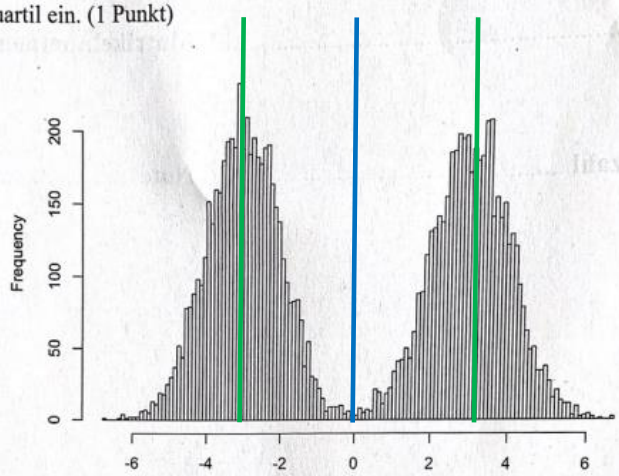
Lösung zum 1. Versuch Statistik und Bioinformatik 2024

Das ist keine Musterlösung, also alle Antworten ohne Gewähr. Für einige Aufgaben gibt es auch mehrere akzeptierte Lösungen, also es muss nicht alles andere falsch sein.

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)



1. Quartil Median/MW 3. Quartil

Verteilung symmetrisch → Median ist ungefähr der Mittelwert

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)

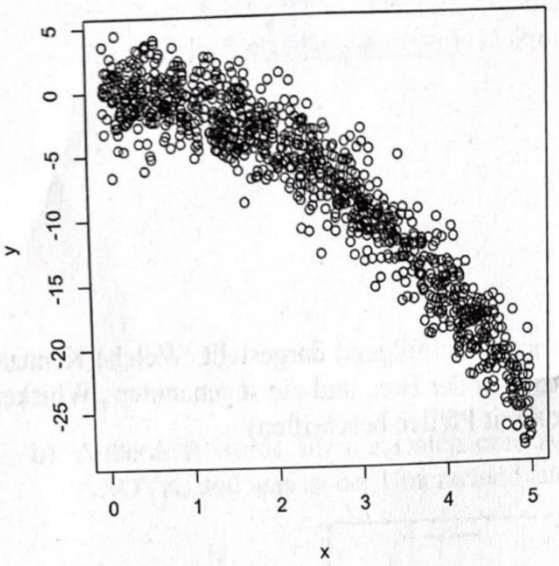
Varianz bleibt näherungsweise gleich / ist unabhängig von Stichprobengröße

Standardfehler wird mit größerer Stichprobe kleiner

Varianz ist unabhängig von Stichprobengröße, also vernachlässigbar

$$(se = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{var}}{\sqrt{n}} \sim \frac{1}{\sqrt{n}})$$

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)

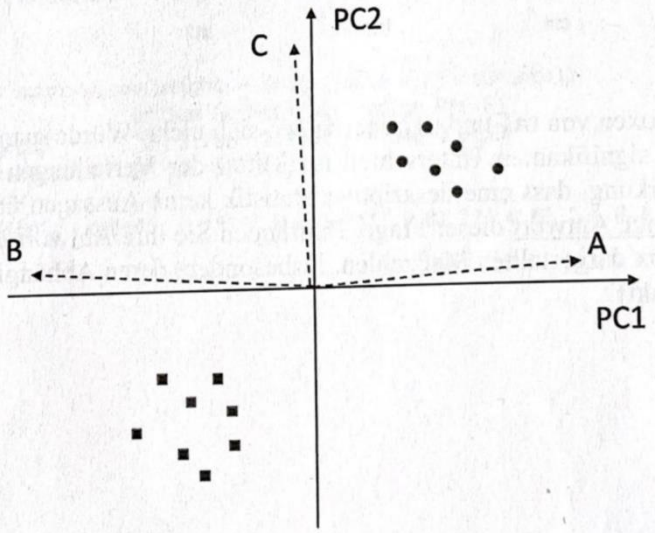


(Rang-)Korrelation ist negativ

Absolutwert der Korrelation würde für Pearson kleiner werden (weil Pearson lineare Zusammenhänge misst und deswegen nicht-lineare Beziehungen unterschätzt).

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)

w a. Die Variablen A,B sind negativ korreliert
f b. Die Variablen B,C sind stark positiv korreliert
w c. Die Variablen A,C sind fast unabhängig (nicht korreliert)
f d. Wenn 2 Beobachtungen einen ähnlichen Wert von PC2 haben, dann ähneln sie sich stark in allen Werten der Variablen A,B,C
f 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

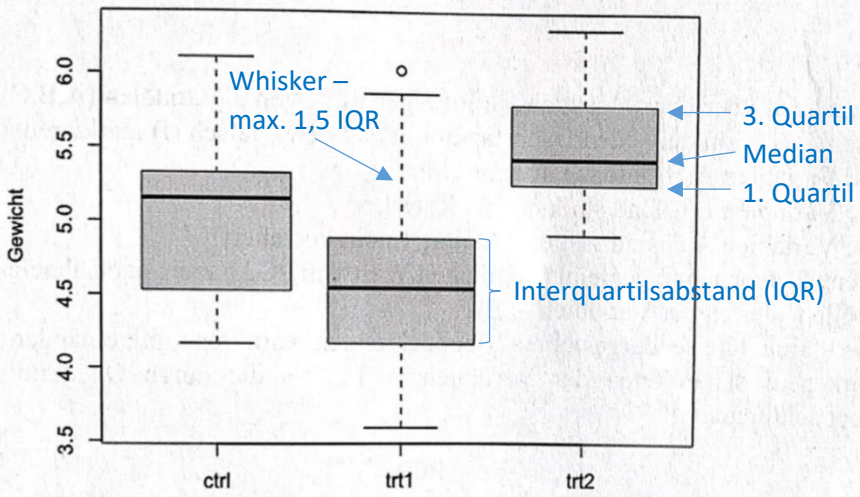


(Variablen, die in dieselbe Richtung zeigen → positiv korreliert
Variablen, die in die Gegenrichtung zeigen → negativ korreliert
Variablen, die senkrecht aufeinander stehen → nicht korreliert
Beobachtungen mit ähnlichem PC2-Wert ähneln sich in C, können aber für A und B unterschiedlich sein)
Reihenfolge der Antwortmöglichkeiten in der Klausur evtl. anders

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: numerisch/metrisch
- Variable 2 = Behandlung: nominal oder ordinal (müsste hier beides gehen)

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) Was stellt eine Kontingenztabelle oder Kreuztabelle dar (1 Punkt)?

Vorkommenshäufigkeit der möglichen Kategoriekombinationen zweier kategorischer Variablen.

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)

H0: Mittelwerte von trt1 und trt2 sind gleich

Verteilungsannahme: Normalverteilung

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)

H0: Mittelwerte aller Gruppen sind gleich / Variable (Behandlungen) hat keinen Effekt

T-Test testet immer zwischen 2 Gruppen, ANOVA zwischen allen Gruppen (auch mehr als 2)

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
```

1) Ja, ist signifikant (weil p < 0,05)
2) Ja, H0 wird abgelehnt
3) Die Mittelwerte der Gruppen unterscheiden sich signifikant (ANOVA, p = 0.0159)

d) Definieren Sie den p-Wert? (1 Punkt)

Der p-Wert ist die Wahrscheinlichkeit, den beobachteten oder einen extremeren Wert für die Teststatistik zu erhalten, wenn H0 wahr ist.

e) Wie Sie wahrscheinlich wissen, gibt der p-Wert nicht die Wahrscheinlichkeit der Nullhypothese p(H0) an. Falls wir an dieser Wahrscheinlichkeit interessiert sind - mit Hilfe welcher Methode, die in der Vorlesung besprochen wurde, könnten wir diese berechnen? (1 Punkt)

Bayes'sche Posterior

(FDR geht hier nicht. FDR ist die Rate der falsch positiven unter den signifikanten Ergebnissen und setzt damit voraus, dass der Wert schon signifikant war – was nicht unbedingt der Fall ist)

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)

$$FDR := \frac{p(H0) \cdot \alpha}{p(H0) \cdot \alpha + p(!H0) \cdot (1 - \beta)}$$

Ja ist möglich

Bedingungen:

- Ein großer Teil der getesteten Hypothesen haben wirklich einen Effekt, also kleines p(H0) → im Extremfall p(H0) = 0 wäre die FDR immer 0
- Zusätzlich möglichst hohe Power (oder niedriges alpha)

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
```

1) Linear (sonst würde da sowas wie Temp ~ Wind + Wind^2 stehen)
2) Ja, am p – Wert < 0,05 → signifikant
3) Negative Korrelation → Schätzer für den Effekt von Wind negativ
4) Standardfehler → entspricht einem 68% CI

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)

Maximierung der Wahrscheinlichkeit, die beobachteten Daten unter bestimmten Parametern zu erhalten.

Parameter sind hier die Parameter der Geradengleichung, Steigung und Intercept

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
```

1) Logistische Regression
   Verteilung: Binomial
   Logit-Link → link $^{-1}$ = exp (x) / 1 + exp (x)
2) Kann man nicht sagen… Die p-Werte für die 2te und 3te Klasse beziehen sich beide auf den Unterschied zum Intercept, also der ersten Klasse. Für den Unterschied zwischen 2. und 3. Klasse müsste man die 2. oder die 3. Klasse als Referenz nehmen

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)

Störfaktoren beeinflussen die abhängige und erklärende Variable und können so Scheinkorrelationen (Korrelation, der kein kausaler Zusammenhang zu Grunde liegt) erzeugen.

Die Scheinkorrelation kann dazu führen, dass die tatsächliche Korrelation verdeckt oder sogar umgedreht wird. Dadurch können Variablen scheinbar in beliebiger Richtung korrelieren, unabhängig davon, ob tatsächlich ein Effekt existiert oder nicht.

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)

Rauchen beeinflusst nicht beide Variablen. Es besteht zwar ein Zusammenhang zwischen Krebs und Rauchen, aber die Glyphosatexposition wird durch Rauchen nicht beeinflusst. Also ist Rauchen in diesem Fall kein Störfaktor

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)

Z. B. Landwirte rauchen generell mehr als andere Personengruppen, dann wäre Rauchen mit Landwirten assoziiert.

Oder Zigaretten enthalten Glyphosat, dann würde Rauchen direkt die Glyphosatexposition beeinflussen.

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)

Für ein valides Experiment sollte die Kontrolle bis auf die untersuchte Bedingung in allen anderen Bedingungen gleich zur Behandlung sein → wäre nicht der Fall, wenn man Werte des Vorgängers übernimmt

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

Bei einem Blockdesign wird immer Behandlung und Kontrolle in einen Block zusammengebracht, also paarweise nebeneinander platziert. Durch die Nähe von Kontrolle und Behandlung zueinander wird sichergestellt, dass die Bedingungen für beide gleich sind und sich unbeobachtete Effekte gleichermaßen auf beide auswirken. Die Blöcke sollten möglichst weit entfernt vom nächsten Block sein, um Unabhängigkeit sicherzustellen

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 0
- Effektstärke +
- Effektrichtung 0
- Varianz / Stochastizität -
- Stärkere Balance +

Bioinformatik, insgesamt 25 Punkte

Sequenzvergleich (7 Punkte)

a) Für welche Aufgabe wurde der Needleman-Wunsch-, für welche der Smith-Waterman-Algorithmus entwickelt? (0,5 Punkte)

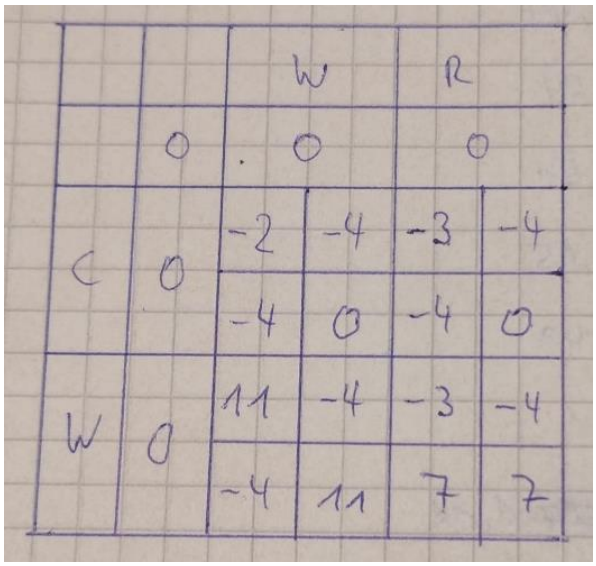
Needleman-Wunsch: Globale Alignments

Smith-Waterman: Lokale Alignments

b) Wofür steht das N in der Blosum N Matrix? (0,5 Punkte)

Maximal N% Sequenzähnlichkeit der Blöcke, aus denen die Matrix erstellt wurde

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) zu Multiplen Sequenzalignments und Scoring-Systemen angegeben ist. Tragen sie alle Teilergebnisse in die folgende Matrix ein. Achtung die BLOSUM Matrix steht auf der nächsten Seite! (5 Punkte)



Smith-Waterman geht wie Needleman-Wunsch, außer:

- Statt Lücken kommen ganz außen Nullen statt den Gap-Scores rein
- Bei negativen Zahlen stattdessen 0 in das max. Score-Feld schreiben
- Scores für (Mis-)Match unten in der Matrix ablesen

d) Bei einem Proteindesign-Experiment müssen Sie ein Gln-Residuum ersetzen. Sie wollen einen möglichst konservativen Austausch vornehmen, d. h. eine Aminosäure einbauen, die Gln 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)

	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
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

Am besten Glu, dann Arg oder Lys, da sie die höchsten Scores haben. Die BLOSUM-Matrix beruht auf biochemischer Ähnlichkeit der AS, d.h. ein höherer Score bedeutet, dass 2 AS ähnlicher zueinander sind.

Strukturvorhersage (5 Punkte)

a) Beschreiben Sie kurz das Grundprinzip der Proteinfaltung, auf dem die theoretischen Methoden der Strukturvorhersage beruhen. (1 Punkte)

3D-Struktur eines Proteins ist von der Aminosäuresequenz festgelegt → ähnliche Sequenz bedeutet i.d.R. auch ähnliche Struktur/Funktion des Proteins

Bei der Faltung folgen Proteine Pfaden durch eine Energielandschaft, die als Trichter dargestellt werden kann. Die eingenommene Konformation entspricht dabei dem globalen Energieminimum, also dem niedrigsten Punkt des Trichters. Die Energielandschaft kann durch eine Energiefunktion physikalisch beschrieben werden, dadurch lässt sich die Faltung eines Proteins vorhersagen.

b) Nennen Sie vier wichtige Schritte des Arbeitsablaufs zur Homologiemodellierung und beschreiben Sie diese kurz. (2 Punkte)

- Finden eines geeigneten Template: Man sucht ein möglichst ähnliches, bereits aufgelöstes Protein als Vorlage (z. B. BLASTen)
- Alignment (Das finale Template wird mit der zu modellierenden Sequenz nochmal korrekt aligned)
- Modellierung: Backbone (Grundgerüst), dann die fehlenden Loops, dann die Seitenketten
- Optimierung: Strukturverfeinerung durch Energieminimierung und Moleküldynamik
- Validierung mit experimentellen Methoden, z. B. Spektroskopie

c) Welches zweite Konzept neben der Homologiemodellierung gibt es zur Strukturvorhersage? Benennen Sie dieses und beschreiben Sie es kurz. (1 Punkt)

Ab-initio-Strukturvorhersage
Berechnen der Struktur über die Energielandschaft nur aus der Aminosäuresequenz und physikalischen Parametern, ohne auf bekannte Strukturen ähnlicher Proteine zurückzugreifen.

d) Welche beiden Datenbanken wurden zum Training des neuronalen Netzes des Strukturvorhersagealgorithmus AlphaFold vewendet? (1 Punkt)

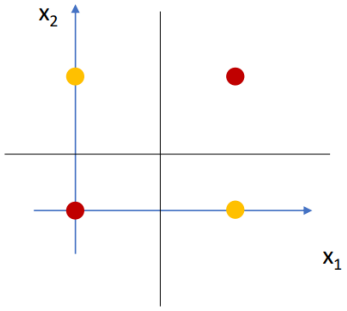
Genetische Sequenzdatenbank und Strukturdatenbank

Neuronale Netze (3 Punkte)

Die XNOR-Funktion (Exklusive-Nicht-Oder-Funktion) soll als neuronales Netz dargestellt werden. Skizzieren Sie die Lösbarkeit und die damit verbundene Architektur des minimal notwendigen neuronalen Netzes. Geben Sie außerdem die notwendigen weiteren Schritte in Stichpunkten an, die zur vollständigen Bestimmung das neuronalen Netzes notwendig sind. (3 Punkte)

Wahrheitstabelle:

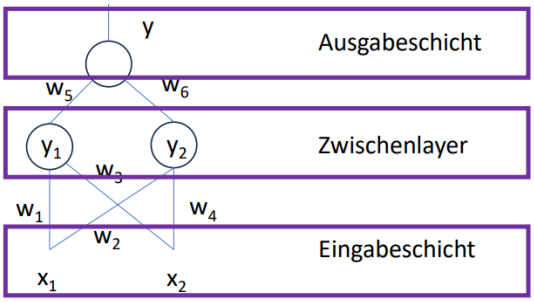
x1	x2	y
0	0	1
0	1	0
1	0	0
1	1	1



Y-Werte werden durch Farbe gekennzeichnet
0: rot
1: gelb

=> Nicht linear separierbar
=> NN mit nur einem Layer **nicht** möglich

Nur mit einem Layer nicht möglich => Zwischenlayer notwendig, Architektur wäre dann so:



Weitere Schritte:

- Festlegen einer geeigneten Schwellenwertfunktion
- Bestimmen der Gewichte w1 – w6.

Molekulardynamik (4 Punkte)

a) Wie ist die bioinformatische/physikalische Sichtweise eines Proteins im Kontrast zur biochemischen Sichtweise eines Proteins als Biomolekül, das aus Aminosäuren aufgebaut ist? (0,5 Punkte)

Bioinformatisch gesehen ist ein Protein ein Vektor aus x-, y- und z-Koordinaten

b) Ist die folgende Aussage richtig oder falsch? (0,5 Punkte)

Die Energieminimierung von Proteinsystemen führt normalerweise zu einem globalen Energieminimum

Falsch (meistens landet man erstmal nur in einem lokalen Minimum)

c) Erläutern Sie kurz die konzeptionellen Näherungen (nicht mathematische Formel) mit der eine Bindung zwischen zwei Atomen und die Atome selbst in einer klassischen molekularmechanischen Simulation beschrieben werden. (0,5 Punkte)

Bindung wird als Feder (verhält sich als harmonischer Oszillator), Atome als weiche Kugeln mit einer Ladung betrachtet

d) Nennen Sie vier Wechselwirkungspotentiale, die die Energielandschaft eines Proteins beschreiben. (2 Punkte)

- Bindungslänge
- Bindungswinkel
- Torsion
- Coulomb-Kraft
- Van-der-Waals-Wechselwirkungen

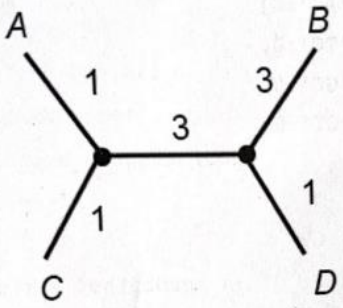
e) Ist die Folgende Aussage richtig oder falsch? (0,5 Punkte)

Im Rahmen einer quantenmechanischen Simulation können Bindungen aufbrechen.

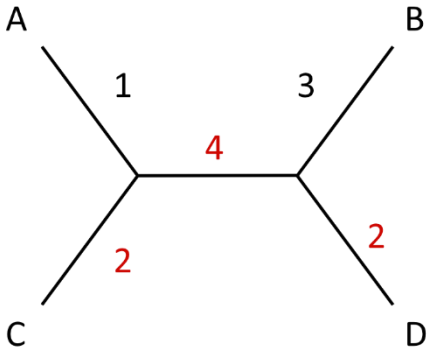
richtig

Phylogenie (2 Punkte)

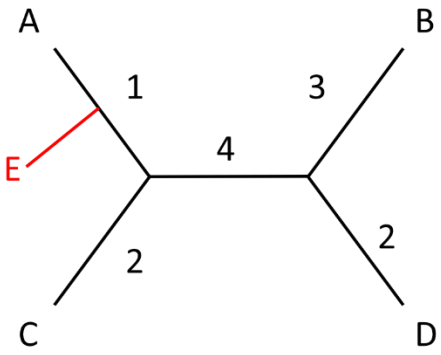
a) Für die Spezies A, B, C, D, E ist ein Phylogeniebaum mithilfe des Quartett-Puzzle-Ansatzes zu konstruieren und es sei die Spezies E einzufügen. Begonnen wurde mit dem Teilbaum AC || BD und es wurden bereits mehrere Relationen bearbeitet. Einzutragen ist noch BE || CD. Wie sind die Kanten des Baumes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



BE || CD eintragen:



E an Kante mit niedrigstem Score eintragen:



b) Beschreiben Sie kurz die Funktion von Outgroups. Wie ist eine solche zusammengesetzt? (1 Punkt)
Eine Outgroup ist eine Gruppe aus Sequenzen, die sich deutlich von den untersuchten Taxa unterscheidet (weniger eng mit ihnen verwandt sind). Sie wird verwendet, um die Wurzel eines phylogenetischen Baumes zu finden. Die Outgroups sollten dann auf einer Seite, die untersuchten Taxa auf der anderen Seite des Baumes stehen. Wenn die Outgroup nicht zusammen auf einer Seite, sondern verstreut liegt, ist der Baum nicht interpretierbar.

Programmieren mit Python (3 Punkte)

a) Welche Funktion hat dieses Programm? Wie lautet die erste Zeile der Ausgabe (2 Punkte)

```
def unknown(dna):
    site = 'TCGA'
    for i in range(len(dna) - 3):
        if dna[i:i+len(site)] == site:
            print("Prot site an Pos " + repr(i+1).rjust(2) + " : " +
dna[i:i+len(site)])
# Hauptprogramm
my_dna='TCGAGACGCTGGTCGACTGGATCGA'
unknown(my_dna)
```

Der Code sucht in einer gegebenen DNA-Sequenz (myDNA) nach der site ‚TCGA' und gibt die Positionen aus, an denen sie vorkommt.

Erste Zeile Ausgabe: Prot site an Pos 1 : TCGA

b) Korrigieren Sie die beiden Syntaxfehler dieses Programmes (1 Punkt)

```
A = "ATGCGCTGAGGCTGGTGA"
codons_A = []
i = 0

while i < len(A):
    cdn=A[i:i+4]
    codons_A.append(cdn)
    i==i+3 ← nur = statt dem ==
while i < len(codons_A):
    print("Codons at position", i+1, codons_A[i])
    i=i+1
```