

# Bioinformatik Prüfung, 18.01.2008 (Gruppe B)

## 1) Berechnen der Laplace Wahrscheinlichkeit und des Erwartungswert

a) Wahrscheinlichkeit dass 1 und 2 bei einem Würfelexperiment (Würfelseiten bezeichnet mit 1, 2, 3, 4, 5, 6, 7, 8) zutrifft?

Laplace Wahrscheinlichkeit:

$$P(E) = \frac{\sum \text{günstigen Fälle (E)}}{\sum \text{aller möglichen Fälle } (\Omega)}$$

$$P(E) = \frac{2}{8} = \frac{1}{4}$$

b) dem zugehörigen Erwartungswert bei 10 Würfeln?

Erwartungswert:

$$E = \sum n_i \times P_i(E)$$

$$E = 10 \cdot \frac{1}{8} + 10 \cdot \frac{1}{8} = 2,5$$

2) Zeichnen eines Dotplots für A:= DDCCDDCD, B:= CDDCDDC codieren  $a_i = b_j = 2$  und  $a_i \neq b_j = 0$ .

Markieren Sie den längsten Teilstring.

	D	D	C	C	D	D	C	D
C	0	0	2	2	0	0	2	0
D	2	2	0	0	2	2	0	2
D	2	2	0	0	2	2	0	2
C	0	0	2	2	0	0	2	0
D	2	2	0	0	2	2	0	2
D	2	2	0	0	2	2	0	2
C	0	0	2	2	0	0	2	0

Längster Teilstring: 5

3) Eine Substitutionsmatrix für den Vergleich von Aminosäuresequenzen beruht auf dem Prinzip der/des:

- Buchstabencodes der AS
- relativen Isotopenzusammensetzung von AS
- Sequenzabfolge der AS
- chemischer Ähnlichkeit von AS

4) Was versteht man unter SNPs, wie ist deren Verteilung im Genom und wofür werden sie verwendet?

(Single Nucleotide Polymorphisms)

- Sind DNA Punktmutationen (Basenaustausche, Insertionen oder Deletionen) die als die häufigste Variation in Genomen auftreten.
- Deren Frequenz ist messbar und liegt bei +/- 1000 bp.
- Sie existieren an definierten Positionen im Genom; STSs („Sequence Tagged Sites“)

Die wesentlichen Anwendungsgebiete sind:

- „Mapping“ von Genen
- Definieren von Populationsstrukturen
- Funktionelle Studien von Genen
- Untersuchung des Zusammenhangs von Sequenzvariationen und erblichen Phänotypen
- juristische bzw. forensische Anwendungen

Detektion von SNPs möglich durch direktes Sequenzieren oder z.B. DNA Chips (für den Nachweis einer Vielzahl von SNPs in einem Ansatz)

Vorteile von SNPs gegenüber anderen Genetischen Markern:

- Einfache und unaufwendige Nachweistechiken
- Stabile Vererbung nach Mendelschen Gesetzen
- Geringe Spontanmutationen

5) Erklären Sie die Funktionsweise von DNA Chip Microarrays, nennen Sie die wichtigsten Anwendungen?

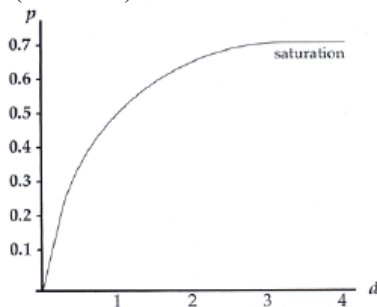
- Eine Vielzahl von cDNA Fragmenten oder Oligonukleotiden (derzeit 100.000 bald 1.000.000 und mehr) werden auf Glas Objektträger in einer Reihe gebunden (Chip).
- Eine fluoreszenzmarkierte RNA Probe wird mit dem Chip hybridisiert.
- Die relative Menge an exprimierter mRNA wird über die Messung der Fluoreszenz des jeweiligen Auftragepunkts gemessen.

Anwendungen: suche nach Stoffwechselwegen, Suche nach genetisch verursachten Krankheiten, Sequenzieren von DNA Molekülen.

6) What is the difference between  $p$ (observed) and  $d$ (estimated) genetic distances? Draw a scatter plot showing how these values correspond to each other. Explain why it is worthwhile to use both measures.

$p$ (observed): Anzahl an, zur Zeit, abzählbaren Punktmutationen

$d$ (estimated): Anzahl an Punktmutationen gesamt (mit hin- und rückmutation)



Auf Grund von Homoplasie (Paralell- oder Rückmutationen) unterschreitet die beobachtete Anzahl ( $p$ ) deutlich die geschätzte Zahl ( $d$ )

**Warum beide? ACHTUNG sehr unsicher!**

observed: daran kann man den Unterschied bzw. die Ähnlichkeit feststellen.

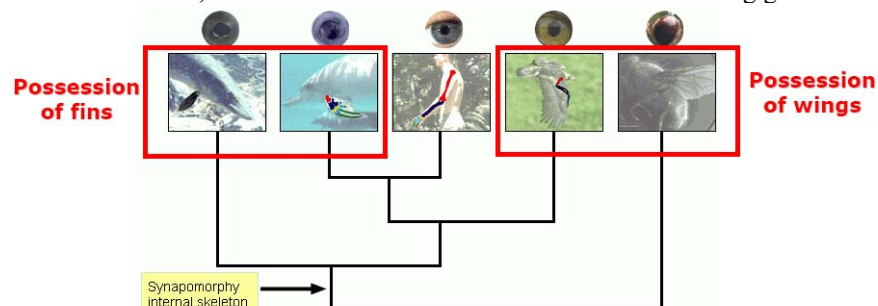
estimated: (inkl. Clock Theorie) An der Gesamtzahl kann man feststellen wie lange es her ist, dass sich die beiden unterscheiden, bzw. welche Zwischenstufe früher da gewesen sein muss.

7) Explain homoplasie in molecular data. Simulate an example nucleotide alignment (minimum 3 steps of series of evolutionary transformations t0ancestral -> t1 -> t2derived) showing homologous and homoplasious characters.

**ACHTUNG: Diese Antwort könnte auch falsch sein. Bin mir nicht sicher.**

HOMOLOGY: Übereinstimmungen von Organen, Organsystemen, Körperstrukturen oder Verhaltensweisen aufgrund eines gemeinsamen evolutionären Ursprungs bei unterschiedlichen systematischen Taxa.

HOMOPLASY: Übereinstimmungen von Organen, Organsystemen, Körperstrukturen oder Verhaltensweisen, die bei mehreren unterschiedlichen Taxa unabhängig voneinander entstanden sind.



Homoplasie: Die Ausbildung von Flügeln und Flossen ist in diesem Beispiel eindeutig nicht aufgrund der Abstammung zu erklären.

Homologie: Die ähnliche Struktur der Knochen der Extremitäten von Delphin, Mensch und Vogel lässt sich dadurch erklären, dass sie gemeinsame Vorfahren hatten.

8) How the reliability of phylogenetic tree can be estimated? Whether the repeated analysis would help you or there are analytical methods available?

Mit einem Bootstrap-Test lassen sich Quantile und/oder  $p$ -Werte (Irrtumswahrscheinlichkeiten) schätzen ohne die Verteilung der Grundgesamtheit kennen zu müssen.

Dieses Verfahren ist ein sog. Computerintensives Verfahren, bei dem durch zufälliges Ziehen (mit Zurücklegen) aus der Stichprobe neue Stichproben generiert und darauf die Teststatistiken berechnet werden.

# Prüfungsfragen vom 18.01.08 (Gruppe A)

## 1. dotplot

s.o.

## 2a. wahrscheinlichkeit, 8seitiger würfel

s.o.

## 2b. erwartungswert für 1 bei 100 würfen

Erwartungswert:

$$E = \sum n_i \times P_i (E)$$

$$E = 100 \times 1/8 = 12,5$$

## 3. Welche der Folgenden Begriffe sind keine heuristische Suchfunktion .. (oder so)

[a) Blast; b) KEGG; c) ka mehr; d) ..]

Heuristische Verfahren: FASTA, BLAST, ISS, PSI-BLAST, SAM-T98

## 4. EST-Banken, Funktion + Anwendung

EST (Expressed Sequence Tags) sind cDNA Stücke. (d.h. sie entstehen bei der Vervielfachung der kodierenden Teile der DNA in einer Zelle) → keine Repeats, nur wirklich kodierende Teile (Exons).

Da die Probleme mit Wiederholungen und nicht-kodierenden Stellen wegfallen lassen sich ESTs sehr viel einfacher sequenzieren.

Gene, die nur selten oder nur in Ausnahmen (Krankheit) exprimiert werden sind oft nicht in EST-Banken.

Anwendung: Entdecken von Mitgliedern einer Genfamilie, neue Sequenzdaten können mit EST-Banken verglichen werden, Genexpressionsniveau kann abgeschätzt werden.

## 5. Comparative Genomics + welche Modelltypen

Comparative Genomics versucht die Funktion von Genen in „einfachen“ Organismen zu erforschen und die Erkenntnisse dann auf andere Organismen (z.B. Mensch) umzulegen. Die DNA-Sequenz eines bekannten Gens wird dann zur Suche nach ähnlichen Teilsequenzen (Ähnlichkeitssuche) in der neu sequenzierten DNA verwendet.

Mensch: Keine „Gen knock-outs“, keine kontrollierte Züchtung, keine Mutantenkollektion

Modellorganismen: Maus, Ratte, Hamster, Zebrafisch, Hefe, ...

Programme: COG (Cluster of Orthologous Groups), PEDANT (Protein Extraction Description and Analysis Tool), WIT (What Is There)

## 6. Eng: Kombination von Bäumen bei 3 Taxas zeichnen .. wieviele Möglichkeiten?

## 7. Eng: Evolution .. schreibe 10 freie .. ka

## 8. Eng: minimum parsimony or evolution .. oder so in der art

s.u.

# Prüfungsfragen vom 18.04.07

## 1. Dotplot

s.o.

## 2. Wahrscheinlichkeit Erwartungswert 8 seitigen Würfel

s.o.

## 3. EST's + Anwendung

s.o.

## 4. Comparative GENomcs + Welche Modelltypen

s.o.

## 5. MC Frage irgendwas über Datenbanken

## 6. Alle möglichen Kombinationen von Bäumen zeichnen bei 3 Taxas, wie viele Möglichkeiten gibt es ..

## 7. Keine ahnung mehr, irgendwas mit 10 freie Parameter bei Evolution,

## 8. Parismony + minimum evolution...

THE LAW OF PARSIMONY: Die einfachste der plausiblen (gleichrichtigen) Theorien ist vorzuziehen.

PARSIMONY (MINIMUM EVOLUTION): Der phylogenetische Baum, bei dem die Individuen die wenigsten Änderungen durchmachen soll bevorzugt werden.

# Prüfungsfragen aus dem VoWiki

- ESTs sind... (es folgte eine Auflistung von Fakten, von denen man die richtigen ankreuzen musste, darunter auch wofür die Abkürzung stehen kann), Anwendungsbeispiele der ESTs  
s.o.
- Anwendungsbeispiele der DNA Chips  
s.o.
- Wie nennt man dieses Diagramm? (+ Abbildung eines speziellen Baumtyps),  
• Ordnen Sie A,B,C und D zu (es lag eine Grafik bei, die in einem speziellen Baum gewisse Objekte mit diesen Buchstaben markierte, sowie 5 (sic!) Begriffe(Clade, Branch, Node, Ancestor etc.), denen man diese Buchstaben zuordnen konnte)
- Nennen Sie 4 Modellorganismen  
s.o. (Maus, Ratte, Hamster, Zebrafisch, Hefe, ..)
- was gehört zu sekundär Struktur eines Proteins (Coiled Coil, Helix Turn Helix...)  
Sekundärstruktur:  $\alpha$ -Helix,  $\beta$ -Faltblatt, Turn;  
Zusätzlich (auf selber Ebene): Transmembrane Domänen, Coiled Coil Regionen, Helix-Turn-Helix Motive, Signalpeptide
- Nennen Sie 3 wichtige Bioinformatikdatenbanken  
NCBI (GenBank, SWISS-PROT, HTGS), EBI (SWISS-PROT, EMBL-Datenbank), GenomeNet (KEGG), ....., PIR, PRF, GENES, PDB
- Nennen Sie 4 Einsatzmöglichkeiten der Genomanalyse  
Pharmakologische Tests auf Genbasis, Abstammungs- und Verwandtschaftsnachweise,
- Berechnen Sie die Levensteindistanz zwischen TATAATAGA und TATATAGA, wobei sie als Kosten 1 für eine Ersetzung und 2 für Einsetzung oder Löschen verwenden initialisieren sie die Matrix mit 0,2,4,6,...

		T	A	T	A	A	T	A	G	A
	0	2	4	6	8	10	12	14	16	18
T	2	0 4 4 0	3 6 2 2	4 6 4 4	7 10 6 6	9 12 8 8	10 14 10 10	13 16 12 12	15 18 14 14	17 20 16 16
A	4	3 2 6 2	0 4 4 0	4 6 2 2	4 8 4 4	6 10 6 6	9 12 8 8	10 14 10 10	13 16 12 12	14 18 14 14
T	6	4 4 8 4	3 2 6 2	0 4 4 0	3 6 2 2	5 8 4 4	6 10 6 6	9 12 8 8	11 14 10 10	13 16 12 12
A	8	7 6 10 6	4 4 8 4	3 2 6 2	0 4 4 0	2 6 2 2	5 8 4 4	6 10 6 6	9 12 8 8	10 14 10 10
T	10	8 8 12 8	7 6 10 6	4 4 8 4	3 2 6 2	1 4 4 1	2 6 3 2	5 8 4 4	7 10 6 6	9 12 8 8
A	12	11 10 14 10	8 8 12 8	7 6 10 6	4 4 8 4	2 3 6 2	2 4 4 2	2 6 4 2	5 8 4 4	6 10 6 6
G	14	13 12 16 12	11 10 14 10	9 8 12 8	7 6 10 6	5 4 8 4	3 4 6 3	3 4 5 3	2 6 5 2	5 8 4 4
A	16	15 14 18 14	12 12 16 12	11 10 14 10	8 8 12 8	6 6 10 6	5 5 8 5	3 5 7 3	5 4 5 4	2 6 6 2

TATAATAGA

TATAATAGA (+ ein A)

- Welche ist die richtige Reihenfolge? (Es folgten mehrere Reihenfolgen der Klassifizierung in Kingdom, Species, Order usw...)  
Kingdom→Phylum→Class→Order→Family→Genus