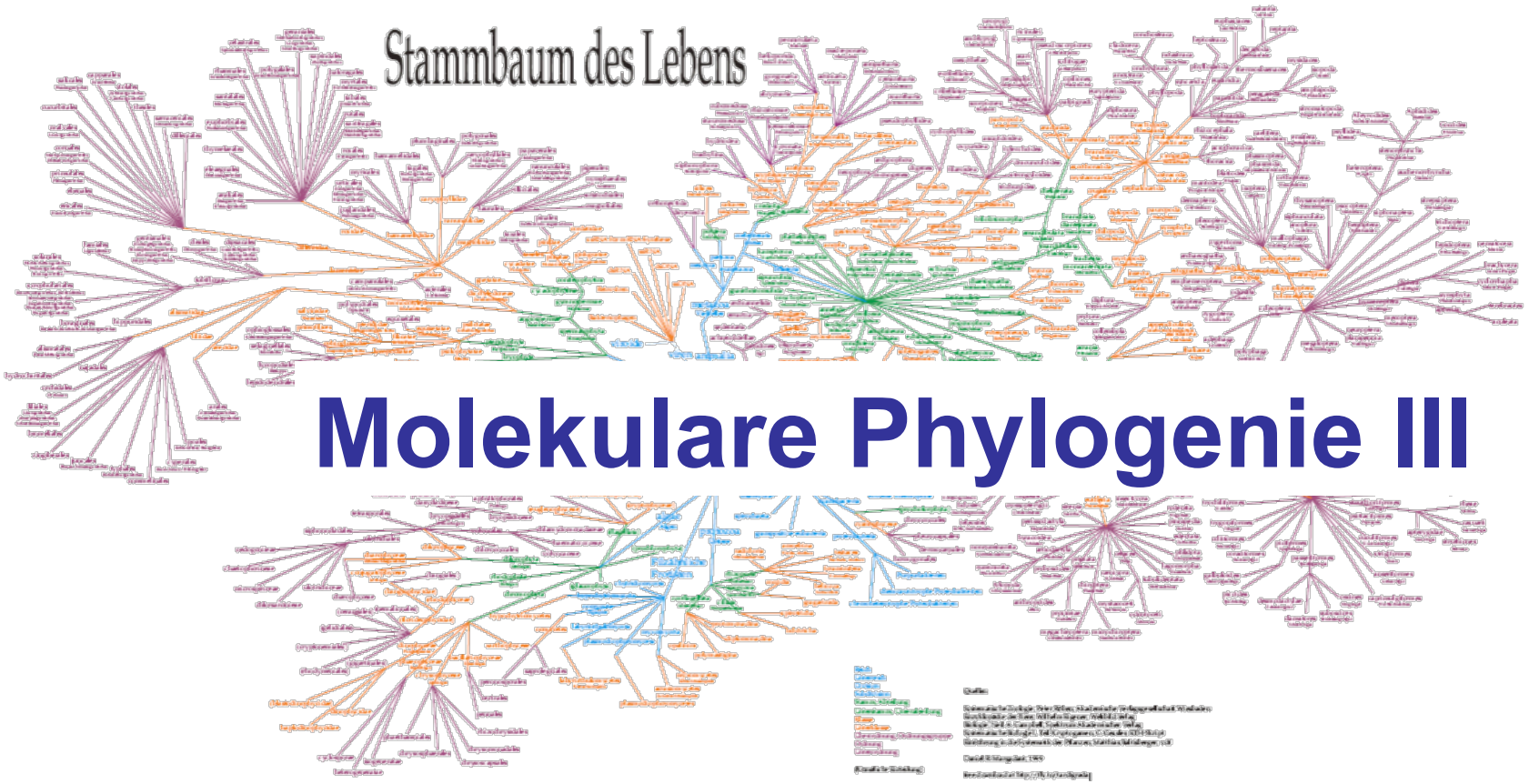
WS 2016/2017

„Genomforschung und Sequenzanalyse
- Einführung in Methoden der Bioinformatik- “

Thomas Hankeln



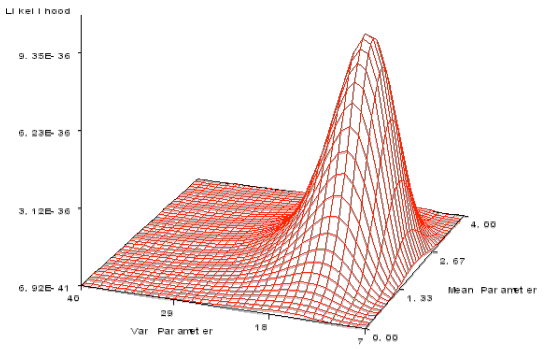
Molekulare Phylogenie III

Charakter-orientierte Methoden

1. Maximum Parsimony (MP) ✓

2. Maximum Likelihood (ML)

3. Bayes



Probability vs. Likelihood

Bedingte Wahrscheinlichkeit
> Wahrscheinlichkeit eines Ereignisses A, gegeben das Ereignis B: **P(A|B)**.

probability > ermittelt unbekannte Wahrscheinlichkeit eines Ereignisses aufgrund bekannter Parameter
> **P(Hypothesis|Data)**

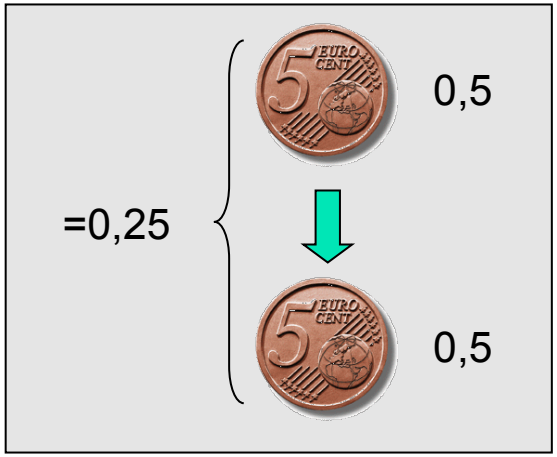
likelihood > bestimmt Wahrscheinlichkeit, mit der ein Datensatz ein parametrisiertes Modell unterstützt
> **P(Data|Hypothesis)**

Probability

gegeben...

0,5

0,5



Wie wahrscheinlich ist,

dass „zweimal Zahl“ kommt?

$P(H|D)$ =

$P(ZZ|p_Z=0,5) = 0,25$

Likelihood

Wie hoch ist „Kopfwahrscheinlichkeit“ allgemein?

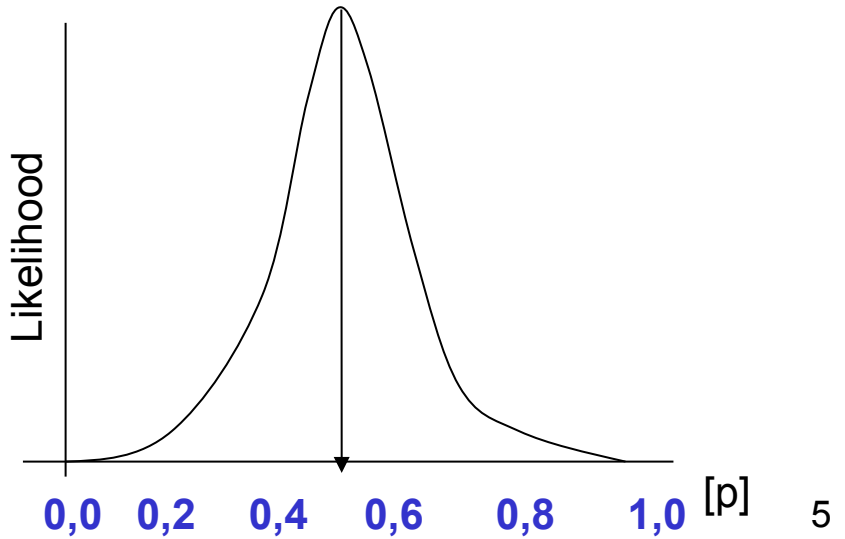
KKZKZKKZZZ

$$L = P(\textit{Daten} \mid \textit{Hypothese})$$

=> Likelihood L= P(D|H) = pp(1-p)p(1-p)p(1-p)pp(1-p)(1-p)(1-p)

Plot der Daten (KKZKZKKZZZ) gegen verschiedene Werte von p (Hypothese)

=> mit welcher Kopfwahrscheinlichkeit *p* bekomme ich am ehesten diese Daten?



Maximum Likelihood

$$L = P(data|hypothesis)$$

- Wahrscheinlichkeit der beobachteten **Daten** (Alignment) im Lichte der **Hypothese** (Stammbaum).

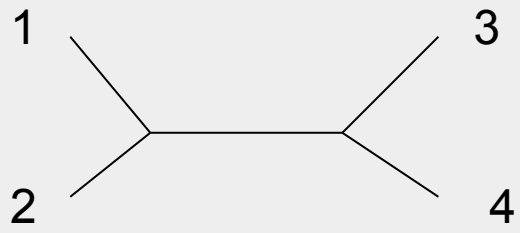
- d.h, es wird der **Stammbaum** (ML tree) ermittelt, der die beobachteten Daten (also das **Sequenz-Alignment**) am besten **unter der Annahme eines Evolutionsmodells** erklärt.

Maximum Likelihood

Sequenzalignment

1	A A G A C
2	A G C C C
3	A G A T A
4	A G A G G

Ein möglicher Baum (von dreien):



L = P(data|tree)

Wie hoch ist die Wahrscheinlichkeit, dass der gezeigte Stammbaum für das Sequenz-Alignment verantwortlich ist?

Finde unter allen mögliche Bäumen denjenigen mit dem höchsten Likelihood-Wert L!

Maximum Likelihood

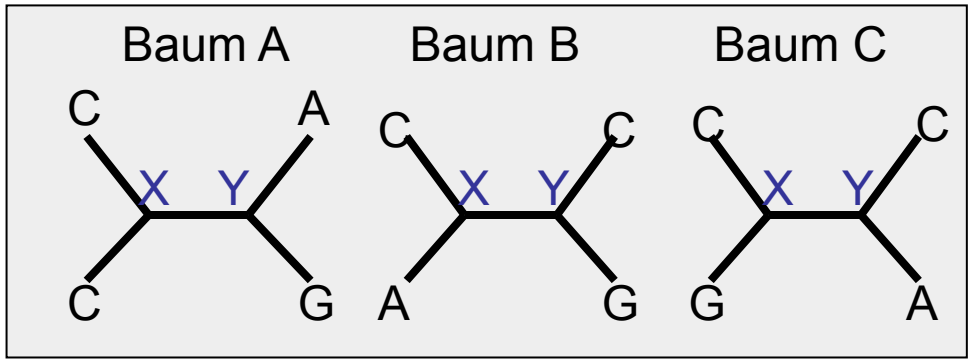
Daten

z. B. Pos.5
1. C
2. C
3. A
4. G

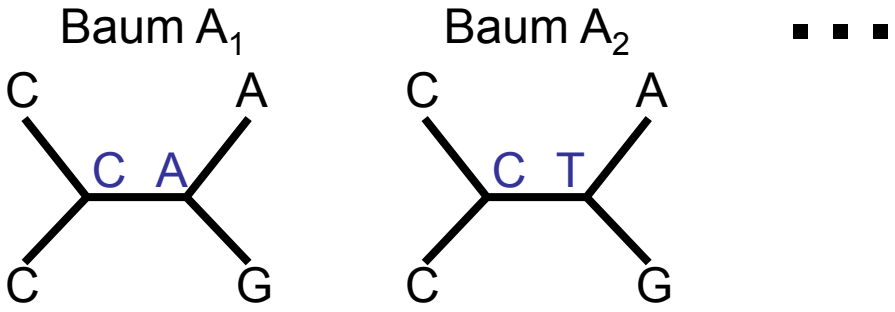
Evolutionsmodell
(nicht real!! Nur zum Rechnen!)

	A	T	G	C
A	1	0,1	0,1	0,1
T		1	0,1	0,1
G			1	0,1
C				1

vier OTUs > drei mögliche Bäume



16 Möglichkeiten anzestraler Nukleotide



$P(A_{1)} + P(A_{2)} + ... + P(A_{16})$

= L (Baum A an Pos. 5)

$P(A_1)$=1x1x0,1x1x0,1
= 0,01

$P(A_2)$=1x1x0,1x0,1x0,1
= 0,001

Maximum Likelihood

$P(A_{1)} + P(A_{2)} + ... + P(A_{16})$ = Likelihood L von Baum A an Alignment-Pos. 5

Dann Likelihood des Baumes A für alle Alignmentpositionen berechnen:

$L_{(Pos1)} \times L_{(Pos2)} \times L_{(Pos3)} \times ....$ = L (Baum A)

Mathematisch einfacher:

$lnL_{(Pos1)} + lnL_{(Pos2)} + lnL_{(Pos3)} + ...$ = **lnL** (Baum A) ***logLikelihood***

Dann Bäume B, C usw genauso berechnen....

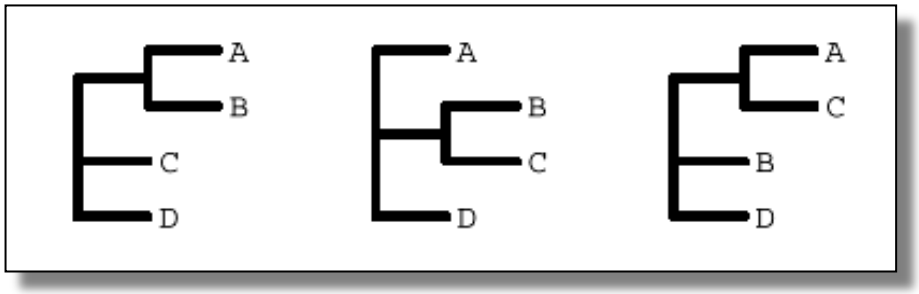
höchster lnL-Wert = ML-Tree

Vieles spricht für ML !

- komplette Sequenzinfo wird genutzt; selbst nicht-Parsimony-informative Orte! Bsp:

A	acgcaa
B	acataa
C	atgtca
D	gcgtta

Keine MP-informativen Orte!



In MP sind alle Bäume gleich gut!

ML kann dennoch die vorhandene phylogenetische Information nutzen!

- wenn Evolutionsmodell OK > wenig Probleme mit LBA

Nur Weniges spricht dagegen...

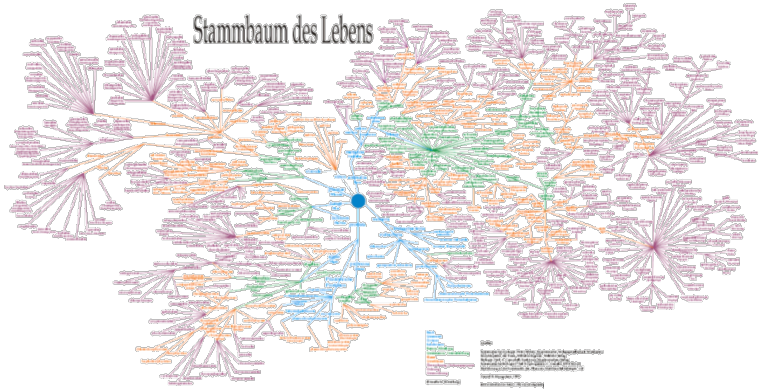
- falsches Evolutionsmodell > falscher Stammbaum!

- aber: Modell muss nicht blind angenommen werden, sondern kann aus den Daten selbst berechnet werden!!

  Man macht likelihood-Analysen mit verschiedenen Werten für Parameter (z. B. T_i/T_v, α-Parameter etc), vergleicht die L-Werte und stellt dann den Parameter letztendlich so ein, dass lnL am größten ist.

- leider ist ML extrem rechenaufwändig

Maximum Likelihood



Für n=50 Sequenzen gibt es
$2{,}84 \times 10^{76}$ mögliche Bäume...

Effiziente Algorithmen erforderlich! Z. B...

- Quartet puzzling
- RaXML
- Bayes/MCMCMC

$L = P(data|hypothesis)$

Bayesian MCMCMC

...kombiniere **Bayes-**Statistik und schnelle Computeralgorithmen

Bayes who?

Reverend Thomas Bayes
1702-1761

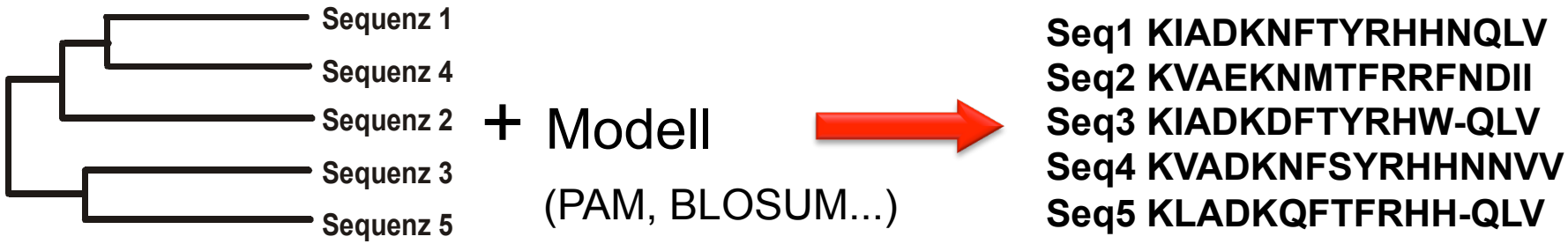
Bayes, T. 1763. An essay towards solving A problem in the doctrine of chance. Philosoph. Transact. Royal Soc. London

$$P(\text{model}|\text{data}) = \frac{P(\text{data}|\text{model}) \times P(\text{model})}{P(\text{data})}$$

P(data)	> „Randbedingungen“
P(data|model)	> Likelihood
P(model)	> „Prior“, Erwartung an Modell

Maximum Likelihood

"Wahrscheinlichkeit" (Likelihood) der Sequenzdaten, gegeben die Topologie des Stammbaums und ein Evolutionsmodell.



Anders formuliert:

Welcher Stammbaum (und welches Evolutionsmodell) erklärt am besten meine Sequenzdaten?

$$L = P(data|tree)$$

Bayes

Bayes´sche Statistik beruht auf dem Satz von Bayes.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

A ist die Hypothese (Baum).
B ist das beobachtete Ereignis (Sequenzalignment).
P(A) ist die *A-Priori*-Wahrscheinlichkeit von A.
P(B | A) ist die bedingte Wahrscheinlichkeit von B, unter der Bedingung dass die Hypothese A wahr ist (Likelihood-Funktion)
P(B) ist die unbedingte Wahrscheinlichkeit von B.

P(A | B) ist die *A-Posteriori*-Wahrscheinlichkeit von A gegeben B.

Satz von Bayes: Beispiel

Berechne Wahrscheinlichkeit der Diagnose Bronchitis bei dem Befund ‚Husten ‘:

$$P(\text{Bronchitis}|\text{Husten}) = \frac{P(\text{Husten}|\text{Bronchitis}) \times P(\text{Bronchitis})}{P(\text{Husten})}$$

P(Bronchitis) = 0.05
P(Husten) = 0.2
P(Husten|Bronchitis) = 0,8

= 0.05 x 0.8 / 0.2 = 0.2

d.h., die posteriore Wahrscheinlichkeit, dass ein Patient mit Husten wirklich Bronchitis hat, ist 20%
(und damit 4 x so hoch wie die *a priori*-Wahrscheinlichkeit von Bronchitis)

Bayes und Bäume

• **Bayes Theorem:** **P (tree | data) ≈ P (data | tree) x P (tree)**

P (tree | data) > **posterior probability gibt die Wahrscheinlichkeit an, mit der ein Baum korrekte Topologie und Astlängen besitzt**

P (tree) > „prior probability“, daß eine Phylogenie korrekt ist (= Annahme, ohne die Daten zu kennen). (Einfachste Annahme: alle Bäume sind gleich wahrscheinlich)

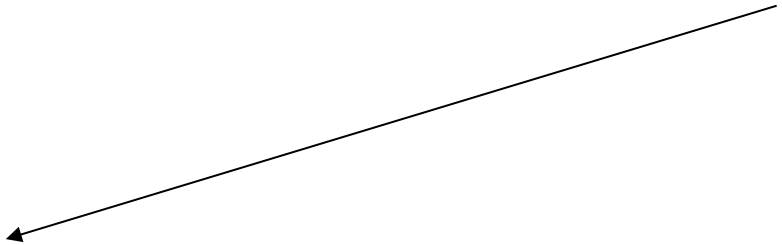
P (data | tree) > Likelihood-Wert der Daten, auf Grundlage eines Baumes und eines Substitutionsmodells (wird berechnet)

Die Bäume mit der höchsten „posterior probability“ werden gesucht!

... und zwar bei allen möglichen Kombinationen von Topologien und Astlängen, sowie den verschiedenen Parametern von Substitutionsmodellen.

MCMCMC

Metropolis-coupled Markov chain Monte Carlo



Zufällige Stichprobe aus der **posterior probability-Verteilung** der Bäume (tree space) ziehen.

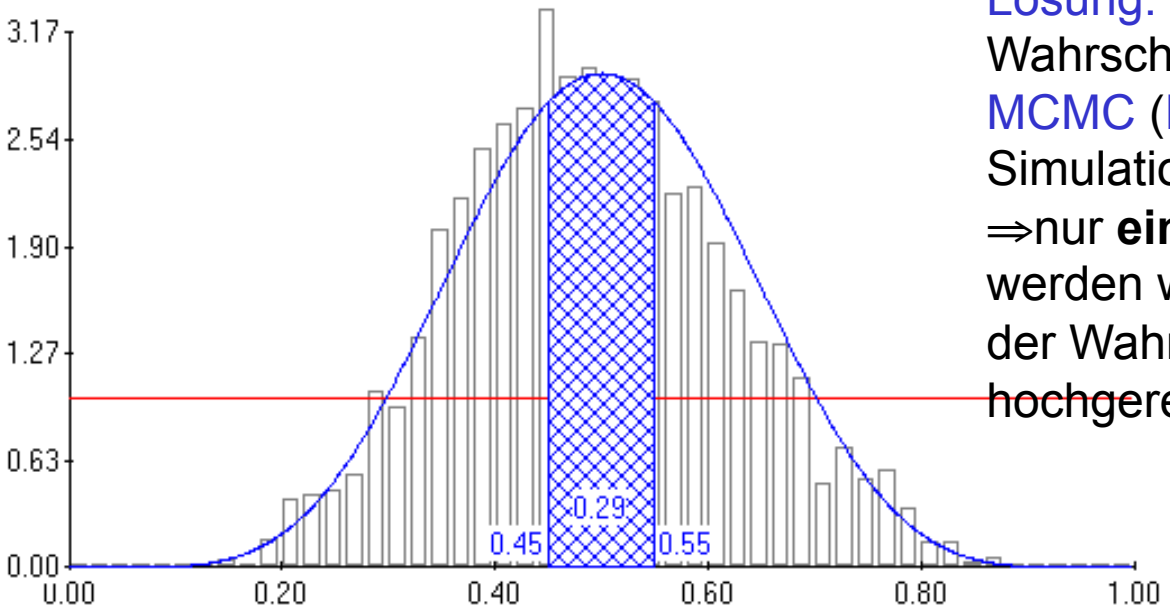
Stichprobe muss nur groß genug sein.
(„Die Bank gewinnt immer!“)

Prozentsatz, mit dem eine ‚Clade‘ bei den Bäumen auftritt, wird als Wahrscheinlichkeit interpretiert, dass ‚Clade‘ korrekt ist.

MCMCMC

Metropolis-coupled Markov chain Monte Carlo

Problem: Wie ermittelt man die Verteilung der Wahrscheinlichkeiten mit einer endlichen Anzahl* von Versuchen?



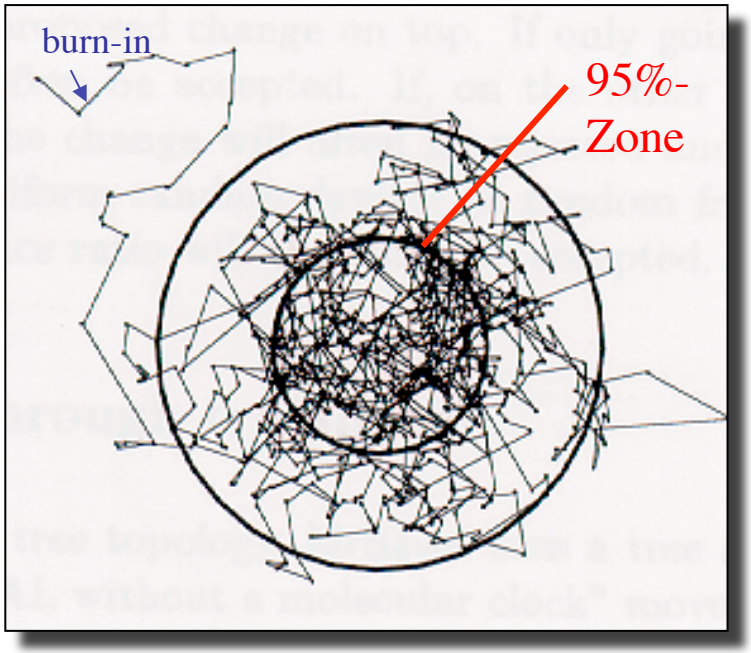
Lösung: Ermittlung der Wahrscheinlichkeitsdichte mittels MCMC (Markov Chain Monte Carlo) Simulation

⇒nur **einige Wahrscheinlichkeiten*** werden wirklich ermittelt, die Verteilung der Wahrscheinlichkeiten wird hochgerechnet.

*z. B. „nur“ 1 - 10 Mio.

MCMCMC

Metropolis-coupled Markov chain Monte Carlo



Kette von Zufallsereignissen,
bei denen die Wahrscheinlichkeit von Änderungen
nur vom gegenwärtigen Zustand abhängt

➢Kette „sucht“ nicht DEN optimalen Baum, sondern konvergiert, d.h. „sammelt“ die Bäume mit der höchsten posterior probability („Gipfel in der Baumlandschaft“).

➢Anhand dieser Baum-Sammlung wird ein Konsensus-Baum erstellt.
➢Die Güte der Verzweigungen wird durch die Höhe der pp-Werte gekennzeichnet.

MCMCMC

Metropolis-coupled Markov chain Monte Carlo

Metropolis, N. et al. (1953)
Equations of state calculations
by fast computing machines.
J. Chemical Physics

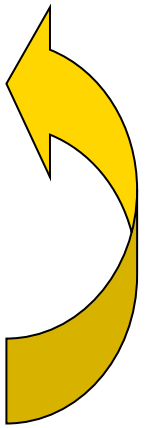
Vorgegebener (Zufalls-)Baum **Tj**
mit Topologie, Astlängen und Evolutionsmodell

↓

Neuer Baum (**Ti**), neue Parameter

↓

Akzeptieren oder Verwerfen?



$$Q_{\text{posterior probabilities}} = \frac{\frac{P(B \mid T_i) \times P(T_i)}{P(D)}}{\frac{P(B \mid T_j) \times P(T_j)}{P(D)}} = \frac{P(B \mid T_i) \times P(T_i)}{P(B \mid T_j) \times P(T_j)}$$

↓

Likelihood ausrechnen

MCMCMC

Metropolis-coupled Markov chain Monte Carlo

Metropolis, N. et al. (1953)
Equations of state calculations
by fast computing machines.
J. Chemical Physics

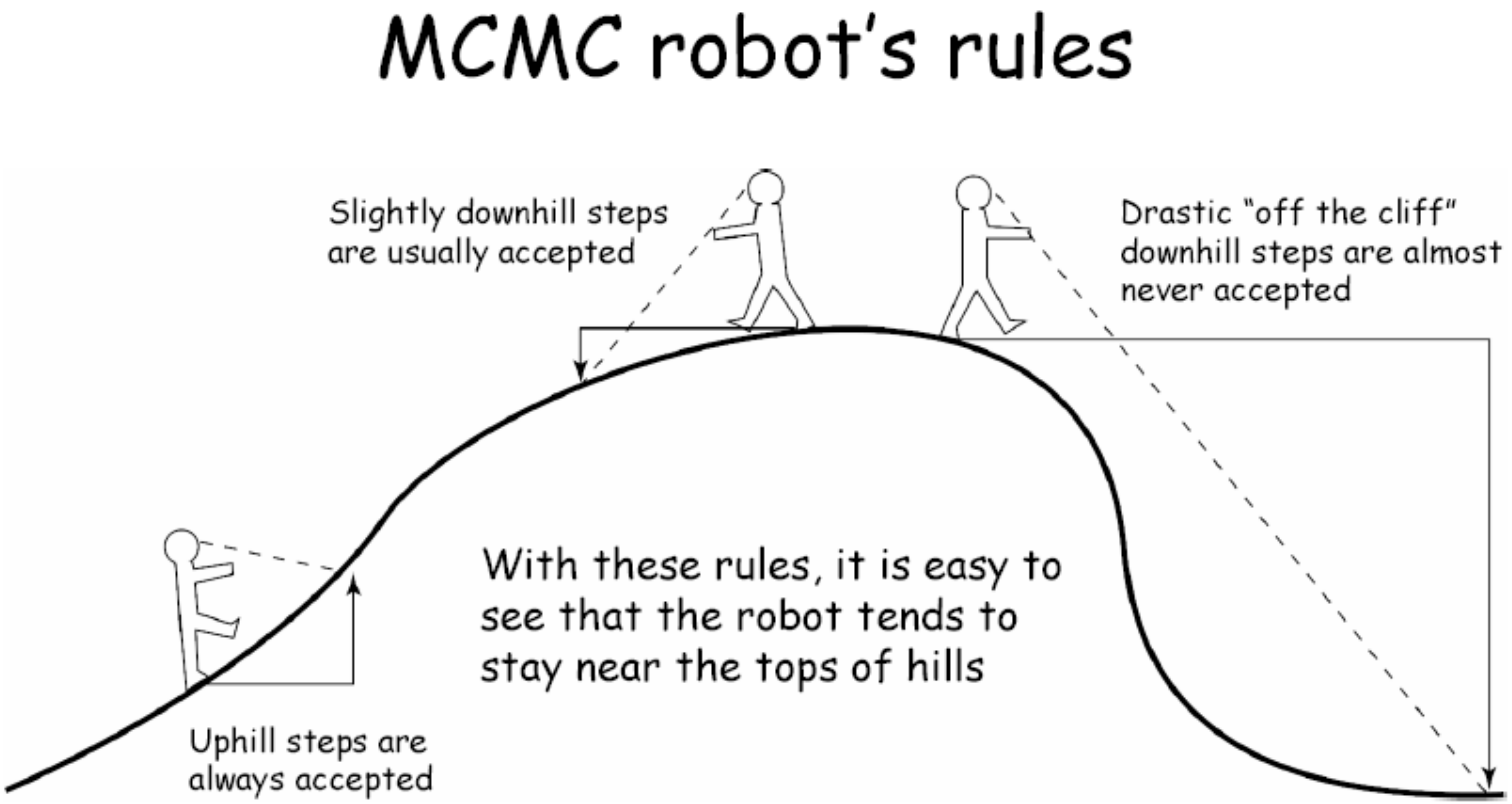
$$Q_{\text{posterior probabilities}} = \frac{0{,}5}{0{,}1} = 5$$ ⇒ neuen Baum akzeptieren

$$Q_{\text{posterior probabilities}} = \frac{0{,}1}{0{,}5} = 0{,}2$$ ⇒ neue Zufallszahl

- 0,1 ⇒ alte behalten
- 0,3 ⇒ neuer Baum

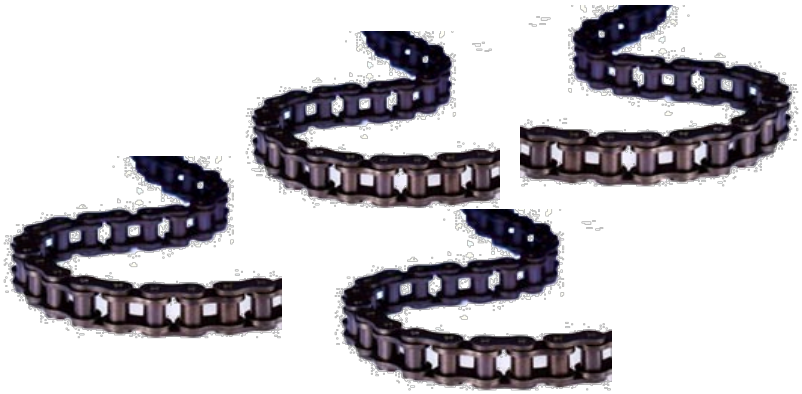
Im Gegensatz zur ‚Maximum Likelihood ' kann hier ‚likelihood ' rauf und runter

Und wieder: Lokale Maxima



MCMCMC

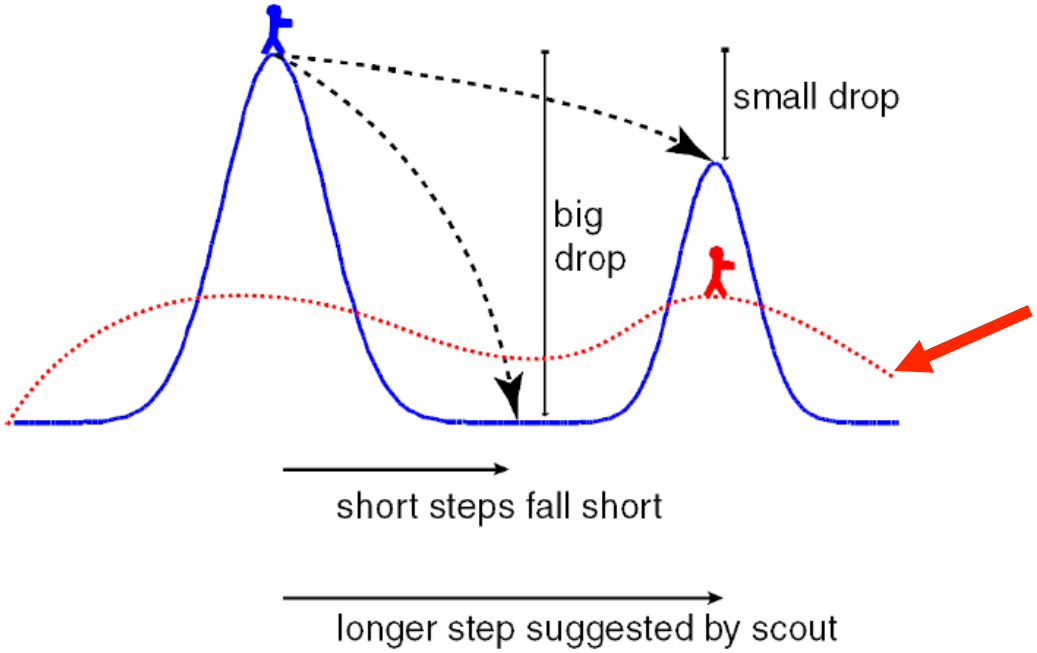
MC ^{3} - MCMCMC - „Metropolis-Coupled Markov Chain Monte Carlo“



=> mehrere MCMC-Ketten laufen parallel und ‚kommunizieren‘

MCMCMC

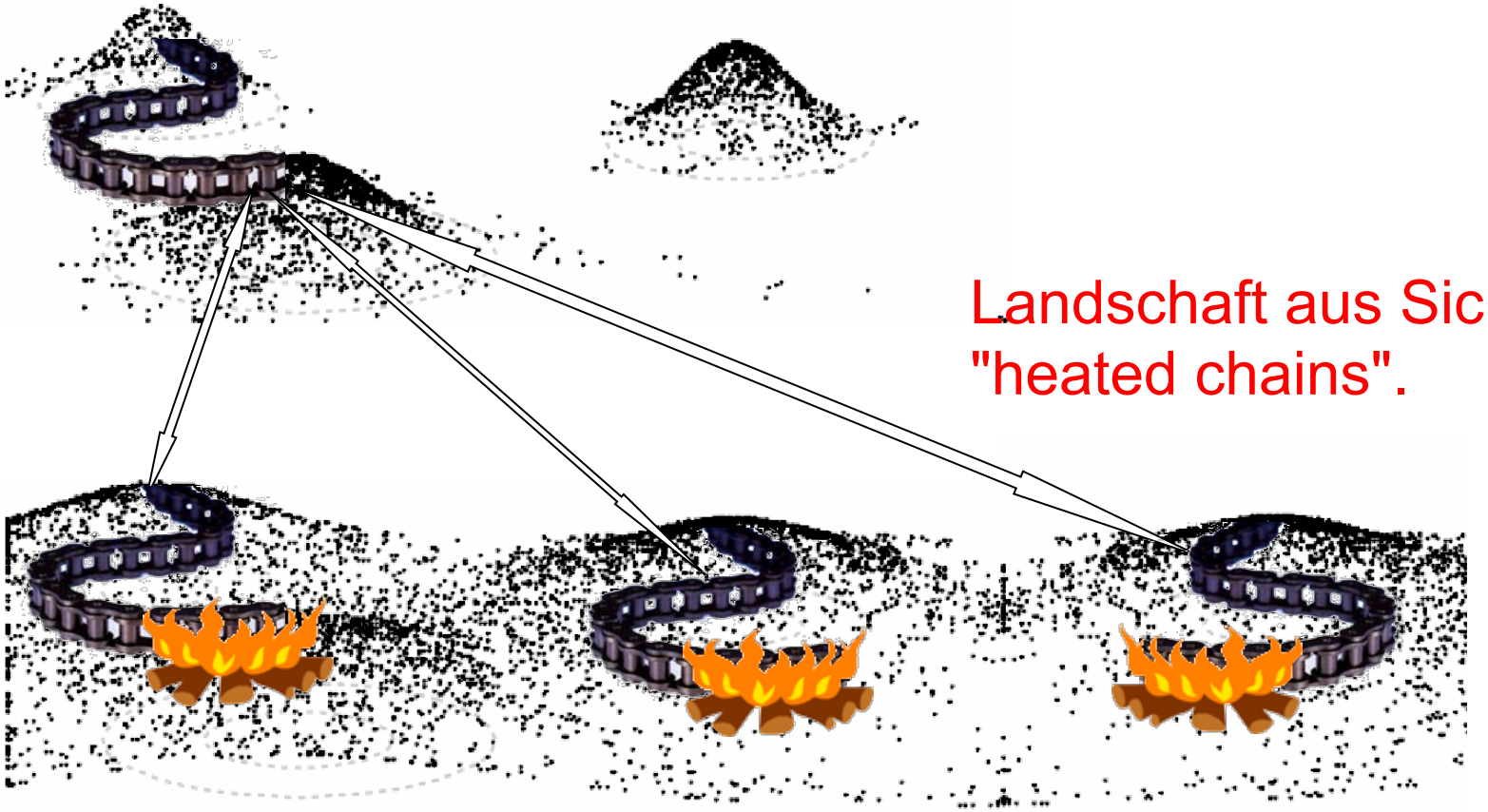
- MC^3 läßt mehrere "chains" suchen
- Die "cold chain" zählt, die "heated chains" scouten



Landschaft wird für heated chain "geebnet" => Übergang zu einem anderen Optimum ist leichter möglich.

MCMCMC

Landschaft aus Sicht der "cold chain".



Landschaft aus Sicht der "heated chains".

Bayes und Bäume

Vorteile:

- Vorabinformation wird berücksichtigt.
- Sehr schnelle ‚Lösung ‘ komplexer phylogenetischer Probleme möglich!
- Diskrete Wahrscheinlichkeitswerte werden für jeden Ast gegeben.

Nachteile:

- Vorabinformation wird berücksichtigt.
- Wahrscheinlichkeitstheoretisch umstritten.

Methodenübergreifende Fragen...

Wo ist die Wurzel in meinem Baum?

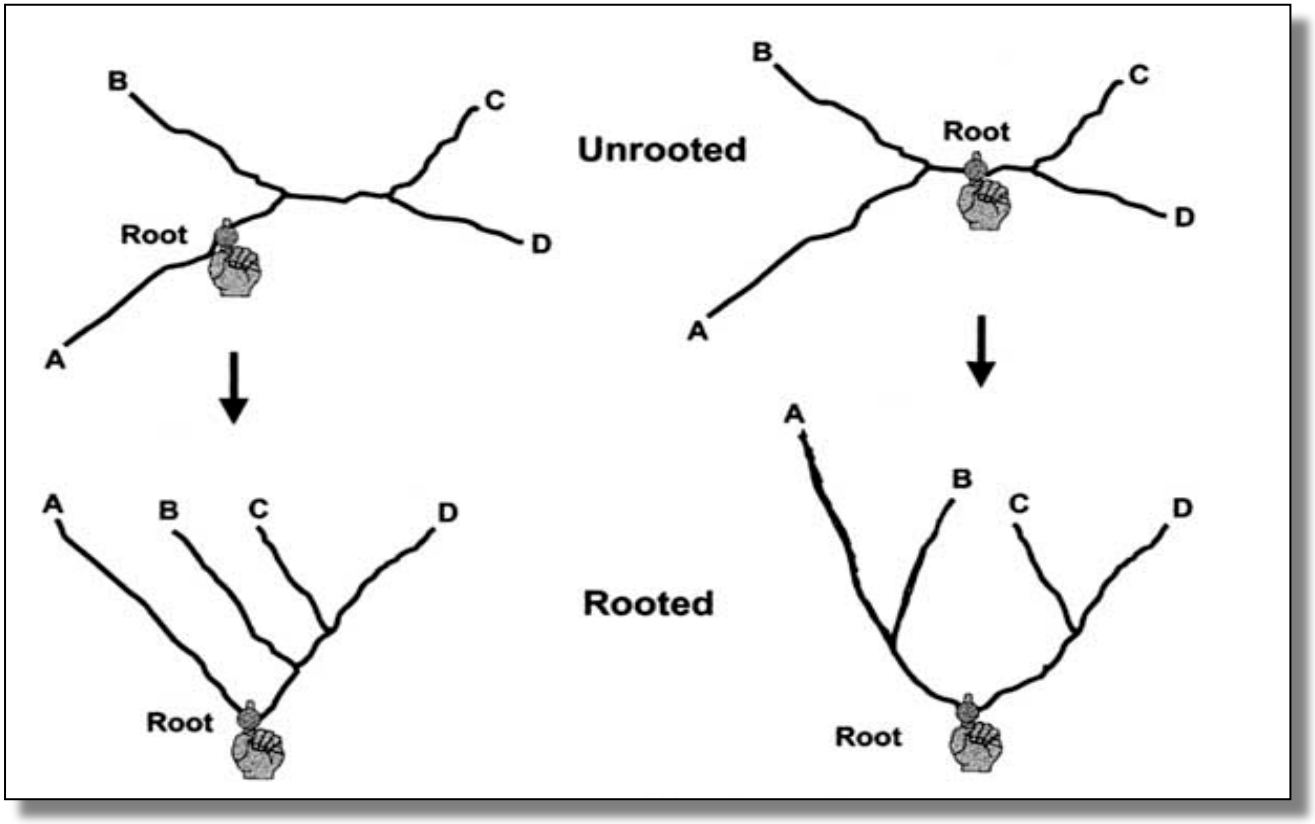
Wie vergleiche ich Bäume miteinander?

Wie bewerte ich die Verläßlichkeit von Bäumen?

Welche Programme?

Welche Methoden funktionieren am besten?

‚Rooting ‘ –
Wo ist die Wurzel des Baums?

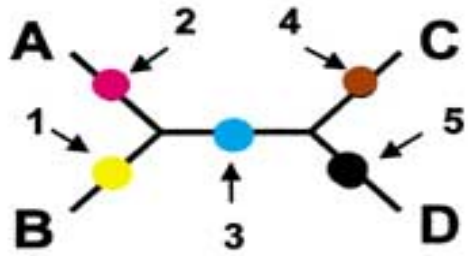


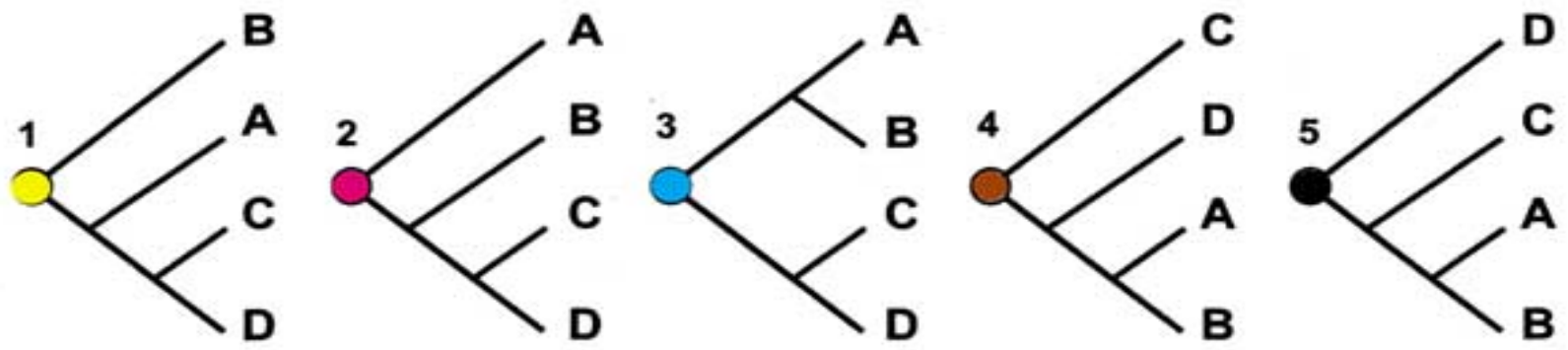
Achtung: die meisten Rekonstruktionsmethoden produzieren zunächst unrooted trees

‚Rooting ‘ –
Wo ist die Wurzel des Baums?

Fünf Möglichkeiten, die Wurzel zu setzen....

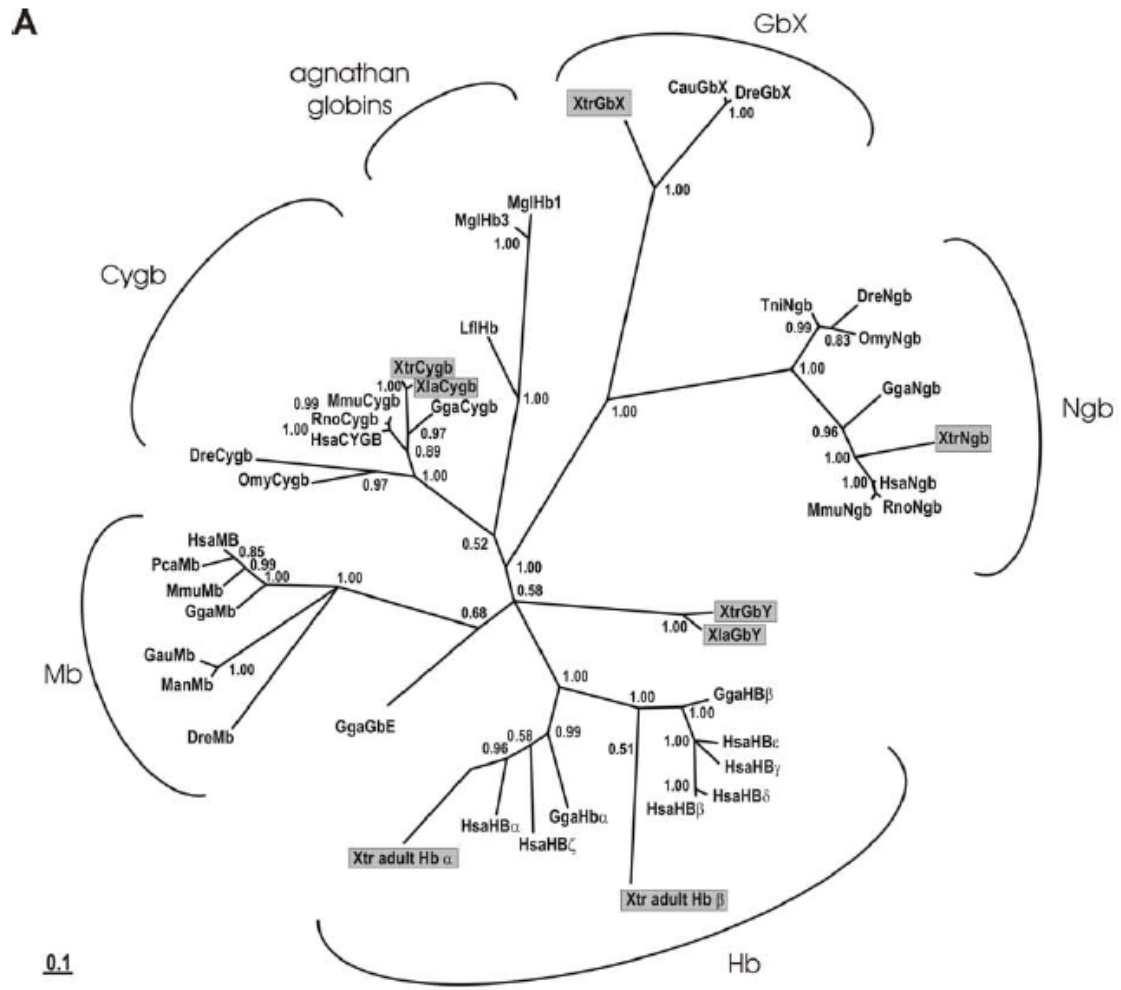
Baum ohne Wurzel





Unterschiedliche Aussage!

„Rooting ‘ –
Ist ein unrooted tree ausreichend?



Globingen-Stammbaum (Bayes) zur Festlegung der Verwandtschaft neu gefundener Gene

Ein Baum ist eine Evolutionshypothese!

Wie wissen wir, ob der rekonstruierte Baum korrekt ist?

1. **Wie verlässlich ist der Baum?**
2. **Welche Verzweigungen sind verlässlich?**
3. **Ist *der* Baum signifikant besser als ein anderer?**

Qualitätsbewertung von Bäumen

- MP, ML, NJ u.a. Distanzmethoden:

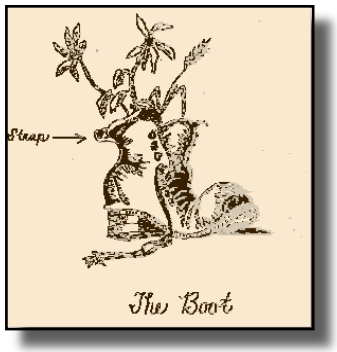
  Bootstrapping

- ML: likelihood ratio test (LRT)
  QuartetPuzzling (QP)-Werte

- Bayes : Posteriore Wahrscheinlichkeiten

…und andere

Bootstrapping



Wie gut ist die Gruppierung zweier OTUs zu einer Clade im Baum statistisch abgesichert?

• Erstellung von 500-1000 Teildatensätzen (Pseudosamples) durch „resampling with replacement“

D. h., manche Positionen des Sequenzalignments werden mehrfach ausgewählt, andere dafür gar nicht!

• Baumrekonstruktion für diese z.T. artifiziellen Teil-Datensätze

• Bootstrap-Wert = 80% bedeutet: in 80% der Fälle werden die OTUs 1 und 2 einander zugeordnet

Bootstrapping

Orginalsequenzen

Sequence	Position 1	2	3	4	5	6	7	8	9
A	A	A	A	A	G	T	G	C	A
B	A	G	C	C	G	T	G	C	G
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	G

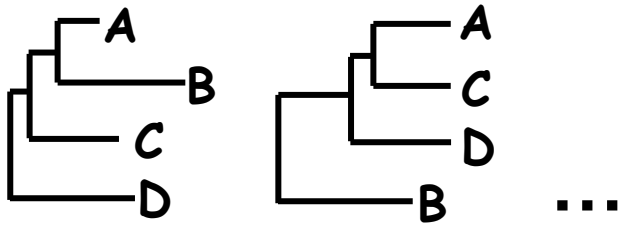
Pseudosample 1

Sequence	Position 1	2	2	4	5	5	7	8	8
A	A	A	A	A	G	G	G	C	C
B	A	G	G	C	G	G	C	C	C
C	A	G	G	T	A	A	C	C	C
D	A	G	G	G	A	A	C	C	C

Pseudosample 2

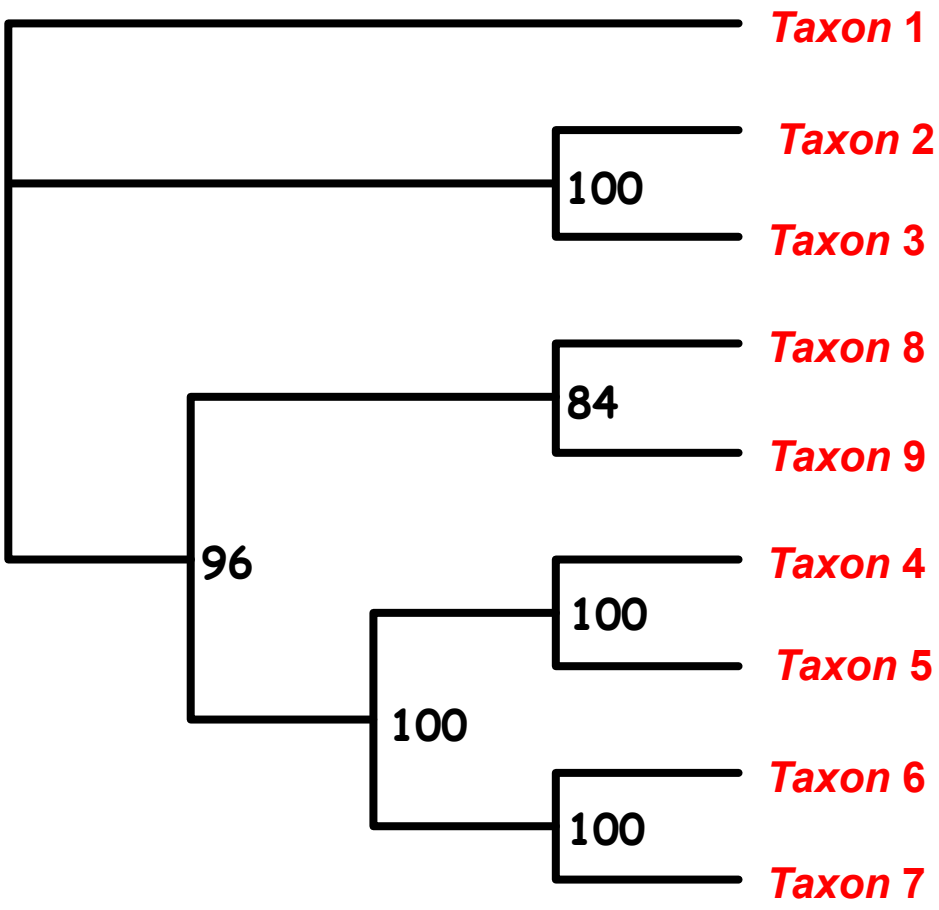
Sequence	Position 1	1	1	4	4	6	7	7	7
A	A	A	A	A	A	T	G	G	G
B	A	A	A	C	C	T	G	G	G
C	A	A	A	T	T	T	C	C	C
D	A	A	A	G	G	T	C	C	C

z.B. 100 Wiederholungen



➡ 100 Stammbäume

Ergebnis eines Bootstrappings

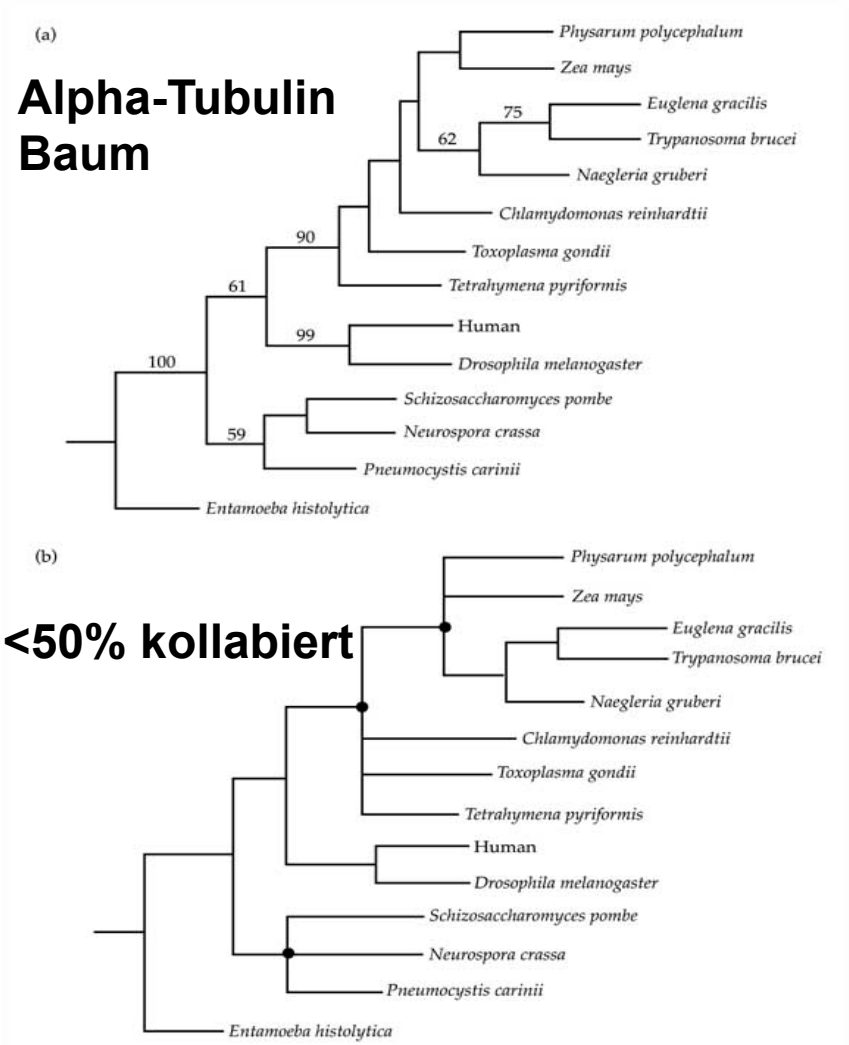


Majority-rule consensus tree

123456789	Freq
.**.....	100.00
...**....	100.00
.....**..	100.00
...****..	100.00
...******	96.00
.......**	84.00
...****.*	13.00
...*****.	5.00
.*******	3.00
.**....*.	1.00
.**.....*	1.00

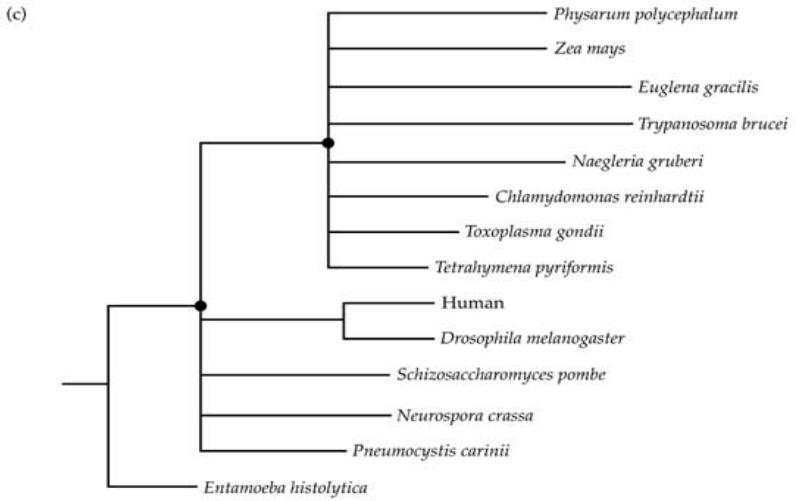
Umgang mit Bootstrap-Werten

Alpha-Tubulin Baum



<50% kollabiert

<90% kollabiert



Interpretation von Bootstrap-Werten

• hohe Bootstrap-Werte (>70%) zeigen eine gute Unterstützung der Gruppierung durch die Daten an

• Verzweigungen mit B-Werten unter 50% sollten „auf eine gemeinsame Linie kollabiert werden“ > Polytomie

• niedrige Werte bedeuten nicht, dass die Gruppierung falsch ist! Sie ist nur von den vorliegenden Daten nicht ausreichend unterstützt.

• Bootstrapping kann als Versuch gesehen werden, die Robustheit einer phylogenetischen Rekonstruktion zu testen gegenüber Störungen in ihrer „Balance“ für und wider die Zueinandergruppierung von Taxa.

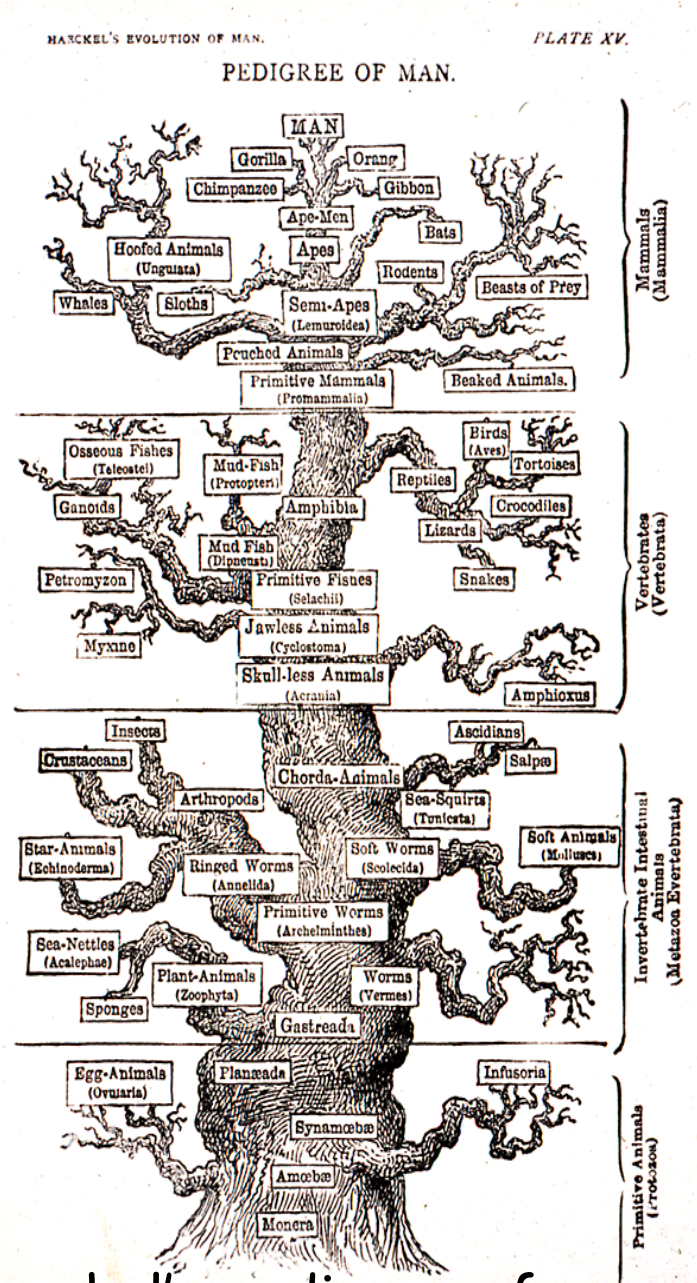
• Wissenschaftliche Journale machen Verwendung dieser Methode zur Pflicht beim Zeigen eines Stammbaums!

Allgemeine Hinweise

- **stets mehrere Verfahren ausprobieren**
- **je größer der Datensatz, desto besseres Ergebnis (jedenfalls theoretisch)**
- bei Distanzen: adäquate Substitutionsmodelle und Korrektur für multiple Austausche
- bei ML: zu Grunde liegende Modelle relativ zueinander testen („so kompliziert wie nötig, so einfach wie möglich“)
- bei MP, aber auch bei anderen: schnell evolvierende Taxa (Homoplasien! LBA!) entfernen oder mehr Taxa hinzufügen

Darwin's letter to Thomas Huxley 1857

"The time will come I believe, though I shall not live to see it, when we shall have fairly true genealogical (*phylogenetic*) trees of each great kingdom of nature."



Haeckel's pedigree of man

Phylogenetic methods

Molecular Ecology (2006) 15, 3669–3679 doi: 10.1111/j.1365-294X.2006.03036.x

Worldwide phylogeography of the blacktip shark (*Carcharhinus limbatus*) inferred from mitochondrial DNA reveals isolation of western Atlantic populations coupled with recent Pacific dispersal

D. B. KEENEY and E. J. HEIST
Fisheries and Illinois Aquaculture Center, Department of Zoology, Southern Illinois University Carbondale, Carbondale, IL 62901-6511, USA

Schwarzspitzenhai

Evolutionary relationships among unique mtDNA haplotypes were reconstructed using the **maximum-parsimony (MP)** optimality criterion with all mutations weighted equally and indels treated as a fifth state. A two-nucleotide indel at positions 1045 and 1046 was treated as one event by omitting the second nucleotide from analyses. **Heuristic tree searches** were performed for all MP analyses with 1000 random-addition replications, saving a maximum of 1000 trees per replicate, and **tree-bisection–reconnection** (TBR) branch swapping. Statistical support for nodes was determined via 1000 nonparametric **bootstrap replicates** (Felsenstein 1985) with 10 random-addition sequences per replicate, saving a maximum of 1000 trees per replicate, and **nearest neighbour interchange** (NNI) branch swapping. Haplotype trees were initially rooted using blacktip reef shark (*C. melanopterus*) and Australian blacktip shark, *C. tilstoni* , sequences as outgroups. Although the relationships of species within the genus *Carcharhinus* are not fully resolved (Lavery 1992; Naylor 1992), *C. melanopterus* and *C. tilstoni* were the closest relatives to *C. limbatus* for which tissue samples were available. *C. melanopterus* was used as the sole outgroup after *C. limbatus* was found to be paraphyletic to *C. tilstoni* in the MP analyses.