Biostatistik (25 Punkte)(24 Punkte erreicht)

1a) Um Spam aus ihrer Mailbox zu fischen, installieren Sie einen Spam-Filter, der verschiedene Kriterien verwendet um Nachrichten als Spam zu erkennen und in den Mülleimer zu befördern. Nachrichten die nicht als Spam Mails erkannt werden, werden dagegen in die Mailbox weitergeleitet.

(i) Formulieren Sie das inhaltliche Hypothesenpaar bei diesem Testvorgang.

H_0: Nachricht ist kein Spam

H_1: Nachricht ist Spam

(ii) Beschreiben Sie inhaltlich die Konsequenzen eines Alpha/ Beta Fehlers beim Testen der Hypothesen

Alpha Fehler: Nachricht ist kein Spam, wird aber als Spam eingestuft

Beta Fehler: Nachricht ist Spam, wird aber nicht so eingestuft

(iii) Welcher der beiden Fehlerarten sollte ihrer Meinung nach in diesem Fall minimiert werden?

Alpha Fehler, weil sonst evtl. wichtige Emails verloren gehen

[2,5 P]

(b) Sie führen einen klinischen Test eines neuen Medikaments gegen eine bereits gut therapierbare Krankheit durch. Falls das neue Medikament zu einer signifikanten Verbesserung der Messwerte führt soll es auf den Markt. Falls die Ergebnisse nicht signifikant sind, soll die Forschung am Präparat eingestellt werden.

(i) Formulieren Sie das inhaltliche Hypothesenpaar bei diesem Testvorgang.

H_0: neues Medikament führt nicht zu einer signifikanten Verbesserung der Messwerte

H_1: neues Medikament führt zu einer signifikanten Verbesserung der Messwerte

(ii) Beschreiben Sie inhaltlich die Konsequenzen eines Alpha/ Beta Fehlers beim Testen der Hypothesen

Alpha Fehler: Neues Medikament ist nicht besser, wird aber als besser angesehen

Beta Fehler: Neues Medikament ist besser, wird aber als schlechter angesehen

(iii) Welcher der beiden Fehlerarten sollte ihrer Meinung nach in diesem Fall aus wissenschaftlicher Sicht (nicht aus wirtschaftlicher) minimiert werden?

Beta Fehler

[2,5 P]

2.) Kals & al. 2007 untersuchte die Auswirkung von unterschiedlichen dienstlichen Tätigkeiten von Feuerwehrmännern auf das Auftreten von tödlichen Herzinfarkten u. stellte sich die Frage ob diese Todesfälle vor allem bei der Bekämpfung von Bränden auftreten. Dazu wurden 449 Todesfälle untersucht.

Unterteilung dienstlicher Tätigkeiten: Feuerbekämpfung 2% Rüstzeit u. Aufräumarbeiten 16% physisches Training 8% Sonstiges 74%
Die Todesfälle verteilten sich wie folgt: Feuerbekämpfung 144; Rüstzeit u. Aufräumarbeiten 138; physisches Training 56; Sonstiges 111.

Überprüfen Sie die Nullhypotese, dass die Todesfälle von der dienstlichen Aktivität unabhängig sind. Geben Sie dabei (a) den zu verwendenden Test, (b) den Wert der Teststatistik (c) den kritischen Wert der Teststatistik (α=0,05) und (d) Ihre Testentscheidung an.

a) X^2-Test

b)X^2=2025+61+11+147=2244

c) $X^2_{n-1,1-\alpha} = X^2_{3;\ 0,95}$ =7,815

d) H_0 ablehnen → Todesfälle sind von der Tätigkeit abhängig

3.) In folgender Tabelle sind Informationen zu den Messungen der Körperlänge an Männchen und Weibchen einer Tierart:

Männchen	Weibchen
n=6	n=8
$\bar{x}_1$=74,8	$\bar{x}_2$=72.99
S_1=1,04	S_2=1,48

Prüfen Sie ob die beiden Gruppen bezüglich ihrer Körpergröße signifikant voneinander abweichen.

a) Verwendeter Test: t-Test zweiseitig
b) Wert für Teststatistik: t= 74,8-72,99 / 1,314 * sqrt[6*8 / 6+8] = 2,55
c) den kritischen Wert der Teststatistik (α=0,05): $t_{n1+n2-2,1-\alpha} = t_{12,\ 0,975}$ = 2,179
d) Ihre Testentscheidung: 2,55 > 2,179 → Ablehnen von H_0 → signifikanter Unterschied der Körpergröße zwischen den Geschlechtern
[4 P]

4.)Für eine Stichprobe von Kleinsäugern wurden folgende Messwerte in Gramm ermittelt:

19,4 21,4 22,3 22,1 20,1 23,8 24,6 19,9 21,5 19,1

Berechnen Sie:

a) Arithmetisches Mittel: $\bar{x}$=(19,4+21,4+22,3….19,1)*0,1 = 21,2
b) Varianz: 1/9 * $\sum_{i=1}^{10} (x_i-\bar{x})^2 = 30{,}536*1/9 = 3{,}393 = s^2$
c) 95% Konfidenz Intervall des Mittelwerts
   S= sqrt[s^2] = 1,842 s_x = 0,582
   $T_{9,\ 0{,}975}$ = 2,262
   G_u= x- $T_{9,\ 0{,}975}$ *s_x = 20,10
   G_o= x+ $T_{9,\ 0{,}975}$ *s_x = 22,74
   [3 P]

5.) Ein Hundezüchter ist an der Anzahl weiblicher Welpen in einem Wurf interessiert. Berechnen sie die Sie die Wahrscheinlichkeitsverteilung für x(Weibchen im Wurf; Wurfgröße=5 wenn die Wahrscheinlichkeit für Weibchen=Männchen ist.

P(X=0) = $\binom{5}{0}$ $0{,}5^5$ = 1/32

P(X=1) = $\binom{5}{0}$ $0{,}5^5$ =5/32

P(X=2) = $\binom{5}{0}$ $0{,}5^5$ =10/32

P(X=3) = $\binom{5}{0}$ $0{,}5^5$ =10/32

P(X=4) = $\binom{5}{0}$ $0{,}5^5$ =5/32

P(X=5) = $\binom{5}{0}$ $0{,}5^5$ =1/32

[3 P]

6.) Die Wahrscheinlichkeit an Porphyrie zu erkranken, ist 1 zu 10^4. Ein biochemischer Test ergibt bei 500 Erkrankten in 410 der Fälle ein positives Ergebnis. Bei 500 gesunden ergibt sich allerdings auch in 19 Fällen ein positives Ergebnis. Geben Sie die Wahrscheinlichkeit an bei einem positiven Testergebnis tatsächlich erkrankt zu sein.

P(krank)=1/10.000 P(gesund)=9.999/10.000

P(positiv|krank)= 410/500 P(positiv)=429/500

P(positiv|gesund)= 19/500

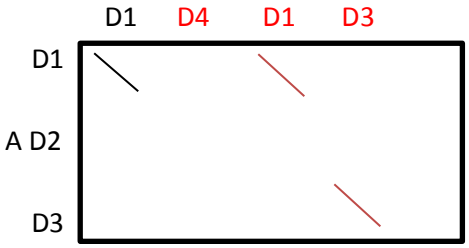
P(krank|positiv)= $\frac{P(positiv|krank)*P(krank)}{P(positiv)}$

[4 P]

Statistik/ Boinformatikklausur 2013 Versuch1

Bio Informatik

a) Sie vergleichen 2 Proteine A u. B mit Hilfe eines Dotplots. In beiden Proteinen kommen nur 4 Domänen D1 D2 D3 u. D4 vor die untereinander keine Sequenzähnlichkeit aufweisen.
In A: D1 D2 D3
In B: D1 D4 D1 D3
Skizzieren Sie den Dotplot



b) Gegeben sei der folgende Algorithmus. Bestimmen Sie seine Funktion am Beispiel A=RSTWLL u. B=WTS

Eingabe: Zwei Zeichenketten
A:= a1, a2, … an u. B:= b1, b2, … bm

Algorithmus:
Beginn
Für i=1,2, … n führe aus
Beginn
Mismatch<-0
Für j=0,1,2, … m-1 führe aus
Beginn
Falls ai+j != bm-j dann
Mismatch ++
Ende
Falls mismatch ==0: gib i aus
Ende
Ende

I=1 j=0 → $a_{i+j}=a_1=R$
$B_{m-j}=b_3=s$
Mismatch→mismatch=1
i=1; j=1
$a_{i+j}=a_2=S$
$B_{m-j}=b_2=t$
Mismatch→mismatch=2
…
Alinieren von beiden Sequenzen schrittweise bei mismatch wird der mismatch wert um 1 erhöht (gesamte Antwort gab einen von 2 Punkten)

[2 P]

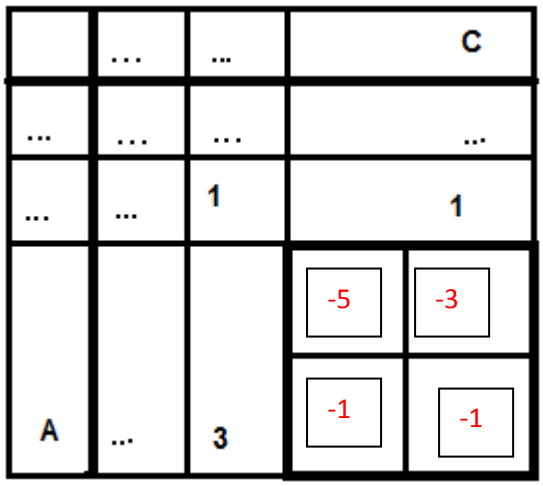
c) Sie haben mit Hilfe von Blast eine Sequenzdatenbank durchsucht. Einen Treffer TR_1 der Blast-Ausgabe wurde ein E-Wert von 15, einem 2. Treffer TR_2 einen E-Wert von 5x10^-20 zugewiesen. Welcher der beiden Treffer weist auf die Funktion der Querry hin? **[1 P]**

´TR_2, weil ein Treffer mit kleinem Wert e-value signifikanter wird

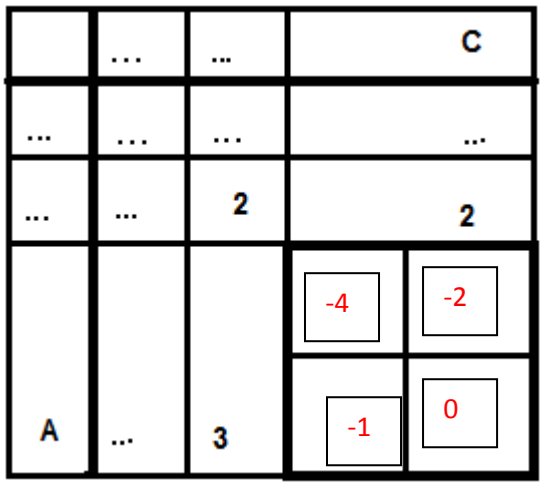
d) Weshalb werden beim Sequenzverglech affine Kostenfunktionen verwendet **[1 P]**
Um Lücken adäquat bewerten zu können (0,5)

e) Konstruieren Sie einen Suffixbaum für die Sequenz A= RTRT **[2 P]**

f) Berechnen Sie den Wert für die dick umrandete Zelle gemäß den Algorithmus von Needleman u. Wunsch. TragenSie Ihr Ergebniss bei „…“ ein. Geben Sie sämtliche Teilergebnisse an, benutzen Sie hierfür die anderen Zellen. Verwenden Sie folgende Scores: Match=3; Mismatch=-6; Gap=-4 **[2 P]**



a) Berechnen Sie den Wert für die dick umrandete Zelle gemäß dem Algorithmus von Smith u. Waterman. Geben Sie die Ergebnisse wiederum wie oben eingeführt an. Scores: Match=3; Mismatch=-6; Gap=-4
[2 P]



Sequenzalignment, Scoring-Systeme (3P)

Sie wollen ein MSA homologer Proteinsequenzen dazu benutzen, Mutationsexperimente zu planen, mit denen sie die Funktion wichtiger Residuen charakterisieren wollen. Welche Positionen des MSAs wählen sie aus? (1P)

Stark konservierte Positionen

Für eine Menge von Promotoren wurden die folgenden Häufigkeiten ermittelt: p(A)=0.3, p(T)=0.33, p(C)=0.1, p(G)=0.27. Für die positionspezifischen Häufigkeiten an Position k gilt: p(A,k)=0.2, p(T,k)=0.32, p(C,k)=0.2, p(G,k)=0.28. Welches Nucleotid ist an Position k am stärksten überrepräsentiert? Begründung! (1P)

A=log10 $(\frac{p(A,k)}{p(A)})$ = -0,176

T=log10 $(\frac{p(T,k)}{p(T)})$ = -0,0133 → Nucleotid C, weil log10-odd-score am höchsten

C=log10 $(\frac{p(C,k)}{p(C)})$ = 0,301

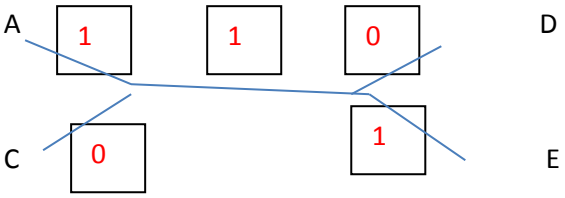
G=log10 $(\frac{p(G,k)}{p(G)})$ = 0,0157

Bei einem PRoteindesign-Experiment müssen sie eine Aminosäure αi ersetzen, wobei sie möglichst konservativ vorgehen wollen um die Funktion des Proteins zu erhalten. Sie haben drei Alternativen αj , αk, und αl zur Auswahl, wobei der BLOSUM-Wert s(αi , αj)=3 und der von s(αi , αk)=-3 und s(αi , αl)=0 ist. Welche Aminosäure wählen sie? Begründen sie ihre Wahl (1P)

αj , da die Aminosäuren umso ähnlicher sind, je höher der Score ist

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 B einzufügen. Begonnen wird mit dem Teilbaum A,C,II D,E. Die erste Nachbarschaftsrelation, die für das Einfügen von B bewertet wird, sei AE II BD. Wie sind die Kanten des Teilbaums nach diesem Schritt markiert? (1 P)



b) Sie haben einen phylogenetischen Baum konstruiert und hierbei ein Bootstrapping-Verfahren verwendet. Eine Kante des Baumes, die zwei Teilbäume TB1 und TB2 verbindet, ist mit dem Wert 95 markiert. Eine weitere, die Teilbäume TB3 und TB4 verbindet, mit dem Wert 10. Was schließen sie aus diesen Werten für die Gruppierung der Teilbäume? (1P)

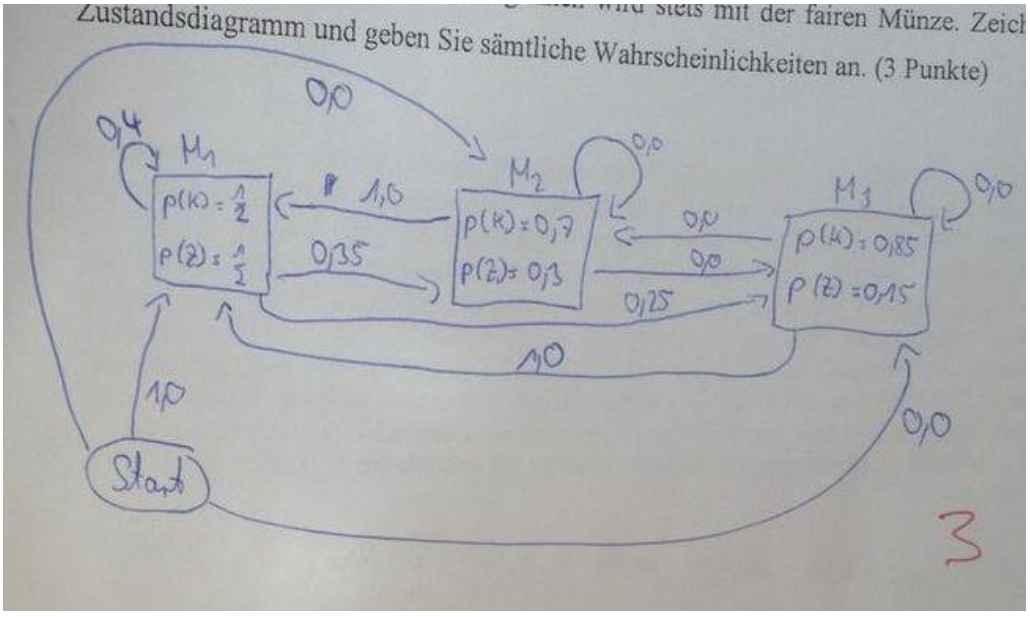
TB1 und TB2 sind ziemlich sicher richtig gruppiert, TB3 und TB4 nicht.

c)Die Sequenzen, die sie als Outgroup bei der Berechnung eines phylogenetischen Baumes verwendet haben, sind an verschiedenen Stellen in den Baum eingefügt worden. Was schließen sie aus diesem Ergebnis im Hinblick auf die Qualität des Baumes? (1P)

Schlechter Baum, weil die Sequenzen der Outgroup bei guten Bäumen zusammen stehen sollte

Hidden-Markov-Modelle

a)In einem unehrlichen Kasino werden drei Münz en M1, M2 und M3 im Wechsel verwendet. M1 ist eine faire Münze, bei M2 tritt Kopf zu 70% und bei M3 Zahl zu 15% Wahrscheinlichkeit auf. Auf den Wurf mit einer unfairen Münze folgt stets ein Wurf mit der fairen Münze. Auf den Wurf mit der fairen Münze folgt mit den Wahrscheinlichkeiten 40%, 35% und 25% ein Wurf mit M1, M2 bzw. M3. Begonnen wird stets mit der fairen Münze. Zeichnen Sie ein Zustandsdiagramm und geben sie sämtliche Wahrscheinlichkeiten an. (3Punkte)



b) Welche der Folgen F1, F2 ist wahrscheinlicher: F1= Start, M1, „Z“, M2, „Z“, M1, „Z“ oder F2= Start, M1, „Z“, M1, „Z“, M1, „Z“? Durch Rechnung begründen. (1P)

F1= $1{,}0 \cdot 0{,}5 \cdot 0{,}35 \cdot 0{,}3 \cdot 1{,}0 \cdot 0{,}5 = 0{,}02625 = 2{,}625\ \%$

F2=$1{,}0 \cdot 0{,}5 \cdot 0{,}4 \cdot 0{,}5 \cdot 0{,}4 \cdot 0{,}5 = 0{,}02 = 2\%$ →F1

c) Beschreiben sie kurz den Aufbau der PFAM-Datenbank (1P)

PFAM-A (gut charakterisierte Domänen von Proteinen)+PFAM-B (unbekannte Domänen)
→Mustererkennung durch Hidden Markov-Modelle (Wikipedia)

Homologie-Modellierung (3P)

1a) Begründen sie weshalb für Proteine 3D-Modellierungen per Homologie-Modellierung berechnet werden können. Was ist das entscheidende Argument? (1P)

Struktur ist stärker konserviert als Sequenz, ähnliche Sequenz lässt auf ähnliche Strukturen schließen

b) Sie wollen für ein Enzym, dessen Funktion nicht bekannt ist, herausfinden, welche Liganden (Substrate) möglicherweise umgesetzt werden. Ein BLasten der Enzymsequenz von ENZ gegen die Sequenzen der in der PDB abgelegten Proteine ergab drei Treffer, die als Template für eine Homologiemodellierung in Frage kommen. STRUC_1 besitzt die höchste Auflösung, allerdings ist kein Ligand gebunden. Das Alignment von ENZ und STRUC_1 weist eine Lücke am N-Terminus auf. In STRUC_2 und STRUC_3 ist jeweils ein Substratanalogon gebunden. Das Sequenzalignment von ENZ mit STRUC_2 wird durch zwei Lücken unterbrochen, die N-Terminal bzw. C-terminal liegen. Das Sequenzalignment von ENZ mit STRUC_3 weist nur eine Lücke auf, die in der Nähe des katalytischen Zentrums liegt. Für welche der drei Strukturen STRUC_1-STRUC_3 entscheiden sie sich als Templat? Begründung (1P)

STRUC_2, weil STRUC_1 kein Liganden besitzt und STRUC_3 am katalytischen Zentrum→Substratbindestelle anders ist

c) Sie suchen für ein Protein PROT_1 eingeegnetes Templat. Ein BLasten hat keinen Treffer in der PDB-Datenbank gegeben. Welche Verfahren probieren sie als nächstes aus, um möglicherweise dennoch ein geeignetes Protein mit hinreichend ähnlicher Sequenz zu finden? Begründen sie ihr vorgehen. (1P)

PSI-Blast, da dieser sensitiver für entfernt verwandte Proteine ist