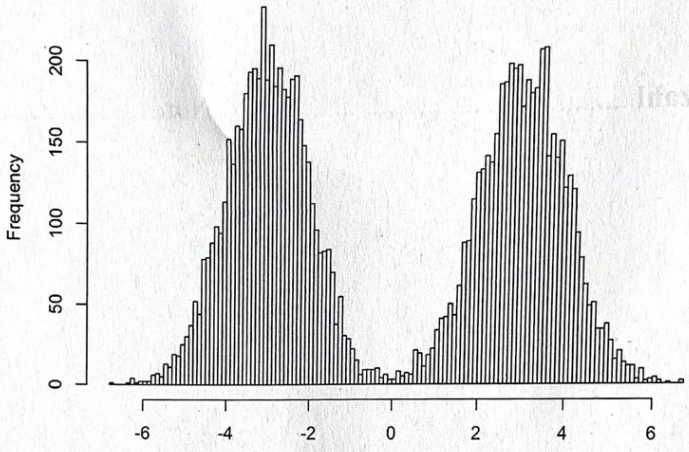
Biostatistik, insgesamt 25 Punkte

Deskriptive Statistik (8 Punkte)

a) Sie haben eine metrische Variable gemessen und die unten abgebildete praktisch symmetrische Verteilung erhalten. Zeichnen Sie per Hand geschätzte Werte für i) Mittelwert, ii) Median, iii) 1. und 3. Quartil ein. (1 Punkt)

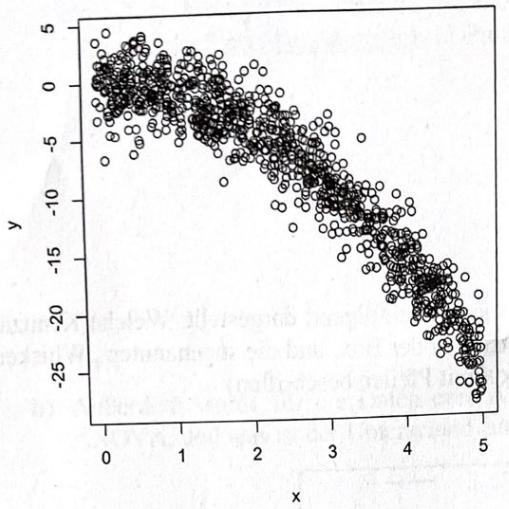


b) Stellen Sie sich vor, wir ziehen wiederholt zufällig x% der Beobachtungen aus den oben dargestellten Daten. Mit welchem funktionalen Zusammenhang ändert sich i) die Varianz ii) der Standardfehler des Mittelwerts mit der Größe x der Stichprobe? Anmerkung: zufällige Schwankungen und eventuelle Fehler durch eine kleine Stichprobengröße können hier vernachlässigt werden, es geht um die grundliegende Abhängigkeit der beiden Maßzahlen von der Stichprobengröße. (1 Punkt)

- Varianz:

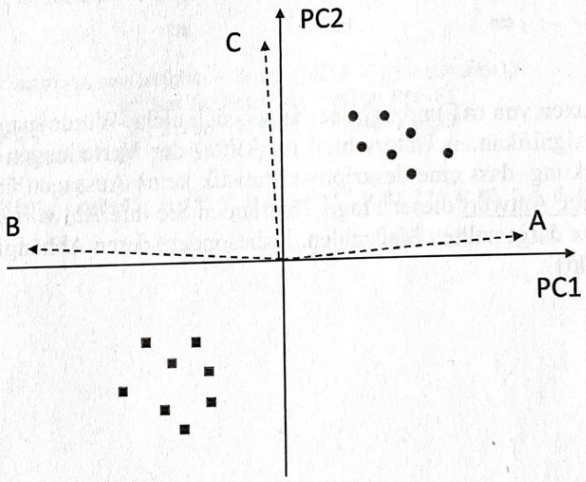
- Standardfehler:

c) Ist die Rangkorrelation nach Spearman zwischen den beiden Variablen x,y in der folgenden Abbildung eher positiv, negativ, oder Null? Wie würde sich der Absolutwert (Betrag) der Korrelation ändern, wenn man stattdessen die Pearson Korrelation berechnet? (1 Punkt)



d) Welche Aussagen über den folgenden Biplot einer PCA von 3 Variablen (A,B,C) sind richtig, und welche falsch? Einfach jeden Buchstaben mit wahr (w), falsch (f) markieren. (2 Punkte)

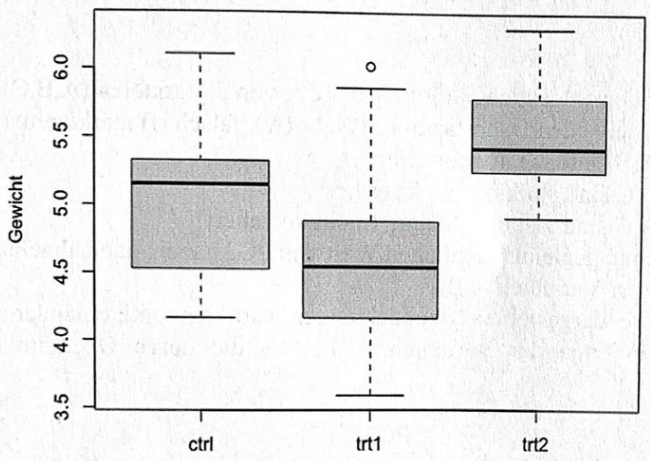
a. Die Variablen A,B sind negativ korreliert
b. Die Variablen B,C sind stark positiv korreliert
c. Die Variablen A,C sind fast unabhängig (nicht korreliert)
d. Wenn 2 Beobachtungen einen ähnlichen Wert von PC2 haben, dann ähneln sie sich stark in allen Werten der Variablen A,B,C
e. Die durch Kreise dargestellten Beobachtungen variieren untereinander mehr als 5x stärker in den Werten der Variablen A,B,C als die durch Quadrate dargestellten Beobachtungen



e) In einem Experiment wurde das Wachstum einer Pflanzenart (Gewicht nach 3 Monaten) unter 3 verschiedenen Bedingungen (ctrl, trt1, trt2) gemessen (siehe nachfolgender Boxplot). Was sind die Skalenniveaus der 2 Variablen in diesem Datensatz (Variable 1 = Gewicht, Variable 2 = Behandlung-> ctrl, trt1, trt2) (1 Punkt).

- Variable 1 = Gewicht:
- Variable 2 = Behandlung:

f) Die Ergebnisse dieses Experiments sind nachfolgend dargestellt. Welche Kennzahlen werden durch die Box selbst, den dicken Strich in der Box, und die sogenannten „Whisker" dargestellt (1 Punkt)? (Hinweis: einfach die Box mit Pfeilen beschriften)



g) Die Boxen von trt1 und trt2 überlappen sich nicht. Würde man deshalb erwarten, dass ein t-test einen signifikanten Unterschied im Mittel der Verteilungen ergibt? Hinweis: die pauschale Bemerkung, dass eine deskriptive Statistik keine Aussagen über Signifikanz zulässt, ist keine zulässige Antwort dieser Frage. Begründen Sie ihre Antwort über die Eigenschaften der durch die Box dargestellten Maßzahlen, insbesondere deren Abhängigkeit von der Stichprobengröße. (1 Punkt)

Schließende Statistik (11 Punkte)

a) Um den Unterschied in trt1 und trt2 (vorherige Frage) formal auf Signifikanz zu testen, wurde ein t-Test durchgeführt. Nennen Sie die vollständige Nullhypothese H0 für diesen Tests, inklusive der Verteilungsannahme (1 Punkt)

b) Außerdem wurde für die Daten eine ANOVA gerechnet. Was ist die Nullhypothese der ANOVA, und was ist der Unterschied zum t-test? (1 Punkt)

c) Nachfolgende das Ergebnis der ANOVA i) Ist das Ergebnis signifikant? ii) Wird die Nullhypothese abgelehnt? iii) Wie würden Sie das Ergebnis des Tests in einen Satz in ihrer Bachelorarbeit angeben? (1 Punkte)

```
> summary(aov(weight ~ group, data = PlantGrowth))
            Df Sum Sq Mean Sq F value Pr(>F)
group        2  3.766  1.8832   4.846 0.0159 *
Residuals   27 10.492  0.3886
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

d) In einer klinischen Studie taucht der folgende Satz auf: "Der p-Wert (t-test zwischen Änderung des Gewichts Kontrolle gegen Behandlungsgruppe) war 3%, also besteht eine 97% Wahrscheinlichkeit, dass das Medikament das Gewicht der Patienten beeinfluss" – was ist an dieser Aussage falsch? Begründen Sie die Antwort, indem Sie die Definition des p-Wertes erklären (1 Punkt).

e) Definieren Sie kurz die Typ I Fehlerrate und die False Discovery Rate (FDR). Was ist der Unterschied zwischen den beiden Konzepten? (1 Punkt)

f) Geben Sie die Formel für die False Discovery Rate (FDR) an. Ist es möglich, eine FDR < 5% zu erreichen? Wenn ja, unter welchen Bedingungen? Wenn nein, warum nicht? (1 Punkt)

g) Nachfolgend die Ergebnisse einer linearen Regression in R. In dem Regressionsmodell wurde eine mögliche Abhängigkeit zwischen Lufttemperatur und Wind untersucht. Beantworten Sie die folgenden Fragen: i) welche Form der Abhängigkeit wurde hier unterstellt? ii) Würde man aus den Ergebnissen schließen, dass es eine Abhängigkeit gibt, und woran sieht man das? iii) In welche Richtung geht die Abhängigkeit, und woran sieht man das? iv) Wo in der Tabelle wird das sogenannte Konfidenzintervall angegeben? (2 Punkte)

```
Call:
lm(formula = Temp ~ Wind, data = airquality)

Residuals:
    Min      1Q  Median      3Q     Max
-23.291  -5.723   1.709   6.016  19.199

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  90.1349     2.0522  43.921  < 2e-16 ***
Wind         -1.2305     0.1944  -6.331 2.64e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

h) Die Parameter der Regression werden über den Maximum-Likelihood-Schätzer (MLE) bestimmt. Bei dieser Methode sucht man die Parameter, die eine bestimmte Wahrscheinlichkeit maximieren. Um welche Wahrscheinlichkeit handelt es sich hier? (1 Punkt)

i) Im Folgenden ein generalisiertes lineares Modell, in dem die Abhängigkeit des Überleben auf der Titanic von Geschlecht und Passagierklasse untersucht wird. i) um welches GLM handelt es sich hier? Geben Sie Verteilungsannahme und Linkfunktion an (1 Punkt) ii) gibt e einen signifikanten Unterschied zwischen Passagierklasse 2 und 3, d.h. hat ein 3. Klass Passagier eine signifikant höhere Mortalitätswahrscheinlichkeit als ein 2. Klasse Passagier (1 Punkt)

```
Call:
glm(formula = survived ~ sex + passengerClass, family = "binomial",
    data = TitanicSurvival)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1089  -0.6984  -0.4741   0.7167   2.1173

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         2.1091     0.1728  12.203  < 2e-16 ***
sexmale            -2.5150     0.1467 -17.145  < 2e-16 ***
passengerClass2nd  -0.8808     0.1977  -4.456 8.34e-06 ***
passengerClass3rd  -1.7231     0.1715 -10.047  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Versuchsplanung (6 Punkte)

a) Erklären Sie das Konzept eines Störfaktors. Warum sind Störfaktoren so gefährlich, d.h. welche Folgen kann ein unbeachteter Störfaktor in einer Datenanalyse haben? (1 Punkt)

b) In einer Beobachtungsstudie soll untersucht werden, ob Glyphosat krebserregend ist. Hierzu wurde eine repräsentative Stichprobe deutscher Landwirte bzgl. ihrer Glyphosatexposition und Krebserkrankungen befragt. Warum ist Rauchen (Tabakkonsum der Landwirte) wahrscheinlich KEIN Störfaktor bzgl. dieser Frage? (1 Punkt)

c) Welche Annahme müsste man treffen, damit Rauchen ein Störfaktor für die oben genannte Frage wird? Erfinden Sie einen nachvollziehbaren Grund für diese Annahme, auch wenn er unwahrscheinlich ist. (1 Punkt)

d) Stellen Sie sich vor, Sie wollen in Ihrer BSc Arbeit den Effekt eines RNA Präparats auf da Wachstum von Mais untersuchen. Sie planen ein Design mit 10 Replikaten (jeweils Behandlun / Kontrolle). Da fällt ihnen auf: der Student vor Ihnen hat ja auch schon ein ähnliches Experimen gemacht, und hatte auch 10 Maispflanzen als Kontrolle. Könnten Sie nicht einfach seine Wert nehmen und sich so die Kontrolle sparen? Ihr Betreuer rät Ihnen davon ab. Warum? (1 Punkt)

e) Was ist ein Blockdesign, und was ist die Idee / der Nutzen dieses Designs? (1 Punkt)

f) Schreiben Sie hinter jeden der folgenden Punkte, ob dieser die Teststärke (Power) einer typischen Analyse (z.B. t-test, Regression) beeinflusse würden, und wenn ja, in welche Richtung. Es reicht ein + für eine größere Teststärke, und ein - für eine geringere Teststärke, und eine 0 für einen Faktor, der keinen Einfluss hat (1 Punkt)

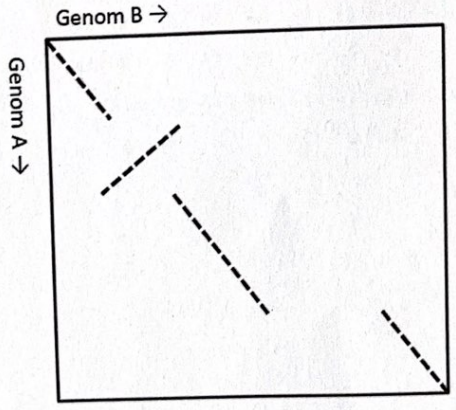
- Stichprobengröße
- Datum des Experiments
- Effektstärke
- Effektrichtung
- Varianz / Stochastizität
- Stärkere Balance

Bioinformatik, insgesamt 25 Punkte

Sequenzvergleich und Datenbanken (9 Punkte)

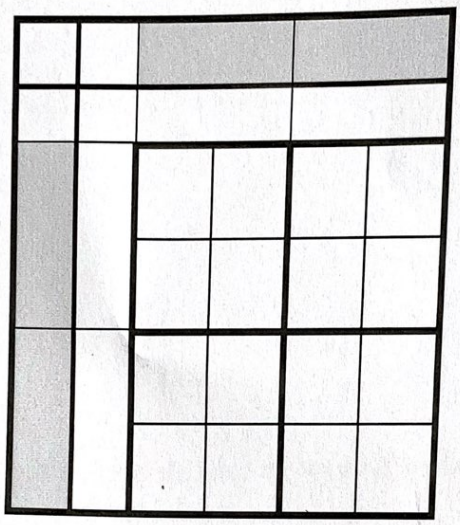
a) Konstruieren Sie einen Suffixbaum für die Sequenz A = CTCTCTTC. Markieren Sie bitte die Wurzel. (1 Punkt)

b) Beim Vergleich der Genome A, B zweier nahe verwandter mikrobieller Arten mithilfe eines Dotplots entstand das unten angegebene Muster. Sie wissen, dass B in einer größeren Pathogenitätsinsel Gene enthält, die im Genom A nicht vorkommen. Wo liegen diese? Bitte markieren. (1 Punkt)



c) Berechnen Sie für die beiden Sequenzen A = CW und B = WR sämtliche Scorewerte entsprechend dem Smith-Waterman-Algorithmus. Gaps werden mit -4, bewertet, für den Vergleich der

Aminosäuren benutzen Sie bitte die Scores aus der BLOSUM-Matrix, die in der Aufgabe a) z
Multiplen Sequenzalignments und Scoring-Systemen angegeben ist. Übertragen Sie die Sequenze
in die grau markierten Zellen. Tragen Sie alle Teilergebnisse in die folgende Matrix ein. (5 Punkte



d) Sie sollen ein additives Scoring-Schema entwickeln, das aus 64 L_j-Termen besteht. Damit sollen in der DNA eines *Bacillus*-Stammes mithilfe eines Klassifikators protein-codierende Sequenzen von zufällig zusammengesetzten unterschieden werden. Welche Daten benötigen Sie für die Berechnung und welches Berechnungsverfahren schlagen Sie für die L_j-Terme vor? Geben Sie eine Formel an und begründen Sie deren Aufbau. (2 Punkte)

Multiple Sequenzalignments, Scoring-Systeme (3 Punkte)

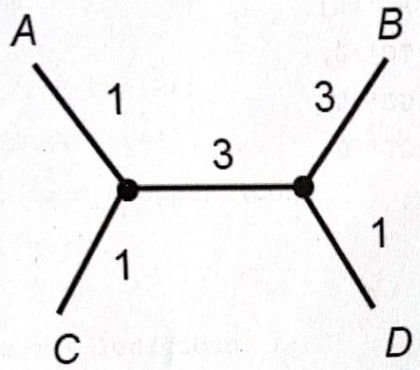
a) Bei einem Proteindesign-Experiment müssen Sie ein Asp-Residuum ersetzen. Sie wollen einen möglichst konservativen Austausch vornehmen, d. h. eine Aminosäure einbauen, die Asp möglichst ähnlich ist. Benutzen Sie für Ihre Entscheidung die unten angegebene BLOSUM-Matrix und wählen Sie zwei Aminosäuren, 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

b) Gegeben seien die folgenden Sequenzen: AGTT, GATT, AGTT, CTTA, TCTT. Wie lautet der $\log_{10}$-odds-score für A in der letzten Spalte des zugehörigen Profils? Geben Sie bitte die Zwischenschritte der Berechnung an. Skizzieren Sie für diese Spalte qualitativ ein Sequenzlogo. (2 Punkte)

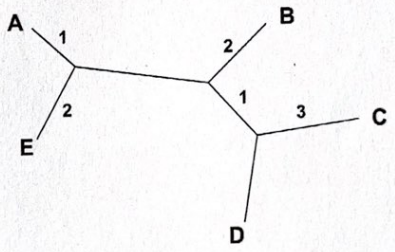
Phylogenie (2 Punkte)

a) Für die Spezies A, B, C, D, E ist ein Phylogeniebaum mithilfe des Quartett-Puzzle-Ansatzes zu konstruieren und es sei die Spezies E einzufügen. Begonnen wurde mit dem Teilbaum AC || BD und es wurden bereits mehrere Relationen bearbeitet. Einzutragen ist noch BE || CD. Wie sind die Kanten des Baumes markiert, nachdem die genannte Relation verarbeitet ist und wo würde E dann platziert werden? (1 Punkt)



b) Gegeben sei für fünf Sequenzen A – E die links gezeigte Distanzmatrix. Können Sie nachweisen, das dies eine additive Matrix ist? Tragen Sie die zwei fehlenden Gewichte im rechts gezeigten Baum ein. (1 Punkt)

	A	B	C	D	E
A	0	7	9	7	3
B		0	6	4	8
C			0	4	10
D				0	8
E					0



Programmieren mit Python (4 Punkte)

a) Welche Funktion hat dieses Programm? Wie lauten die Zahlenwerte, die mit der letzten print-Anweisung ausgegeben werden? (1 Punkt)

Wie ist die Laufzeit? Begründen Sie Ihre Aussage. (1 Punkt)

```
dna = "AATGCTAGTAAT"
dn = {'AA':0,'AT':0,'AG':-1,'AC':0}
#       'TA':0,'TT':0,'TG':0,'TC':0,
#       'GA':0,'GT':0,'GG':0,'GC':0,
#       'CA':0,'CT':0,'CG':0,'CT':0

for i in range(len(dna)):
    key=dna[i:i+2]
    try:
        dn[key]=dn[key]+1
    except:
        print('!')
print(dn)
```

b) Korrigieren Sie mindestens 2 Syntaxfehler dieses Programmes. (2 Punkte)

```
A="ATGCGCTGAGGCTGGTGA

codons_A={}

i=0
while i < len(A):
    cdn=A[i:i+4]
    codons_A.append(cdn)
    i==i+3

while i < len(codons_A)
    print("Codons at position",i+1,codons_A[i])
    i=i+1
```

Hidden-Markov-Modelle (5 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 is p(Kopf) = 1/10. Wird Münze M_1 verwendet, so wird im folgenden Wurf mit $p = 0.95$ M_1 eingesetzt Auf M_2 folgt mit $p = 0.75$ wiederum M_2. Zu Beginn werde mit $p = 0.8$ Münze M_2 gewählt. Zeichnen Sie ein Zustandsdiagramm und geben Sie sämtliche Wahrscheinlichkeiten an. (2 Punkte)

b) Unter Verwendung des oben beschriebenen Modells sind die Viterbivariablen für $t = i$ zu errechnen. Die Werte der Viterbivariablen für $t = i$-1 sind angegeben. Als nächstes Symbol x_i wird „Z" emittiert. Berechnen Sie für die zwei Zustände „M_1" und „M_2" den Wert der Viterbivariablen v_i. Tragen Sie in der Tabelle auch die Produktterme (Zahlenwerte) der zu vergleichenden Teilergebnisse (TE1, TE2) ein. (2 Punkte)

Zustände		**Viterbi-Variablen**	
		$t = i$-1	$t = i$, x_i = Z
M_1	....	0.2	TE1: TE2:
M_2	...	0.8	TE1: TE2:

c) Aus dem unten angegebenem MSA sollen die Parameter für ein Profil-HMM abgeleitet werden. Wie viele Match-Zustände ergeben sich und welche Emissionswahrscheinlichkeiten gelten für den Match-Zustand sieben? Die Korrektur der Werte durch Pseudocounts ist hier nicht gefordert; geben Sie den Rechenweg an. (1 Punkt)

```
LCIEF-PCCH
L-IEFTPCVH
L-YETTPICH
L-IE-TPCCH
```

Homologie-Modellierung (2 Punkte)

a) Begründen Sie, weshalb für Proteine 3D-Strukturen per Homologiemodellierung berechnet werden können. Was ist der entscheidende bioinformatische Befund? (1 Punkt)

b) Bei der Suche nach Templaten für ein Target T wurden zwei geeignete Proteine P1 und P2 gefunden. Im paarweisen Sequenzvergleich ist P1 ähnlicher zu T als P2. Das Alignment T, P1 weist Lücken auf, während das Alignment T, P2 keine Lücken enthält. Welches Templat wählen Sie aus? Begründung! (1 Punkt)