20)^2}{20} = 0.9$

(c) $\chi^2_1 = 3.84$

(d) H_0 nicht verwerfen → gleichverteilt

In welchen der drei Merkmalen unterscheiden sich die beiden Pflanzenarten signifikant voneinander ($\alpha = 0{,}05$)? Geben Sie auch hier (a) den zu verwendenden Test, (b) den Wert für die Teststatistik, (c) den kritischen Wert für die Teststatistik ($\alpha = 0{,}05$) und (d) ihre Testentscheidung an! (**6 Punkte**)

(a) t-Test ✓ zweiseitig ✗

(b) $T = -1.56$ ✗ $T = -2.34$ ✓ $T = -1.216$ ✗

(c) $t_{38,1-\frac{\alpha}{2}} = 2.02$ ✓ $t = 2.02$ ✗ $t = 2.02$ ✓

(d) u.S. ✗ sign. ✗ u.S. ✓

2. Die Zufallsvariable X besitzt eine $N(2,1)$-Verteilung.

(a) Berechnen Sie die Wahrscheinlichkeit $P(2 < X < 3)$ und $P(-1 < X < 3)$ (**4 Punkte**).

$$P(2<X<3) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) = \Phi\left(\frac{3-2}{1}\right) - \Phi\left(\frac{2-2}{1}\right) = \checkmark$$

$$= \Phi(1) - \Phi(0) = 0.8413 - 0.5 = \underline{0.3413} \checkmark$$

$$P(-1<X<3) = \Phi\left(\frac{3-2}{1}\right) - \Phi\left(\frac{-1-2}{1}\right) = \Phi(1) - \Phi(-3) = 0.8413 - (1-0.9986)$$

$$\checkmark \qquad \sim 0.8413 \checkmark$$

(b) Bestimmen Sie die Quantile der Ordnung $\alpha = 0{,}01$, $\alpha = 0{,}05$ und $\alpha = 0{,}95$ (**3 Punkte**).

$$q_\alpha = \sigma \cdot z_\alpha + \mu$$

$$q_{0.01} = 1 \cdot z_{0.01} + 2 = -2.33 + 2 = -0.33$$

$$q_{0.05} = 1 \cdot z_{0.05} + 2 = -1.65 + 2 = 0.35$$

$$q_{0.95} = 1 \cdot z_{0.95} + 2 = 1.65 + 2 = 3.65$$

	Nordsee	Irische See	
+	15 32,7	32 14.3	47
−	86 68,3	12 29.7	98
	101	44	145

$$X^2 = \frac{(15-32.7)^2}{32.7} + \frac{(86-68,3)^2}{68.3} + \frac{(32-14.3)^2}{14.3} + \frac{(12-29.7)^2}{29.7} =$$

$$= 9.6 + 4.6 + 21.9 + 10.5 = \underline{\underline{46.6}}$$

3. Eine Stichprobe auf einem Fischkutter in der Nordsee enthält 86 Flundern ohne Verletzungen und 15 Flundern mit Verletzungen. Ein weitere, von einem Trawler in der Irischen See stammende Stichprobe wies 32 Flundern mit und 12 ohne Verletzungen auf. Gibt es einen statistisch signifikanten Unterschied zwischen den beiden Populationen? Geben Sie dabei (a) den zu verwendenden Test, (b) den Wert für die Teststatistik, (c) den kritischen Wert für die Teststatistik ($\alpha = 0{,}05$) und (d) ihre Testentscheidung an! (**4 Punkte**)

(a) Chi-Quadrat-Test ✓

(b) $\chi^2 = 46.6$ ✓

(c) $\chi^2_{(n-1)(m-1),\,1-\alpha} = \chi^2_{1,\,0.95} = 3.84$ ✓

(d) H_0 verwerfen → signifikante Korrelation. ✓

4. Nehmen Sie an, dass die Petalenlänge in einer Pflanzenpopulation normalverteilt ist und den Mittelwert $\mu = 3{,}2$ cm und die Standardabweichung $\sigma = 1{,}8$ cm aufweist. Geben Sie die Petalenlänge an, die

(a) von 10 % der Pflanzen übertroffen wird,

(b) von 25 % der Pflanzen unterschritten wird.

(c) Geben Sie das 95%- und das 99%-Konfidenzintervall für diese Variable an. (**4 Punkte**)

(a) $z = \frac{x-\mu}{\sigma} \rightarrow x = (z \cdot \sigma) + \mu$ mit $z_{0.9} = 1.29$

$x = (1.29 \cdot 1.8) + 3.2 = 5.5$ cm ✓

(b) $z = \frac{x-\mu}{\sigma} \rightarrow x = (z \cdot \sigma) + \mu$ mit $z_{0.25} = -0.68$

$x = (-0.68 \cdot 1.8) + 3.2 = 2.0$ cm ✓

(c) $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.8}{\sqrt{100}} = 0.18$ ↑ nicht gegeben!

$G_u = \bar{x} - z_{1-\frac{\alpha}{2}} \cdot \sigma_{\bar{x}} = 2.9$ cm

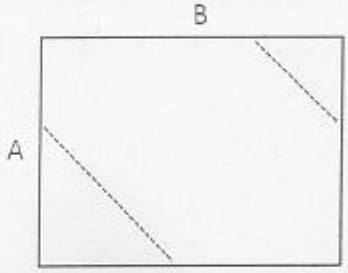
$G_o = \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \sigma_{\bar{x}} = 3.5$ cm } 95%

$G_u = 3.2 - 2.57 \cdot 0.18 = 2.7$ cm

$G_o = 3.2 + 2.57 \cdot 0.18 = 3.7$ cm } 99%

Sequenzvergleich und Datenbanken (15 Punkte)

a) Zwei Proteinsequenzen A und B wurden paarweise aligniert. Der Inhalt der Matrix, aus der das Alignment abgeleitet wurde, ist unten angegeben. Die gestrichelten Linien weisen auf signifikante Scorewerte hin. Leiten Sie aus diesem Alignment den Aufbau der Proteine ab. Beschreiben Sie Gemeinsamkeiten und Unterschiede. (2 Punkte)



b) Sie haben mithilfe von BLAST eine Sequenzdatenbank durchsucht. Ein Treffer (TR_1) der BLAST-Ausgabe hat einen E-Wert von 34, ein zweiter (TR_2) einen E-Wert von 5×10^{-34}. Welcher der beiden Treffer weist auf die Funktion der Query hin? (1 Punkt)

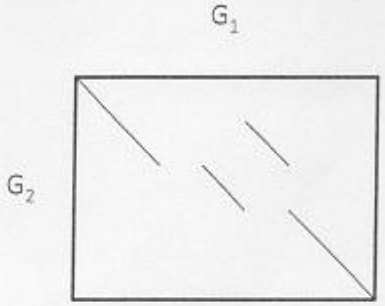
c) Im BLAST-Algorithmus werden anfangs zu jedem Teilwort Δ solche w-mere gesucht, die im Vergleich zu Δ einen Score $> T$ hat. Gegeben sei das Teilwort Δ = RGW. Ist die Sequenz S = MDW ein w-mer, das im Vergleich mit Δ einen Score $T > 8$ aufweist? Begründung! Benutzen Sie für die Lösung die BLOSUM-Matrix, die unten in der Aufgabe b) zu Multiplen Sequenzalignments und Scoring-Systemen angegeben ist. (1 Punkt)

d) Gegeben seien die beiden Sequenzen A = GCCGTGCGGC und B = CCCCGGCGCC. Wie groß ist der Hamming-Abstand und um wie viele PAM-Einheiten unterscheiden sie sich? (2 Punkte)

e) Konstruieren Sie einen Suffixbaum für die Sequenz A = WCCN. (2 Punkte)

f) Der Vergleich zweier Proteine P1 und P2 ergab, dass ihre Sequenzen 50% identische Residuen aufweisen. Allerdings kommen die in P1 bekannten katalytischen Residuen in P2 nicht vor. Welche Schlüsse ziehen sie hieraus für die Funktion und Struktur von P2? (1 Punkt)

g) Dotplots können dazu verwendet werden, zwei Genome G_1, G_2 zu vergleichen. Es wird im Plot dann an Position i,j ein Punkt eingetragen, wenn ein Vergleich der Gene g_i aus G_1 und g_j aus G_2 mit BLAST einen signifikanten E-Wert ergibt. Der folgende Plot zeigt schematisch einen solchen Vergleich. Erläutern Sie, wie sich G_1 von G_2 unterscheidet. Teilen Sie die Genome in Bereiche A, B, C, … auf und diskutieren Sie dann die Zusammensetzung. Wenn G_1 ein pathogener und G_2 ein nicht-pathogener Stamm der selben Art wäre: In welchem Genombereich würden Sie nach Genen suchen, die für die Pathogenität verantwortlich sein können? (2 Punkte)



h) Berechnen Sie den Wert für die dick umrandete Zelle gemäß des Algorithmus von **Needleman und Wunsch**. Tragen Sie Ihr Ergebnis bei "..." ein. Geben Sie sämtliche Teilergebnisse an, benutzen Sie hierfür die anderen Zellen.
Verwenden Sie folgende Scores: Match = 3, Mismatch = -6, Gap = -4. (2 Punkte)

	...	...	G	
...	...	...	...	
...	...	4	1	
G	...	3		...

i) Berechnen Sie den Wert für die dick umrandete Zelle gemäß des Algorithmus von **Smith und Waterman**. Geben Sie die Ergebnisse wiederum wie oben eingeführt an.
Verwenden Sie folgende Scores: Match = 3, Mismatch = -2, Gap = -4. (2 Punkte)

	...	...	C	
...	...	...	...	
...	...	4	2	
G	...	2		...

Multiple Sequenzalignments, Scoring-Systeme (3 Punkte)

a) Gegeben seien die folgenden Sequenzen: AATT, ACTT, GGTT, CTGG, TTTT. Wie lautet der $\log_{10}$-odds-score für A in der zweiten Spalte des zugehörigen Profils? Geben Sie bitte die Zwischenschritte der Berechnung an. **Skizzieren** Sie für diese Spalte **qualitativ** ein Sequenzlogo, das sich ergibt, wenn nur die $f(a_i, 2)$-Werte verwendet werden. (2 Punkte)

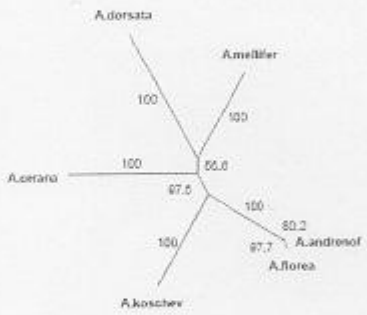
b) Bei einem Proteindesign-Experiment müssen Sie an einer Position im Protein die Aminosäure Isoleucin ersetzen. Hierbei wollen Sie einen möglichst konservativen Austausch vornehmen, d.h. eine Aminosäure einbauen, die Isoleucin möglichst ähnlich ist. Stützen Sie Ihre Entscheidung auf die unten angegebene BLOSUM Matrix und wählen sie zwei Aminosäuren, die hierfür in Frage kommen. Begründen Sie ihre Wahl. (1 Punkt)

A	Ala	4																		
R	Arg	-1	5																	
N	Asn	-2	0	6																
D	Asp	-2	-2	1	6															
C	Cys	0	-3	-3	-3	9														
Q	Gln	-1	1	0	0	-3	5													
E	Glu	-1	0	0	2	-4	2	5												
G	Gly	0	-2	0	-1	-3	-2	-2	6											
H	His	-2	0	1	-1	-3	0	0	-2	8										
I	Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4									
L	Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4								
K	Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5							
M	Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5						
F	Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6					
P	Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7				
S	Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4			
T	Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5		
W	Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	
Y	Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7
V	Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

Phylogenie (3 Punkte)

a) Für die Spezies A, B, C, D, E ist ein Phylogeniebaum mithilfe des Quartett-Puzzle-Ansatzes zu konstruieren und es sei die Spezies A einzufügen. Begonnen wird mit dem Teilbaum BC || DE. Die erste Nachbarschaftsrelation, die für das Einfügen von A bewertet wird, sei AE || CD. Wie sind die Kanten des Baumes BC || DE nach diesem Schritt markiert? (1 Punkt)

b) Sie haben für 6 Ameisenarten einen phylogenetischen Baum berechnet. Das Bootstrapping ergab die unten eingetragenen Werte. Wenn Sie nur Kanten mit Bootstrap-Werten jenseits von 75% trauen: Was schließen Sie aus diesem Baum für die Verwandtschaftsbeziehungen von *A. cerana*, *A. dorsata* und *A. melifer* untereinander und zu den restlichen Arten? (1 Punkt)



c) Sie haben einen phylogenetischen Baum konstruiert und eine Outgroup verwendet. Welche Baum-Topologie erwarten Sie, wenn die Berechnung erfolgreich war? Verwenden Sie eine Skizze zur Beantwortung der Frage. (1 Punkt)

Hidden-Markov-Modelle (4 Punkte)

a) In einem zeitweise unehrlichen Kasino werden zwei Münzen M_1 und M_2 im Wechsel verwendet. Bei Münze M_1 treten Kopf und Zahl jeweils mit gleicher Wahrscheinlichkeit auf, bei Münze M_2 ist p(Kopf) = 1/3. Wird Münze M_1 verwendet, so wird im folgenden Wurf mit $p = 0.91$ M_1 eingesetzt. Auf M_2 folgt mit $p = 0.73$ wiederum M_2. Zu Beginn werde mit $p = 0.2$ Münze M_2 gewählt. Zeichnen Sie ein Zustandsdiagramm und geben Sie sämtliche Wahrscheinlichkeiten an. (2 Punkte)

b) Wie groß ist für obiges Modell die Wahrscheinlichkeit für die Folge "Start, M_2, Kopf, M_1, Kopf"? Geben Sie das Ergebnis als Produkt der Einzelwahrscheinlichkeiten an, ohne den Zahlenwert zu berechnen.
(1 Punkt)

c) Wie werden die Emissionswahrscheinlichkeiten für ein HMM der PFAM-Datenbank bestimmt? Kurze Beschreibung in wenigen Sätzen! (1 Punkt)