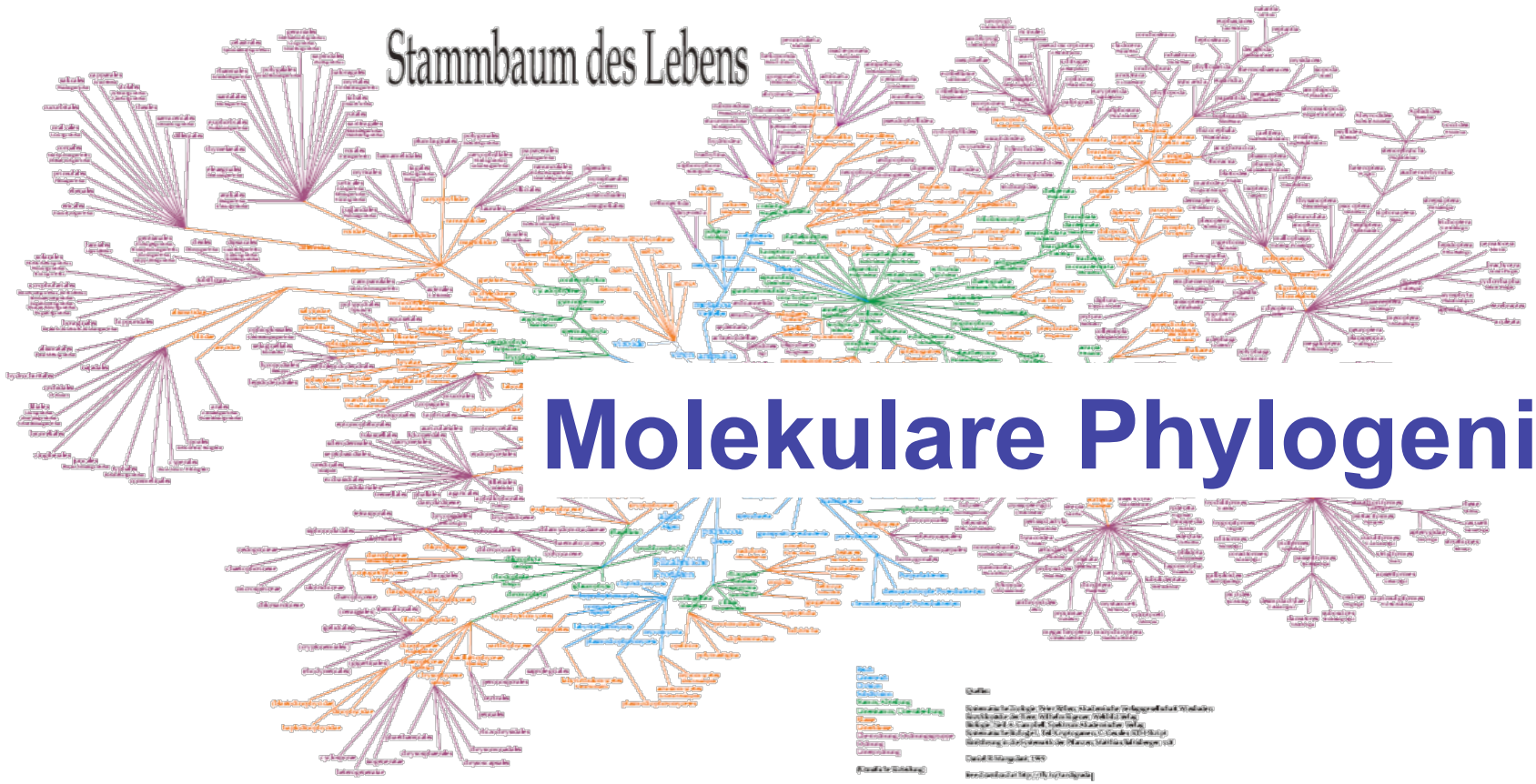
WS 2018/2019

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

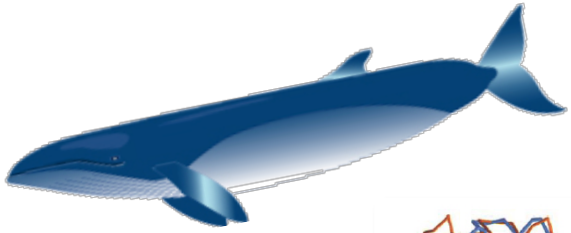
„Genomforschung und Sequenzanalyse
- Einführung in Methoden der Bioinformatik- “

Thomas Hankeln



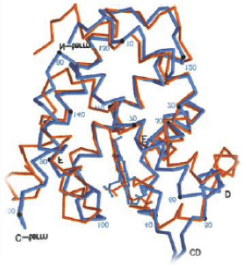
Molekulare Phylogenie

Warum molekulare Phylogenie?



- „Walverwandtschaften“

Systematik



- Wann und wie sind Hämoglobine entstanden?

Funktion

- Wann hat sich der moderne Mensch entwickelt?

Evolution

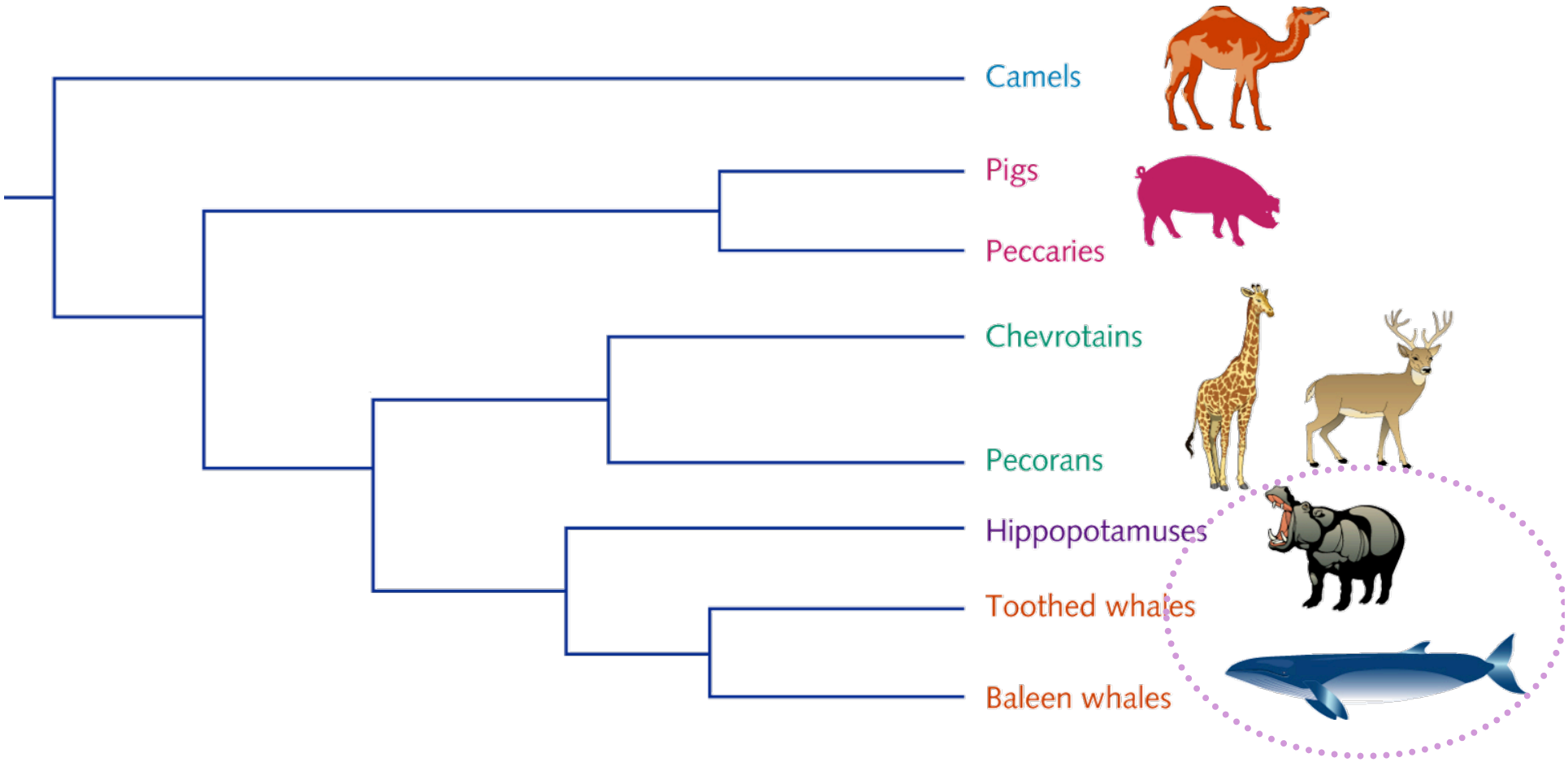
- SARS: woher kommt dieses Virus?

Epidemiologie

- AIDS without sex? - *„The Florida Dentist Case“*

CSI Miami

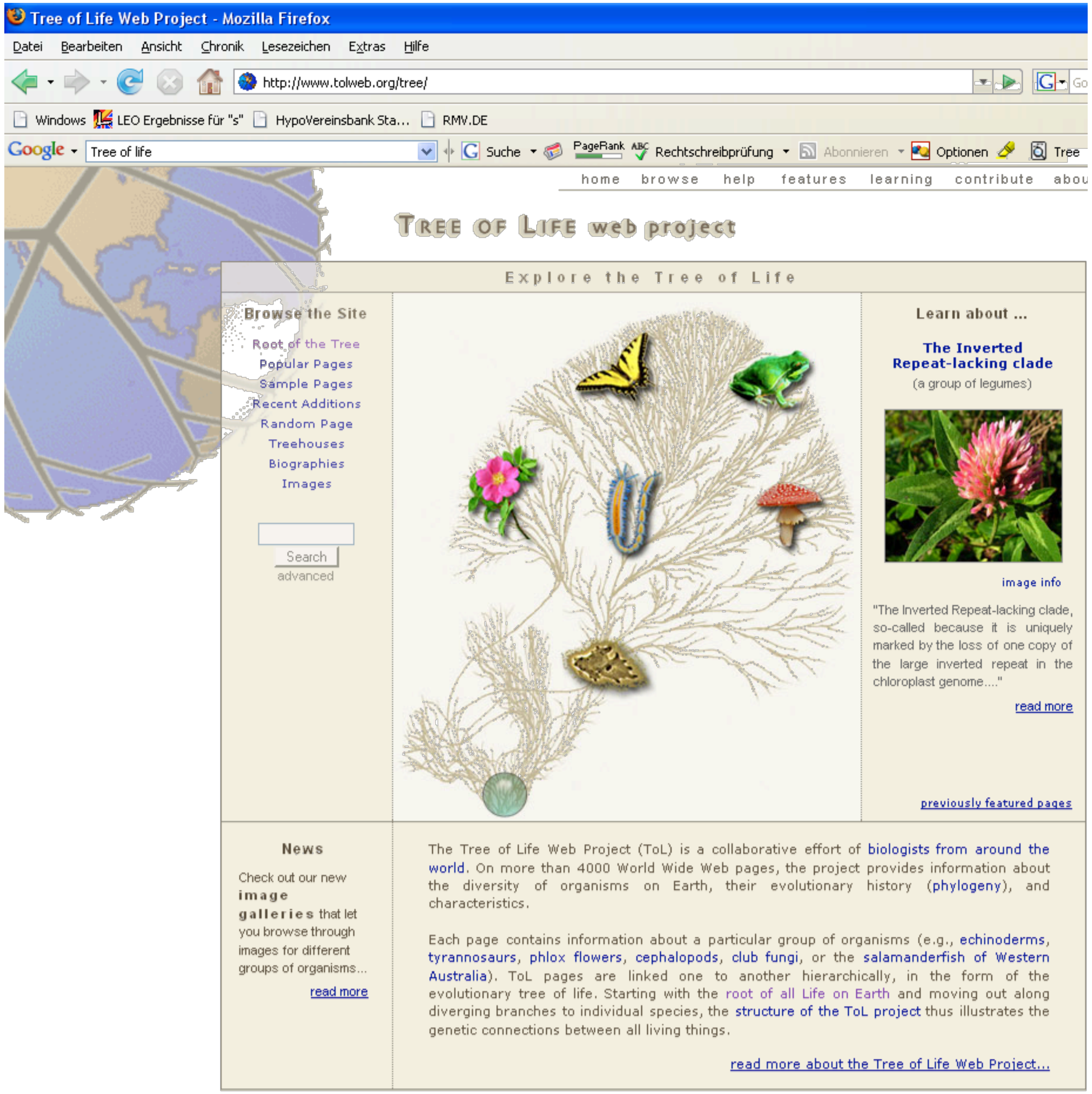
„Walverwandtschaften“



=> Systematik

Stearns & Hoekstra, 2005

Tree of Life



Discussion of Phylogenetic Relationships

Two alternative views on the relationship of the major lineages (c

The "archaea tree":

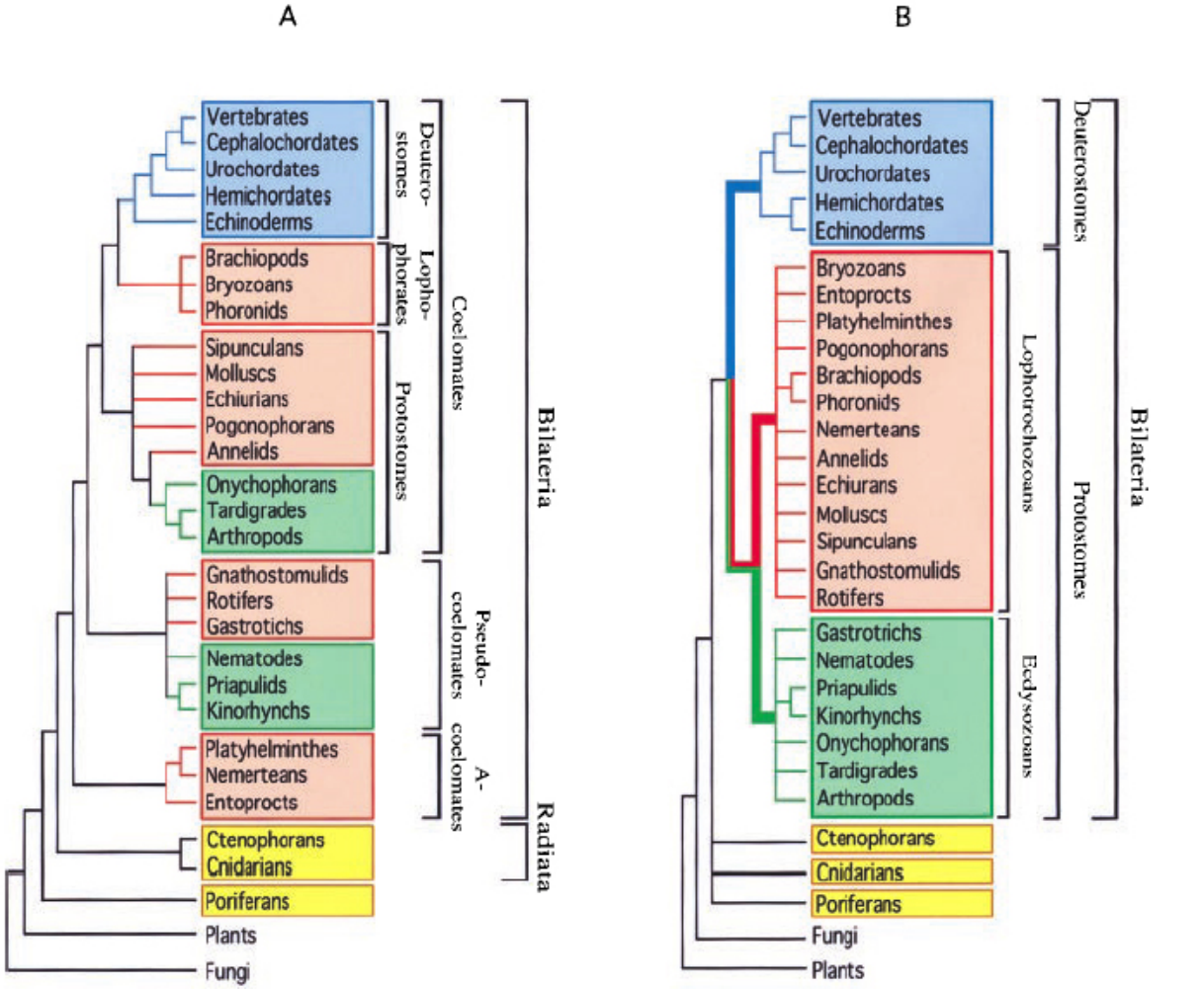
```
       =============== Eubacteria
      |
      |               == Euryarchaeota
=====|     =Archaea=|
      | == |          == Crenarchaeota-Eocytes
           |
            ============ Eukaryotes
```

The "eocyte tree":

```
       ======== Eubacteria
      |
      |     ===== Euryarchaeota
=====|     |
      | == |      == Crenarchaeota-Eocytes
           | == |
                  == Eukaryotes
```

Tree of Life

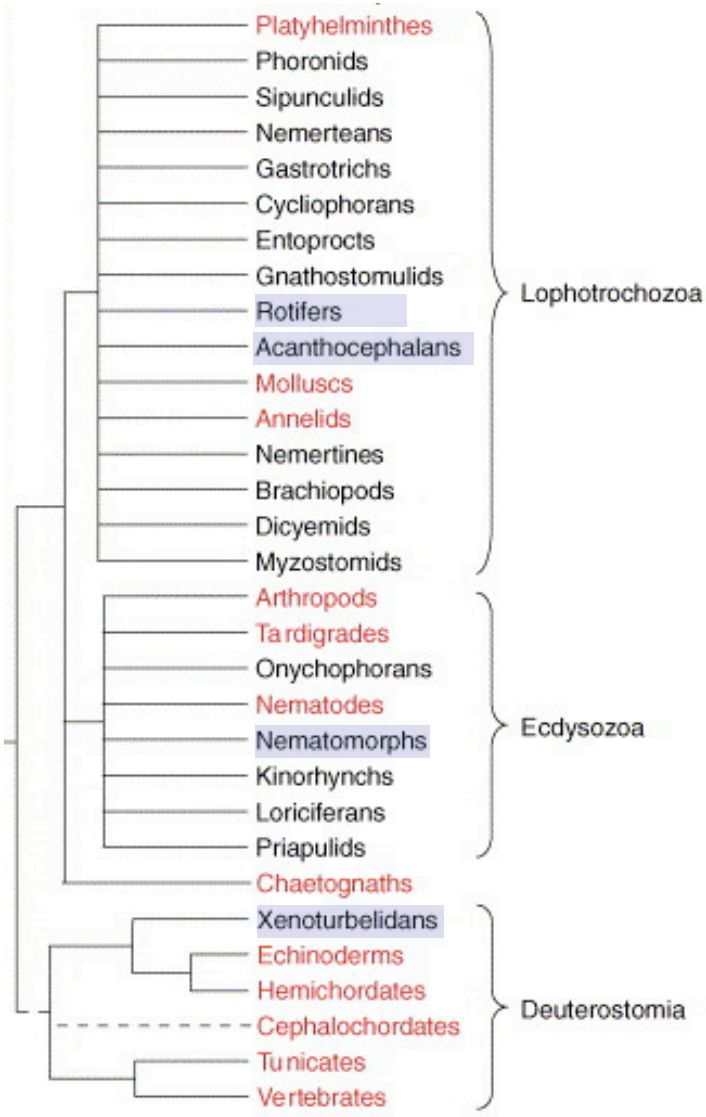
Morphologie

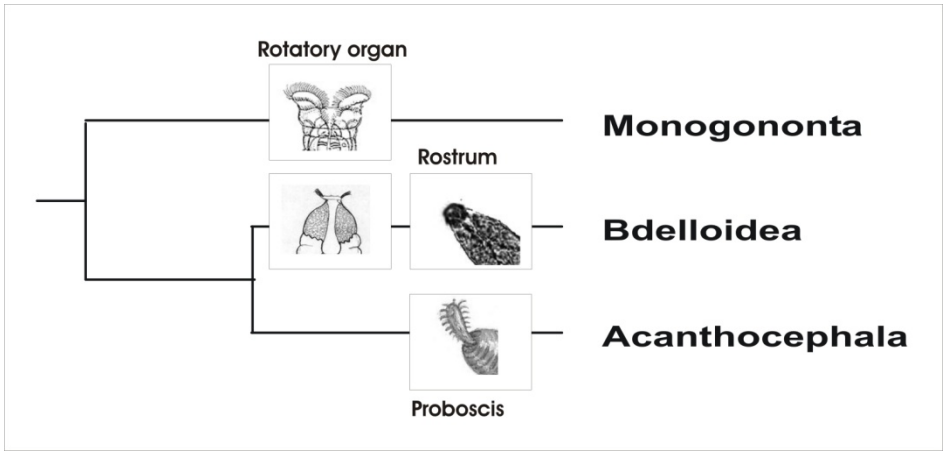


rRNA-Sequenzen

Fig. 1. Metazoan phylogenies. (*A*) The traditional phylogeny based on morphology and embryology, adapted from Hyman (11). (*B*) The new molecule-based phylogeny. A conservative approach was taken in *B*: i.e., some datasets provide resolution within some of the unresolved multifurcations displayed, but we have limited the extent of resolution displayed to that solidly provided by rRNA only.

Tree of Life



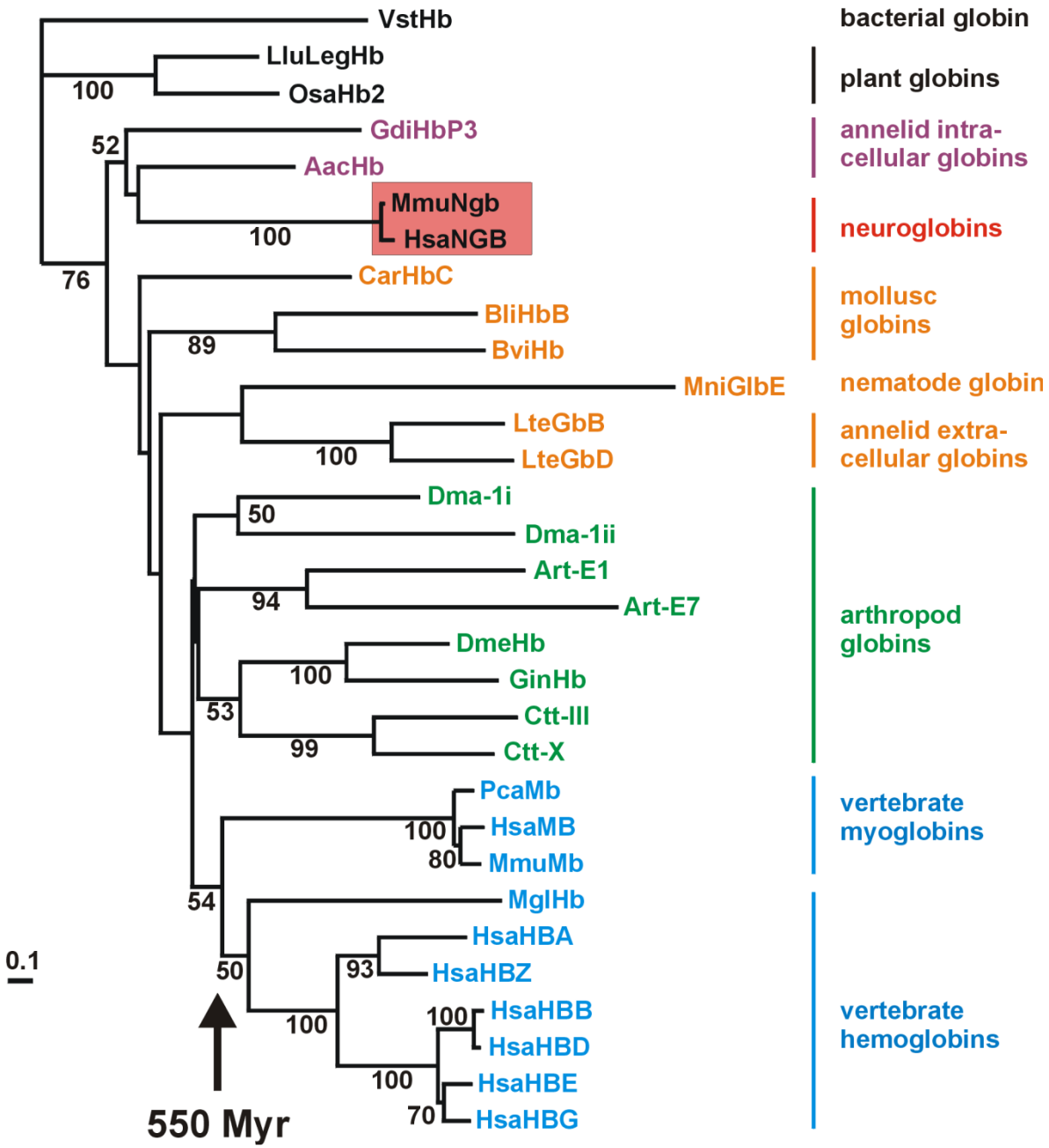


- **Rotatory organ**

  > partially reduced in bdelloids
  > absent in acanthocephalans

- **Retractable anterior end (rostrum/proboscis)**

  > emergence likely before acanthocephalan split
  > key event for the evolution of acanthocephalan endoparasitism



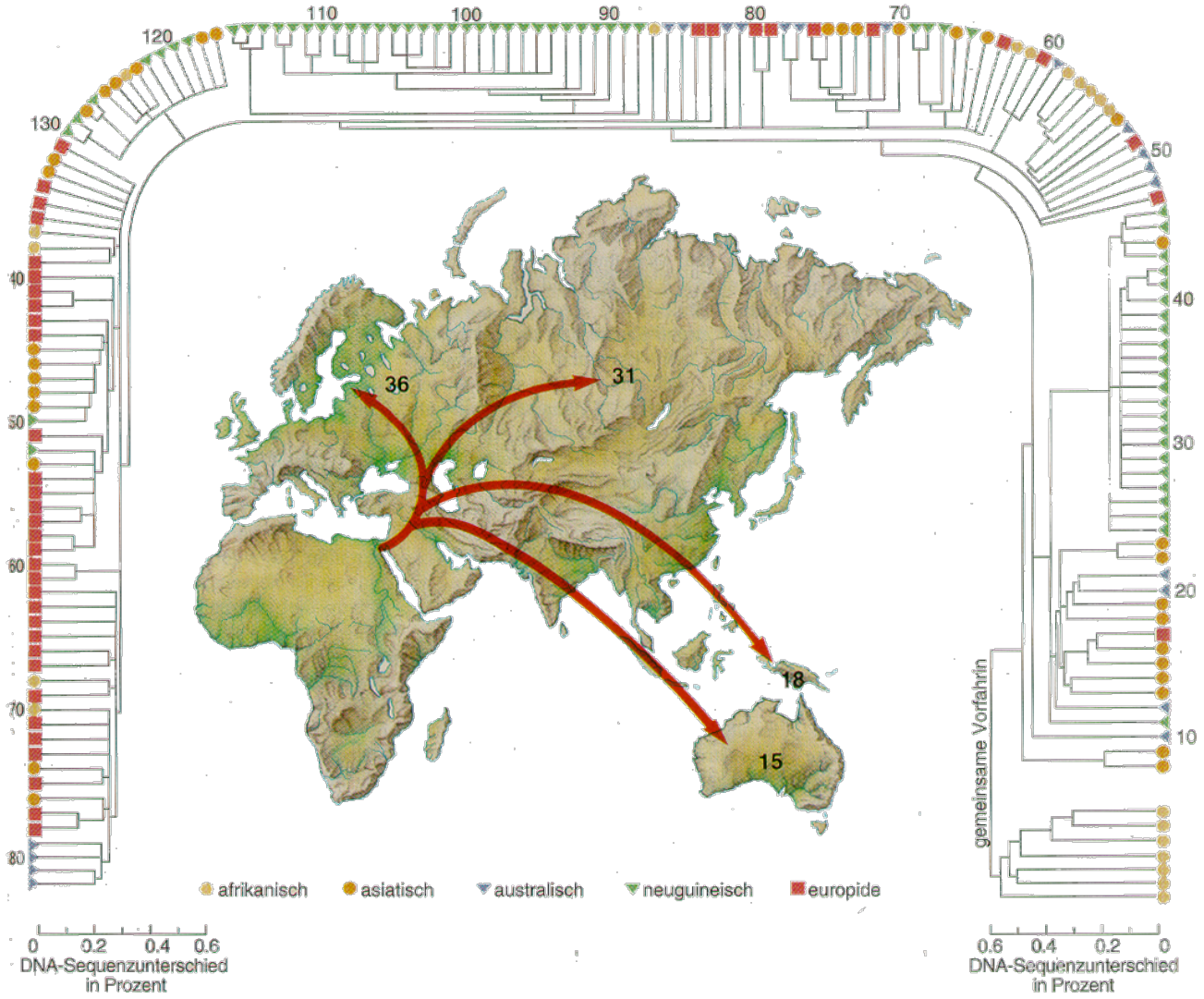
Globin-Evolution

…Neuroglobin muss sehr alte Erfindung sein

- In allen Spezies?
- Ursprüngliche Funktion?

Evolution von Homo sapiens

…**out of Africa**
vor etwa 60.000 bis 100.000 Jahren



=> Populationsbiologie

SARS-Phylogenie

DNA (Komplettgenom)

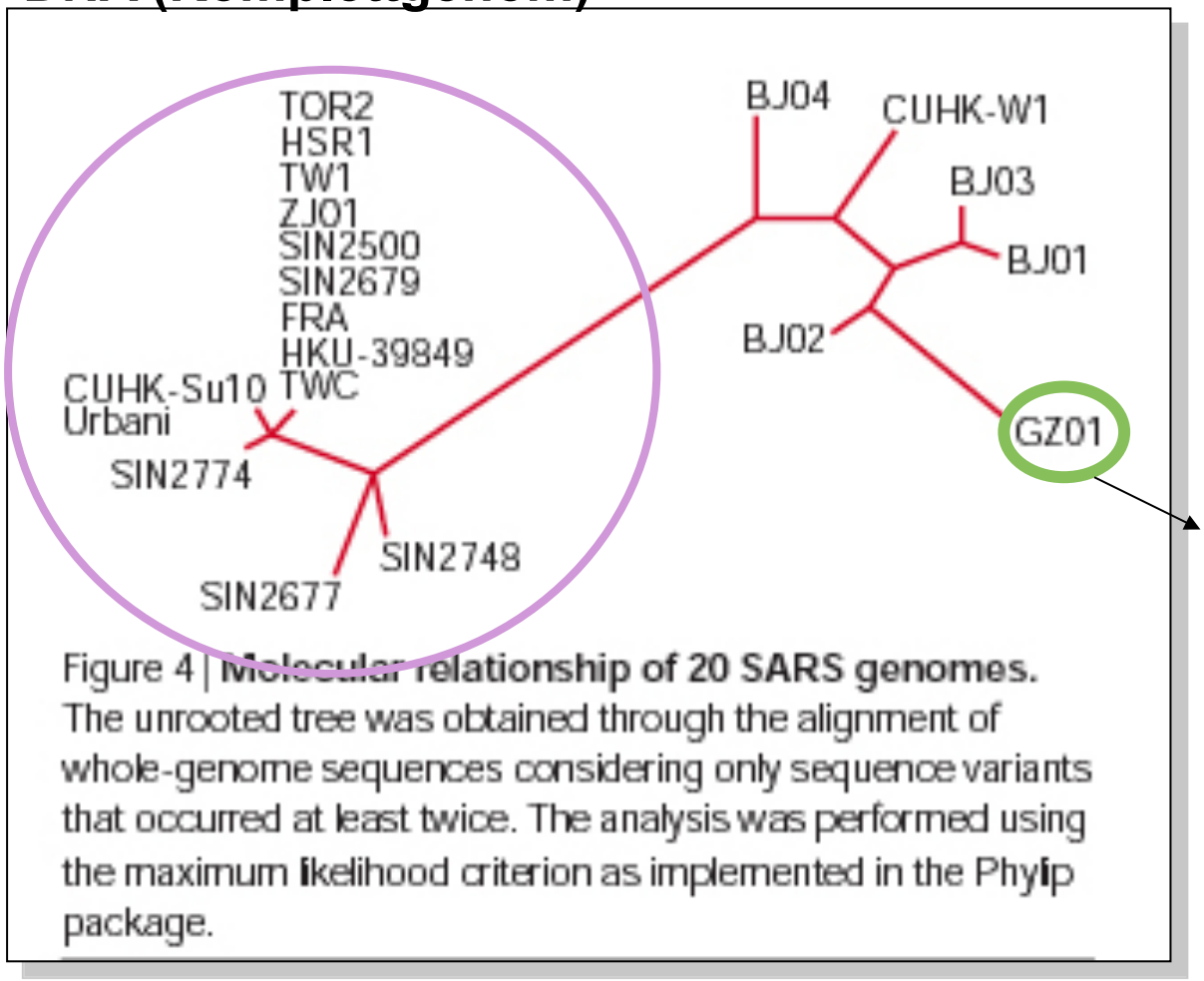


Figure 4 | Molecular relationship of 20 SARS genomes. The unrooted tree was obtained through the alignment of whole-genome sequences considering only sequence variants that occurred at least twice. The analysis was performed using the maximum likelihood criterion as implemented in the Phylip package.

- **Varianten sind >99% identisch. Dennoch ist eine geographische Zuordnung möglich.**

Sequenz zeigt Besonderheit:

Sein Spike-Gen hat 29 Bp zusätzlich, die sonst nur in tierischen SARS-Verwandten gefunden worden sind!

HIV

The New York Times
Thursday, November 2, 2006

Health

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

FITNESS & NUTRITION | HEALTH CARE POLICY | MENTAL HEALTH

Man Infected With H.I.V. by Florida Dentist Dies

Published: June 28, 1993

A man who contracted H.I.V. from his dentist died Saturday of complications of AIDS. He was 33.

The man, Richard Driskill, died of pancreatic failure, said his lawyer, Robert Montgomery.

Mr. Driskill was one of six patients known to have been infected by Dr. David J. Acer, who died of AIDS in 1990. The sixth person on that list, Sherry A. Johnson, 18, was identified last month by the Centers for Disease Control and Prevention.

Kimberly Bergalis was the first of Dr. Acer's infected patients to die. She died in 1991 at the a

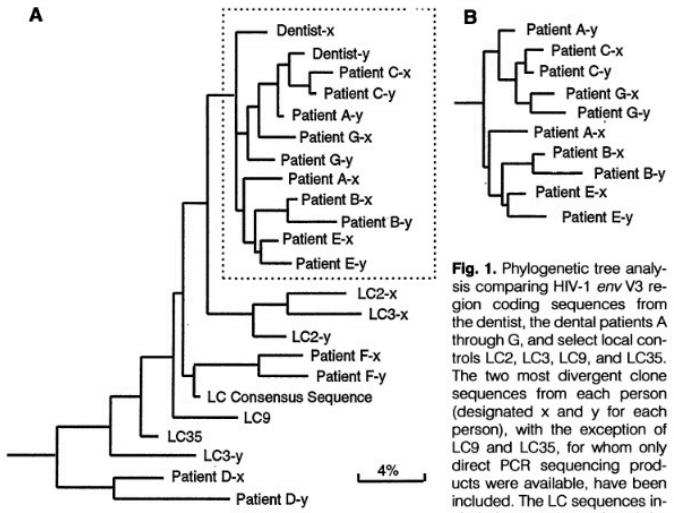


Fig. 1. Phylogenetic tree analysis comparing HIV-1 *env* V3 region coding sequences from the dentist, the dental patients A through G, and select local controls LC2, LC3, LC9, and LC35. The two most divergent clone sequences from each person (designated x and y for each person), with the exception of LC9 and LC35, for whom only direct PCR sequencing products were available, have been included. The LC sequences included were those found by pairwise distance measurement (Table 1) and signature pattern analysis (Table 2) to be the closest control sequences to the dental group sequences, which are enclosed by a box in (**A**). An LC consensus sequence has also been included, and the tree was rooted upon the African sample ELI. The PAUP parsimony algorithm was used to analyze 279 aligned sites (*25*), of which 146 sites were varied. When the dentist's viral sequences were withdrawn from the analysis, or required to cluster with the LC consensus sequence (*25*), the dental clade remained otherwise unaffected, as shown in (**B**). Vertical distances are for clarity only; the lengths of the horizontal branches are proportional to the single base changes and can be read as percentage differences with the scale bar.

SCIENCE • VOL. 256 • 22 MAY 1992 **1167**

Dr. Acer continued to treat patients after learning in 1986 that he was infected with H.I.V. and after developing AIDS symptoms in 1987.

=> Epidemiologie/Forensik…

Phylogenie, aber wie?

- Vergleich der strukturellen Merkmale von Fossilien
- Vergleich der **morphologischen** und **physiologischen** Merkmale rezenter Lebewesen
- Vergleich der **Ontogenese**
- Analyse der **DNA, RNA** und/oder **Proteinsequenzen**

Vorteile molekularer Daten

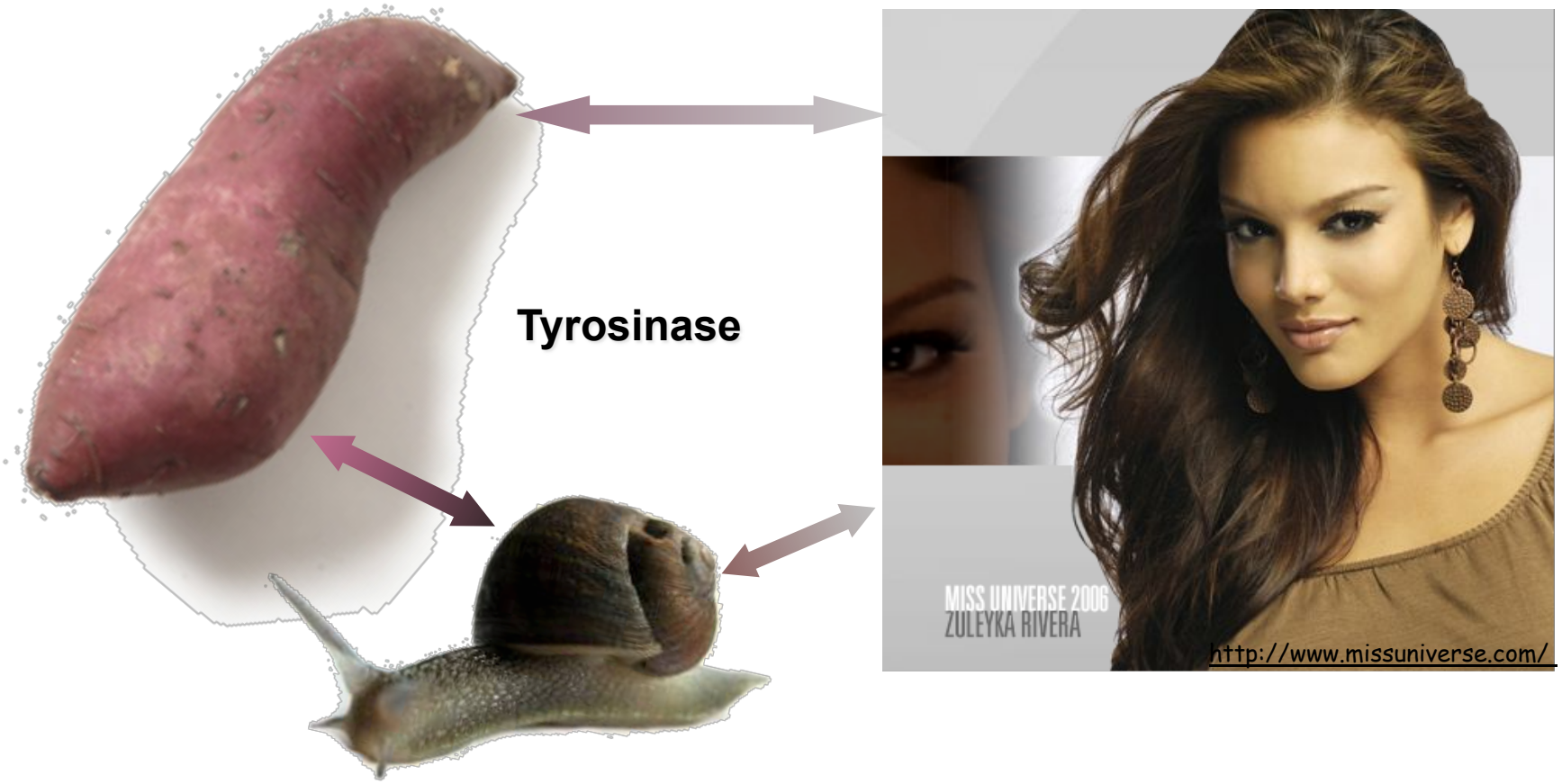
Die Beschreibung der Zustände der molekularen Daten ist stets eindeutig

Die 3. Aminosäure im Präproinsulin des Kaninchens (*Oryctolagus cuniculus*) ist **immer** Serin, und die homologe Position im Präproinsulin des Goldhamsters (*Mesocricetus auratus*) ist **immer** Leucin ("diskrete, diskontinuierliche Charaktere").

Morphologische Beschreibungen arbeiten häufig mit Bezeichnungen wie "dünn", "reduziert", "etwas verlängert", "teilweise geschlossen"etc.

Vorteile der molekularen Daten

Es lassen sich auch entfernt verwandte Spezies und Gene/Proteine vergleichen…



Folie: Bernd Lieb

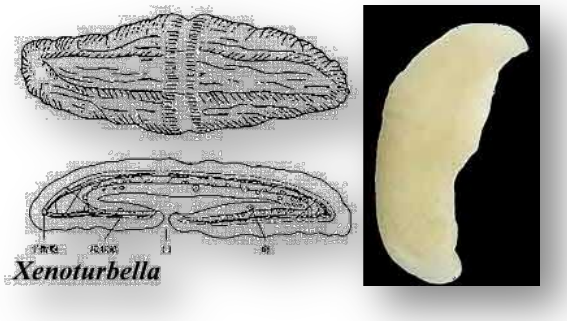
Vorteile der molekularen Daten

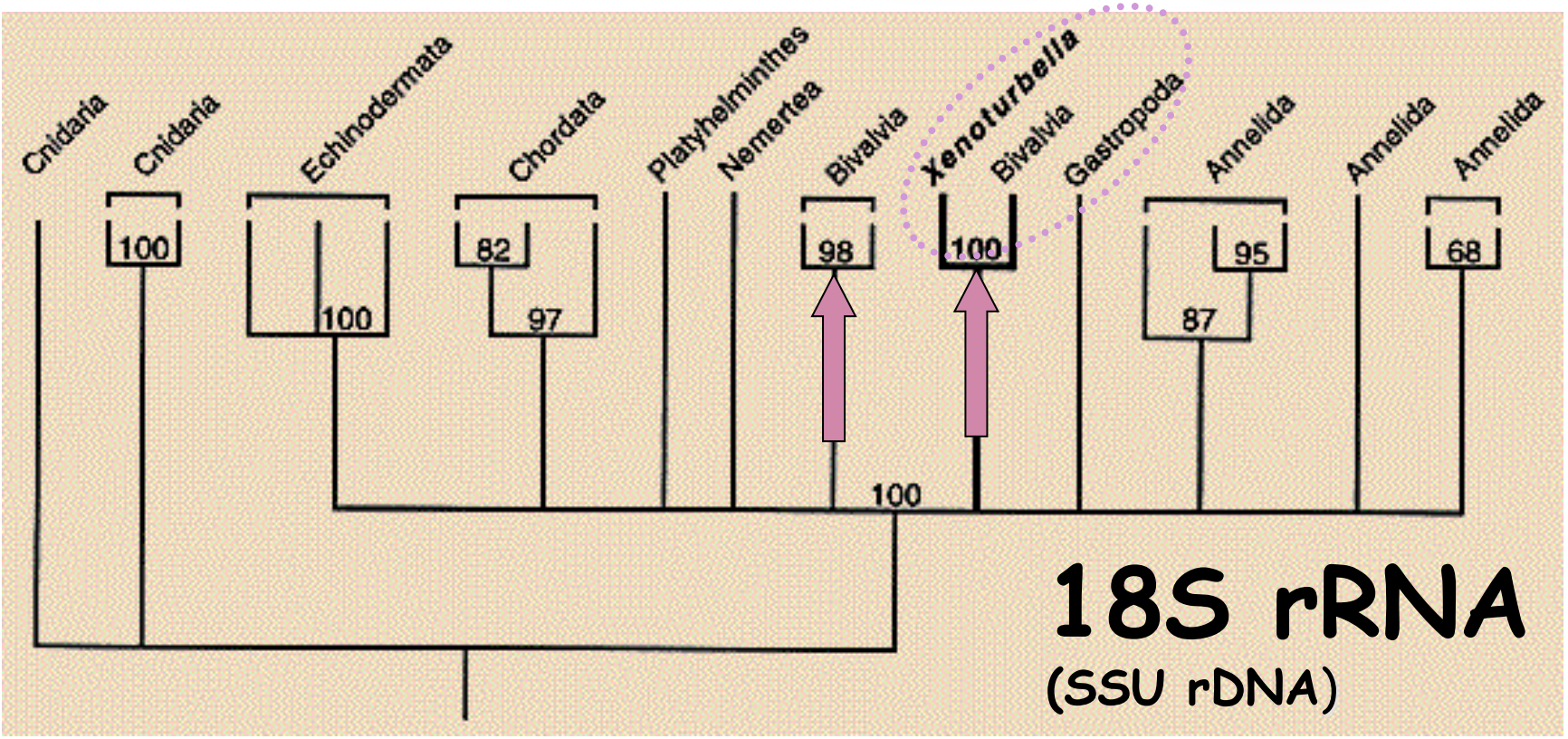
- Evolution der Sequenzen lässt sich mit Modellen beschreiben (PAM, BLOSUM...).
- Molekulare Daten lassen sich relativ einfach quantifizieren.
- Sequenzen zweier Organismen lassen sich ‘einfacher’ homologisieren.
- Molekulare Daten sind in fast beliebiger Menge vorhanden

Aber: Haben molekulare Daten immer recht?

1. Problem: Kontaminationen
2. Problem: ‘Falsche’ Daten
3. Problem: ‘Falsche’ Methoden

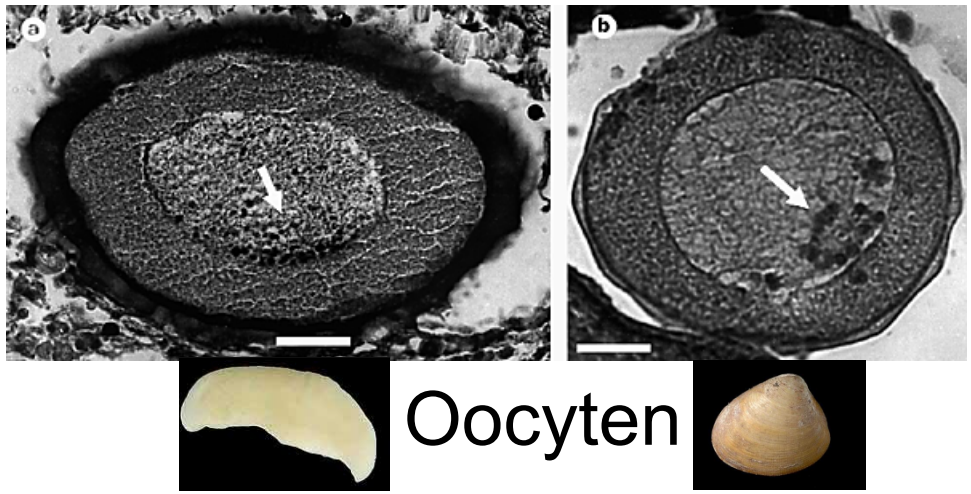
Xenoturbella





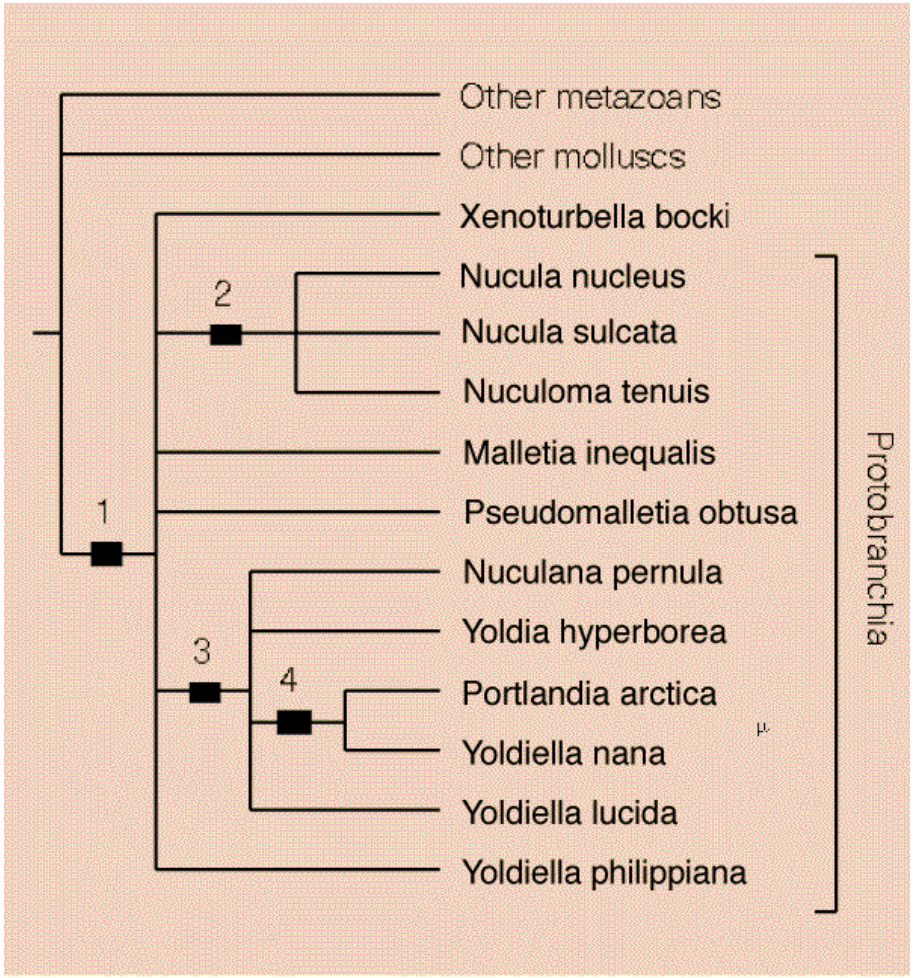
M. Norén, U. Jondelius (1997) Nature 390: 31 – 32

Xenoturbella



Xenoturbella bocki *Nucula sulcata*

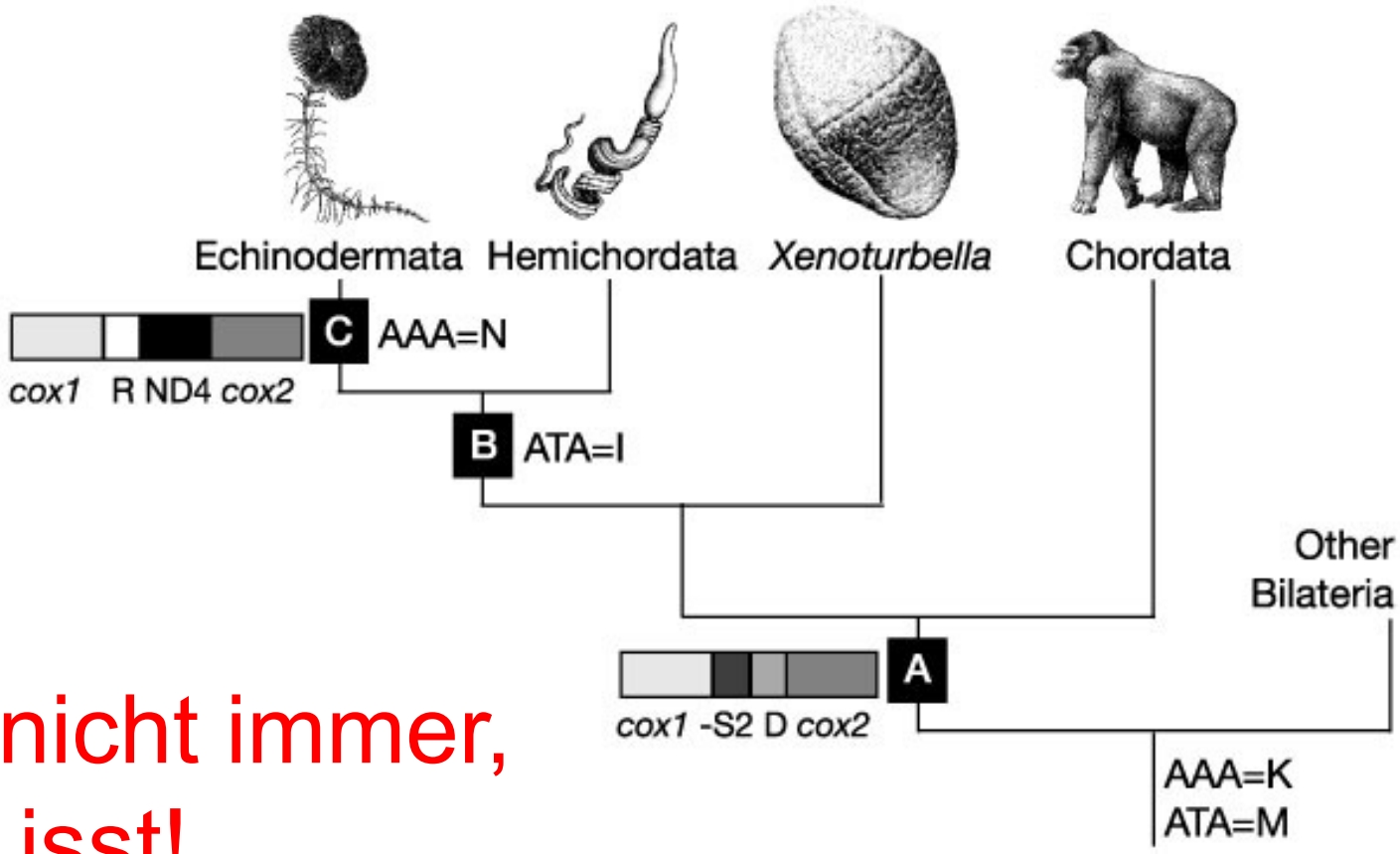
"Cladistic analyses of oogenesis indicates that *Xenoturbella bocki* is a sister group or a subgroup of protobranch bivalves."



Israelsson (1997) Nature 390: 32

Xenoturbella

„….Xenoturbella is a deuterostome that eats molluscs…...“



Du bist nicht immer, was Du isst!

Bourlat et al (2003) Nature 424: 925-928.

Aber: Haben molekulare Daten immer recht?

1. Problem: Kontaminationen
2. Problem: ‘Falsche’ Daten
3. Problem: ‘Falsche’ Methoden

..dazu später mehr

Unterschiedliche Datensätze für verschiedene Fragestellungen

- **auf Protein-Ebene**

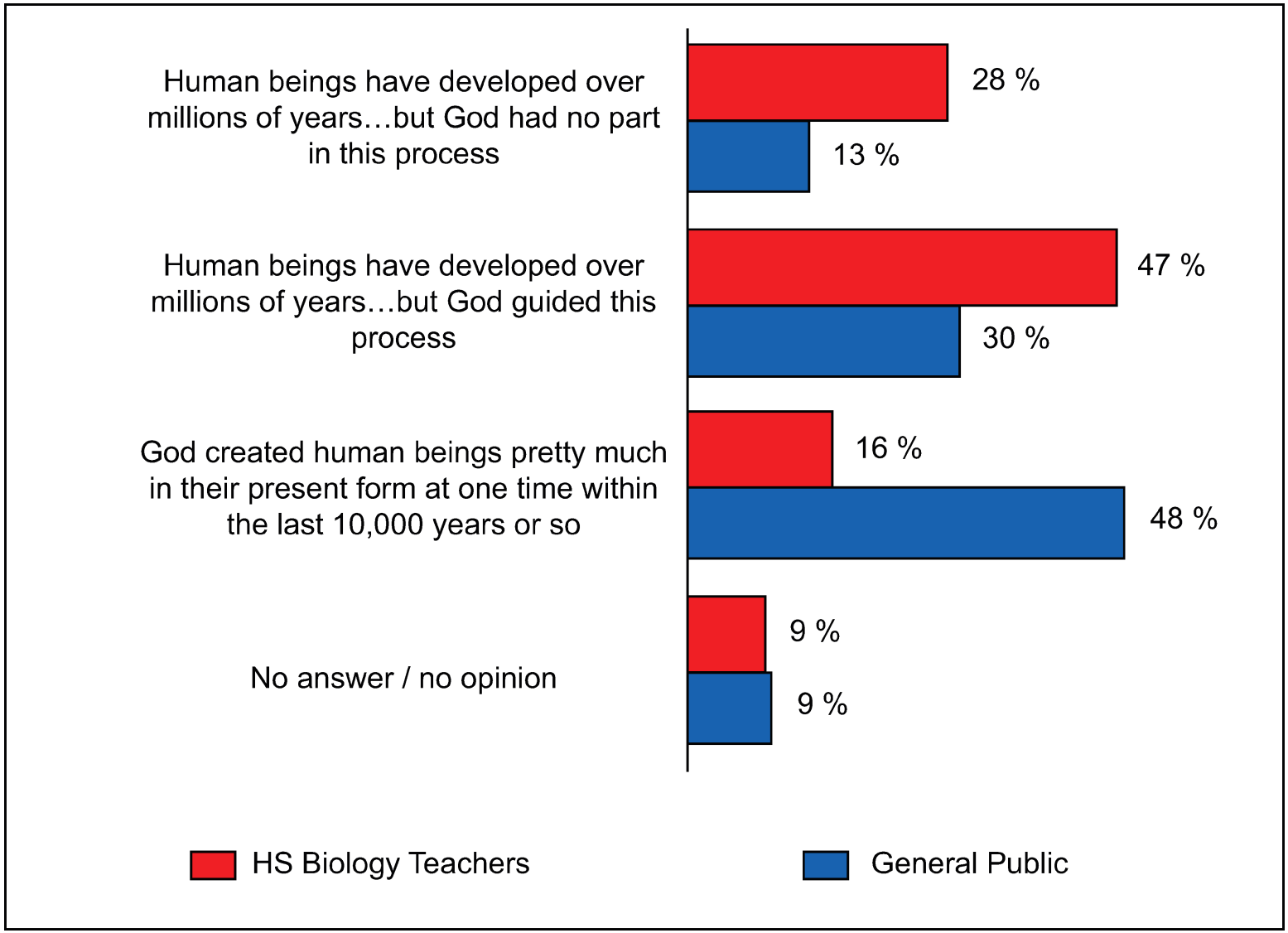
- AS-Sequenzen	**weit entfernte Taxa, Proteinfunktion und –evolution**
- Allozyme*	**nah verwandte Taxa, Populationen**

- **auf DNA-Ebene**

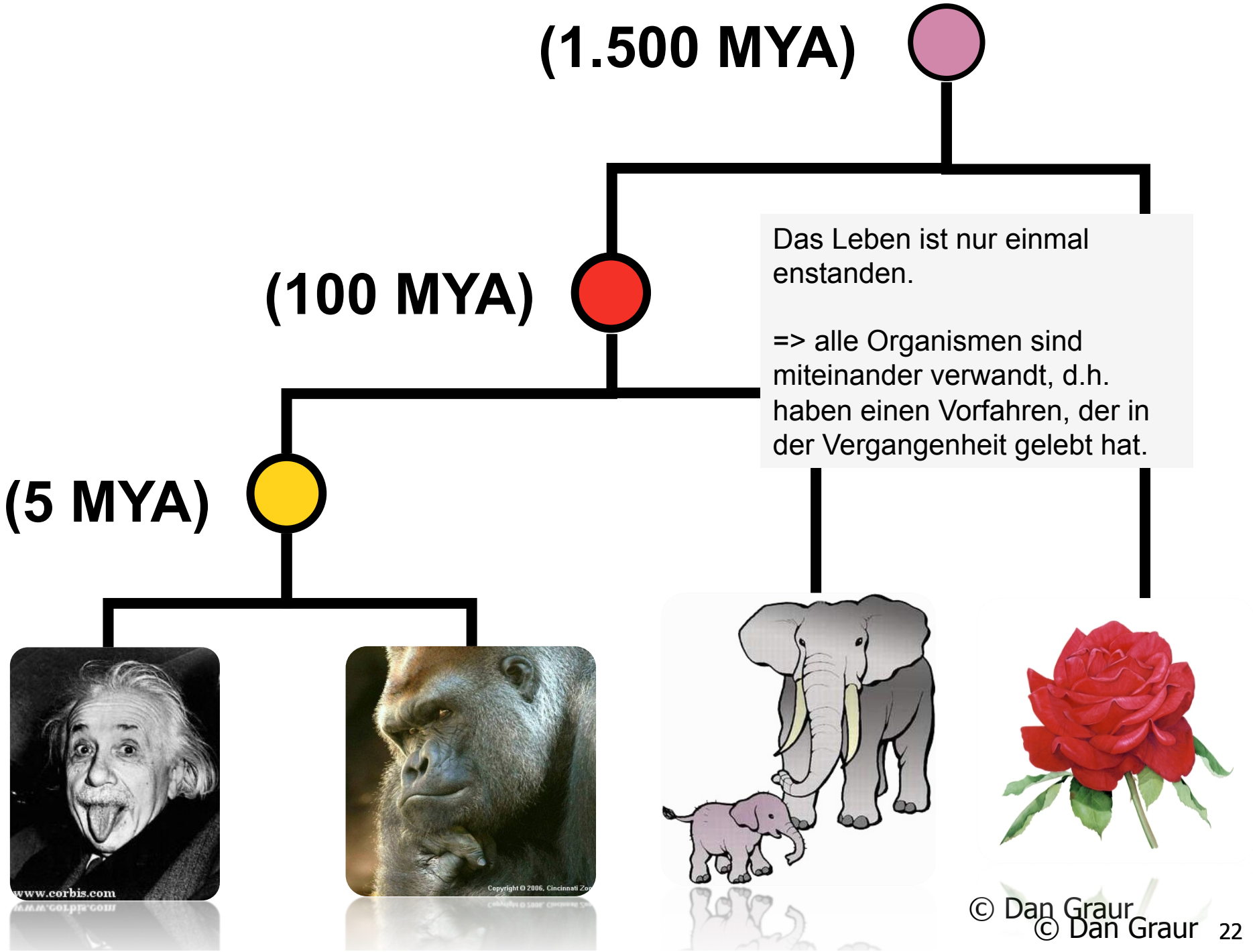
- DNA-Sequenzen !!	**Systematik, Gen(om)evolution**
- polymorphe Marker: (Mikro)satelliten, SNPs RFLPs, AFLPs, RAPDs	**Populationsgenetik, Forensik (DNA-Fingerprinting)**
- DNA-DNA-Hybridisierung*	**Systematik nicht zu weit entfernter Taxa**

* Veraltete Methoden

Die Grundlage...

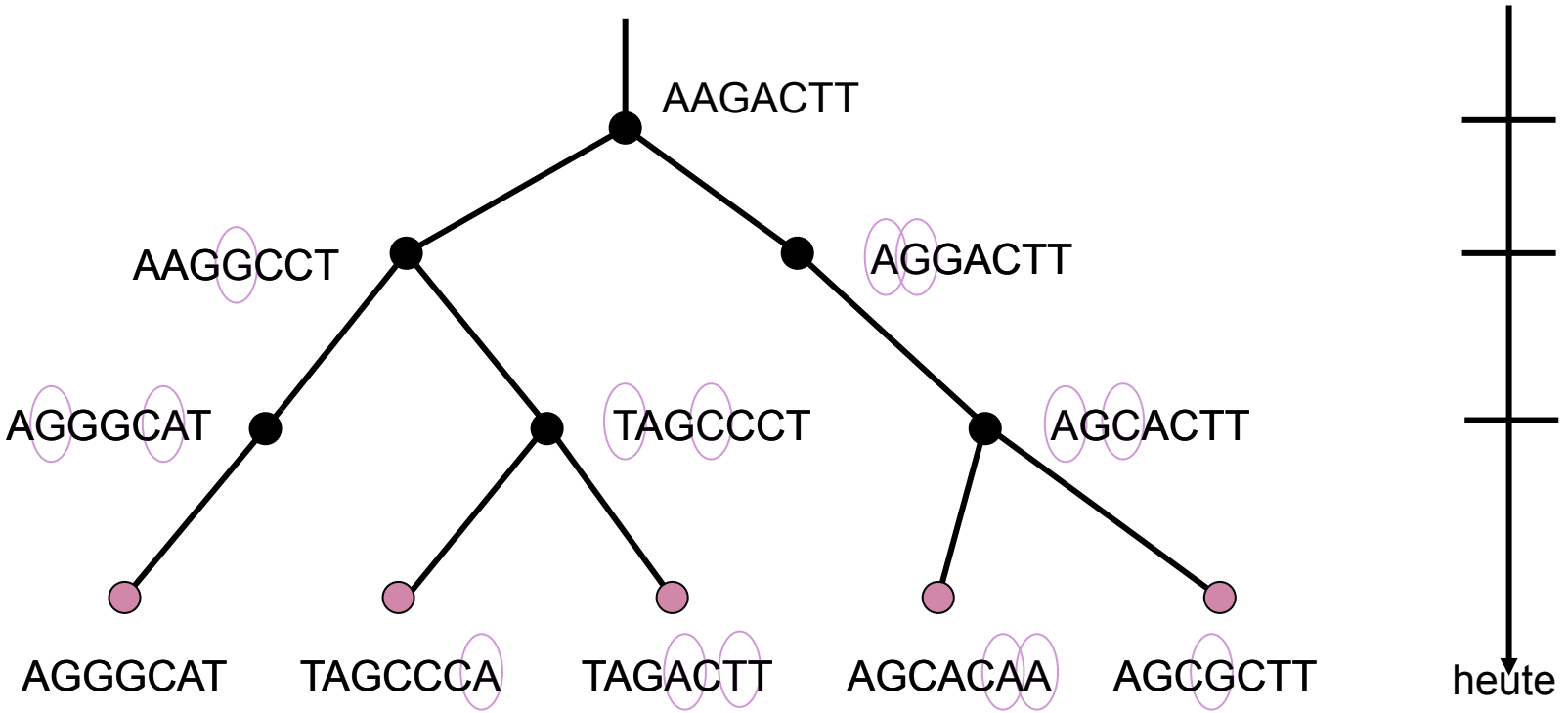


2007



© Dan Graur
© Dan Graur

Evolution vollzieht sich durch Veränderungen

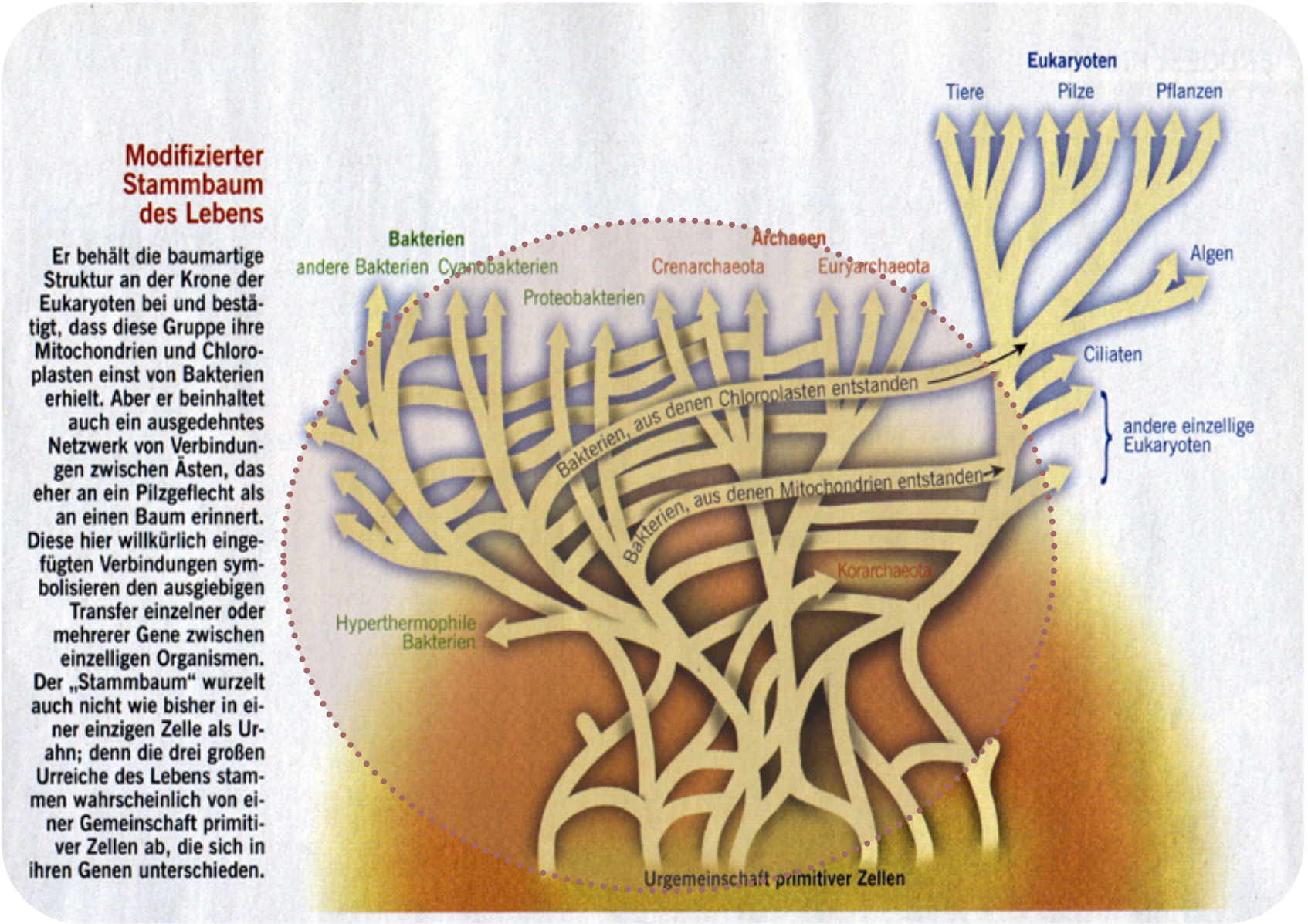


OTU **A** **B** **C** **D** **E**

Voraussetzungen der molekularen Phylogenie

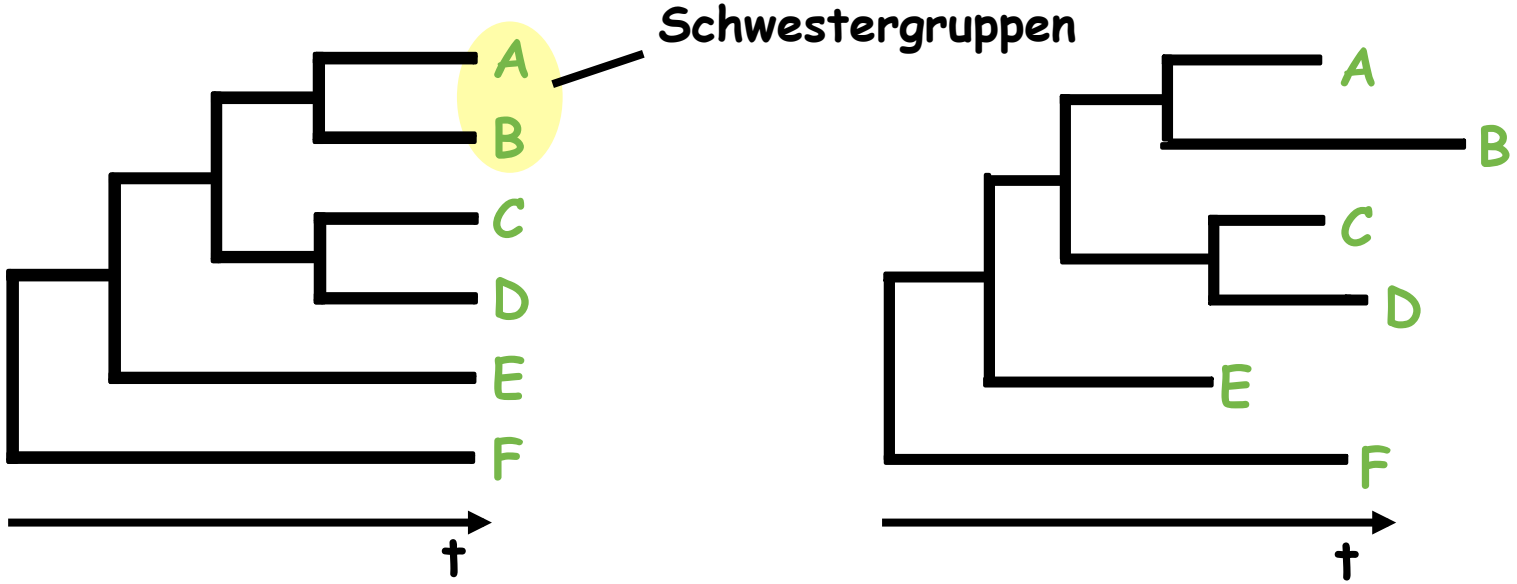
1. Evolution vollzieht sich durch Veränderungen.
2. Verwandte Spezies stammen von einem gemeinsamen Vorfahren ab.
3. Die Speziesbildung vollzog sich durch hierarchische Auftrennung.
4. Deren Verlauf lässt sich durch Stammbäume darstellen.
5. Es gibt nur einen historisch korrekten Stammbaum.
6. Organismen sind historisch. Sowohl die Morphologie als auch die DNA- und Aminosäuresequenzen speichern die Informationen über die Vergangenheit.
7. Die Methoden der molekularen Evolution erlauben die Extraktion der in der DNA bzw. den Proteinen gespeicherten Informationen.

Horizontaler Gentransfer



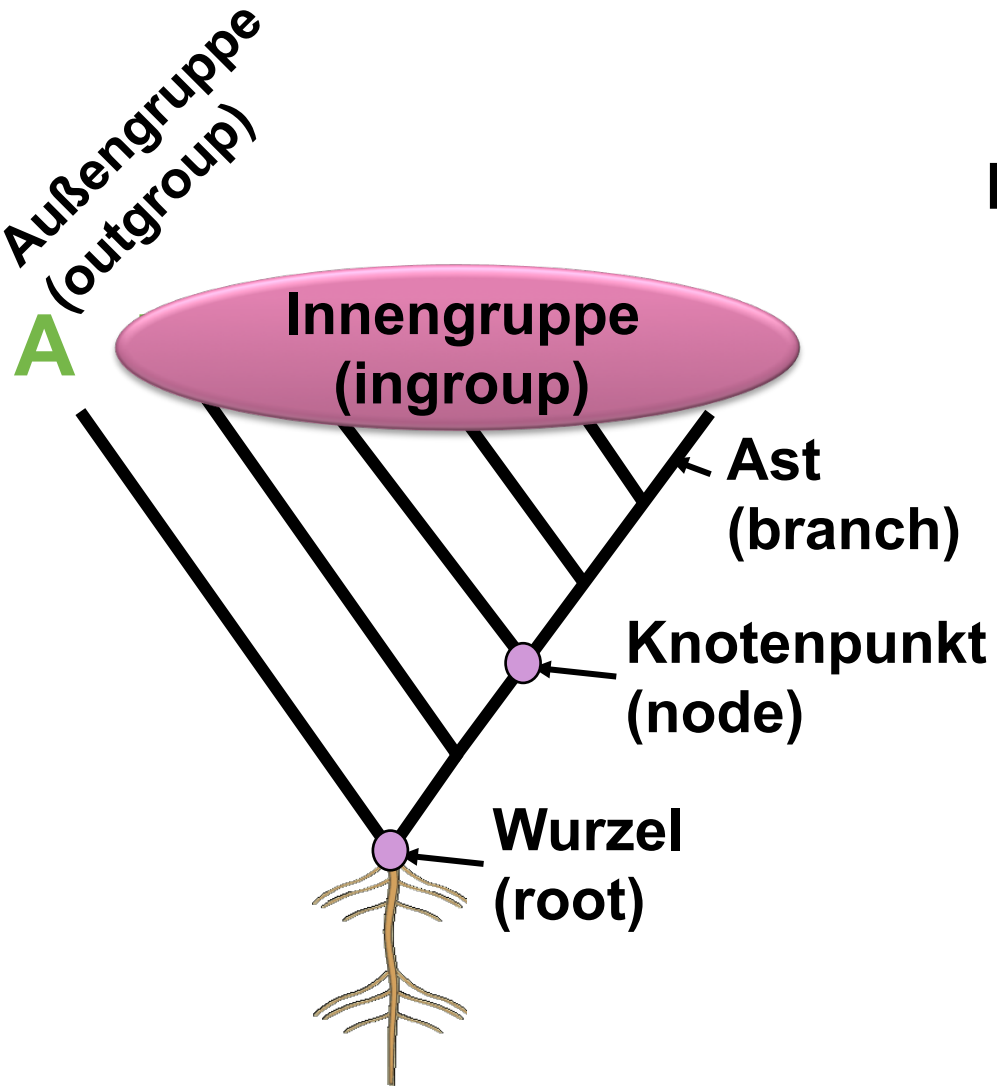
Was ist ein Stammbaum?

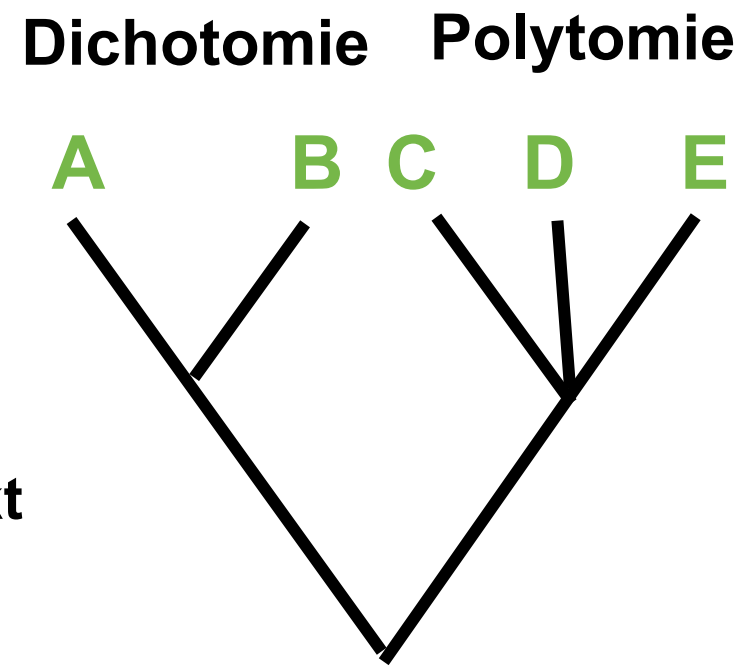
- Darstellung der Verwandtschaftsverhältnisse



A – F sind "operational taxonomic units" (OTUs)

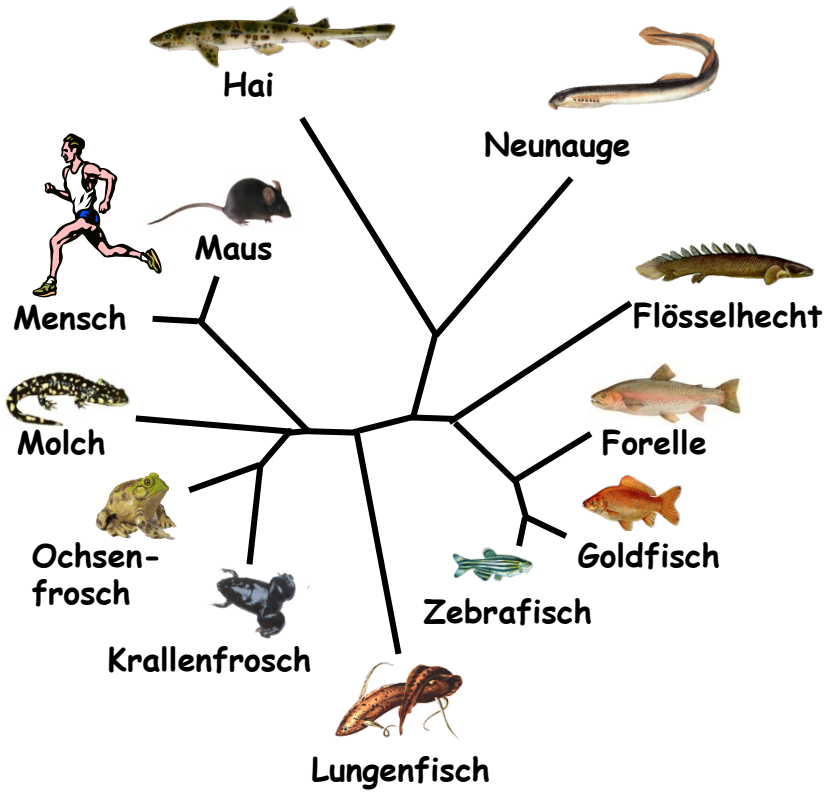
Phylogenetische Grundbegriffe





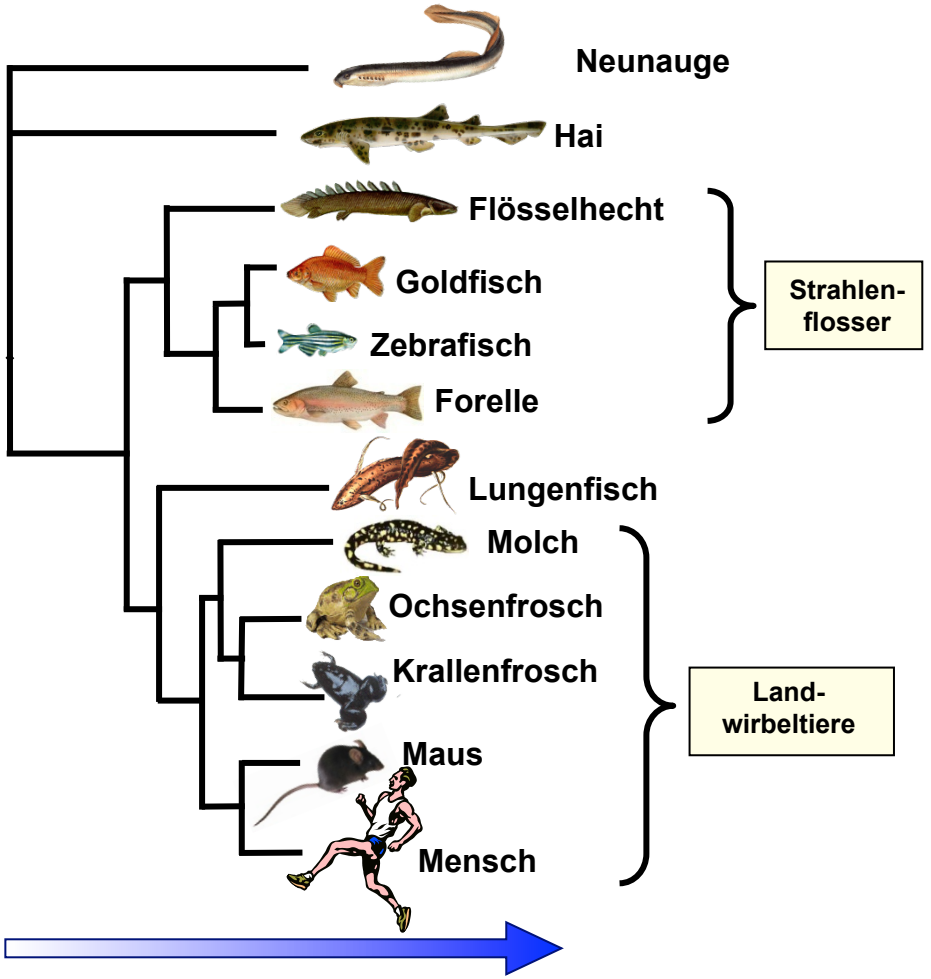
Stammbaum-Typen

Ohne Außengruppe:



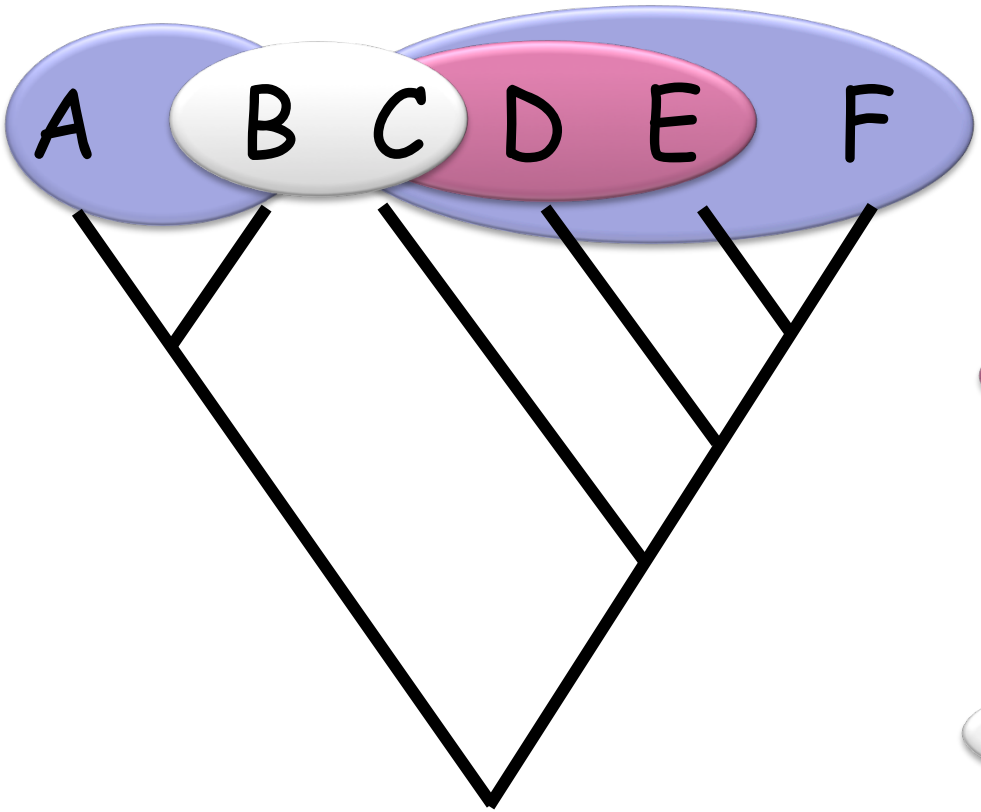
Evolutionsrichtung?

Mit Außengruppe:



Evolutionsrichtung

Mono-, Para- und Polyphylie



Monophyletische Taxa:

<u>Alle</u> Nachkommen einer gemeinsamen Stammform

Paraphyletische Taxa:

<u>Nicht</u> alle Nachkommen einer gemeinsamen Stammform

Polyphyletische Taxa:

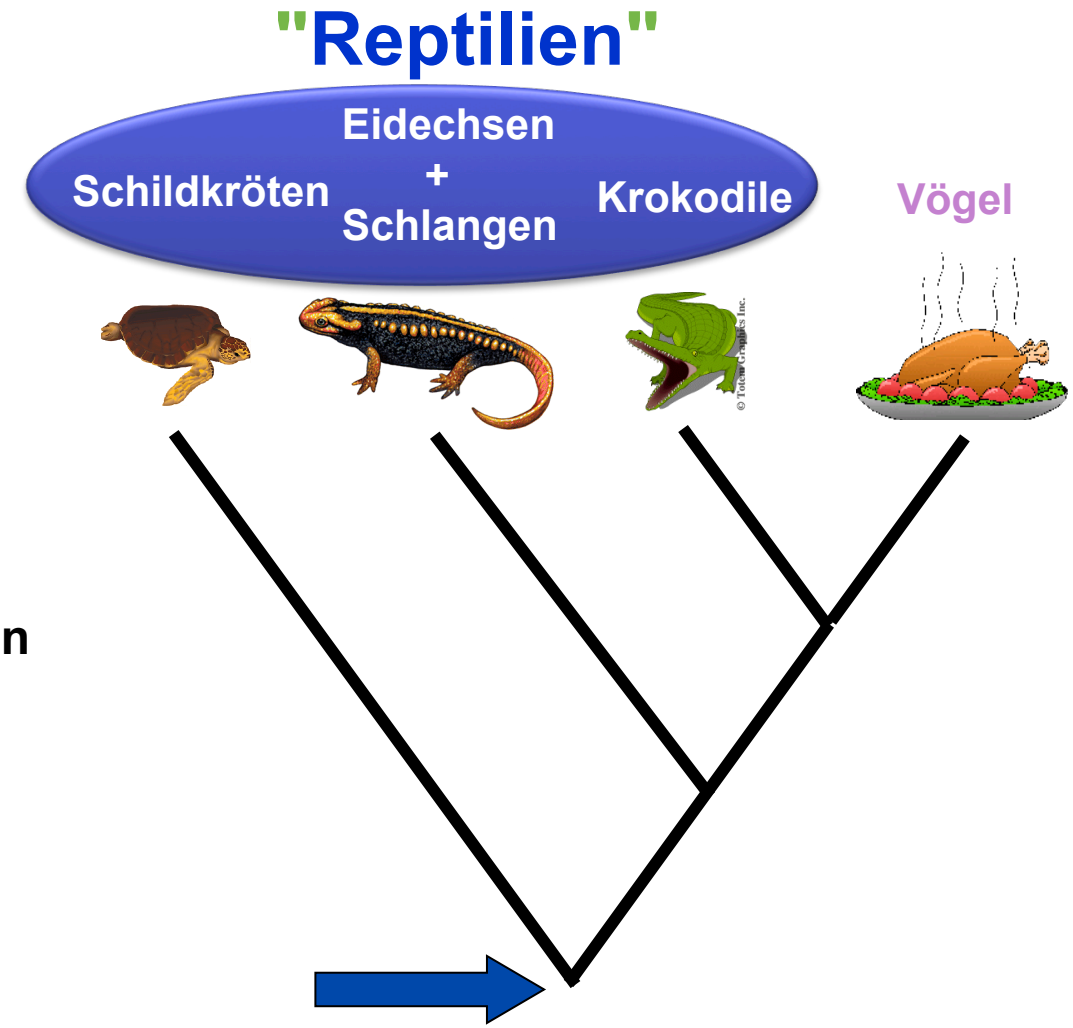
<u>Keine</u> gemeinsame Stammform (unterschiedliche Vorfahren)

Phylogenetische Grundbegriffe

Paraphylum

aufgrund von homologen (ursprünglichen) Merkmalen

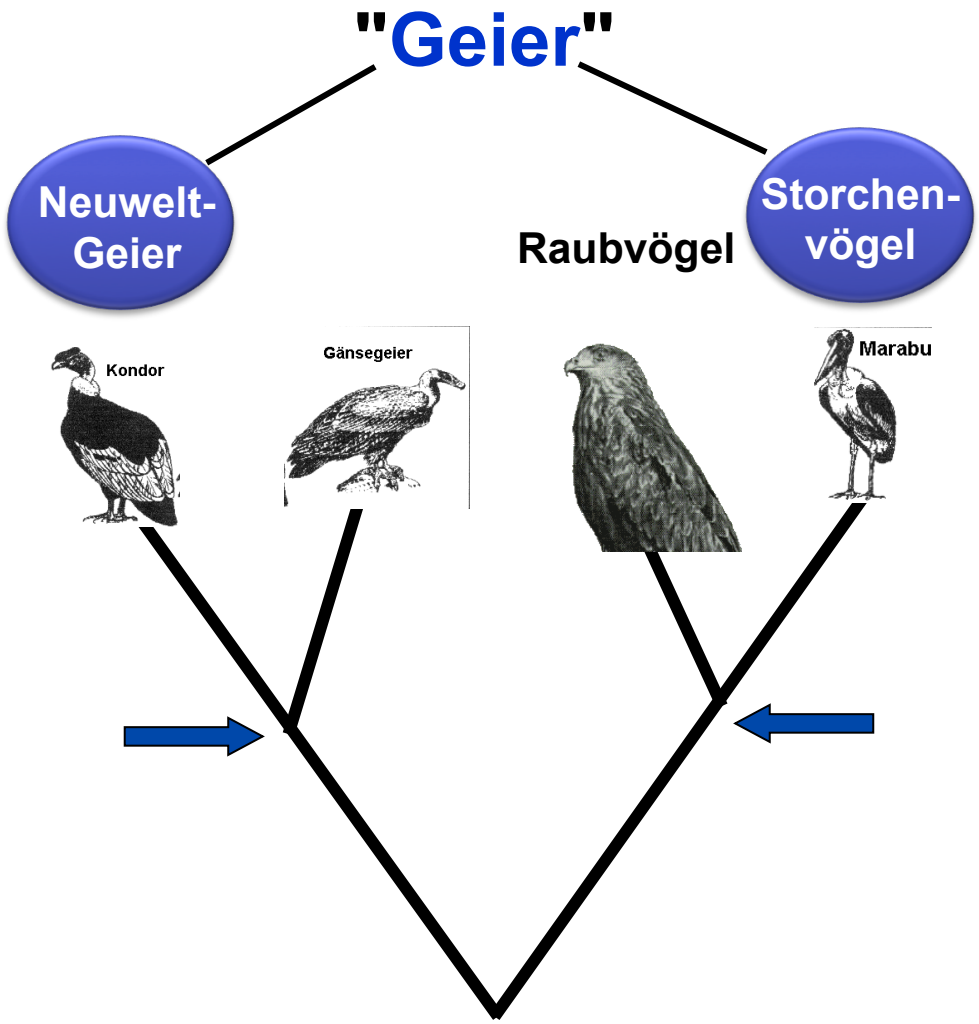
aber nicht alle Nachkommen werden erfasst!



Phylogenetische Grundbegriffe

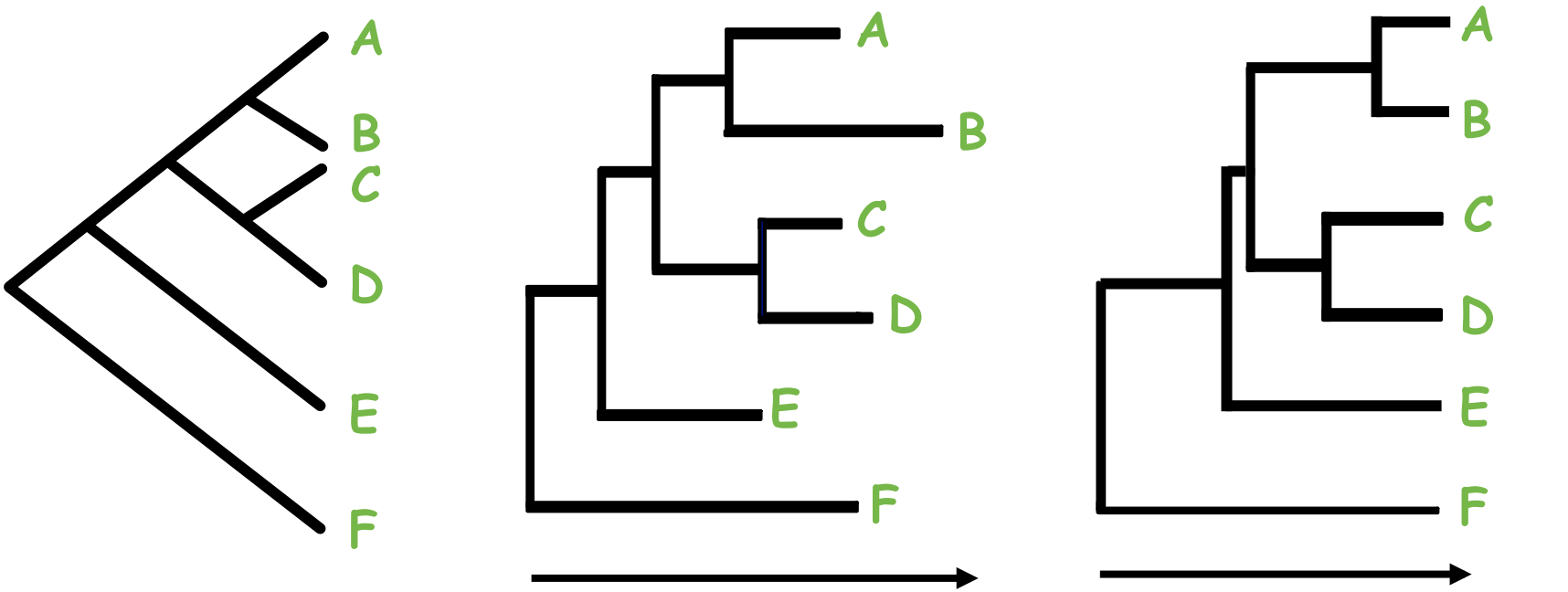
Polyphylie

=> verschiedenen Ursprungs



aufgrund von Homoplasien (Konvergenzen)

Clado-, Phylo- und Dendrogramm



Cladogramm
(Astlängen ohne Bedeutung)

Änderungen
Phylogramm
(metrisch)

Änderungen & Zeit
Dendrogramm
(ultrametrisch)

Additive Phylogramme

Newick-Format

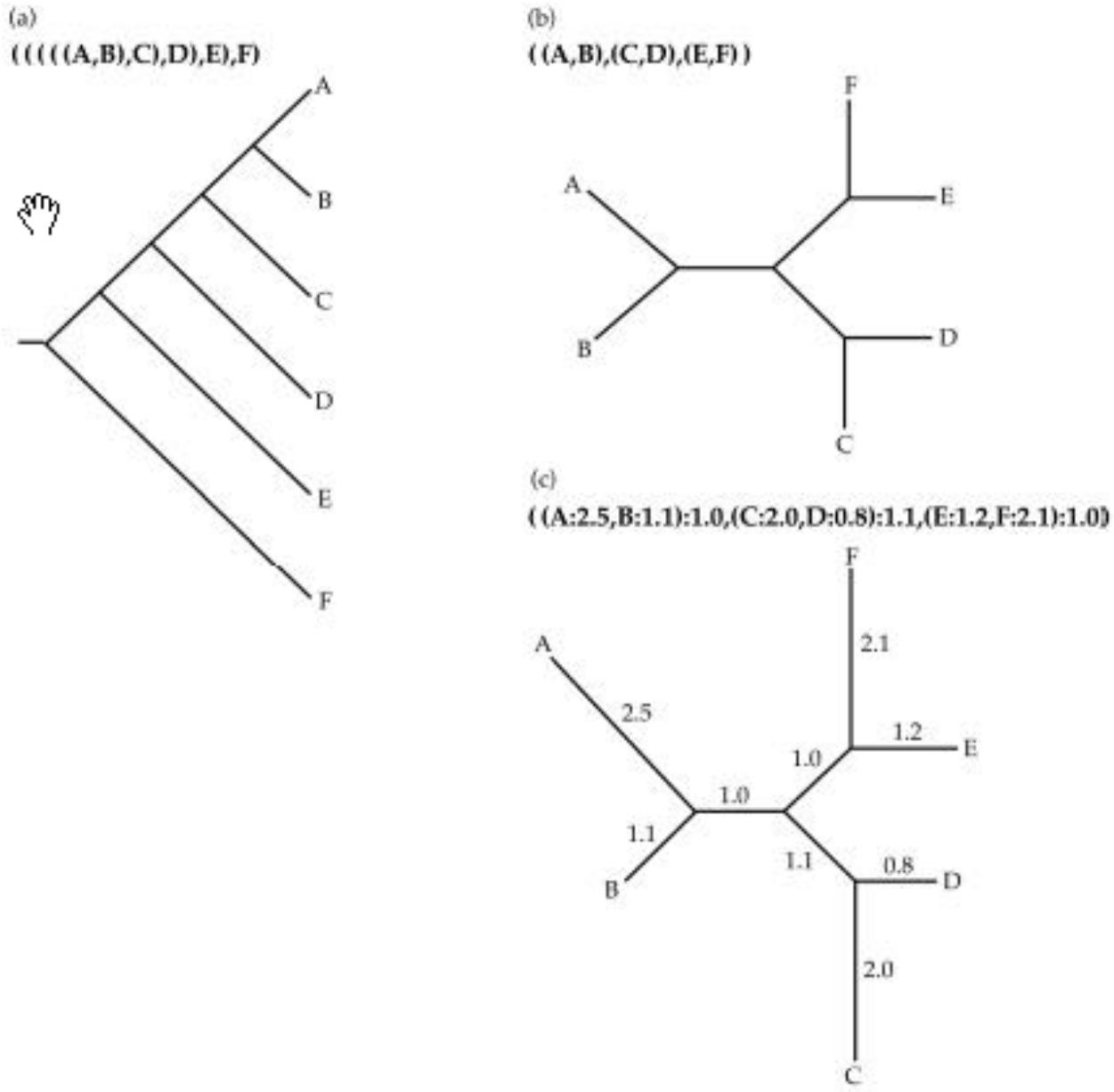
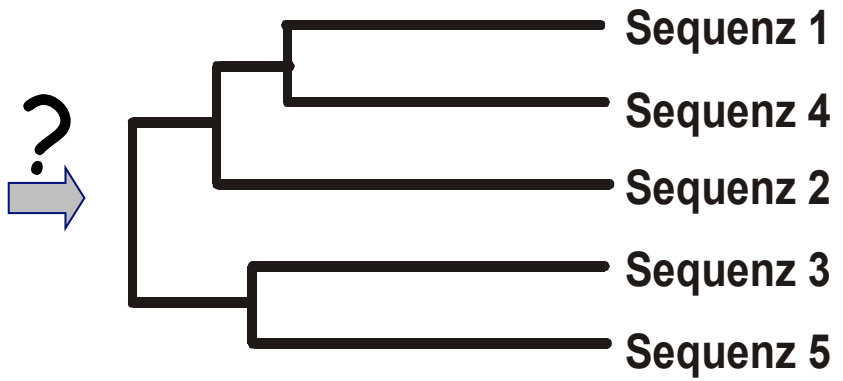


Figure 4. The Newick-format representation for rooted (a), unrooted (b), and scaled trees (c). Numbers in (c) represent branch lengths.

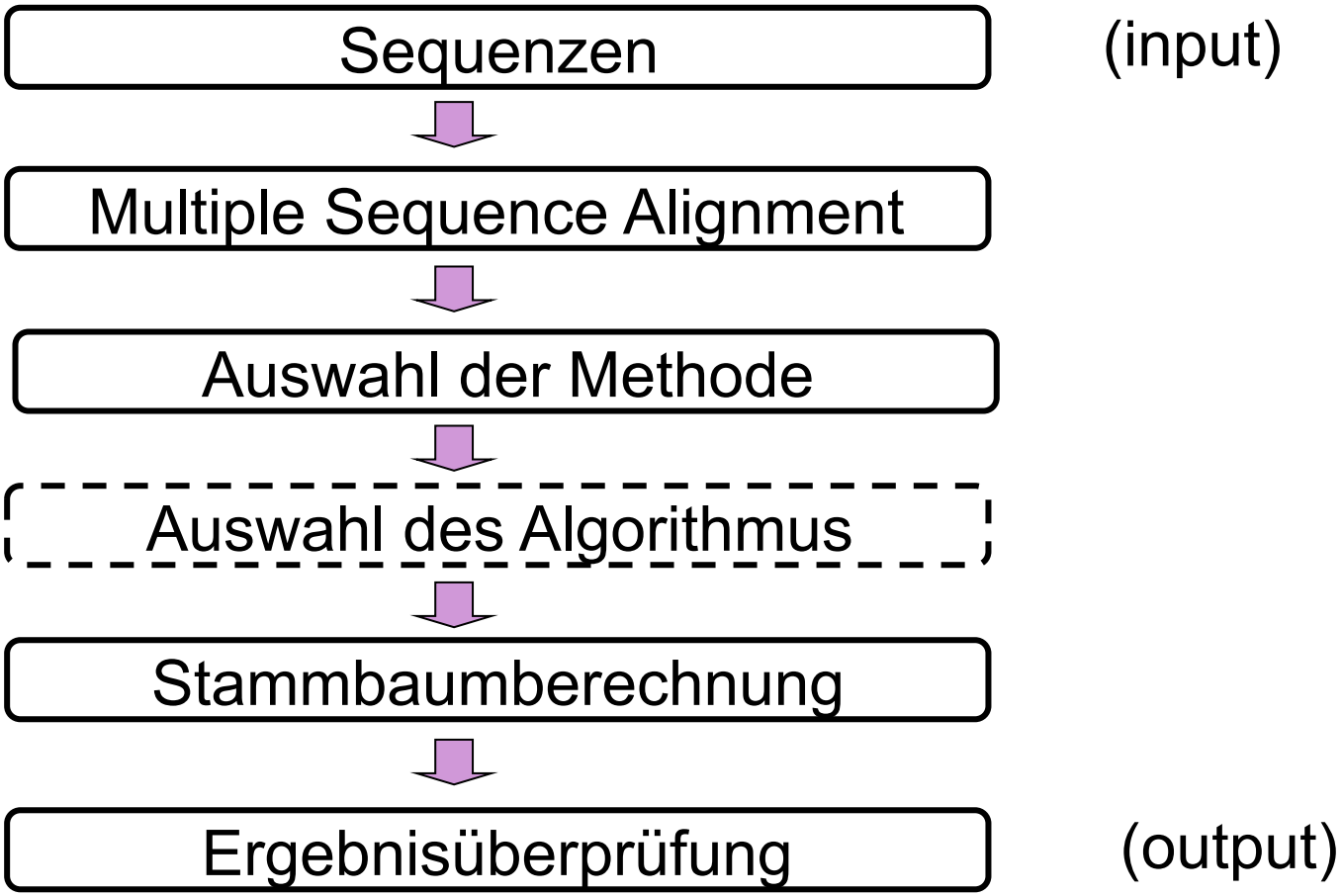
Molekulare Phylogenie

```
Sequenz 1:   KIADKNFTYRHHNQLV
Sequenz 2:   KVAEKNMTFRRFNDII
Sequenz 3:   KIADKDFTYRHW-QLV
Sequenz 4:   KVADKNFSYRHHNNVV
Sequenz 5:   KLADKQFTFRHH-QLV
```

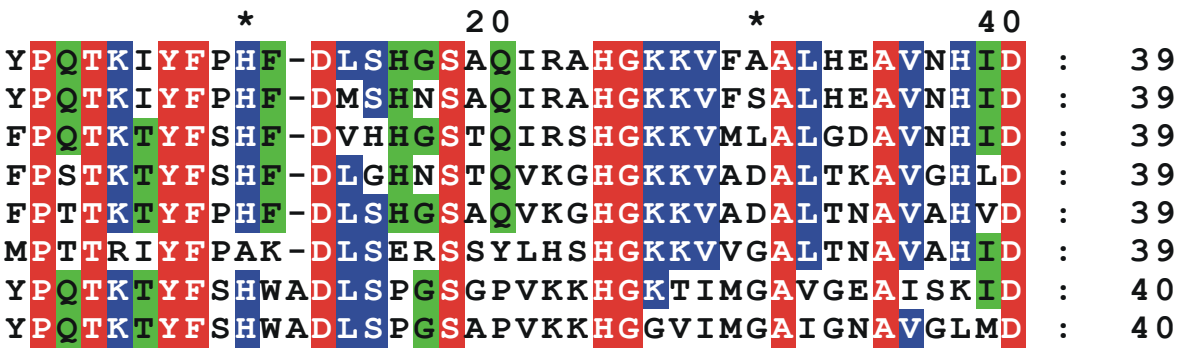


Von der Sequenz zum Baum

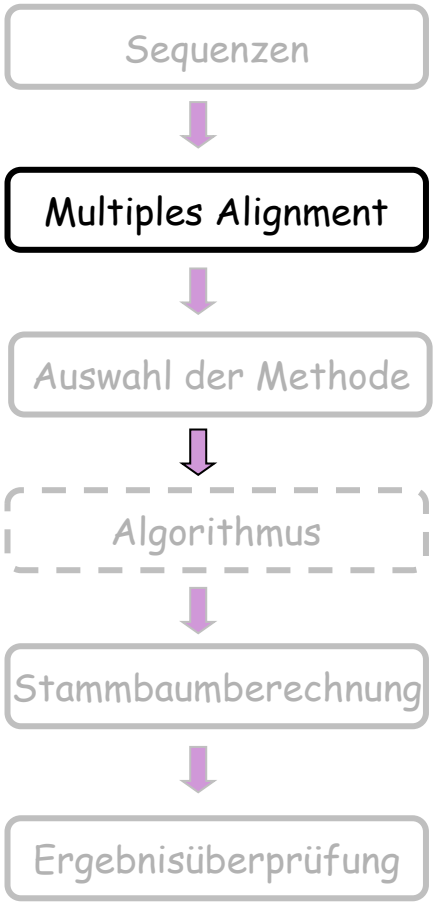
Vorgehensweise



Sequenzalignments



Gutes Alignment ist Voraussetzung für korrekten Stammbaum!



Stammbaumerstellung

Methoden der molekularen Phylogenie müssen erlauben, trotz der vielen möglichen Stammbäumen einen diskreten Stammbaum zu berechnen, der die Evolution der Sequenzen widerspiegelt.

Ab einer gewissen Anzahl von Sequenzen ist es nicht möglich, alle theoretischen Stammbäume zu berechnen!

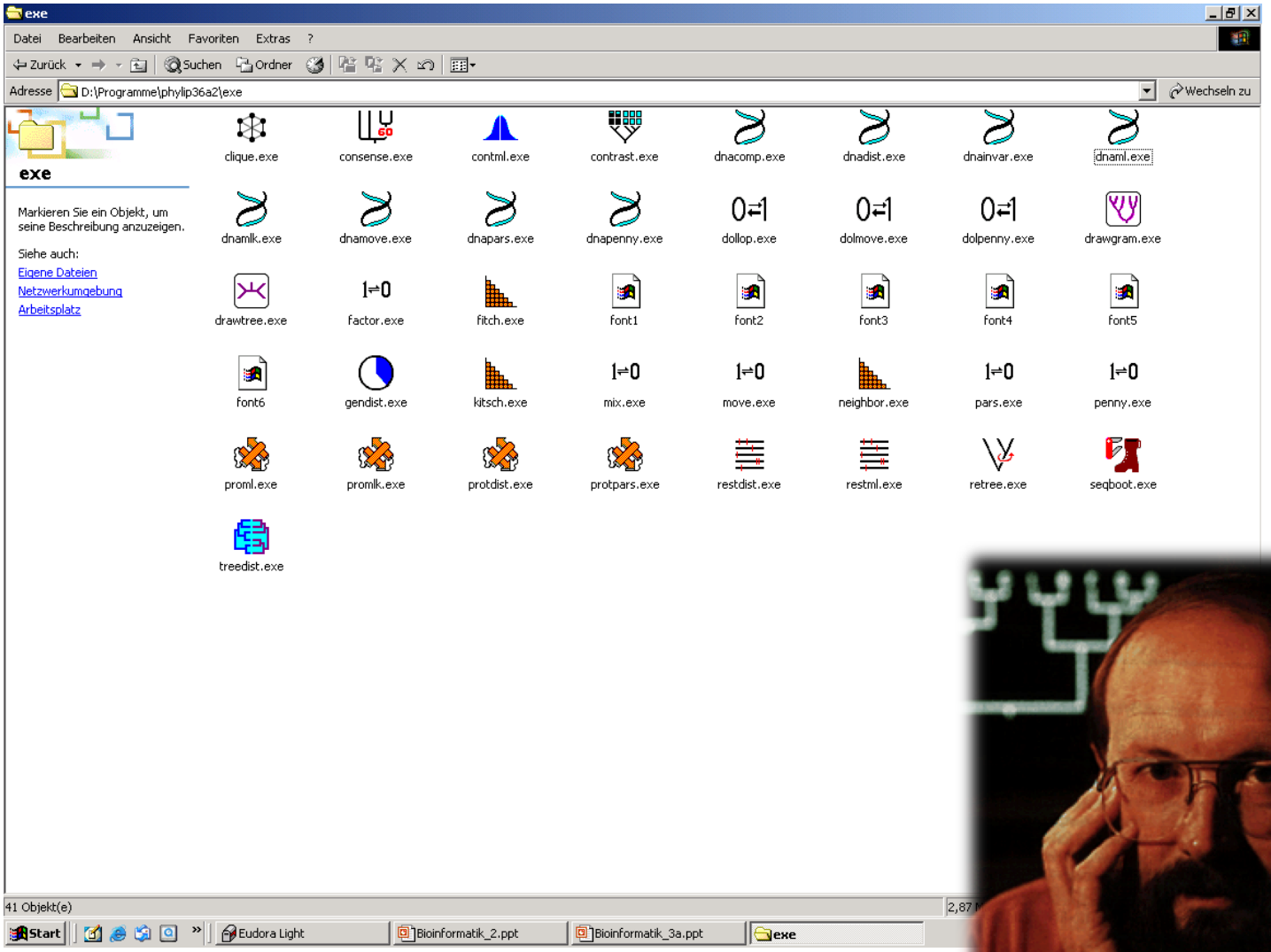
=> d.h., wir brauchen "intelligente" Algorithmen.

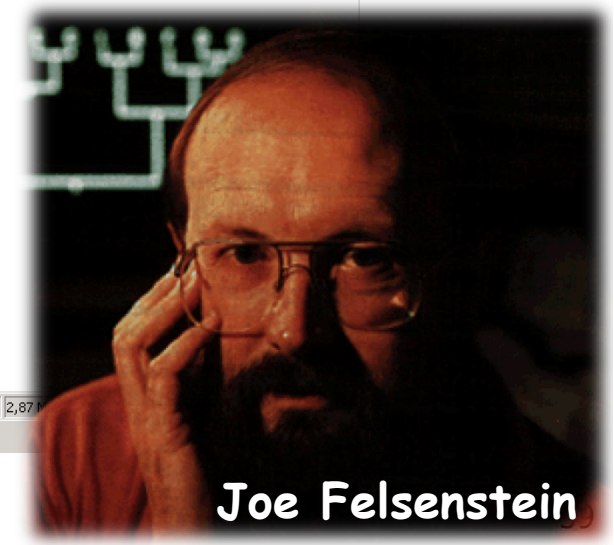
Das Problem…

Zahl der Taxa	Zahl der möglichen gewurzelten Bäume
2	1
3	3
4	15
5	105
6	954
7	10.395
8	135.135
9	2.027.025
10	34.459.425
15	213.458.046. 676.875
20	**8.200.794.532.637.891.559.375**

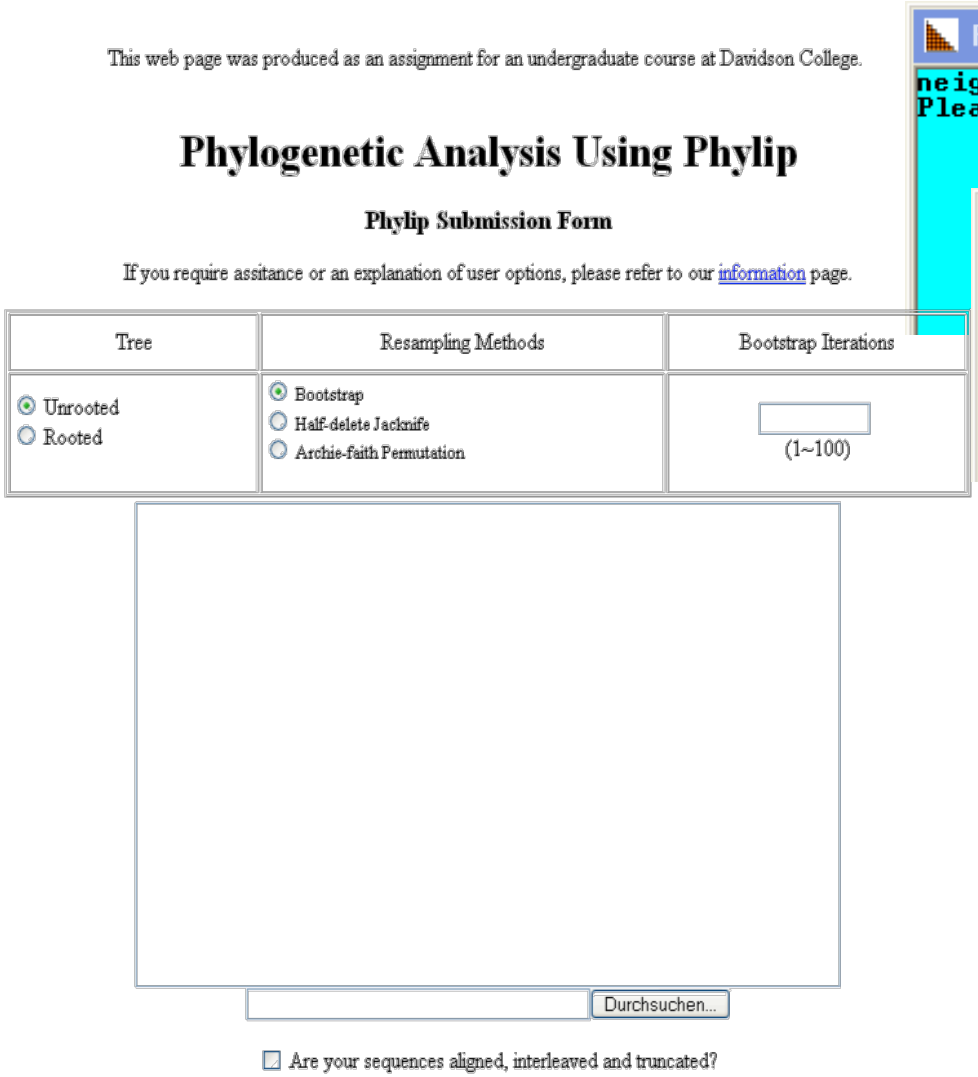
Aber: Es gibt nur einen historisch korrekten Stammbaum:

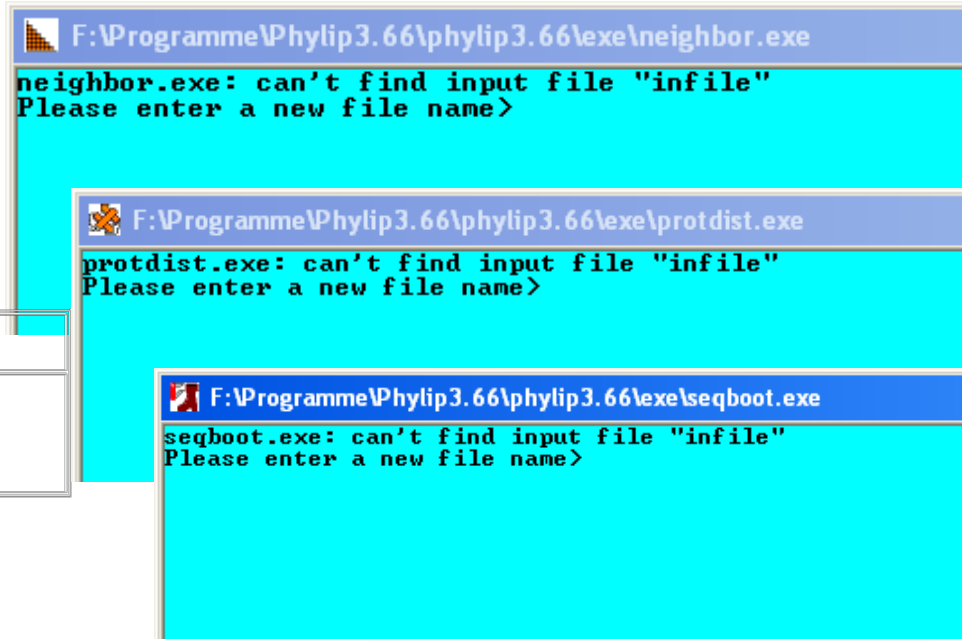
Programmpaket: PHYLIP

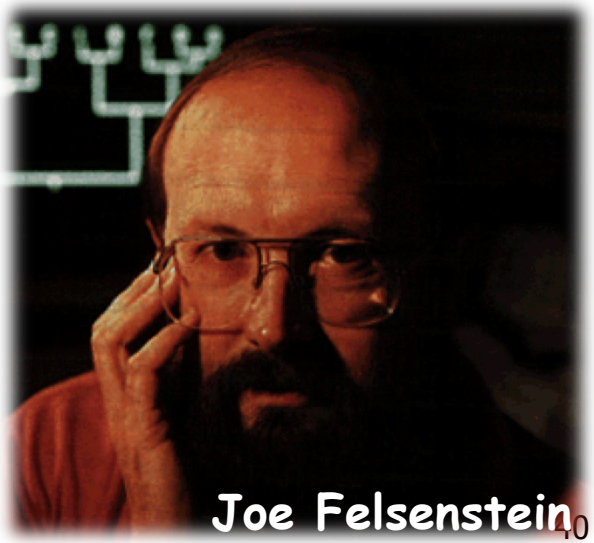




Programmpaket: PHYLIP





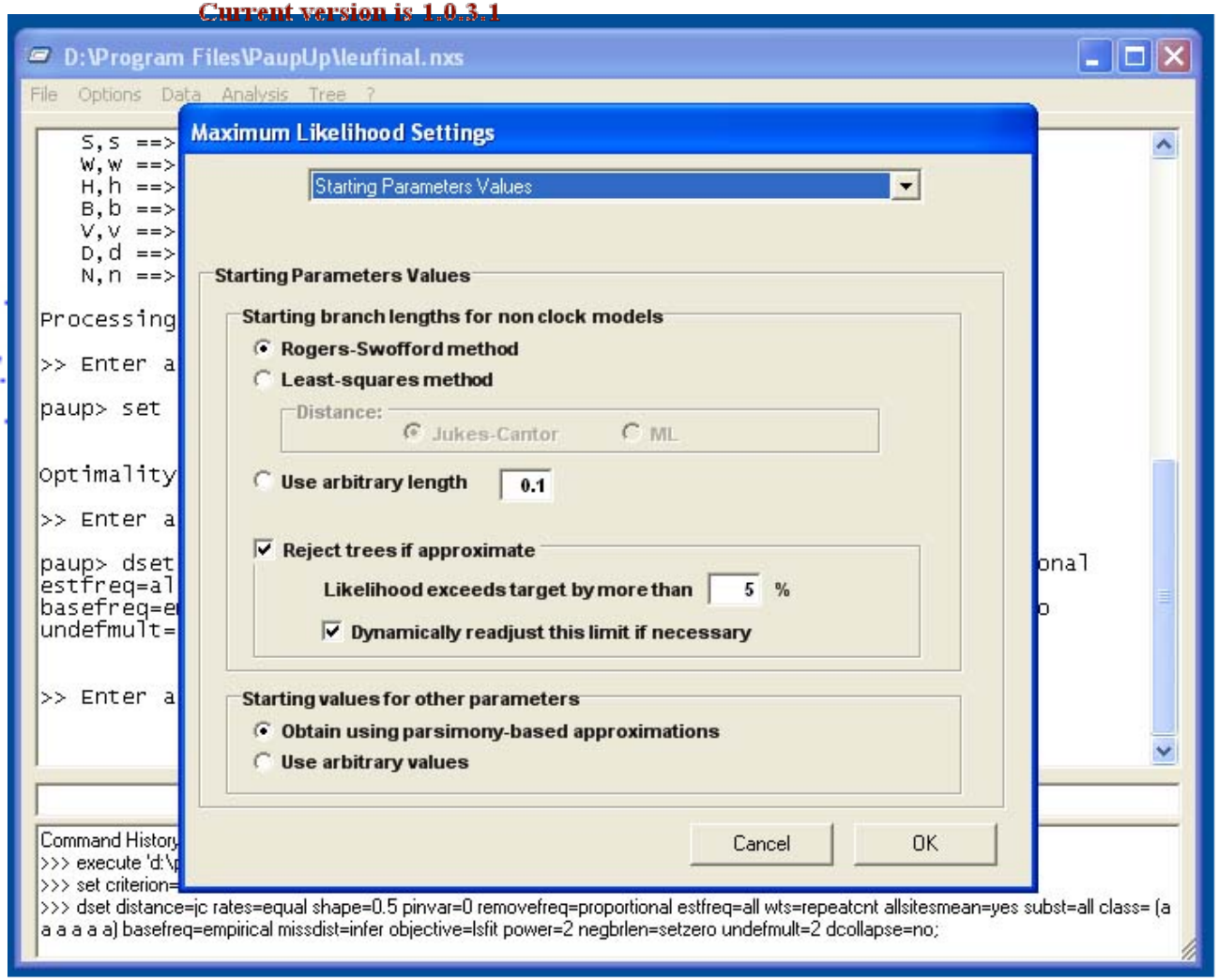


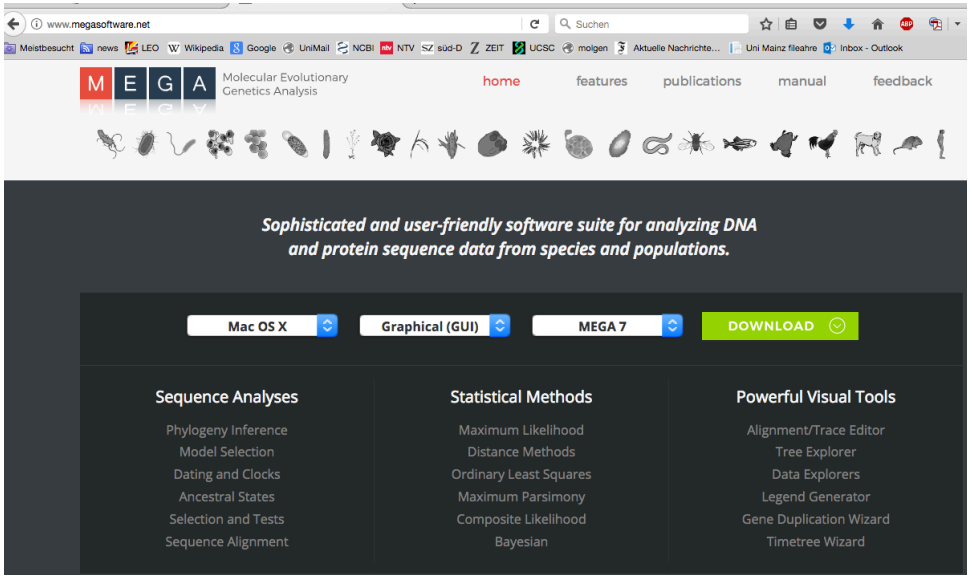
Programm: PAUP* 4.0

PaupUp: A free graphical frontend for Paup* DOS software

© 2005-2007 Frédéric Calendini & Jean-Francois Martin

Current version is 1.0.3.1





Sudhir Kumar

上機嫌で酔ってます。

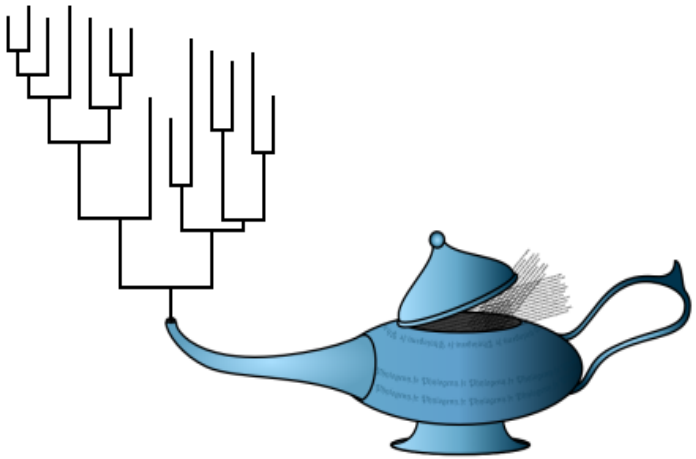
Koichiro Tamura ?

Masatoshi Nei

http://www.megasoftware.net/
Aber auch....
https://www.researchgate.net/post/Why_is_MEGA_NOT_a_good_program_for_Phylogenetics

Zum Einstieg ok…

Phylogeny.fr
Robust Phylogenetic Analysis For The Non-Specialist



Phylogeny.fr is a free, simple to use web service dedicated to reconstructing and analysing phylogenetic relationships between molecular sequences.

Phylogeny.fr runs and connects various bioinformatics programs to reconstruct a **robust phylogenetic tree from a set of sequences**.

If you use this site, please cite:

Dereeper A., Guignon V., Blanc G., Audic S., Buffet S., Chevenet F., Dufayard J.-F., Guindon S., Lefort V., Lescot M., Claverie J.-M., Gascuel O. ***Phylogeny.fr: robust phylogenetic analysis for the non-specialist*** **Nucleic Acids Research. 2008 Jul 1; 36 (Web Server Issue):W465-9. Epub 2008 Apr 19. (PubMed)**

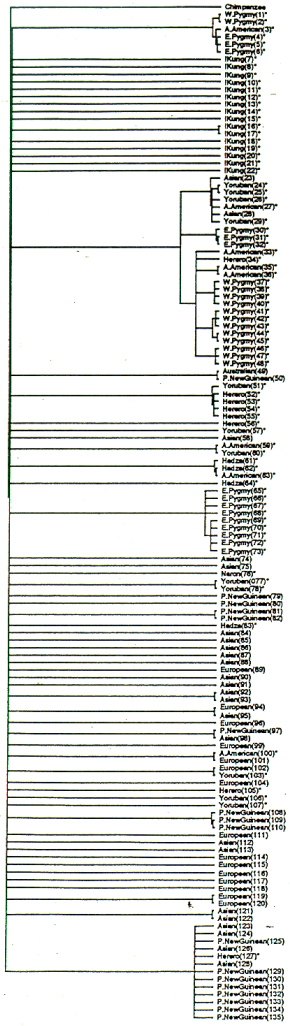
====== Mirror site ======

Nature 1988

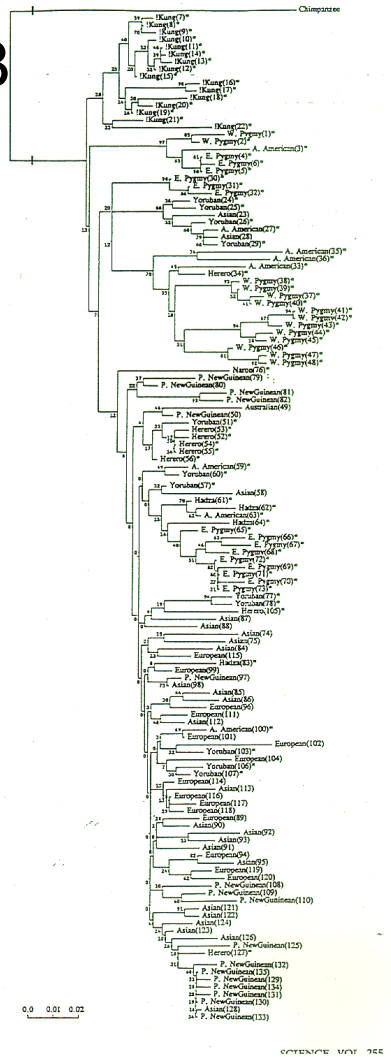
Statistical cloud over African Eden

Reconstructions of the family tree of modern people by phylogenetic analysis based on extant mitochondrial DNA appear unexpectedly difficult. An African Eden seems not yet proven.

A



B



A. Maximum-Parsimony: Konsensusbaum aus 50 000 gleich „guten“ Bäumen

B. neighbour-joining

Der ursprüngliche mt-DNA-Datensatz von Cann et al. (1987) war zu klein für eine abgesicherte Stammbaumrekonstruktion!!!!

Stammbaumerstellung

1. Distanz-orientierte Methoden

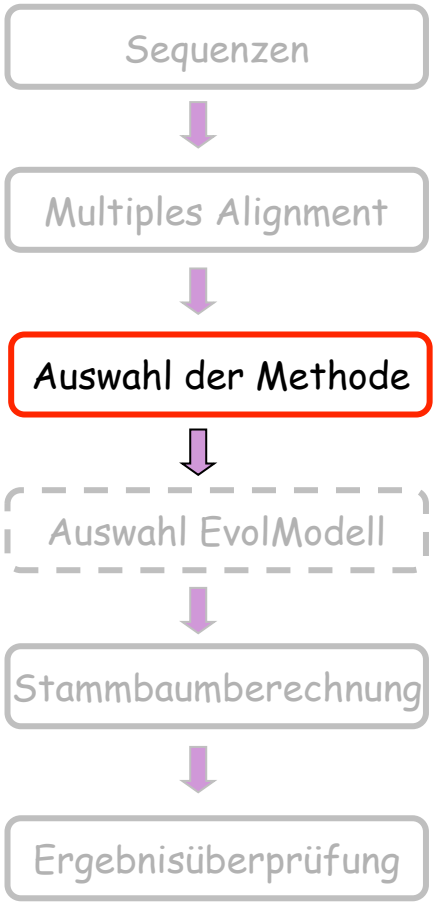
- **UPGMA (Unweighted Pair-Group Method with Arithmetric Means)**
- **Neighbor-joining**
- **Minimal Evolution**

  => Sequenzen werden in Distanzmatrix konvertiert

2. Charakter-orientierte Methoden

- **Maximum Parsimony**
- **Maximum Likelihood**
- **Bayes**

=> jede Position wird als informative Einheit betrachtet



Charakter- vs. Distanz-Methoden

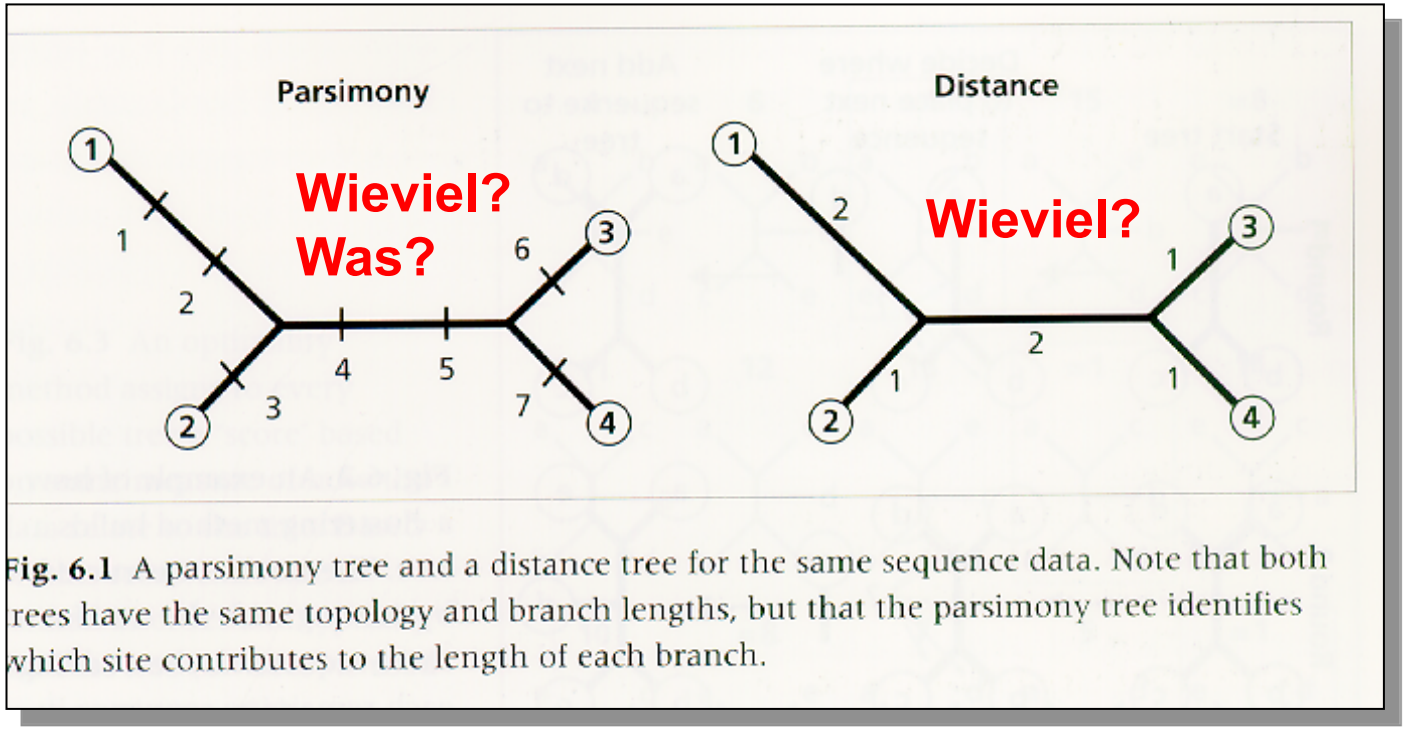


Fig. 6.1 A parsimony tree and a distance tree for the same sequence data. Note that both trees have the same topology and branch lengths, but that the parsimony tree identifies which site contributes to the length of each branch.

sequences

sites

sequences	1	2	3	4	5	6	7
1	T	T	A	T	T	A	A
2	A	A	T	T	T	A	A
3	A	A	A	A	A	T	A
4	A	A	A	A	A	A	T

distances

sequences	**1**	**2**	**3**
2	3		
3	5	4	
4	5	4	2

sequences

Berechnung einer Distanzmatrix

Sequenz 1	TATAAGCATG ACTAGTAAGC
Sequenz 2	TATTAGCATG ACTGGTAACC
Sequenz 3	TATTGGCATG ACTAGCAGGC
Sequenz 4	TGTTGCCACG ATTAGCTACC
Sequenz 5	CGTAGCTATG ACCAACGGGC

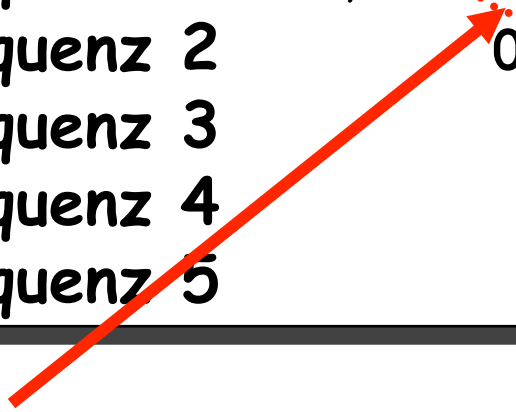
Distanz = Durchschnittliche Änderung pro Position

hier: 20 Positionen; => Wie viele Änderungen?

3 von 20 Positionen verändert => Distanz zwischen 1 und 2 = 0,15

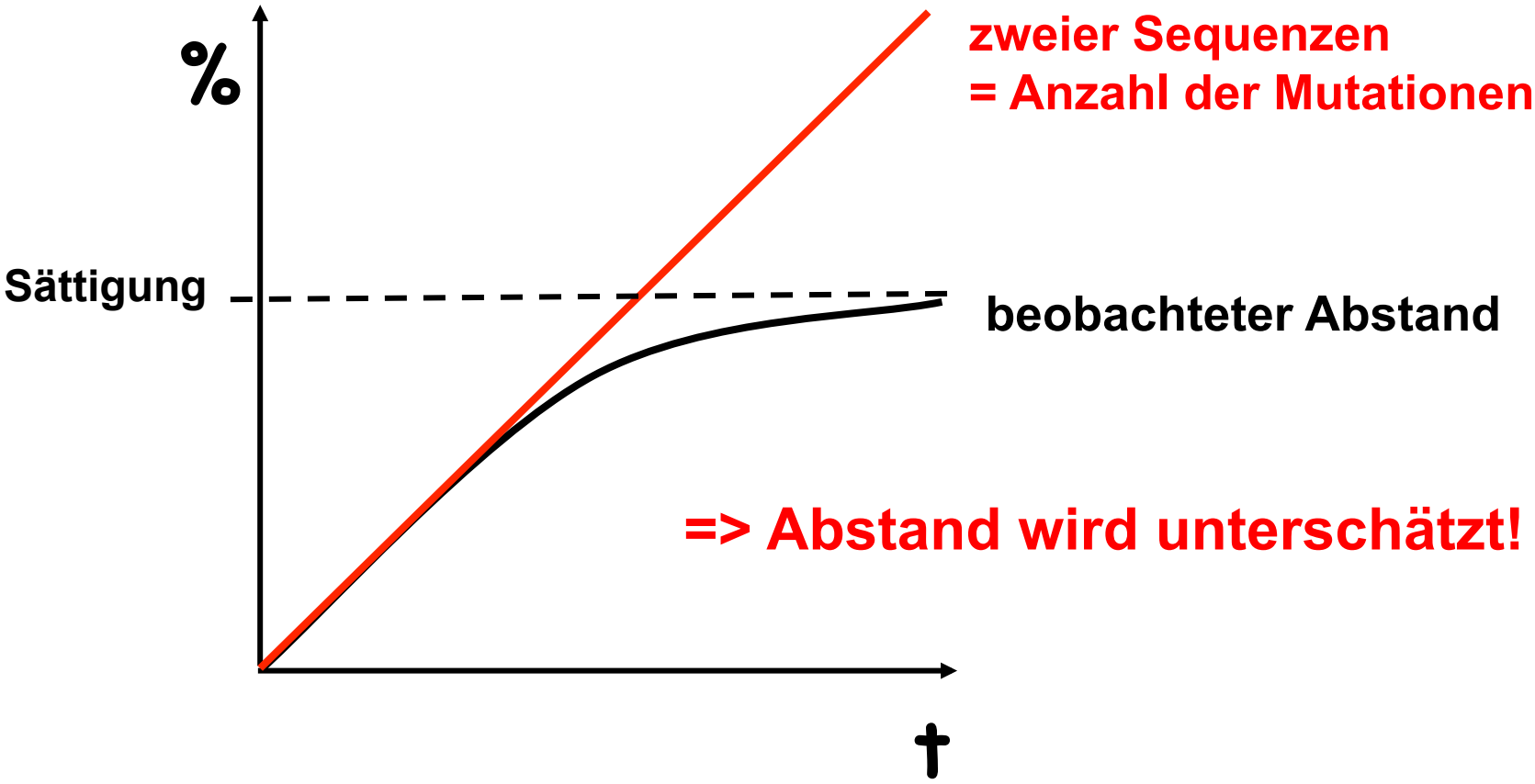
Distanzmatrix

	1	2	3	4	5
Sequenz 1	0,00	0,15	0,20	0,45	0,50
Sequenz 2		0,00	0,25	0,40	0,65
Sequenz 3			0,00	0,35	0,40
Sequenz 4				0,00	0,50
Sequenz 5					0,00



Abstand zwischen Sequenz 1 und Sequenz 2, ausgedrückt in durchschnittlichen Änderungen pro Nukleotidposition (unkorrigierte Hamming-Distanz).

Abstand gegen Zeit!

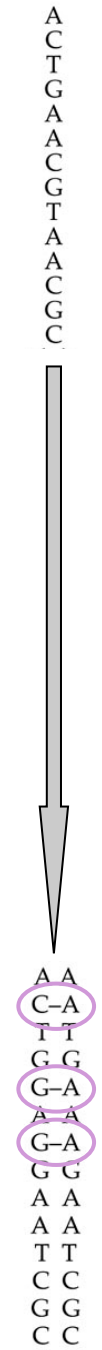


Warum?

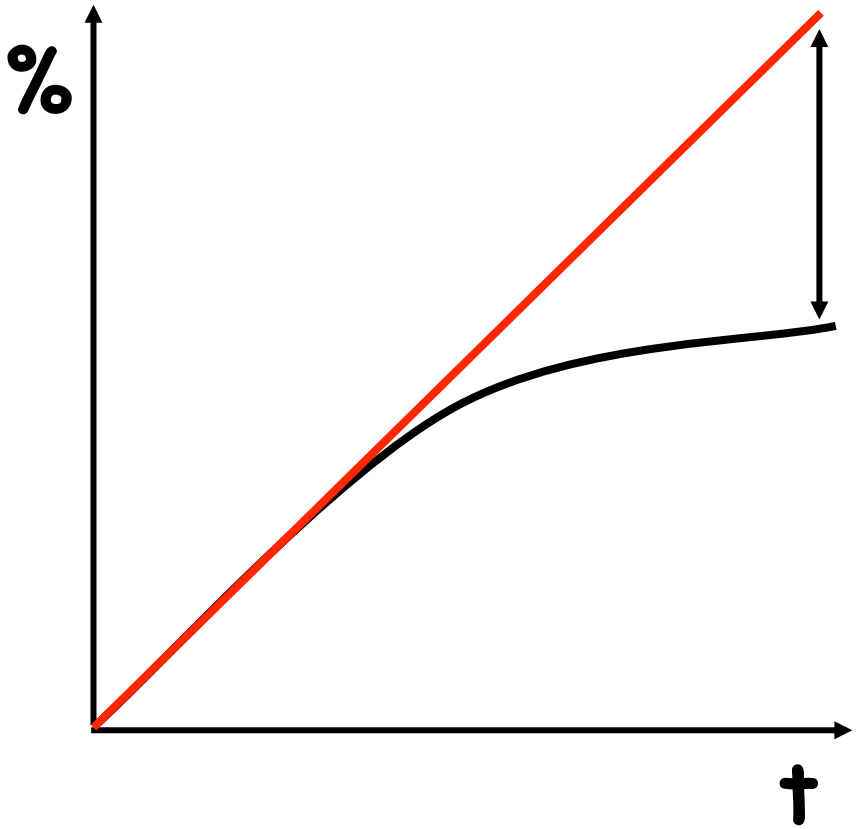
13 Mutationen

=>

3 Unterschiede



Korrektur der Distanzen



tatsächlicher Abstand
= Anzahl der Mutationen

beobachteter Abstand

Korrektur erforderlich!!

Korrektur der Distanzen

Wie korrigieren wir? Regeln?

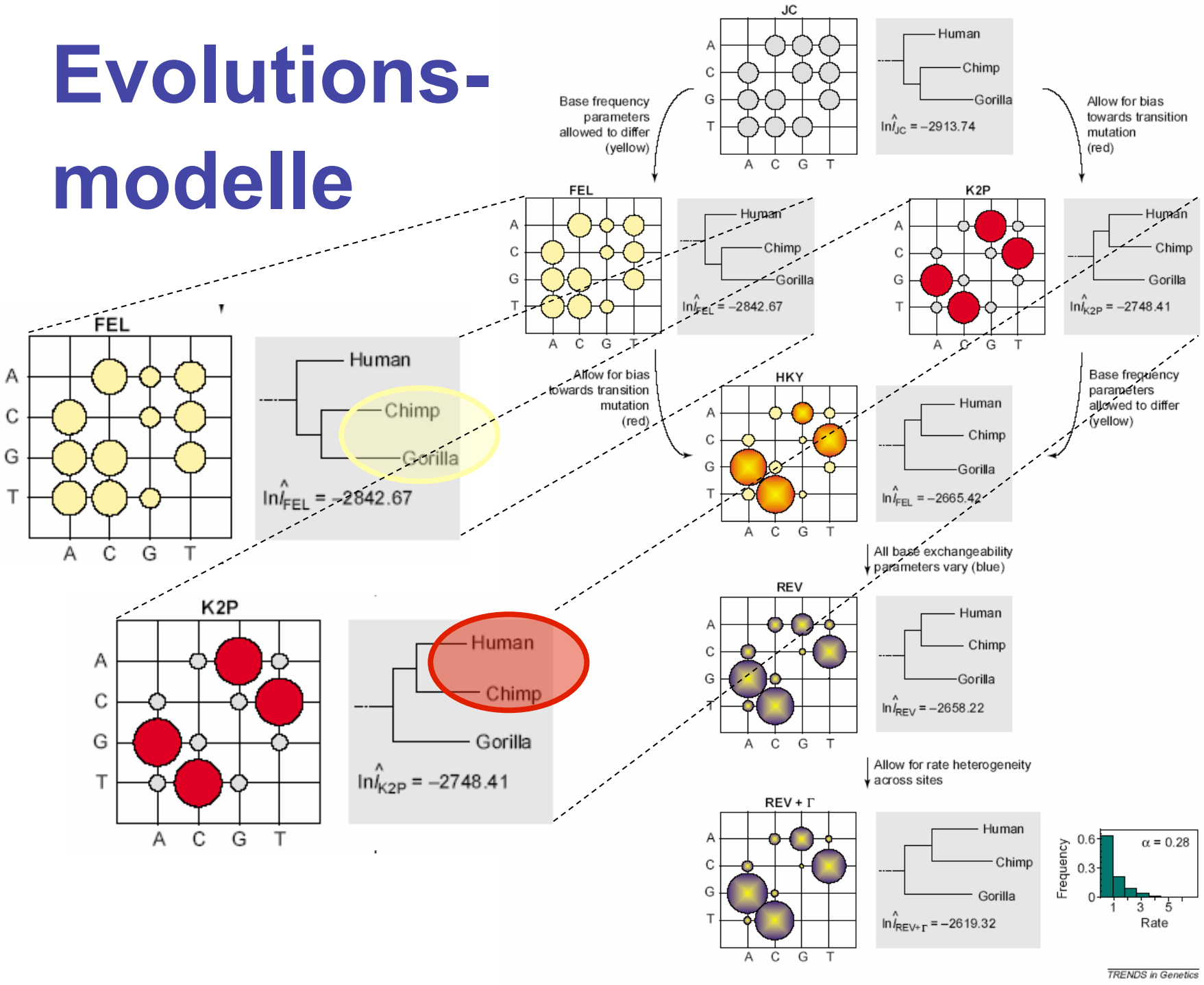
- Wir wollen die tatsächliche Anzahl der evolutiven Ereignisse rekonstruieren.

- Wir brauchen also ein Evolutionsmodell, welches die Wahrscheinlichkeit von multiplen Austauschen, Rückmutationen etc. treffend berücksichtigt.

DNA-Evolutionsmodelle

- 1969: Jukes & Cantor (JC)
- 1980: Kimura 2-Parameter (K2P)
- 1981: Felsenstein 81 (F81)
- 1985: Hasegawa, Koshino & Yano (HKY85)
- 1990: General Reversible Model (REV)
- etc…

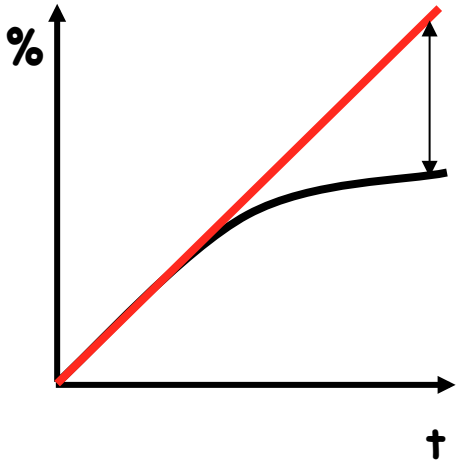
Evolutions-modelle



Evolutionsmodell Jukes & Cantor

Korrigierte Distanz nach Jukes & Cantor:

$$K = -\frac{3}{4}\ln\left(1-\frac{4}{3}p\right)$$



K ist der berechnete Abstand (Anzahl der tatsächlichen Substitutionen), *p* der beobachtete Abstand zwischen zwei Sequenzen.

JC-Modell: Anwendung

$$K = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right)$$

Beispiel:

2 Sequenzen von je 100 bp unterscheiden sich in 50 Positionen

=> p = 0,5
=> K = 0,824

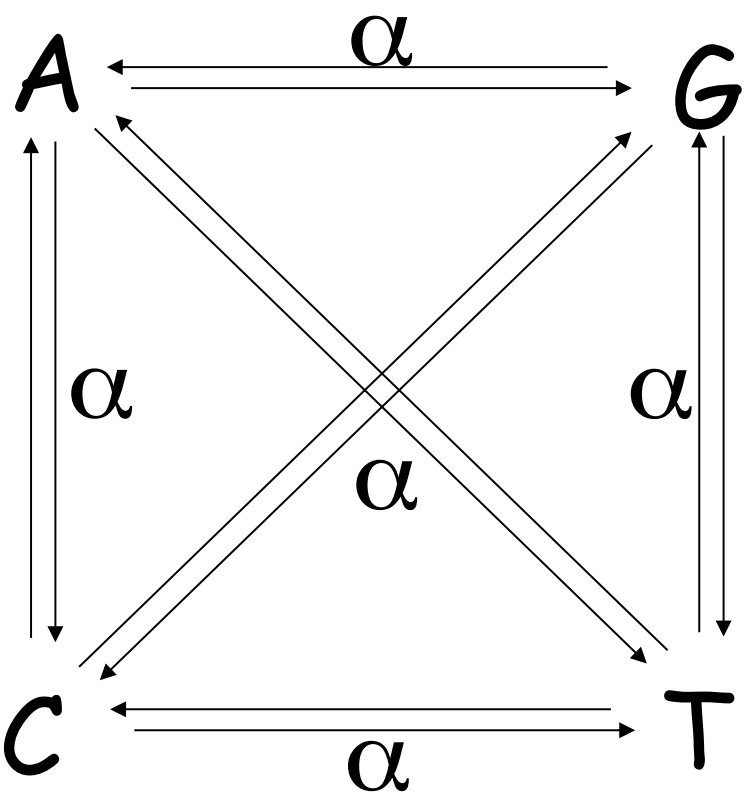
D.h., es haben tatsächlich ~ 82 Mutationen stattgefunden!

JC-Modell: wie abgeleitet?

Modell nach Jukes & Cantor (1969):

=> alle Austausche sind gleich wahrscheinlich!
= Ein-Parameter-Modell

Substitutionsrate = α für alle Änderungen



JC-Modell

Substitutions-Wahrscheinlichkeits Matrix!

$$P = \begin{pmatrix} 1-\sum P_{ij} & P_{AC} & P_{AG} & P_{AT} \\ P_{CA} & 1-\sum P_{ij} & P_{CG} & P_{CT} \\ P_{GA} & P_{GC} & 1-\sum P_{ij} & P_{GT} \\ P_{TA} & P_{TC} & P_{TG} & 1-\sum P_{ij} \end{pmatrix}$$

Gesamtwahrscheinlichkeit =1 => $P_{ii} = 1-\sum_{j \neq i} P_{ij}$

JC-Modell

	A	T	C	G
A	$1-3\alpha$	α	α	α
T	α	$1-3\alpha$	α	α
C	α	α	$1-3\alpha$	α
G	α	α	α	$1-3\alpha$

Ein-Parameter-Modell!

JC-Modell

Beispiel: Position sei "A" zum Zeitpunkt $t=0$

A → A → A

$t=0$ $t=1$ $t=2$

Somit wird klar: Die Wahrscheinlichkeit , dass diese Position zum Zeitpunkt $t=1$ immer noch A (und nicht G, C oder T) ist, beträgt:

$P_{A(1)} = 1 - 3\alpha$

Gesamtwahrscheinlichkeit

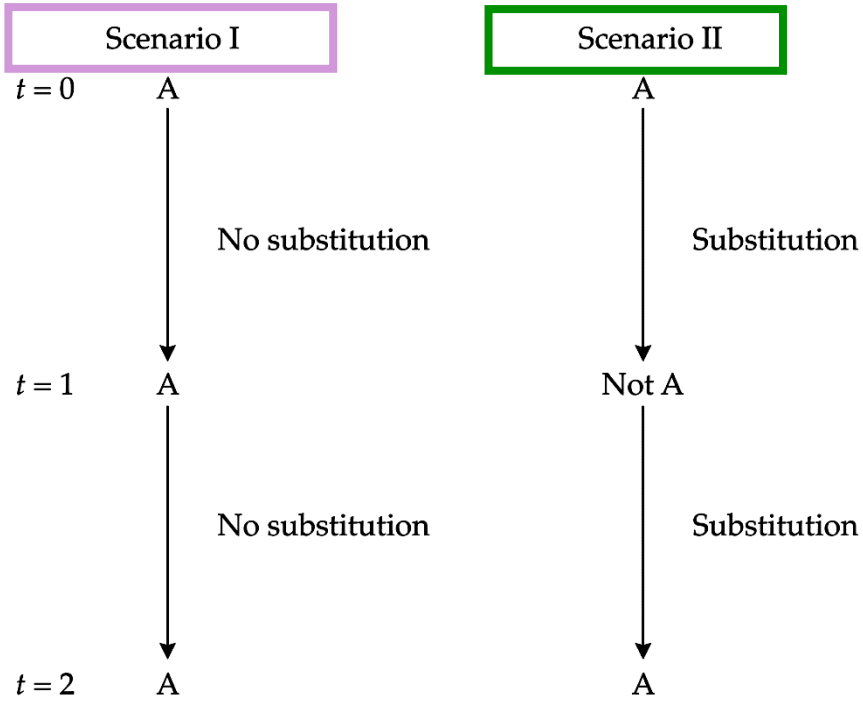
Wahrscheinlichkeit der Mutation nach G, C oder T

	A	T	C	G
A	1-3α	α	α	α
T	α	1-3α	α	α
C	α	α	1-3α	α
G	α	α	α	1-3α

=1

JC-Modell

Aber: Wie wahrscheinlich ist es, dass diese Position auch zum Zeitpunkt $t=2$ A ist?



Antwort:

$$P_{A(2)} = \boxed{(1 - 3\alpha)P_{A(1)}} + \boxed{\alpha(1 - P_{A(1)})}$$

Keine Mutation (I) Rückmutation (II)

	A	T	C	G
A	1-3α	α	α	α
T	α	1-3α	α	α
C	α	α	1-3α	α
G	α	α	α	1-3α

JC-Modell

Beispiel rechnen:

Wahrscheinlichkeit, dass, wenn betrachtetes Nukleotid zum Zeitpunkt t=0 "A" ist, bei t=1 immer noch "A" ist.

Annahme: $\alpha = 0{,}01$

Wie gross ist $P_{A(1)}$? => $P_{A(1)} = 1 - 3\alpha = 0{,}97$

Wie gross ist $P_{A(2)}$? => $P_{A(2)} = (1 - 3\alpha)P_{A(1)} + \alpha(1 - P_{A(1)}) =$

$0{,}97 \cdot 0{,}97 + 0{,}01(1 - 0{,}97) =$

$0{,}9412$

JC-Modell

Verallgemeinerung

$$P_{A(2)} = (1 - 3\alpha)P_{A(1)} + \alpha(1 - P_{A(1)})$$

verallgemeinern

$$\Rightarrow P_{A(t+1)} = (1 - 3\alpha)P_{A(t)} + \alpha(1 - P_{A(t)})$$

ausrechnen

$$\Rightarrow P_{A(t+1)} = P_{A(t)} - 4\alpha P_{A(t)} + \alpha$$

Pro Zeiteinheit:

{Bsp.: P zum Zeitpunkt 2 minus P zum Zeitpunkt 1}

$$P_{A(2)} - P_{A(1)} = -4\alpha P_{A(1)} + \alpha$$

$$\Rightarrow \underbrace{P_{A(t+1)} - P_{A(t)}} = -4\alpha P_{A(t)} + \alpha$$

$$\Rightarrow \frac{dP_{A(t)}}{dt} = -4\alpha P_{A(t)} + \alpha$$ <= Differentialgleichung 1. Ordnung

JC-Modell

$$\frac{dP_{A(t)}}{dt} = -4\alpha P_{A(t)} + \alpha$$

Lösung der Differentialgleichung 1. Ordnung:

$$P_{A(t)} = \frac{1}{4} + \left(P_{A(0)} - \frac{1}{4}\right) e^{-4\alpha t}$$

Da Ursprungssequenz A bei $t = 0$ war, gilt $P_{A(0)} = 1$

$$P_{A(t)} = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t}$$

JC-Modell

Wahrscheinlichkeit „keine Mutation“

Da bei JC gilt: $P_{AA(t)} = P_{GG(t)} = P_{TT(t)} = P_{CC(t)}$,
kann die Wahrscheinlichkeit $P_{ii(t)}$, dass ein Nukleotid i nach t Zeiteinheiten i bleibt beschrieben werden als:

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t}$$

JC-Modell

Wahrscheinlichkeit einer Mutation

Da auch gilt: $P_{AT(t)} = P_{AC(t)} = P_{AG(t)}$ ..., ist die Wahrscheinlichkeit $P_{ij(t)}$, dass ein Nukleotid zum Zeitpunkt t nicht i sondern j ist, gegeben i bei t =0:

$$P_{ij(t)} = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t}$$

$$P_{A(t)} = \frac{1}{4} + \left(0 - \frac{1}{4}\right) e^{-4\alpha t}$$

Annahme: $P_{A(0)} = 0$ (Ursprungssequenz bei $t = 0$)

Zum Vergleich nochmal der Fall „keine Mutation“:

$$P_{ii(t)} = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t}$$

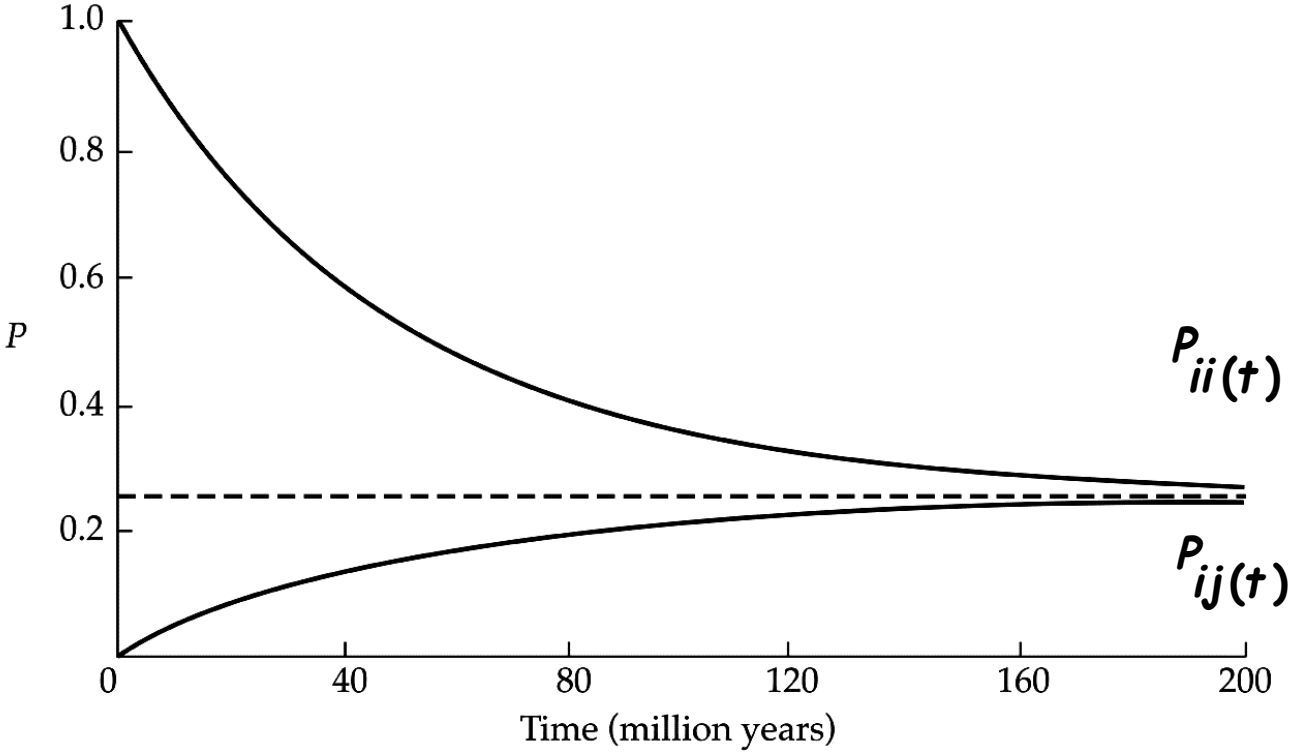
$$P_{A(t)} = \frac{1}{4} + \left(1 - \frac{1}{4}\right) e^{-4\alpha t}$$

Annahme: $P_{A(0)} = 1$ (Ursprungssequenz bei $t = 0$)

JC-Modell

Mit den beiden Formeln der vorangehenden Folie kann man den gesamten Evolutionsprozess nach dem JC-Modell beschreiben.

Im JC-Modell ist nach $t = \infty$ die Wahrscheinlichkeit jedes der Nukleotide an einer betrachteten Position 25%.



$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t} \rightarrow 0$$

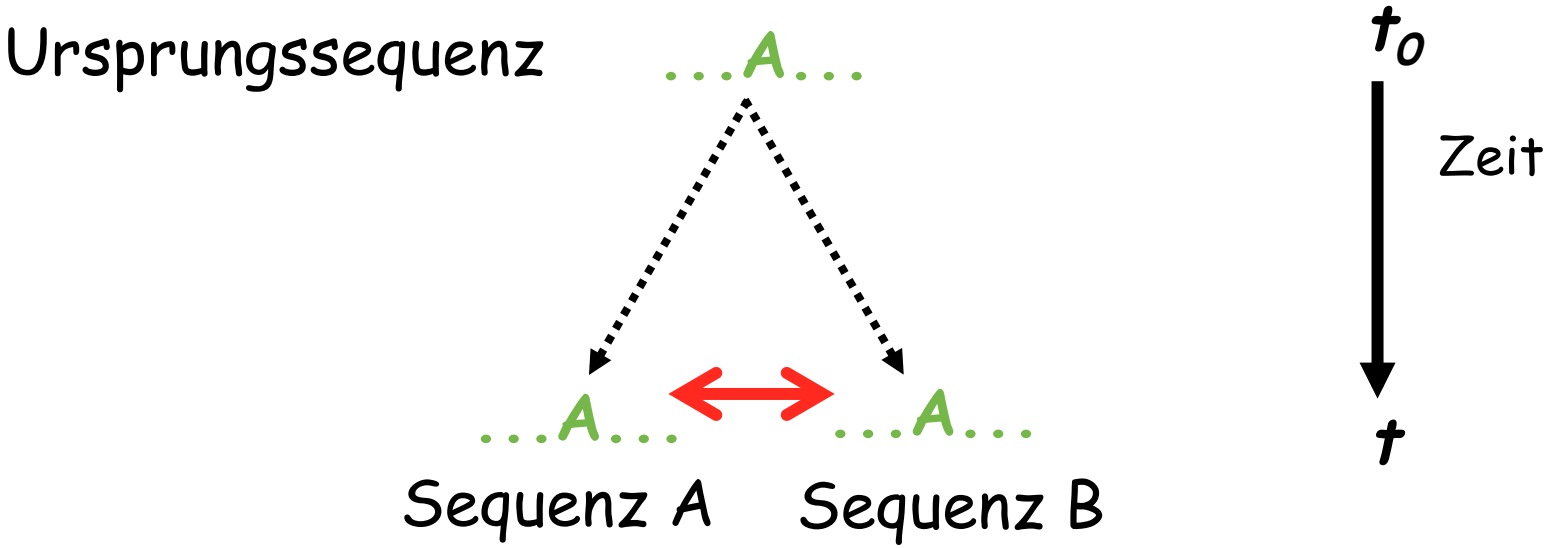
$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} \rightarrow 0$$

(es wurde hier die typische Mut-Rate bei Vertebraten, α = 5 x 10-9, eingesetzt)

JC-Modell

Wie kommen wir vom Evolutionsmodell zur Berechnung der tatsächlichen Distanz?

Bisher: Betrachtung der Evolution einer Sequenz;
Wie sieht es beim Vergleich von 2 Sequenzen aus?



JC-Modell

D.h., wie hoch ist die Wahrscheinlichkeit *I*, dass nach Zeit *t* eine Nukleotidposition in beiden Sequenzen jeweils *A* ist?

$$I_{AA(t)} = P_{AA(t)}^{2} + P_{TA(t)}^{2} + P_{GA(t)}^{2} + P_{CA(t)}^{2}$$

in beiden Seq. "A" geblieben

in beiden Seq. von "T" nach "A" mutiert

JC-Modell

Verallgemeinerung:

$$I_{ii(t)} = \underbrace{\left(\frac{1}{4} + \frac{3}{4}e^{-4\alpha t}\right)^2}_{A->A} + \underbrace{\left(\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right)^2}_{G>A} + \underbrace{\left(\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right)^2}_{T>A} + \underbrace{\left(\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right)^2}_{C>A}$$

Lösung:

$$I_{(t)} = \frac{1}{4} + \frac{3}{4}e^{-8\alpha t}$$

Wobei *I* der Anteil identischer Nukleotide zweier Sequenzen ist, die sich vor *t* Zeiteinheiten getrennt haben.

JC-Modell

$$I_{(t)} = \frac{1}{4} + \frac{3}{4} e^{-8\alpha t}$$

Die Wahrscheinlichkeit identischer Nukleotide zweier Sequenzen, die sich vor t Zeiteinheiten getrennt haben bzw. Anteil ident. Nukleotide zweier Sequenzen.

Wie gross ist jetzt aber nun der Anteil **unterschiedlicher** Nukleotide nach t Zeiteinheiten (Distanz!)?

„Ganz einfach“: $p = 1 - I_{(t)}$. Daraus folgt:

$$p = \frac{3}{4}\left(1 - e^{-8\alpha t}\right)$$

Die Wahrscheinlichkeit unterschiedlicher Nukleotide zweier Sequenzen, die sich vor t Zeiteinheiten getrennt haben bzw.
Anteil unterschiedlicher Nukleotide zweier Sequenzen.

JC-Modell

$$p = \frac{3}{4}\left(1 - e^{-8\alpha t}\right)$$

Wir kennen in der Praxis weder α noch t ! Jedoch können wir p, die beobachtete Distanz zwischen den beiden Sequenzen messen.

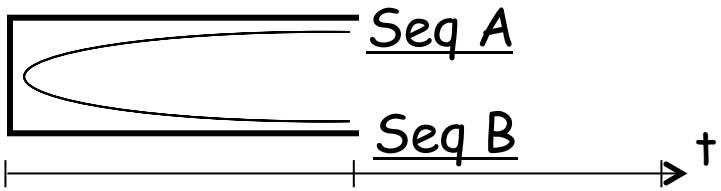
Doch wir wollen: die für multiple Austausche korrigierte Distanz K !!!

JC-Modell

Wir wissen aber auch: Für den korrigierten Abstand K zwischen zwei Sequenzen gilt nach Jukes & Cantor :

$$K = 2t \cdot 3\alpha$$

$2t$: 2 x t, da für jede der zwei Sequenzen die Zeit t vergangen ist



3α: Jedes der Nukleotide kann in drei andere umgewandelt werden mit Substitutionsrate α.

	A	T	C	G
A	1-3α	α	α	α
T	α	1-3α	α	α
C	α	α	1-3α	α
G	α	α	α	1-3α

JC-Modell

Durch mathematisches Umformen beider Gleichungen nach

$$\alpha t$$

Können wir nun eine Beziehung zwischen der gemessenen Distanz p und der gewünschten korrigierten Distanz K herstellen…

JC-Modell

$K = 2t \cdot 3\alpha \Rightarrow K = 6\alpha t$; einsetzen in:

$$p = \frac{3}{4}\left(1 - e^{-8\alpha t}\right) \Rightarrow 8\alpha t = -\ln\left(1 - \frac{4}{3}p\right)$$

$$\alpha t = \frac{-\ln\left(1 - \frac{4}{3}p\right)}{8} \qquad \alpha t = \frac{K}{6} \Rightarrow$$

$$\frac{K}{6} = \frac{-\ln\left(1 - \frac{4}{3}p\right)}{8} \Rightarrow K = \frac{-\ln\left(1 - \frac{4}{3}p\right)6}{8} = -\ln\left(1 - \frac{4}{3}p\right)\frac{6}{8} =$$

$$K = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right)$$

K ist der <u>berechnete</u> Abstand (Anzahl der <u>tatsächlichen</u> Substitutionen), p der <u>beobachtete</u> Abstand zwischen zwei Sequenzen.

JC-Modell

$$K = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right)$$

Beispiel

2 Sequenzen von je 100 bp unterscheiden sich in 50 Positionen

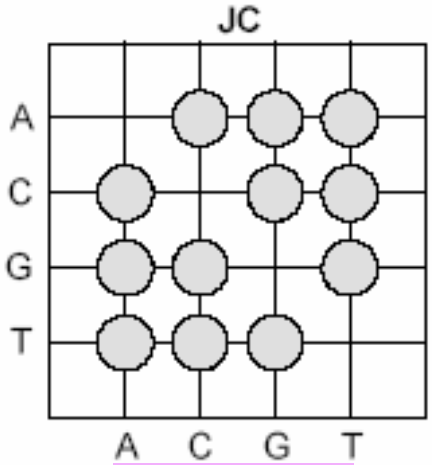
=> p = 0,5
=> K = 0,824

D. h., es haben tatsächlich ~ 82 Mutationen stattgefunden!

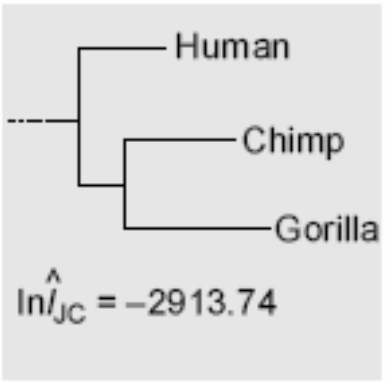
> Korrektur hat erheblichen Einfluss auf Distanzmatrix und (später) den phylogenetischen Baum!!!

Aber:

- Modell ist zu einfach!
- Denn jeder Basenaustausch wird gleich bewertet.
- In der Natur aber nicht so beobachtet.
- In der Praxis sind meist bessere Modelle notwendig.



garbage in

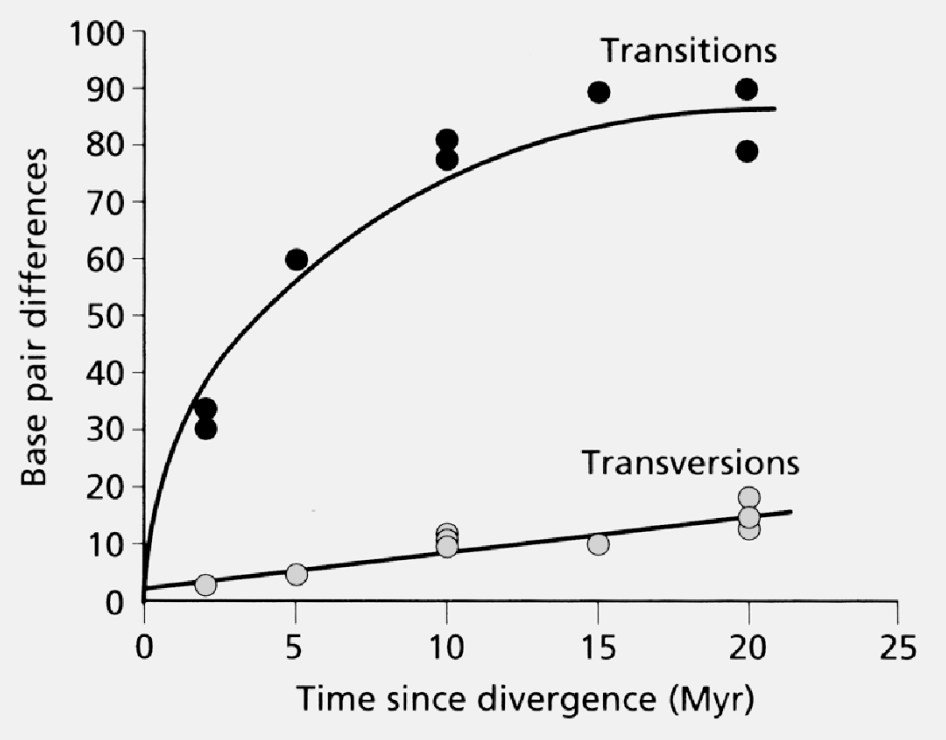


garbage out

DNA-Evolutionsmodelle

- 1969: Jukes & Cantor (JC)
- 1980: Kimura 2-Parameter (K2P)
- 1981: Felsenstein 81 (F81)
- 1985: Hasegawa, Koshino & Yano (HKY85)
- 1990: General Reversible Model (REV)
- uvm…

Transitionen vs. Transversionen



Vergleich der mtDNA verschiedener Huftiere:

In der mtDNA ist der Unterschied zwischen Transitionen und Transversionen besonders krass!

⇒ Transitionen (A <=> G, C <=> T) sind real wesentlich häufiger als Transversionen (A,G <=> C,T)

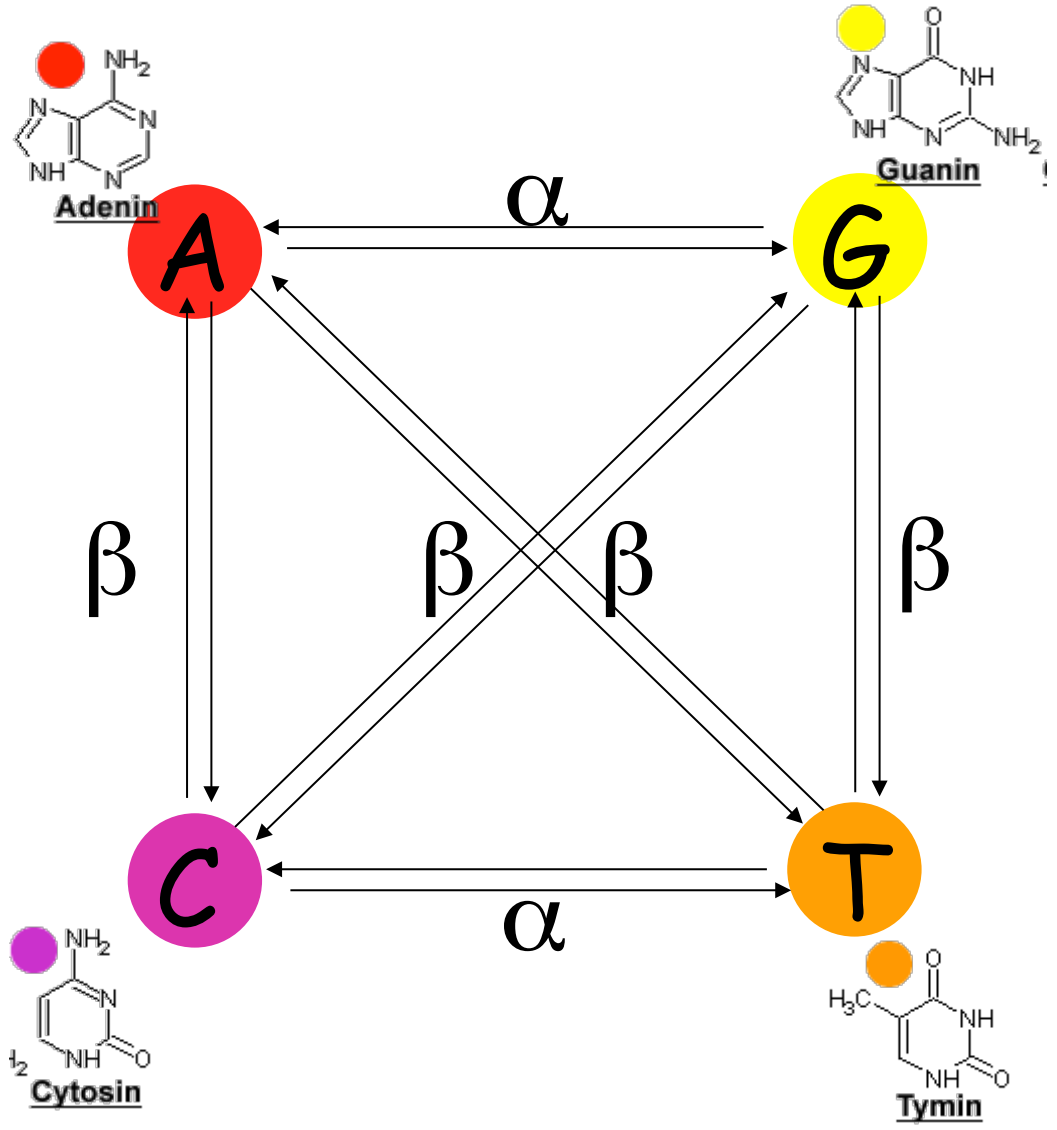
also...

Kimura 2-Parameter-Modell

$$P_{AA(1)} = 1 - \alpha - 2\beta$$

Transition: α

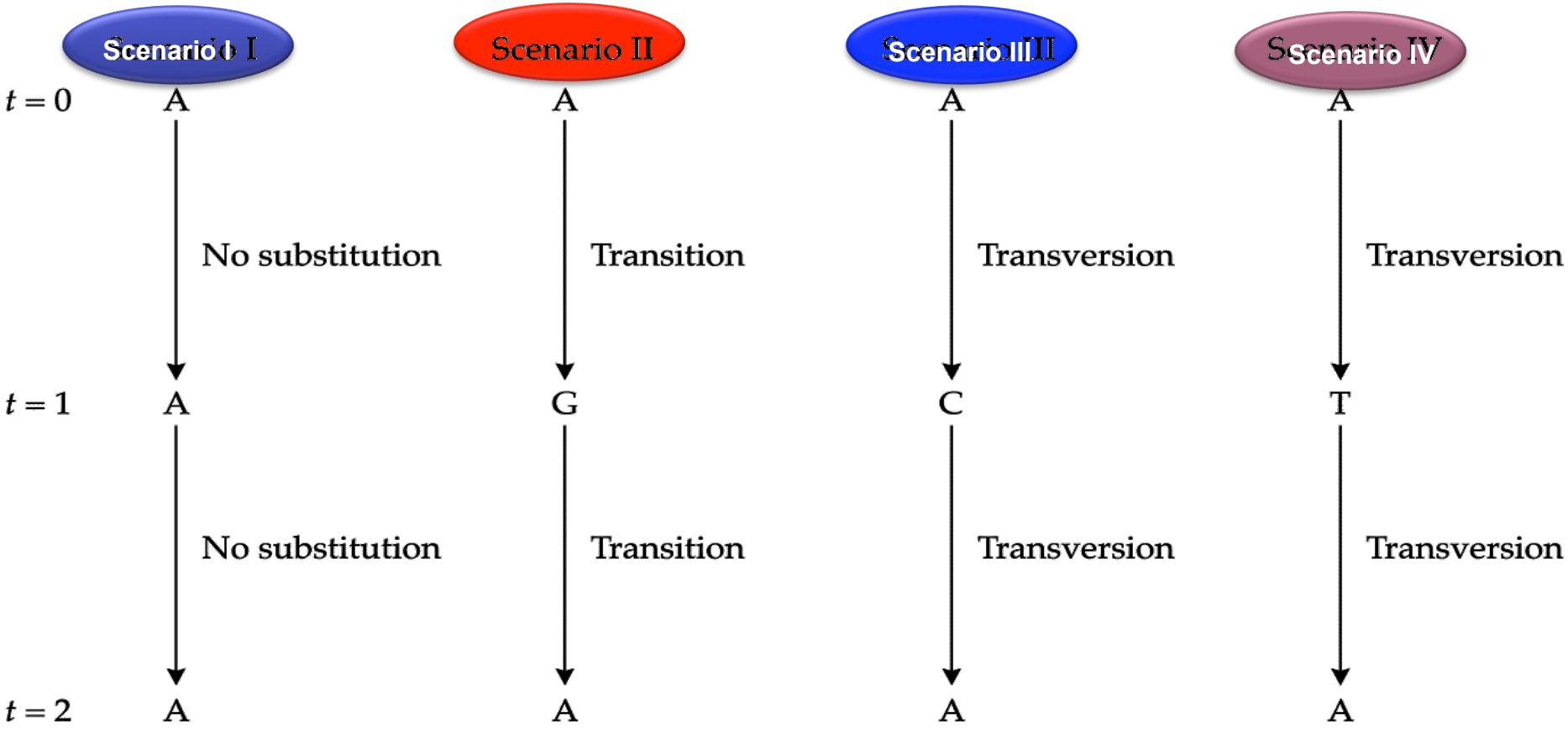
Transversion: β



Kimura 2-Parameter-Modell

	A	T	C	G
A	$1-\alpha-2\beta$	β	β	α
T	β	$1-\alpha-2\beta$	α	β
C	β	α	$1-\alpha-2\beta$	β
G	α	β	β	$1-\alpha-2\beta$

Kimura 2-Parameter-Modell



$$P_{AA(2)} = (1 - \alpha - 2\beta)P_{AA(1)} + \beta P_{TA(1)} + \beta P_{CA(1)} + \alpha P_{GA(1)}$$

Kimura 2-Parameter-Modell

Analog zu Jukes-Cantor ergibt sich nun ...

$$K = \frac{1}{2} \ln\left(\frac{1}{1 - 2P - Q}\right) + \frac{1}{4} \ln\left(\frac{1}{1 - 2Q}\right)$$

P = Anteil der Transitionen (Ti: A <=> G, C <=> T)

Q = Anteil der Transversionen (Tv: A,G <=> C,T)

JC und K2P im Vergleich

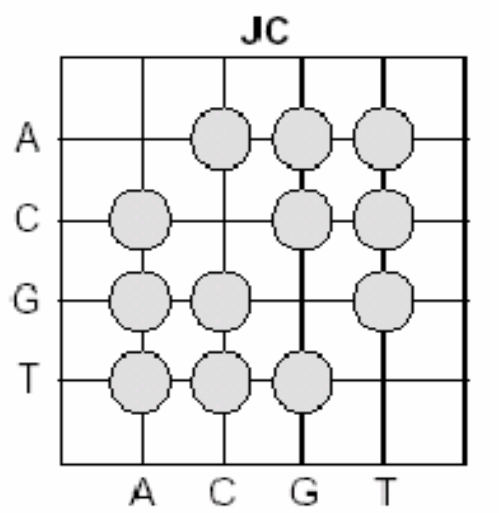
Beispiel 1: 2 Seqenzen mit je 100 bp
Divergenz: 10 Ti, 2 Tv

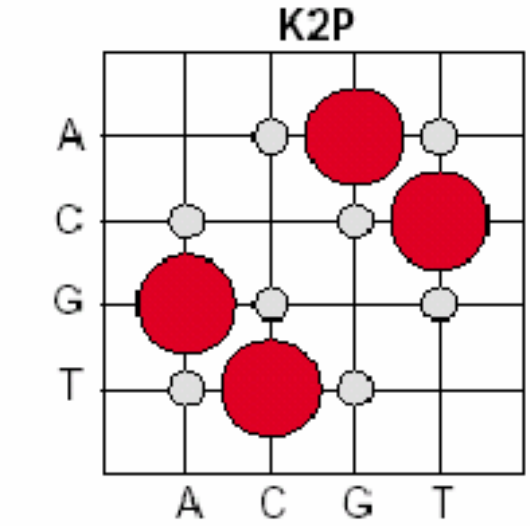
p (unkorrigiert) = 12 / 100 = 0,12
K (JC) = 0,131
K (K2P) = 0,134

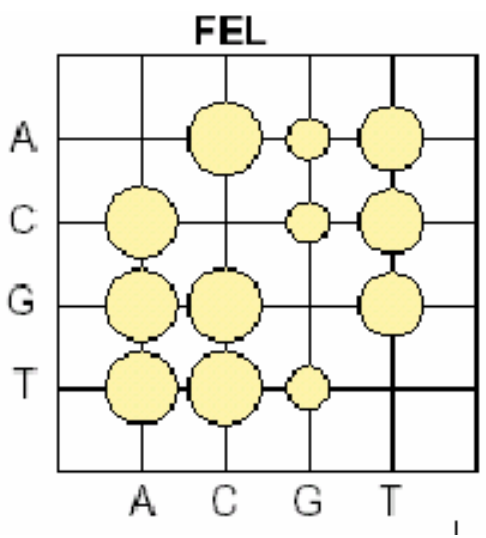
Beispiel 2: 2 Seqenzen mit je 100 bp
Divergenz 25 Ti, 10 Tv

p (unkorrigiert) = 35 / 100 = 0,35
K (JC) = 0,471
K (K2P) = 0,514

JC-, K2P-, F81-Modell





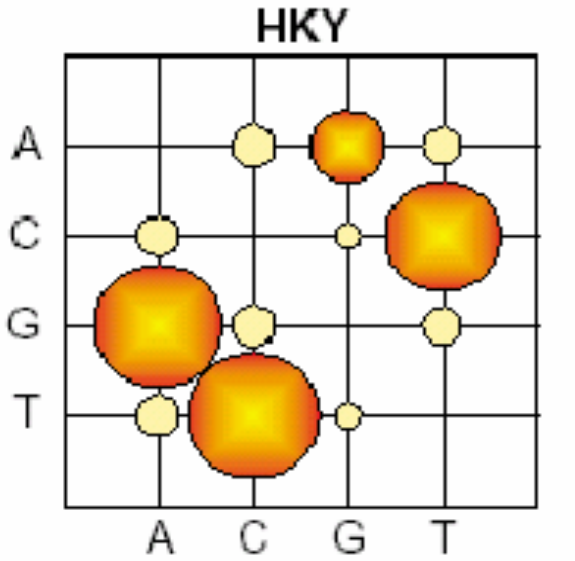


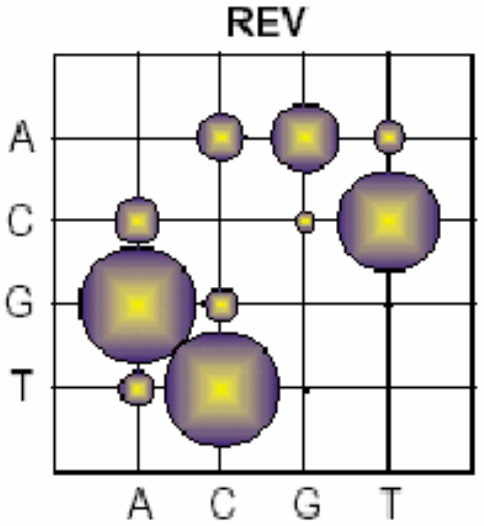
- Alle Substitutionen gleich häufig
- die erwartete Nukleotidzusammen-setzung ist identisch

- Transitionen und Transversionen unterschiedlich häufig
- die erwartete Nukleotidzusammen-setzung ist identisch

- Erwartete Nukleotid-zusammensetzung unterschiedlich.

HKY85 und GTR-Modelle





Transitionen und Transversionen und Nukleotid-zusammensetzung sind unterschiedlich häufig.

Alle Parameter (Austausche und Austauschrichtungen) und Nukleotidzusammensetzung dürfen variieren.

Merke: die einfachen Modelle sind Teilannahmen der komplizierten Modelle

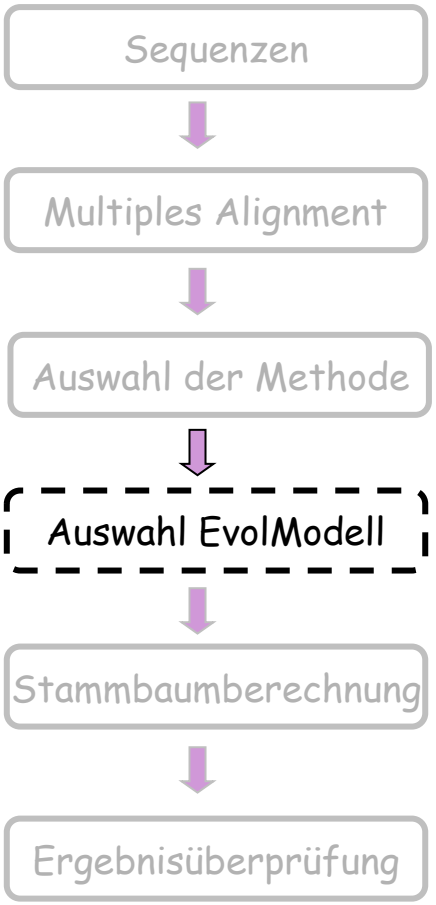
Welches Modell nehmen?

Merke:

die einfachen Modelle sind Teilannahmen der komplizierten Modelle !

...dies erlaubt es uns später, durch statistische Verfahren (z.B. den Likelihood ratio test, LRT) herauszufinden, welches Modell für unseren Datensatz am besten geeignet ist!

Wo stehen wir?



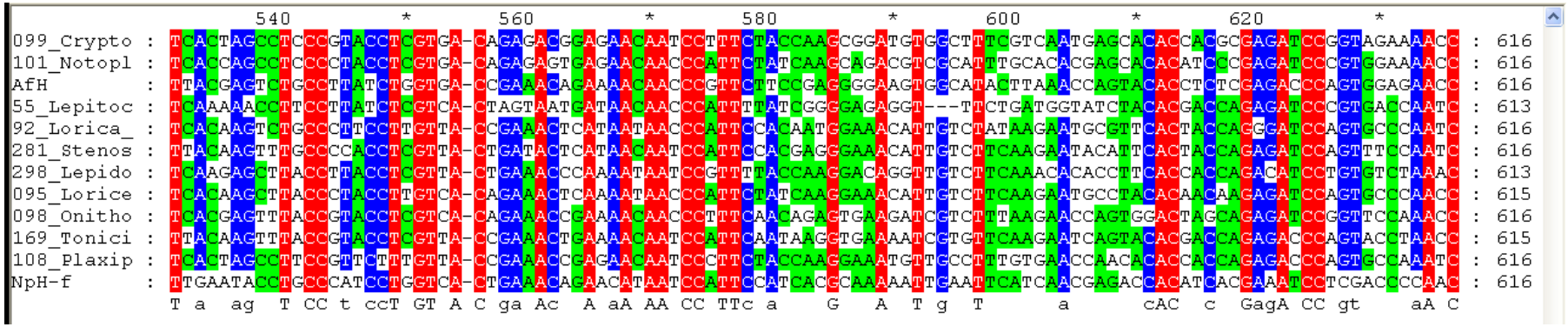
⇨ Distanz!

⇨ Korrektur der Evolutionsraten durch Auswahl des best-geeigneten Evolutionsmodells

Doch zuvor muss noch ein weiterer Parameter unserer Evolutionsmodelle besprochen werden...

Variationen der Substitutionsraten

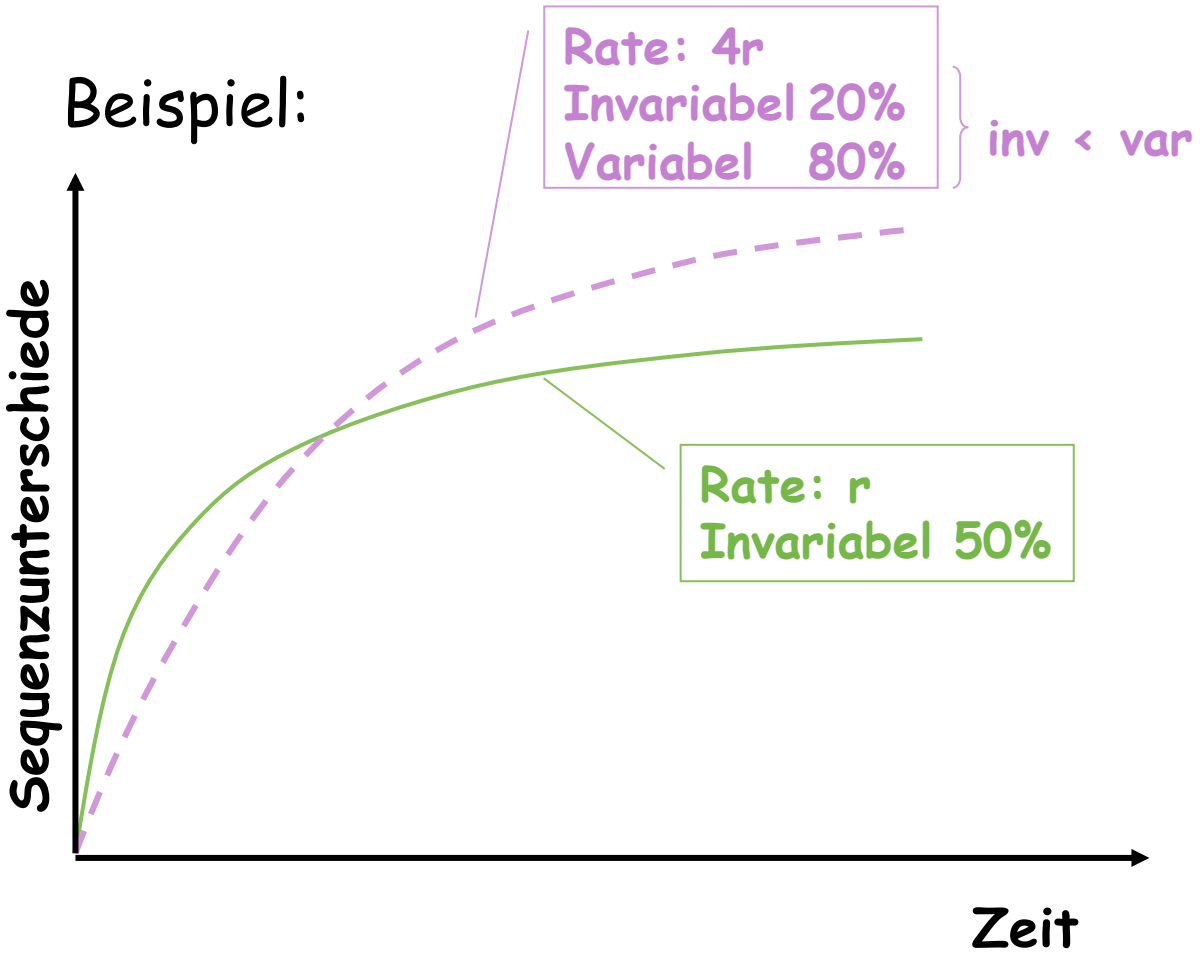
- Bisher nahmen die Modelle an, dass die Evolutions-raten innerhalb einer Sequenz gleich verteilt sind.
- Kann das stimmen?

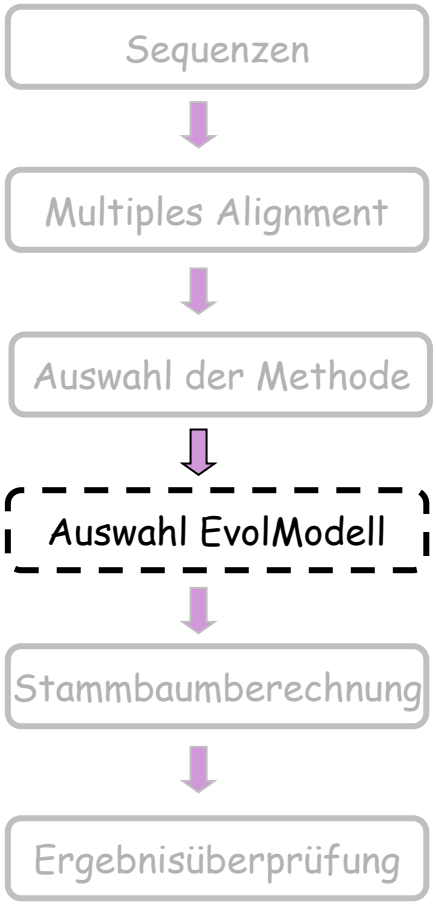


Variationen der Substitutionsraten

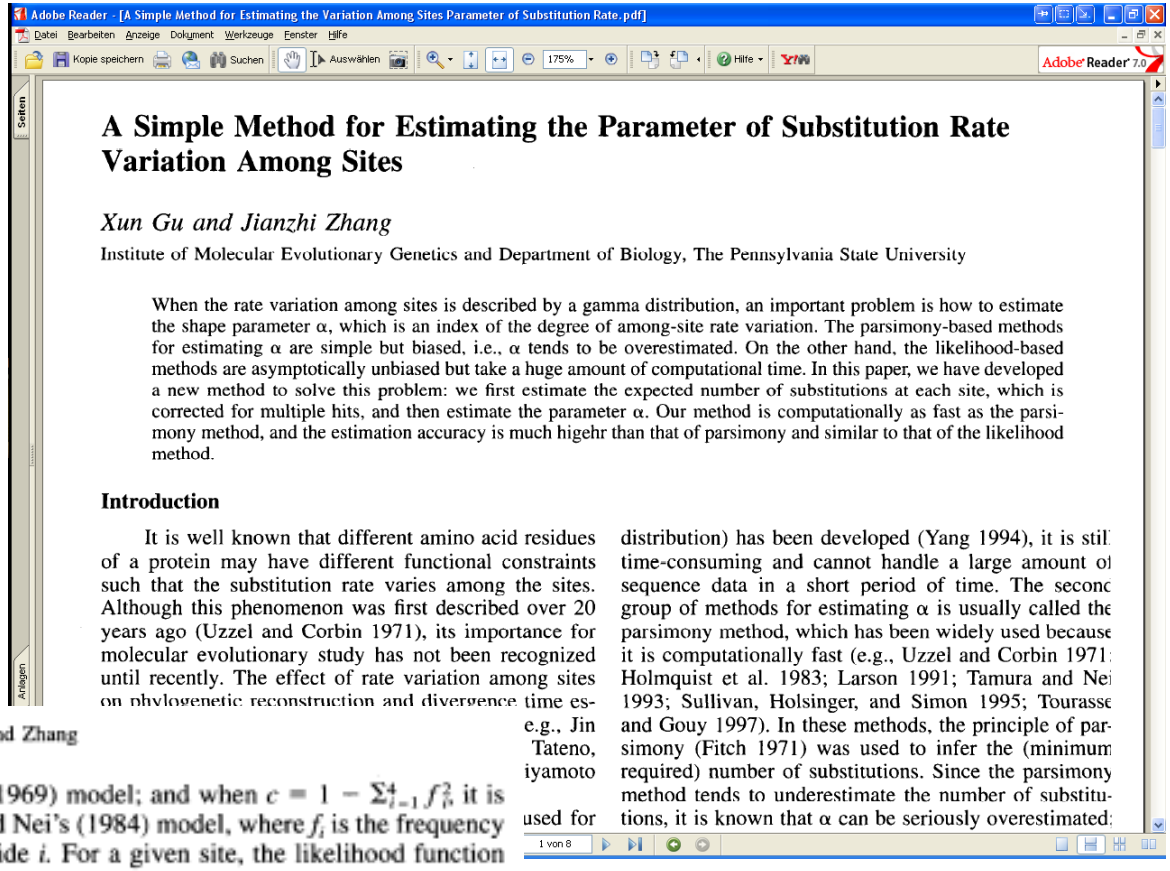
- Bisher nahmen die Modelle an, dass die Evolutionsraten innerhalb einer Sequenz gleich verteilt sind.
- Kann das stimmen?
- ***Nein***! Evolutionsrate ist abhängig von der Funktion: ***synonyme*** Codonpositionen evolvieren z. B. wesentlich schneller als ***nicht-synonyme*** Codonpositionen in Genen!
- Dies hat natürlich Einfluss auf die Abschätzung der Substitutionen.

Variationen der Substitutionsraten





Sequenz GRÜN mutiert langsamer als Sequenz BLAU,
hat aber weniger variable Positionen >> Sättigung schneller erreicht



Γ (gamma)

1108 Gu and Zhang

Cantor's (1969) model; and when $c = 1 - \sum_{i=1}^{4} f_i^2$ it is Tajima and Nei's (1984) model, where f_i is the frequency of nucleotide i. For a given site, the likelihood function in equation (4) can be now modified as follows:

$$L = \prod_{i=1}^{m} c(1 - e^{-\bar{k}h_i}) \prod_{i=m+1}^{M} (1 - c + ce^{-\bar{k}h_i}) \quad (11)$$

Similar to equation (7), $\bar{k}$ at this site can be estimated by letting $\partial \ln L/\partial \bar{k} = 0$, which yields:

$$\sum_{i=1}^{m} \frac{b_i/B}{1 - e^{-\bar{k}h_i}} + (1 - c) \sum_{i=m+1}^{M} \frac{b_i/B}{1 - c + ce^{-\bar{k}h_i}} = c. \quad (12)$$

Clearly, equation (12) will be reduced to equation (7) when $c = 1$.

$$f(k) = \frac{\Gamma(\alpha + k)}{k!\Gamma(\alpha)} \left(\frac{D}{D + \alpha}\right)^{k} \left(\frac{\alpha}{D + \alpha}\right)^{\alpha}, \quad k = 0, 1, \ldots \quad (13)$$

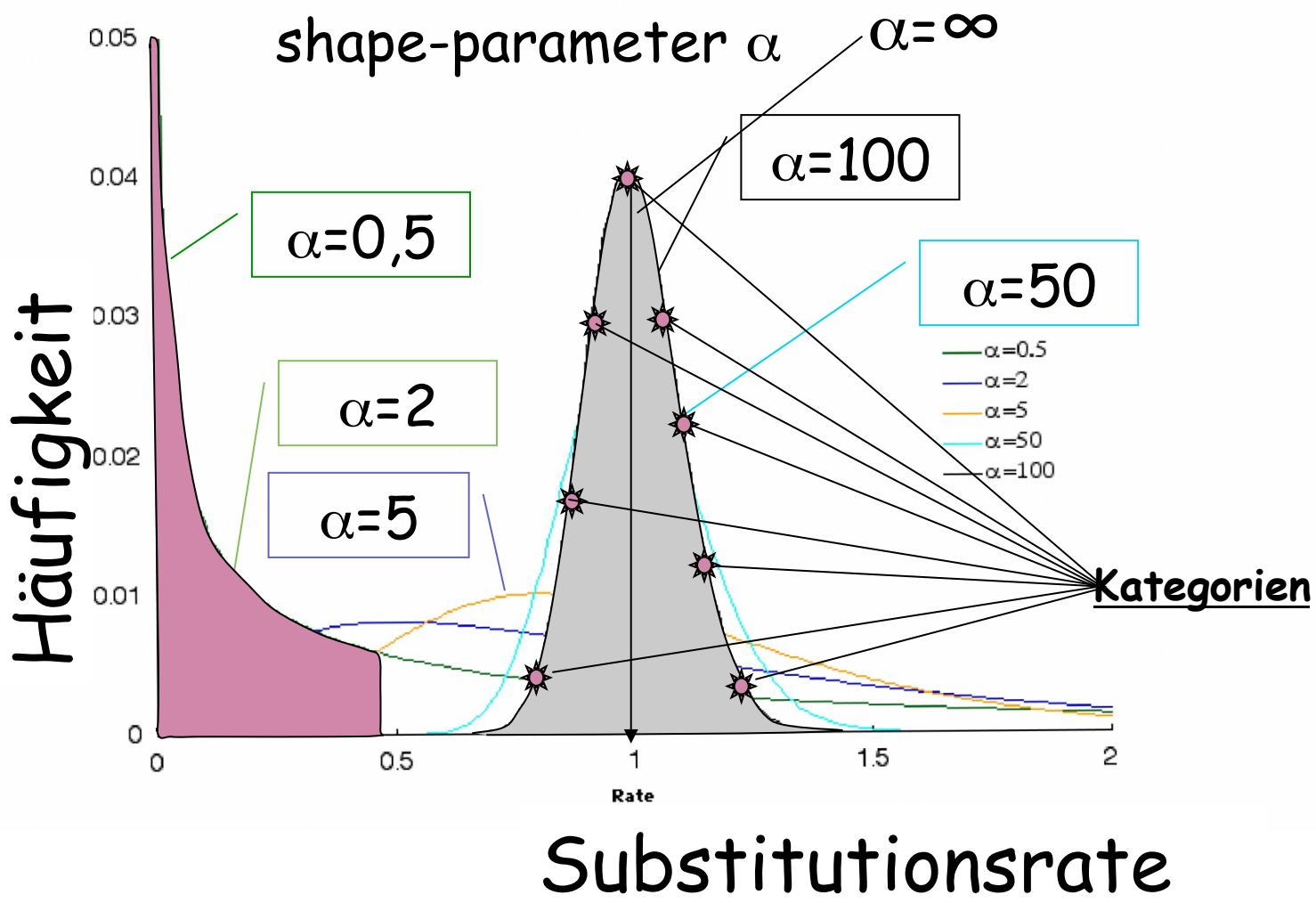
(12). Therefore, the log-likelihood function can

$$\ln L = \sum_{i=1}^{M} \ln f(k_i). \quad (14)$$

Variationen der Substitutionsraten

- Wie kann man solche Variationen der Substitutionsraten einrechnen?
- Jede Stelle im Alignment hat prinzipiell eigene Substitutionsrate („***among site variation***“ => ∞)
- Annahme: Die Variationen lassen sich über eine ***Gamma-*** $\{\Gamma\}$***Verteilung*** beschreiben.
- Der „***shape***“-parameter α gibt die relative Verteilung der unterschiedlichen Substitutionsraten wieder.
- α **groß** => **geringe Streuung** der Substitutionsraten
- α **klein** => **große Streuung** der Substitutionsraten

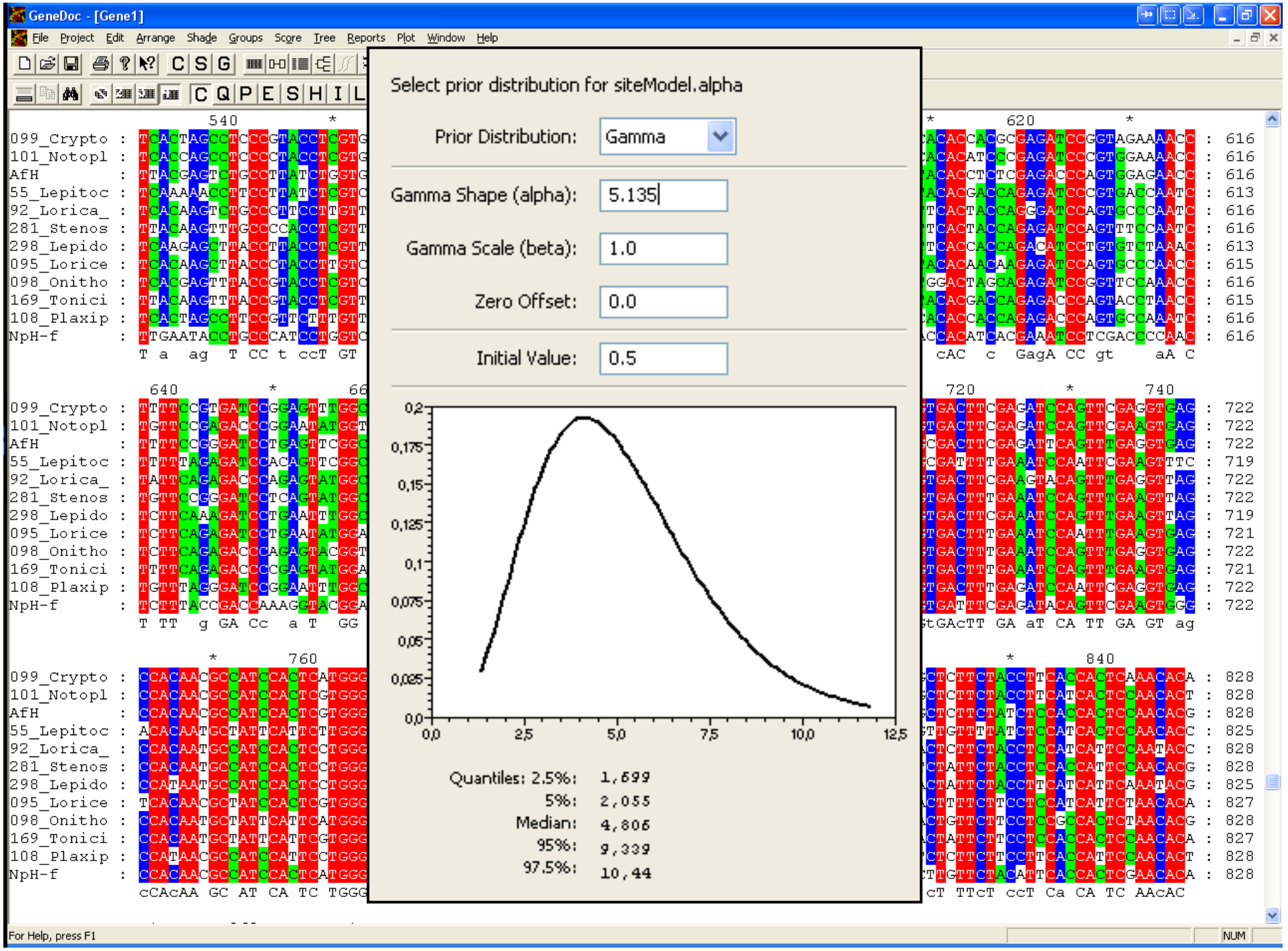
Variationen der Substitutionsraten



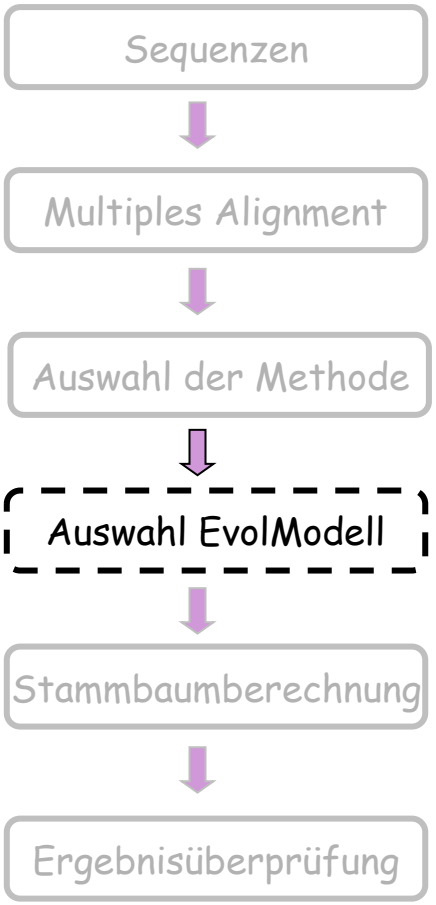
Variationen der Substitutionsraten

Table 5.2 Estimates of α shape parameter of the distribution of rate variation for various genes (after Yang, 1996, Table 1).

Type of sequences	α
Nuclear genes	
Albumin genes	1.05
Insulin genes	0.40
c-myc genes	0.47
Prolactin genes	1.37
16S-like rRNAs, stem region	0.29
16S-like rRNAs, loop region	0.58
ψη-globin pseudogenes	0.66
Viral genes	
Hepatitis B virus genomes	0.26
Mitochondrial genes	
12S rRNAs	0.16
Position 1 of four genes	0.18
Position 2 of four genes	0.08
Position 3 of four genes	1.58
D-loop region	0.17
Cytochrome *b*	0.44


GeneDoc - [Gene1]
File Project Edit Arrange Shade Groups Score Tree Reports Plot Window Help
Select prior distribution for siteModel.alpha
Prior Distribution: Gamma
Gamma Shape (alpha): 5.135
Gamma Scale (beta): 1.0
Zero Offset: 0.0
Initial Value: 0.5
0,2
0,175
0,15
0,125
0,1
0,075
0,05
0,025
0,0
0,0
2,5
5,0
7,5
10,0
12,5
Quantiles: 2.5%: 1,699
5%: 2,055
Median: 4,806
95%: 9,339
97.5%: 10,44
For Help, press F1
NUM

Wo stehen wir?



⇨ Distanz!

⇨ Auswahl des Evolutionsmodells plus Angabe des α-Parameters notwendig

jetzt zur Mutter aller Fragen...

Welches Modell ist das beste?

Je komplexer das Modell (mehr Annahmen), desto genauer und realistischer unsere Berechnung der Substitutionsrate.

***ABER*:**

Zusätzliche Parameter müssen aus den Daten abgeschätzt werden. **Je mehr Annahmen man trifft, desto größer wird der statistische Fehler (Varianz) der erhaltenen Werte!**

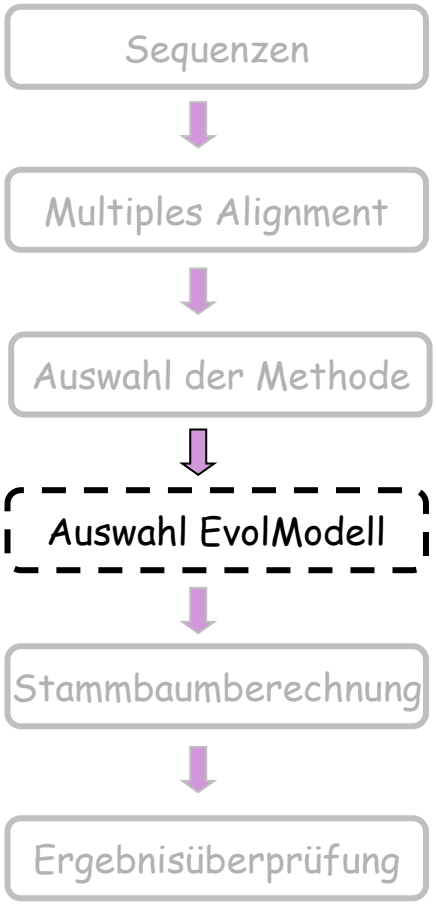
=> möglichst „gute“ Daten
⇒ Modell:
so einfach wie möglich, so kompliziert/exakt wie nötig

Welches Modell ist das beste?

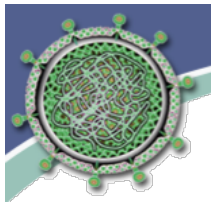
Wer sagt mir, welches Modell das beste für meine Daten ist?

=> Wir können und müssen die Modelle testen!

- ***Modeltest*** :
  Berechnet „vernünftigen“ NJ-Baum und schätzt daraus hierarchisch die Parameter ab, die das Modell spezifizieren



ModelTest, ProtTest, FindModel



http://hcv.lanl.gov/content/hcv-db/findmodel/findmodel.html

Analysen für Proteine ausgehend von einem „*alignment*“

http://darwin.uvigo.es/software/prottest.html

http://darwin.uvigo.es/software/modeltest.html

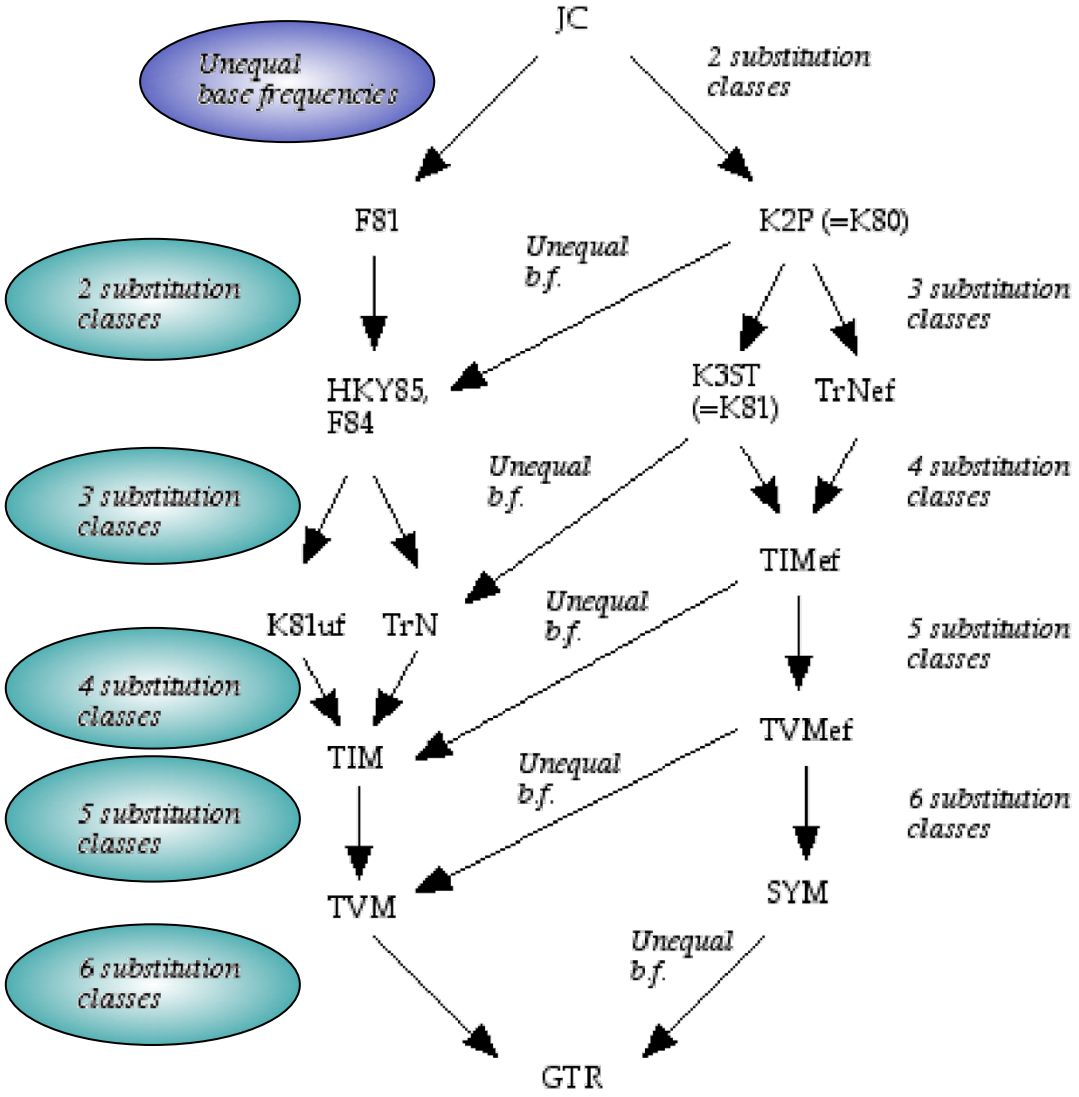
Analyse ausgehend von einer „*score matrix*“

ModelTest

...bewertet Eignung der Modelle für den jeweiligen Datensatz mit folgenden statistischen Tests:

- **Hierarchischer likelihood ratio test (hLRT)**,
- Akaike Information Criterion (AIC = -2 ln*L* + 2*K*; Akaike 1974),
- Korrigiertes AIC (AICc = AIC + 2*K*(*K*+1)/(*N*-*K*-1); Hurvich and Tsai 1989, Sugiura 1978)
- Bayesian Information Criterion (BIC = -2ln*L* + *K*log*N*; Schwarz 1978) [*L* = model likelihood, *K* = number of estimatable parameters, *N* = sample size].

ModelTest



Die einfachen Modelle mit weniger Parametern sind Teil-Annahmen der komplizierten Modelle (Verschachtelung)

hLRT (hierarchischer Likelihood ratio-Test)

- berechnet, ob ein komplizierteres Modell (mit mehr Parametern) signifikant besser zu einem Datensatz passt als ein einfacheres „Nullmodell“.

- nur durchführbar mit hierarchisch verschachtelten Modellen.

- Ziel ist es, das exakteste, gleichzeitig aber auch einfachste Evolutionsmodell zu identifizieren.

- LR = 2* (lnL1-lnL0) folgt der Chi^2-Verteilung

http://www.molecularevolution.org/si/resources/lrt.php

hLRT Beispiel

<u>HKY85 oder GTR-Modell (4 Parameter mehr!) nehmen?</u>

- Baum mit HKY85 -lnL = 1787.08 (Nullmodell)
  Baum mit GTR -lnL = 1784.82

- LR = 2 (1787.08-1784.82) = 4.53
  bei n=4 Freiheitsgraden (wegen 4 zusätzlichen Parametern in GTR)

- Kritischer X^2-Wert (P=0.05) = 9.49 (aus Tabelle)

- GTR ist nicht signifikant besser: HKY85 nehmen!

ModelTest

D.Posada and **K.A.Crandall**

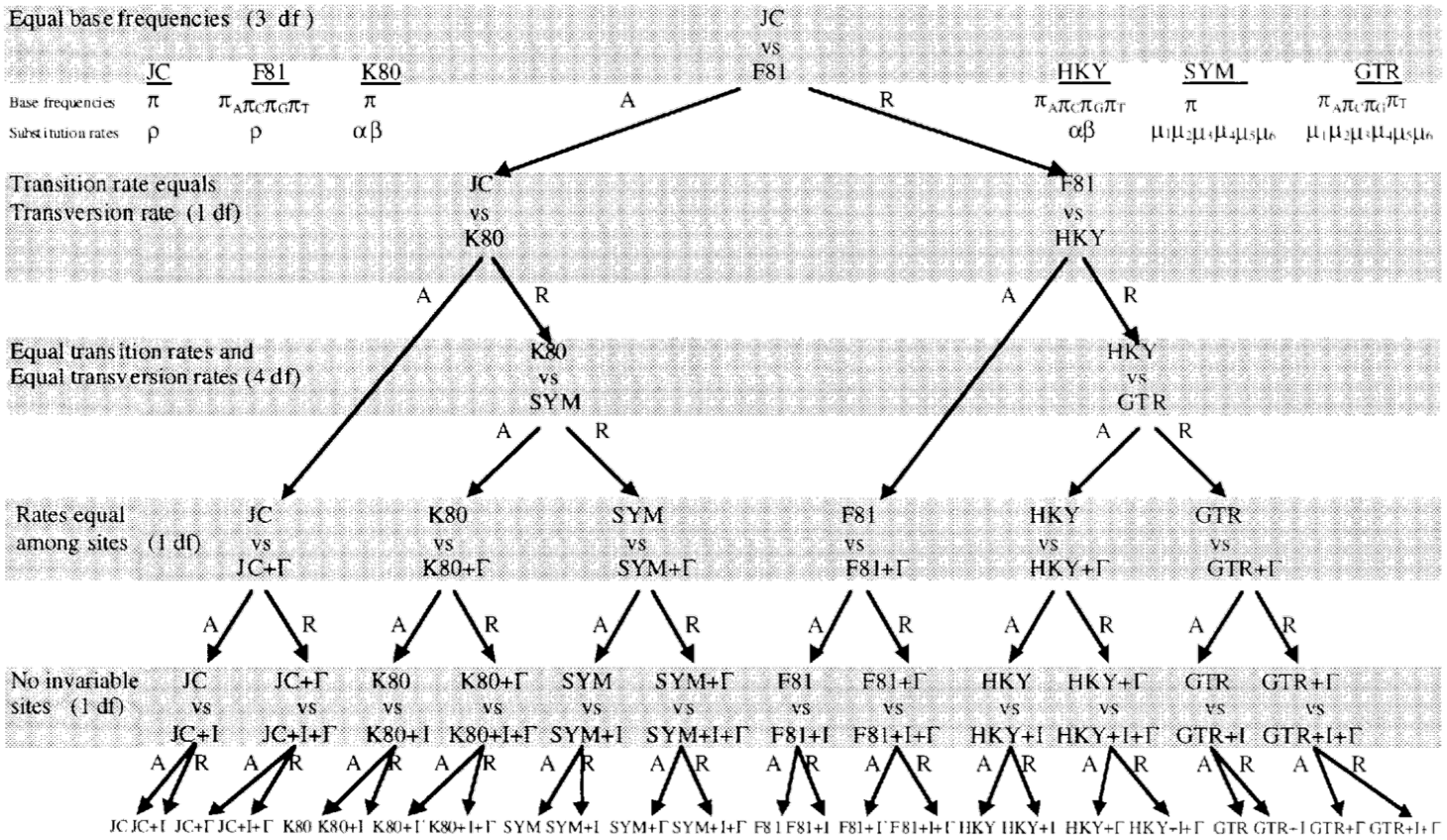


Fig. 1. Hierarchical hypothesis testing in MODELTEST. At each level the null hypothesis (upper model) is either accepted (A) or rejected (R). The models of DNA substitution are: JC (Jukes and Cantor, 1969), K80 (Kimura, 1980), SYM (Zharkikh, 1994), F81 (Felsenstein, 1981), HKY (Hasegawa *et al*., 1985), and GTR (Rodríguez *et al*., 1990). Γ: shape parameter of the gamma distribution; I: proportion of invariable sites. df: degrees of freedom. 1: equal base frequencies (0.25), π_A: frequency of adenine, π_C: frequency of cytosine, π_G: frequency of guanine, π_T: frequency of thymine. ρ: equal substitution rate, α: transition rate, β: transversion rate; μ_1: A⇒C rate, μ_2: A⇒G rate, μ_3: A⇒T rate, μ_4: C⇒G rate, μ_5: C⇒T rate, μ_6: G⇒T rate.

ModelTest output

** Log Likelihood scores **

		+I	+G	+I+G
JC =	9208.0537	8635.4209	8523.8740	8511.6758
F81 =	9189.1426	8605.4951	8492.9854	8480.7363
K80 =	9064.1436	8476.4131	8342.2559	8327.3379
HKY =	9042.0869	8435.8828	8307.2676	8289.9834
SYM =	9026.4961	8452.5811	8305.9883	8293.9131
GTR =	9023.1475	8427.3691	8295.7188	8281.5918

Run settings
Using the standard AIC (not the AICc)
Not using branch lengths as parameters
Running all four hierarchies for the hLRT
Printed parameter values are from the hLRT1 hierarchy

---

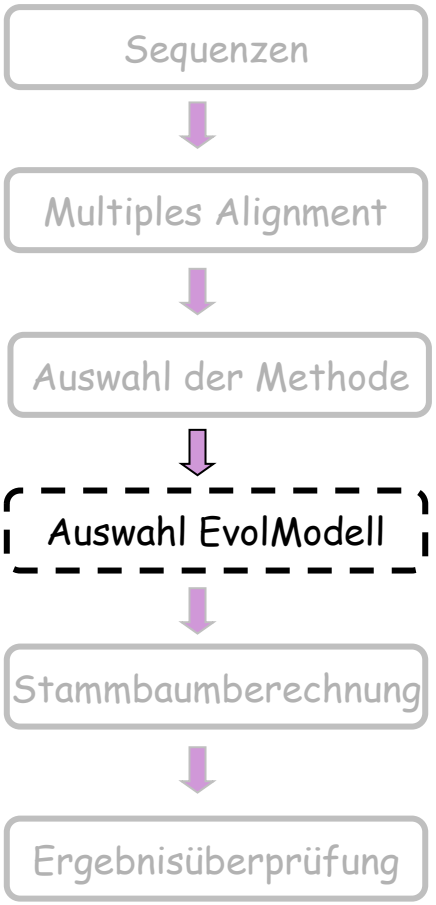
* HIERARCHICAL LIKELIHOOD RATIO TESTS (hLRTs) *

Equal base frequencies
Null model = JC -lnL0 = 9208.0537
Alternative model = F81 -lnL1 = 9189.1426
2(lnL1-lnL0) = 37.8223 df = 3
P-value = <0.000001

Ti=Tv
Null model = F81 -lnL0 = 9189.1426
Alternative model = HKY -lnL1 = 9042.0869
2(lnL1-lnL0) = **294.1113** df = 1
P-value = <0.000001

Unequal Tv and unequal Ti
Null model = HKY -lnL0 = 9042.0869
Alternative model = GTR -lnL1 = 9023.1475
2(lnL1-lnL0) = **37.8789** df = 4
P-value = <0.000001

Wo stehen wir?



⇨ Distanz!

⇨ Evolutionsmodell: z.B. HKY85
gamma shape parameter = z.B 0.5

WIE machen wir mit der korrigierten Distanzmatrix nun den Baum?