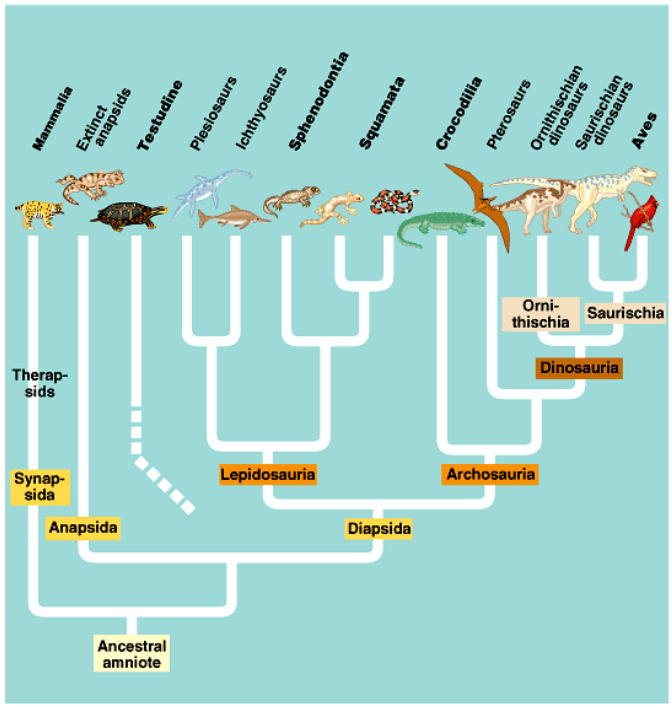
WS2017/18

F1-Praktikum Modul 7A

Genomforschung und Sequenzanalyse: Einführung in Methoden der Bioinformatik

AG Hankeln



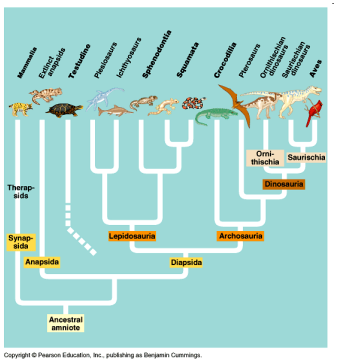
Molekulare Phylogenie

Biologische Systematik umfasst...

1. Taxonomie

→ Bestimmung und Benennung von Lebewesen

2. Phylogenie

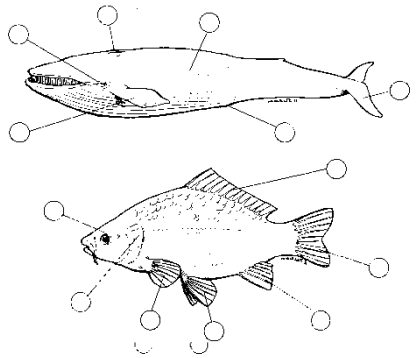


→ Rekonstruktion der Stammesgeschichte
→ auf allen Ebenen möglich

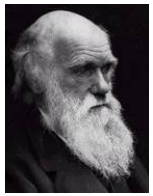
Metazoa, Vertebrata, Mammalia, Rodentia, Cricetinae …

Molekulare Daten - Vorteile

- eindeutiger zu beschreiben
→ Identifizierung von Orthologien ist einfacher



- Vergleich sehr weit entfernt verwandter Taxa möglich



← →

??

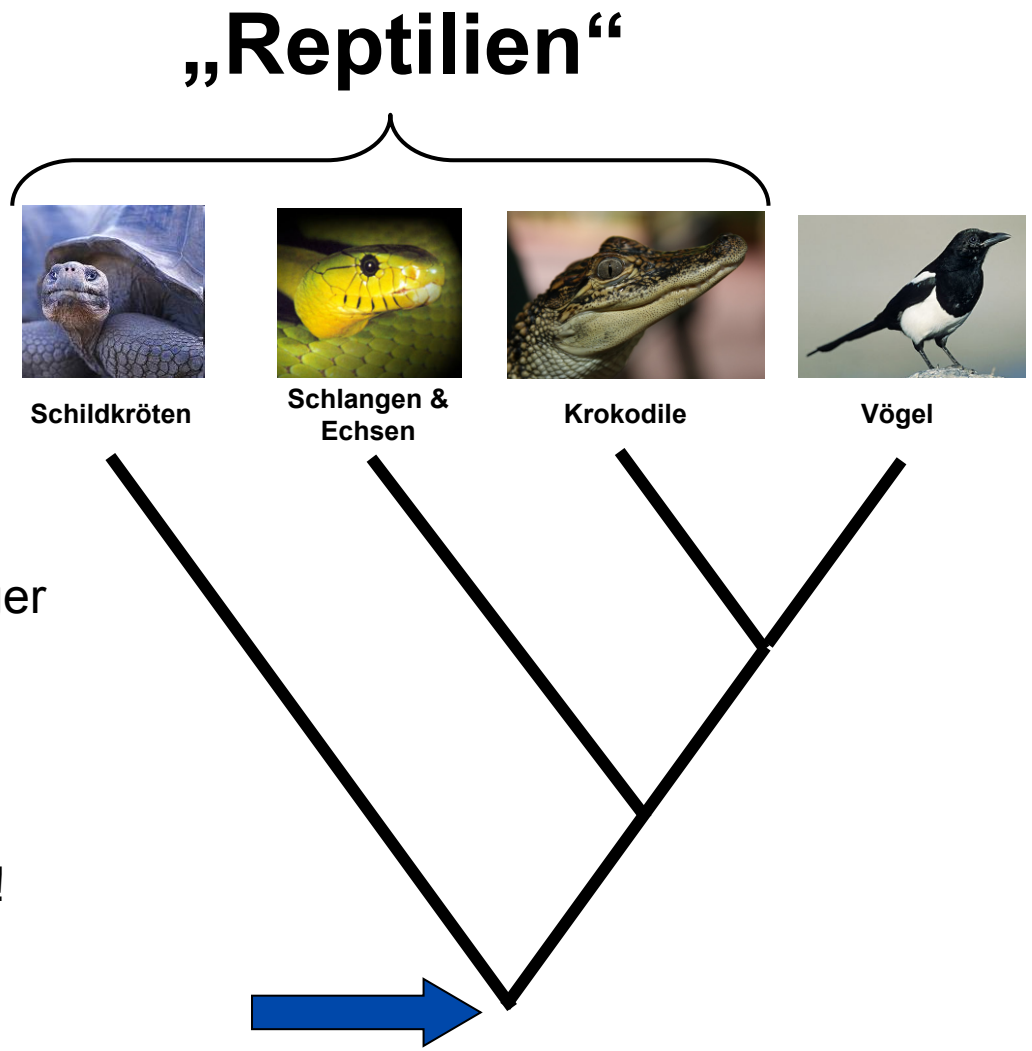
```
                  *        20        *        40        *        60        *        80        *       100        *       120        *       1
Homo_Cox1   : M-FADRWLFSTNHKDIGTLYLLFGAWAGVLGTALSLLIRAELGQPGN--LIGNDHIYNVIVTAHAFVMIFFMVMPIMIGGFGNWLVPLMIGAPDMAFPRMNNMSFWLLPPSLLLLDASAMVEAGAGTGWTVYPPLAGN : 135
Arabidopsi  : MKNLVRWLFSTNHKDIGTLYFIFGAIAGVMGTCFSVLIRMELARPGDQILGGNHQLYNVLITAHAFLMIFFMVMPAMIGGFGNWFVPILIGAPDMAFPRLNNISFWLLPPSLLLLISSALVEVGSGTGWTVYPPISGI : 138
            M    RWLFSTNHKDIGTLY 6FGA AGV6GT  S6LIR EL  PG1  L GN  6YNV66TAHAF6MIFFMVMP MIGGFGNW VP66IGAPDMAFPR6NN6SFWLLPPSLLLL SA6VE G GTGWTVYPPL G

            40        *       160       *       180       *       200       *       220       *       240       *       260       *
Homo_Cox1   : YSHPGASVDLTIFSLHLAGVSSILGAINFITTIINMKPEAMTQYQTPLFVWSVLITAVLLLLSLPVLAAGITMLLTDRNLNTTFFDPAGGGDPILYQHLFWFFGHPEVYILILPGFGMISHIVTYYSGKKEPFGYMGM : 273
Arabidopsi  : ISHSGGAVDLAIFSLHISGVSSILGSINFITTIFNMRPEGMTMHRLPLFVWSVLVTAFLLLLSLPVLAGAITMLLTDRNFNTTFFDPAGGGDPILYQHLFWFFGHPEVYILILPGFGIISHIVSTFSG-KPVFGYLGM : 275
             SH G  VDL IFSLHL GVSSILG INFITTI NM4 P MT    PLFVWSVL6TA LLLLSLPVLA  ITMLLTDRN NTTFFDPAGGGDPILYQHLFWFFGHPEVYILILPGFG6ISHIV3 5SG K  FGY6GM

            280       *       300       *       320       *       340       *       360       *       380       *       400       *
Homo_Cox1   : VWAMMSIGFLGFIVWAHHMFTVGMDVDTRAYFTSATMIIAIPTGVKVFSWLATLHGSNMKWSAAVLWALGFIFLFTVGGLTGIVLANSSLDIVLHDTYYVVAHFHYVLSMGAVFAIMGGFIHWFPLFSGYTLDQTYAK : 411
Arabidopsi  : VYAMISIGVLGFLVWAHHMFTVGLDVDTRAYFTAATMIIAVPTGIKIFSWIATMWGGSIQYKTPMLFAVGFIFLFTIGGLTGIVLANSGLDIALHDTYYVVAHFHYVLSMGAVFAIFAGFYYWVGKIFGRTYPETLGQ : 413
            V5AM6SIG LGF6VWAHHMFTVG6DVDTRAYFT ATMIIA6PTG6K6FSW6AT6 G  6 5   6L5A6GFIFLFT6GGLTGIVLANS LDI LHDTYYVVAHFHYVLSMGAVFA6  GF  W     G T  2T
```

Grundbegriffe

Paraphylie



Gruppierung aufgrund homologer (ursprünglicher) Merkmale

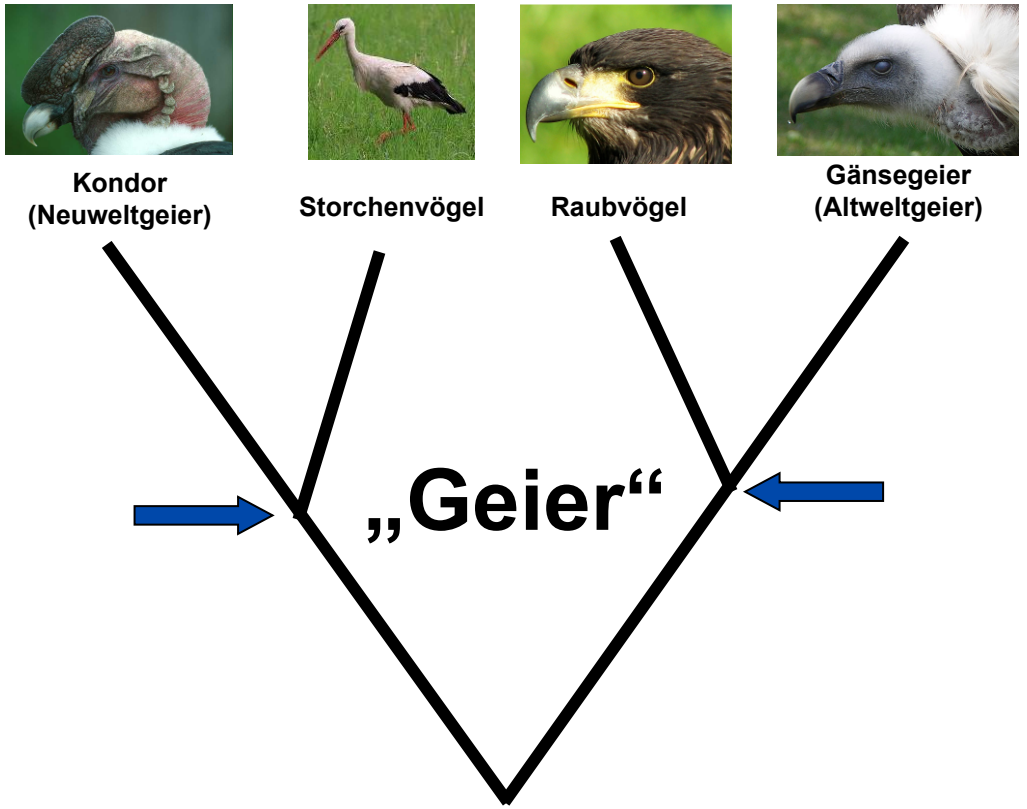
ABER
nicht alle Nachkommen erfasst!

→Monophylum **Sauropsida**

Grundbegriffe

Polyphylie

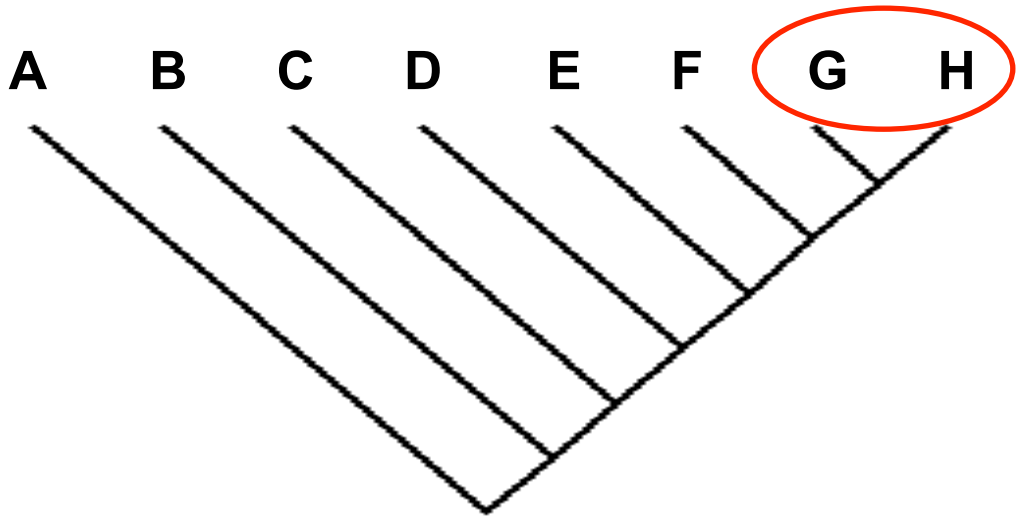
Gruppierung aufgrund konvergent entstandener Merkmale (Homoplasien)
→ Taxa verschiedenen Ursprungs!



Grundbegriffe

Die Taxa (G,H) liegen innerhalb des Monophylums (D,E,F,G,H)

ABER:
Die Taxa (G,H) sind NICHT „abgeleitet“ oder „weiter entwickelt“ und das Taxon D ist NICHT „primitiv“



Grundbegriffe

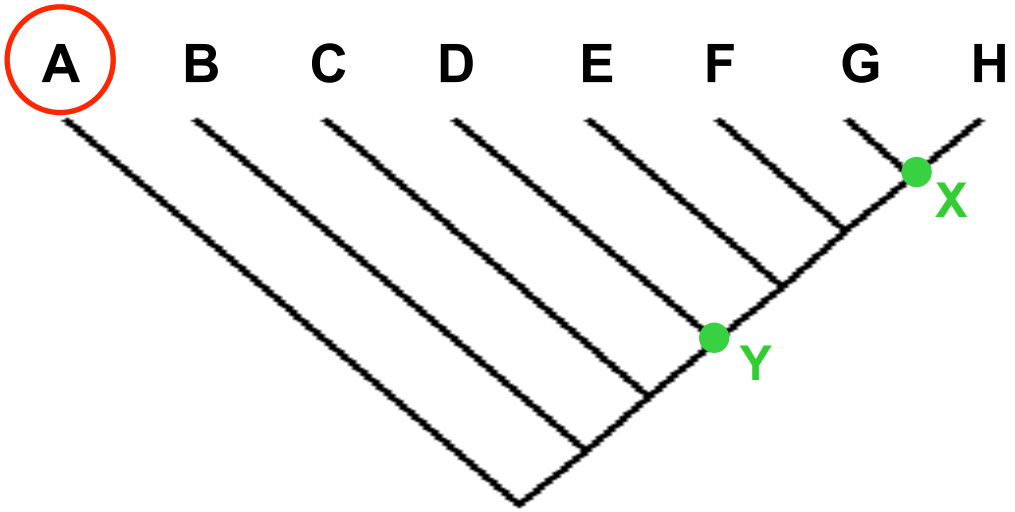
„Basal“

Taxon A ist Schwester zum Monophylum (B,C,D,E,F,G,H)
ABER: Taxon A ist NICHT „basal“

Der Begriff „basal“ wird nur für interne Knoten, NICHT für terminale Taxa Verwendet.

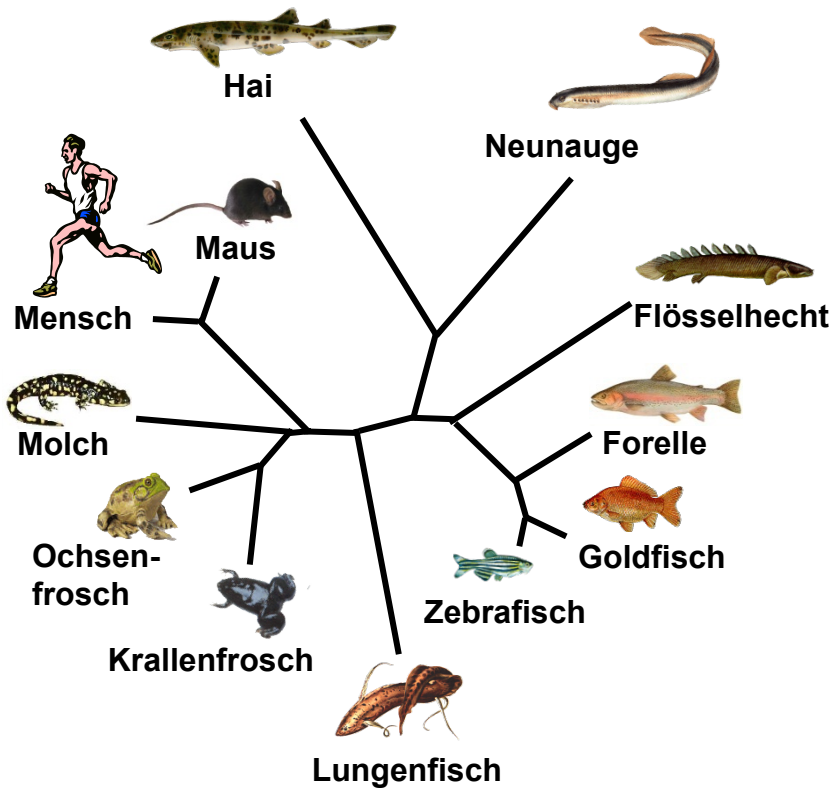
zB. **Y ist basaler als X (aber X ist nicht „weiter entwickelt“!)**

Y = der letzte gemeinsame Vorfahre von (D,E,F,G,H)
X = der letzte gemeinsame Vorfahre von (G,H)



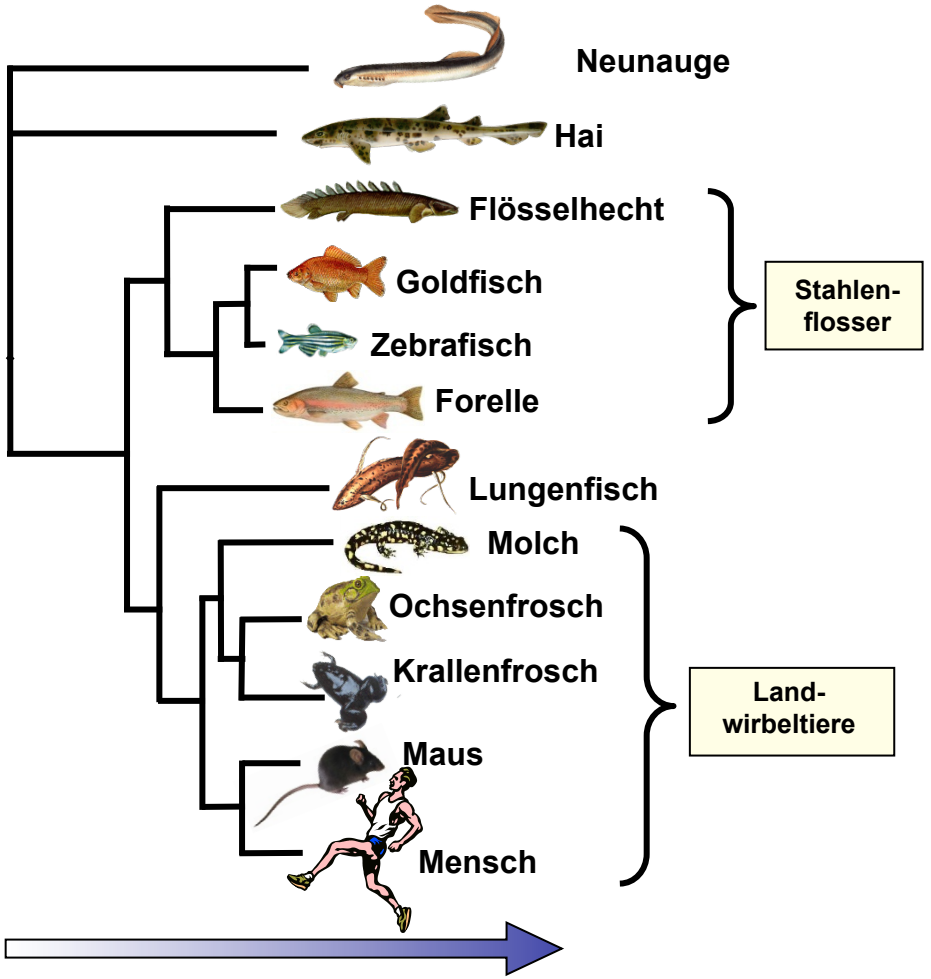
Baumdarstellung

Ohne Außengruppe:



Evolutionsrichtung?

Mit Außengruppe:



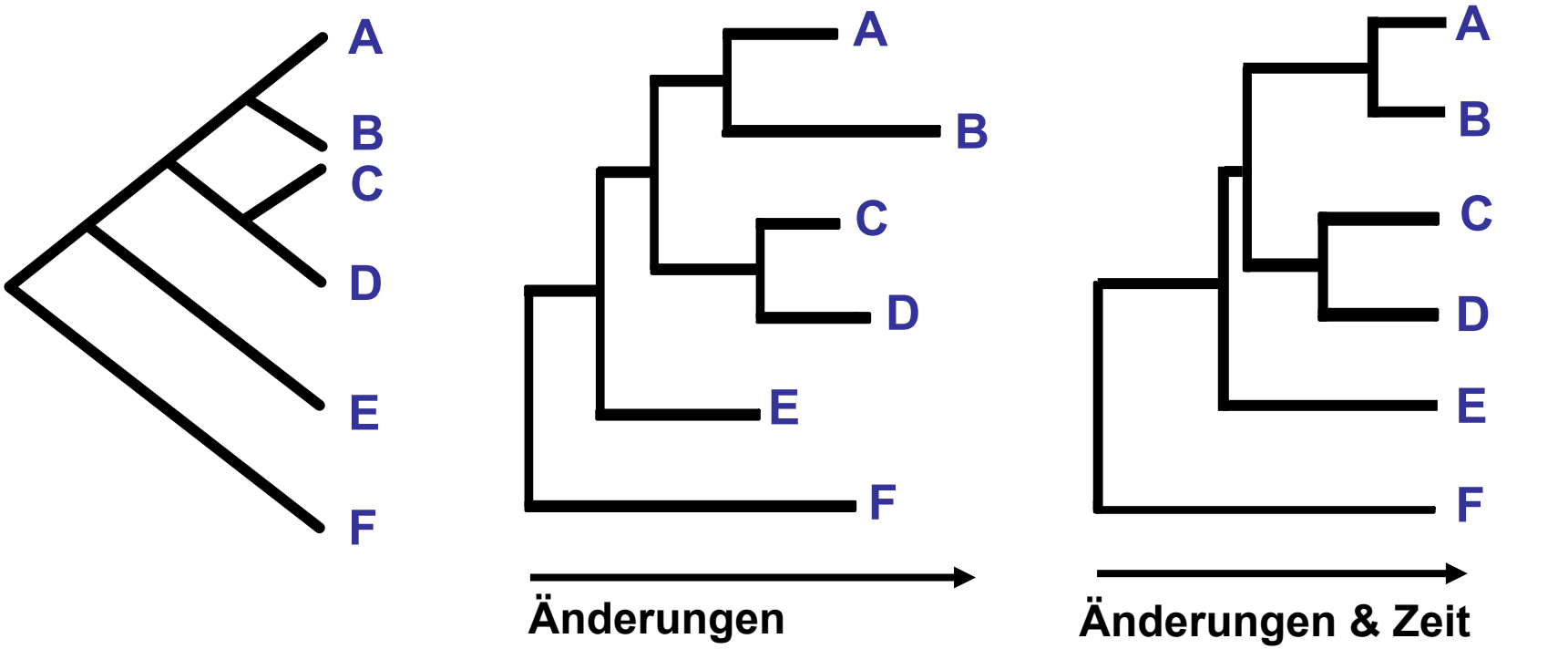
Evolutionsrichtung!

Baumdarstellung

Ungewurzelte Bäume sind keine Phylogenien!

Denn: je nach Wahl der Wurzel ergibt sich eine andere Topologie

Baumdarstellung



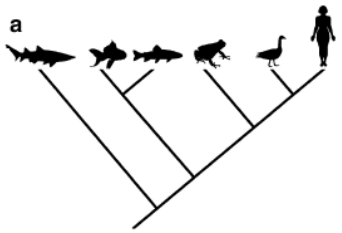
Literaturempfehlung

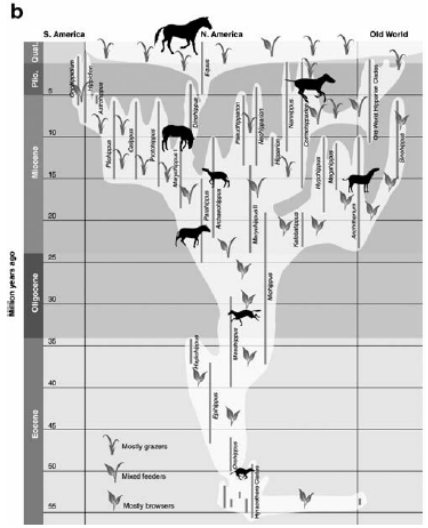
Evo Edu Outreach (2008) 1:121–137
DOI 10.1007/s12052-008-0035-x

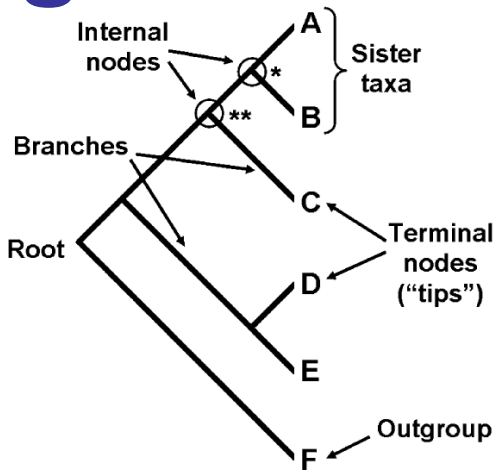
ORIGINAL SCIENCE/EVOLUTION REVIEW

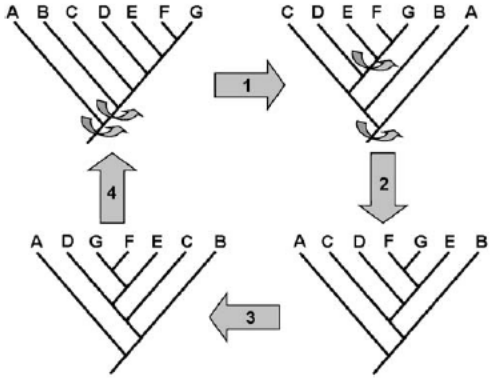
Understanding Evolutionary Trees

T. Ryan Gregory









How to Read Evolutionary Trees

Phylogenies as Family Trees

Although the technical jargon of phylogenetics may be confusing on first pass, achieving a basic understanding of evolutionary trees need not be daunting. Notably, humans in all cultures are skilled at recognizing and understanding relatedness in other contexts, and many of these abilities apply equally well to phylogenies. There are some similarities between species phylogenies and human family pedigrees, and thinking of an evolutionary tree as a "family tree" can be helpful.[5] This

How Not to Read Evolutionary Trees

Misunderstandings of evolutionary trees are pervasive among students, in the media, and among other non-specialists. Even more alarming, they also surface frequently

Von der Sequenz zum Baum

Sequenzen

Multiples Sequenz Alignment

Auswahl eines Evolutionsmodells

Auswahl von Methode & Algorithmus

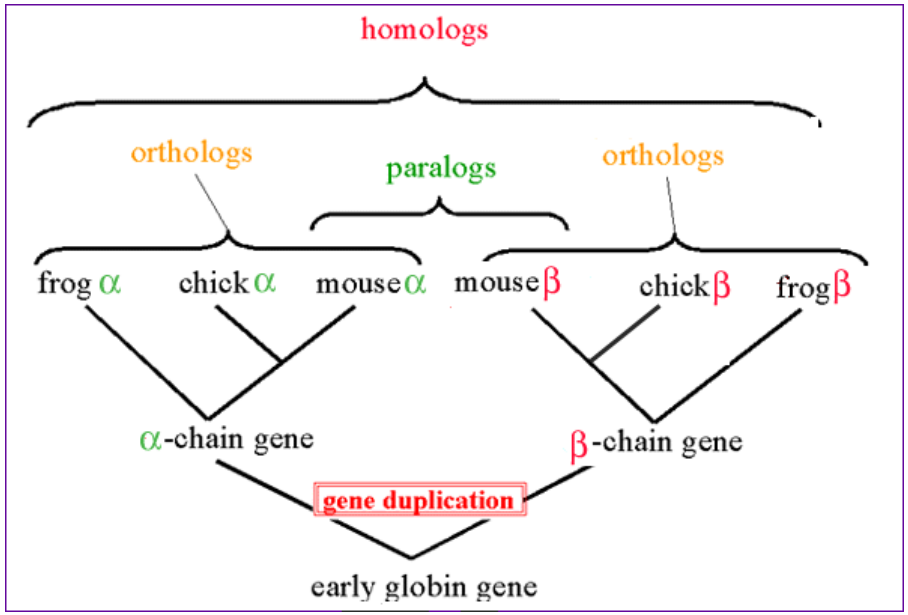
Stammbaumberechnung
(mit/ohne statistischer Auswertung)

Grundbegriffe - Homologie

Orthologie funktional verwandt und von einem gemeinsamen Vorläufer abstammend

Paralogie Verwandtschaft durch Genduplikation entstanden

Beispiel: α- und β-Untereinheiten des Hämoglobins

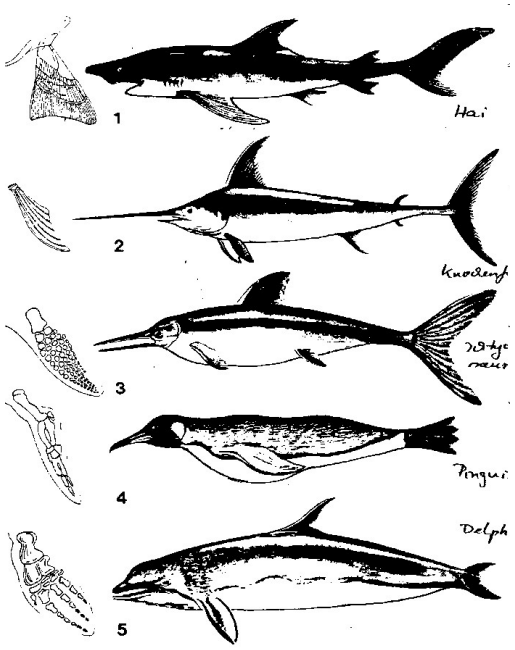


Grundbegriffe - Homoplasie

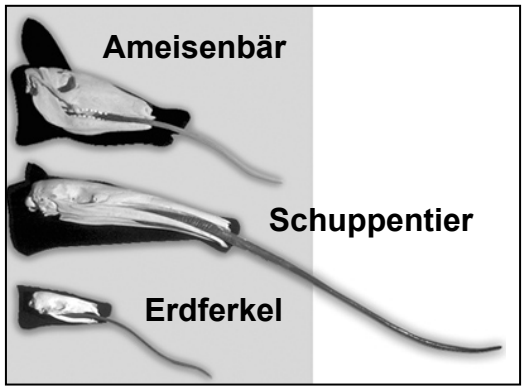
Konvergenz ein Merkmal, das bei mehreren unterschiedlichen Taxa unabhängig voneinander entstanden ist

---

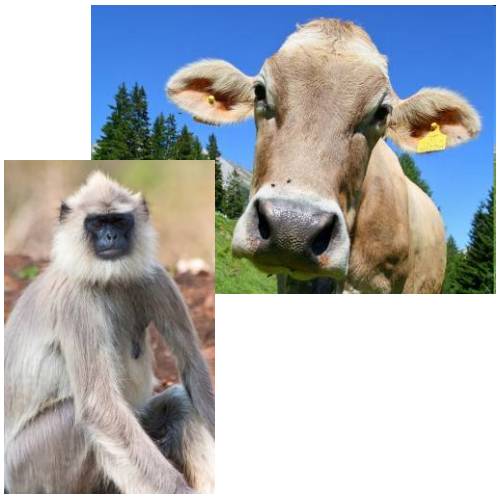
Flossen



Leimrute



Lysozym



Homologie-Interpretation von BLAST-Ergebnissen

Homologe Proteine haben immer strukturelle Ähnlichkeit, aber nicht notwendigerweise auch Ähnlichkeit auf Sequenzebene!

Faustregeln (Proteinebene):

score >45 bits	fast immer Homologe
E-value cutoff	~10^{-5} (bei DNA: ~10^{-10})
% identity	kein gutes Kriterium für Homologie, stark abhängig vom betrachteten Protein

Falsch-Positive

…können z.B. auftauchen wenn Regionen geringer Komplexität (simple sequences) alignieren und den Score erhöhen

Homologie: Interpretation von BLAST-Ergebnissen

Wie kann es sein, dass ich mit weniger % identity einen besseren E-Value bekomme?

1. Die Bewertung hängt nicht nur von der Anzahl sondern auch von der Art der identischen Aminosäuren ab

2. In die Bewertung fließt nicht nur der Anteil identischer, sondern auch der Anteil ähnlicher Aminosäuren ein (% similarity)

Homologie: Interpretation von BLAST-Ergebnissen

Wie kann ich beweisen, dass ein nicht-signifikanter Treffer homolog ist?

Wichtig: Homologie ist transitiv!

(wenn die Treffer den gleichen Teil des Proteins abdecken)

A homolog zu B & B homolog zu C → A homolog zu C !

→ **Lösung:** andere Suchsequenz verwenden! (z.B. *E.coli* statt Mensch)

PS: Nicht-Homologie lässt sich nicht beweisen…

Von der Sequenz zum Baum

Sequenzen

Orthologie!

Multiples Sequenz Alignment

Auswahl eines Evolutionsmodells

Auswahl von Methode & Algorithmus

Stammbaumberechnung
(mit/ohne statistischer Auswertung)

Multiple Sequenzalignments

Gegeben:

```
SeqA  N A F L S
SeqB  N A F S
SeqC  N A K Y L S
SeqD  N A Y L S
```

Indel:

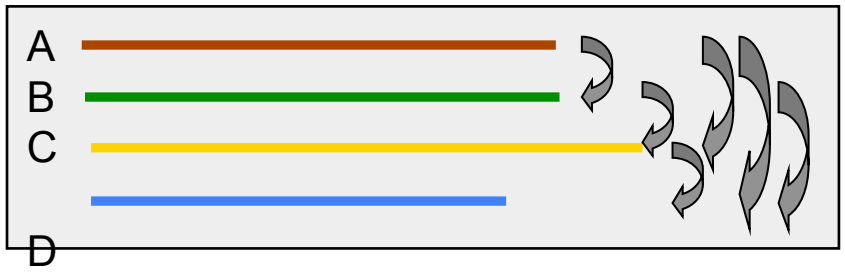
Insertion/Deletion

Gesucht:

```
SeqA  N A - F L S
SeqB  N A - F - S
SeqC  N A K Y L S
SeqD  N A - Y L S
```

Progressives MSA mit ClustalX

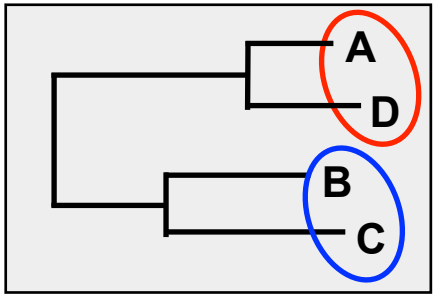
→ Sequenzen paarweise vergleichen



→ Distanzmatrix

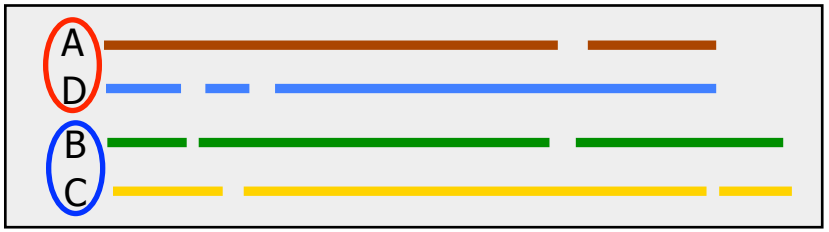
	A	B	C	D
A	-	0.75	0.89	0.27
B		-	0.45	0.82
C			-	0.77
D				-

Guide tree:

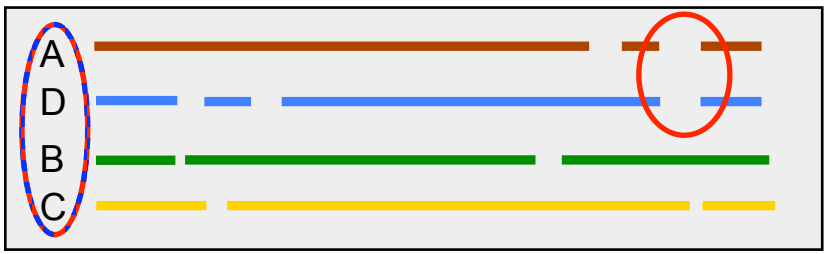


„Cluster“

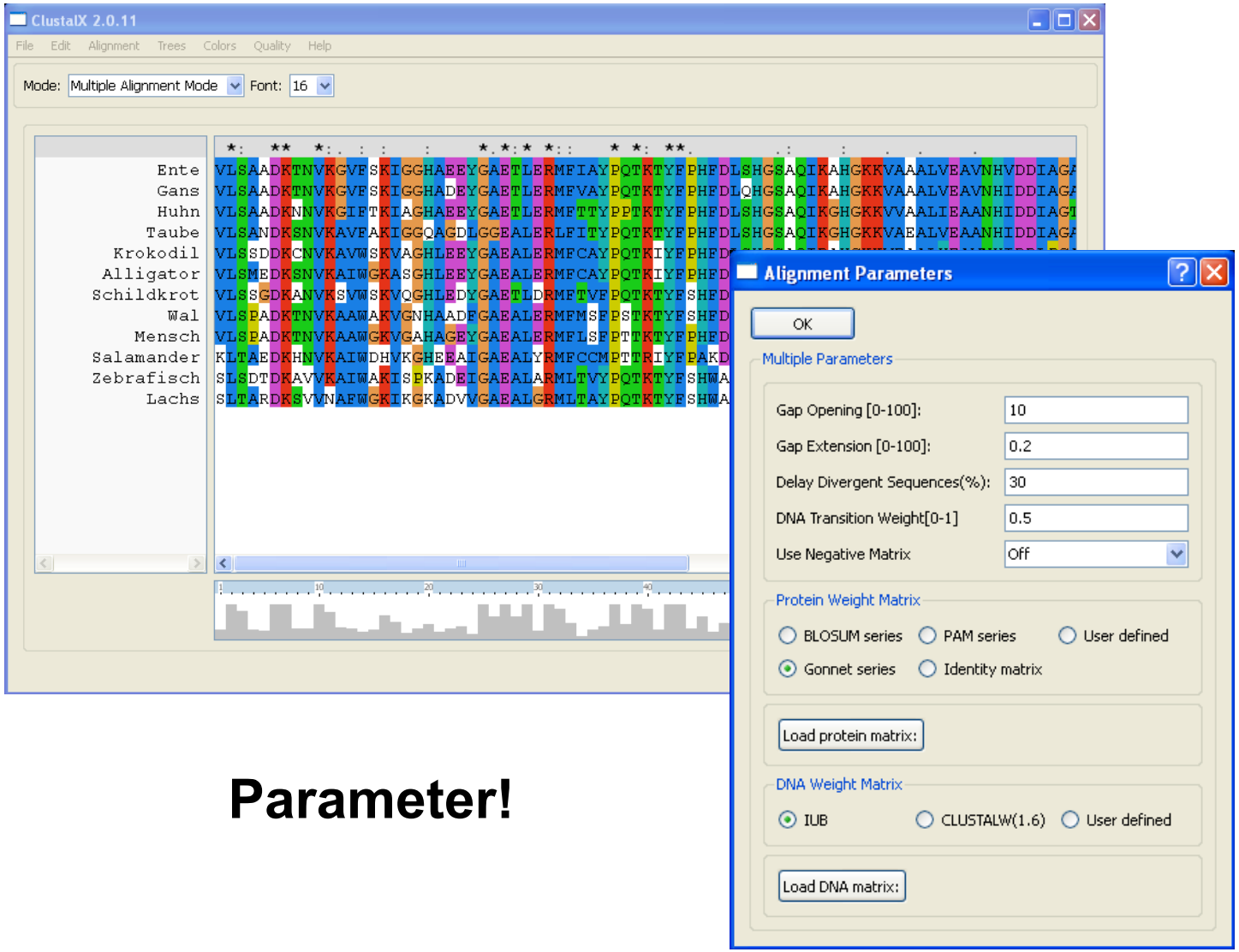
→ Alignment innerhalb der Cluster
Einfügen von Lücken (gaps)



→ Sukzessives globales Alignment
Einfügen neuer Lücken



MSA: ClustalX



Parameter!

Dateiformate - Fasta

Diese Dateien bestehen aus

1. „Headern“ (ein „>“ und die Bezeichnung der Sequenz)
2. den Sequenzen an sich

Die Sequenzen können dabei aus Nukleotiden oder Aminosäuren bestehen. Zeilenumbrüche in den Sequenzen sind erlaubt, aber nicht notwendig.

```
>TaxonXY_Seq1
-GGGGTGTGGATAGGTAGGGAGGCTGTTTATAATGTGTTGGTGACGAGCCATGCTGT
GATGATAGTGTTCTTTCTTGTAATGCCTGTTTTTATAGGAGGGTTTGGTAACTGATT
GACACCAGTTATGTTGGGTTTGAGGGATATGGCTTTACCCCGTTTGAATAACCTTAG
GTTA
>TaxonXY_Seq2
GGGAGTTTGGATGGGTAGAGAGGCCATCTATAATGTATTGGTGACTAGACATGCAGTT
>XY_Seq3
GGGAGTTTGGATGGGTAGAGAGGCCATCTATAACGTATTGGTGACTAGGCATGCAGTT
>XY_Seq4
GGGAGTTTGGATGGGTAGAGAGGCCATCTATAACGTATTGGTGACTAGCCATGCAGTT
>XY_Seq5
GGGAGTTTGGATGGGTAGAGAGGCCATTTATAACGTATTGGTGACTAGACATGCAGTT
>XY_Seq6
AGGGGTTTGGATAGGAAGGGAGGCTGTGTATAATGTTTTAGTAACTAGACACGCTGTT
>XY_Seq7
AGGGGTTTGGATAGGAAGGGAGGCTGTGTATAATGTTTTAGTAACTAGACACGCTGTT
>XY_Seq8
AGGGGTTTGGATAGGGAGGGAGGCTGTGTATAATGTGTTAGTAACTAGACACGCTGTT
```

Dateiformate - Phylip

Das Phylip-Format wird vornehmlich für Alignments verwendet (alle enthaltenen Sequenzen sind gleichlang und enthalten auch Lücken)

Angabe von Taxonzahl
und Anzahl Alignmentpositionen

Header auf 10 Zeichen limitiert!!

```
 8  85
TaxonXY_Se -GGGGTGTGG ATAGGTAGGG AGGCTGTTTA TAATGTGTTG GTGACGAGCC
TaxonXY_Se GGGAGTTTGG ATGGGTAGAG AGGCCATCTA TAATGTATTG GTGACTAGAC
XY_Seq3    GGGAGTTTGG ATGGGTAGAG AGGCCATCTA TAACGTATTG GTGACTAGGC
XY_Seq4    GGGAGTTTGG ATGGGTAGAG AGGCCATCTA TAACGTATTG GTGACTAGCC
XY_Seq5    GGGAGTTTGG ATGGGTAGAG AGGCCATTTA TAACGTATTG GTGACTAGAC
XY_Seq6    AGGGGTTTGG ATAGGAAGGG AGGCTGTGTA TAATGTTTTA GTAACTAGAC
XY_Seq7    AGGGGTTTGG ATAGGAAGGG AGGCTGTGTA TAATGTTTTA GTAACTAGAC
XY_Seq8    AGGGGTTTGG ATAGGGAGGG AGGCTGTGTA TAATGTGTTA GTAACTAGAC

           ATGCTGTGAT GATAGTGTTC TTTCTTGTAA TGCCT
           ATGCAGTTAT GATAGTATTC TTTTTAGTTA TACCA
           ATGCAGTTAT GATAGTATTC TTTTTAGTTA TACCA
           ATGCAGTTAT GATAGTATTC TTTTTAGTTA TACCA
           ATGCAGTTAT GATAGTATTC TTTTTAGTTA TACCA
           ACGCTGTTAT AATGGTGTTT TTTCTAGTAA TACCG
           ACGCTGTTAT AATGGTGTTT TTTCTAGTAA TACCG
           ACGCTGTTAT AATGGTCTTT TTTCTAGTAA TACCG
```

Sequenzen in
Zehnerblöcken,
Umbrüche nach 50
Positionen

Dateiformate - Nexus

```
#NEXUS

BEGIN DATA;
 DIMENSIONS NTAX=8 NCHAR=175;
 FORMAT DATATYPE=DNA INTERLEAVE MISSING=-;
[Name: TaxonXY_Seq1     Len:   175  Check:    0]
[Name: TaxonXY_Seq2     Len:   175  Check:    0]
[Name: XY_Seq3          Len:   175  Check:    0]
[Name: XY_Seq4          Len:   175  Check:    0]
[Name: XY_Seq5          Len:   175  Check:    0]
[Name: XY_Seq6          Len:   175  Check:    0]
[Name: XY_Seq7          Len:   175  Check:    0]
[Name: XY_Seq8          Len:   175  Check:    0]

MATRIX
TaxonXY_Seq1        -GGGGTGTGGATAGGTAGGG AGGCTGTTTATAATGTGTTG GTGACGAGCCATGCTGTGAT GATAGTGTTCTTTCTTGTAA TGCCTGTTTTTATAGGAGGG
TaxonXY_Seq2        GGGAGTTTGGATGGGTAGAG AGGCCATCTATAATGTATTG GTGACTAGACATGCAGTTAT GATAGTATTCTTTTTAGTTA TACCAGTTTTTATAGGGGGG
XY_Seq3             GGGAGTTTGGATGGGTAGAG AGGCCATCTATAACGTATTG GTGACTAGGCATGCAGTTAT GATAGTATTCTTTTTAGTTA TACCAGTTTTTATGGGGGGG
XY_Seq4             GGGAGTTTGGATGGGTAGAG AGGCCATCTATAACGTATTG GTGACTAGCCATGCAGTTAT GATAGTATTCTTTTTAGTTA TACCAGTTTTTATGGGGGGG
XY_Seq5             GGGAGTTTGGATGGGTAGAG AGGCCATTTATAACGTATTG GTGACTAGACATGCAGTTAT GATAGTATTCTTTTTAGTTA TACCAGTTTTTATGGGGGGG
XY_Seq6             AGGGGTTTGGATAGGAAGGG AGGCTGTGTATAATGTTTTA GTAACTAGACACGCTGTTAT AATGGTGTTTTTTCTAGTAA TACCGGTATTTATGGGGGGA
XY_Seq7             AGGGGTTTGGATAGGAAGGG AGGCTGTGTATAATGTTTTA GTAACTAGACACGCTGTTAT AATGGTGTTTTTTCTAGTAA TACCGGTATTTATGGGGGGA
XY_Seq8             AGGGGTTTGGATAGGGAGGG AGGCTGTGTATAATGTGTTA GTAACTAGACACGCTGTTAT AATGGTCTTTTTTCTAGTAA TACCGGTATTTATGGGGGGA

TaxonXY_Seq1        TTTGGTAACTGATTGACACC AGTTATGTTGGGTTTGAGGG ATATGGCTTTACCCCGTTTG AATAACCTTAGGTTA
TaxonXY_Seq2        TTTGGTAACTGACTAATACC TGTTATATTAGGGTTGAGGG ACATGGCCCTACCCCGCTTA AATAATTTGAGGTTG
XY_Seq3             TTTGGTAACTGACTAATACC TGTTATATTAGGGTTGAGGG ACATGGCCCTACCCCGTTTA AATAATTTGAGGTTG
XY_Seq4             TTTGGTAACTGACTAATACC TGTTATATTAGGGTTGAGGG ACATGGCCCTACCCCGTTTA AATAATTTGAGGTTG
XY_Seq5             TTTGGTAACTGATTAATACC TGTTATATTAGGGTTGAGGG ACATGGCCCTACCTCGCTTA AATAATTTGAGGTTG
XY_Seq6             TTCGGTAATTGGCTTATGCC AGTGATGTTAGGGTTAAGGG ATATGGCCCTCCCCCGACTG AACAATTTGAGGCTT
XY_Seq7             TTCGGTAATTGGCTTATGCC AGTGATGTTAGGGTTAAGGG ATATGGCCCTCCCCCGACTG AACAATTTGAGGCTT
XY_Seq8             TTCGGTAATTGGCTCATGCC AGTGATGTTGGGGTTAAGGG ATATGGCCCTTCCCCGATTG AACAATTTAAGGCTT


 ;
 END;
```

Angabe von Taxonzahl
und Anzahl Alignmentpositionen

Sequenzen in Zwanzigerblöcken,
Umbrüche nach 100 Positionen

Von der Sequenz zum Baum

Sequenzen — Orthologie!

Multiples Sequenz Alignment — gap penalties, Matrizen

Auswahl eines Evolutionsmodells

Auswahl von Methode & Algorithmus

Stammbaumberechnung
(mit/ohne statistischer Auswertung)

Evolutionsmodelle

DNA

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

Protein

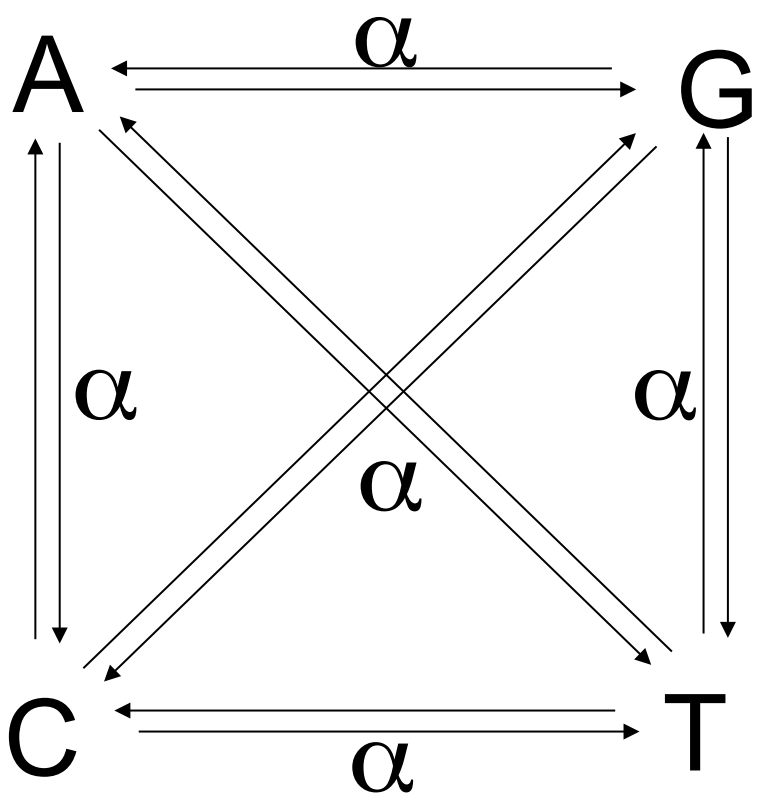
- 1970s: Dayhoff et al. (PAM family)
- 1992: Henikoff & Henikoff (BLOSUM family)
- 1992: Jones, Taylor & Thornton (JTT)
- 1996: Adachi & Hasegawa (mtREV)
- 1998: Yang et al. (mtMam)
- 2000: Adachi et al. (cpREV)
- 2001: Whelan and Goldman (WAG)
- 2002: Dimmic et al. (rtREV)
- 2007: Abascal et al. (mtArt)
- 2008: Le & Gascuel (LG)
- uvm…

Modell nach Jukes & Cantor

→ alle Austausche sind gleich wahrscheinlich!

→ Substitutionsrate = α für alle Änderungen

→ **Ein-Parameter-Modell**

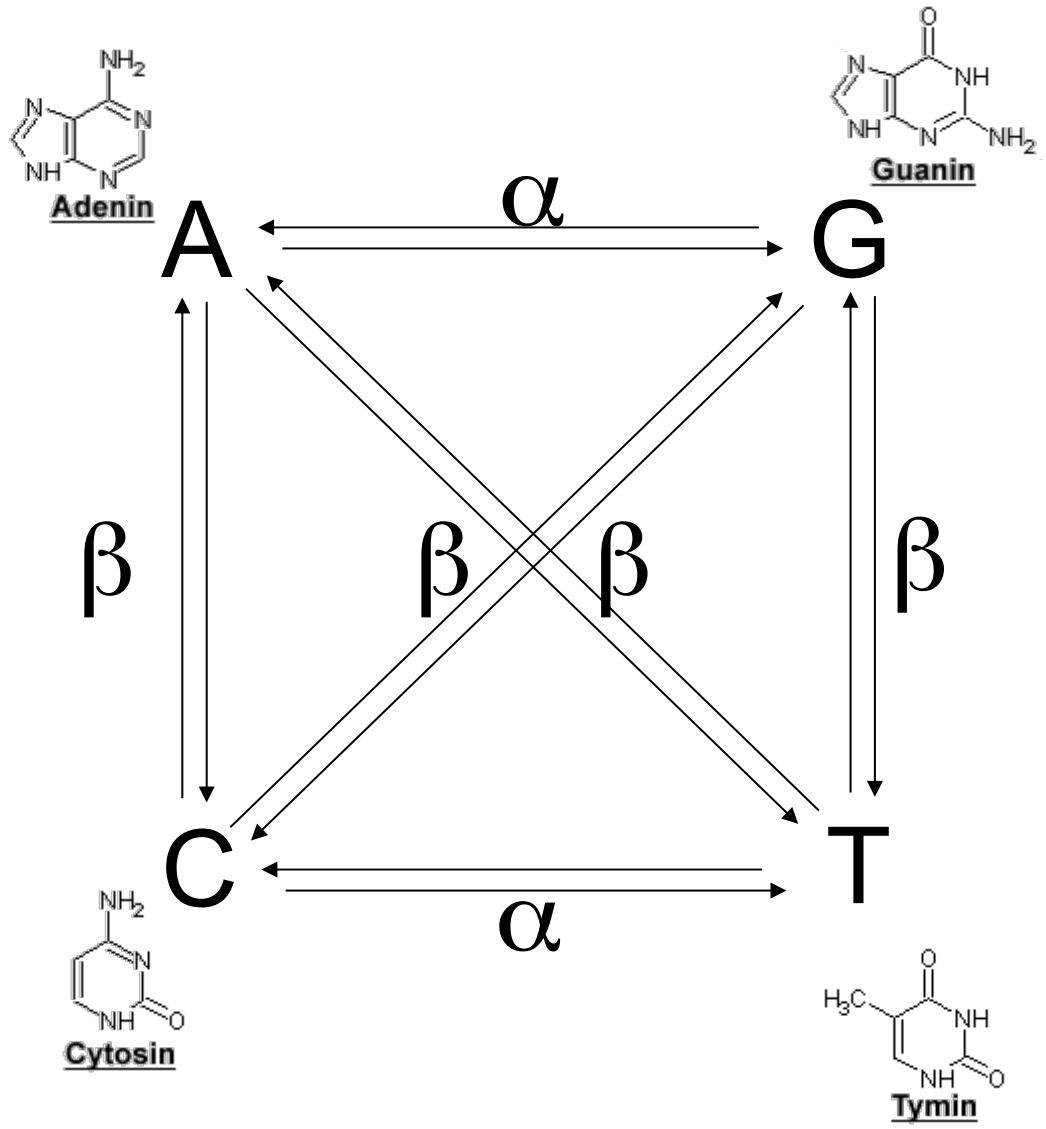


Kimura 2-Parameter-Modell

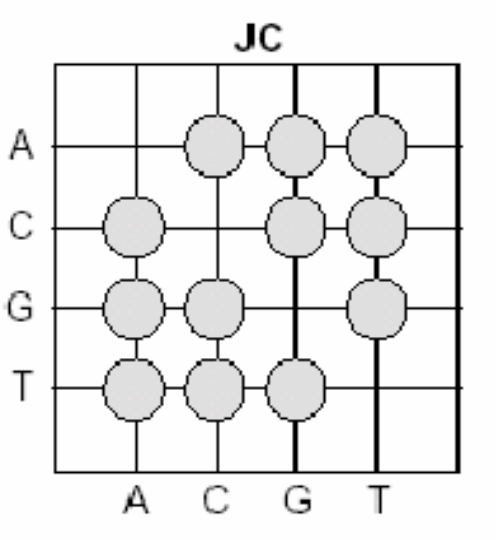
→ unterschiedliche Wahrscheinlichkeiten für Transitionen und Transversionen

Transition: α
Transversion: β

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

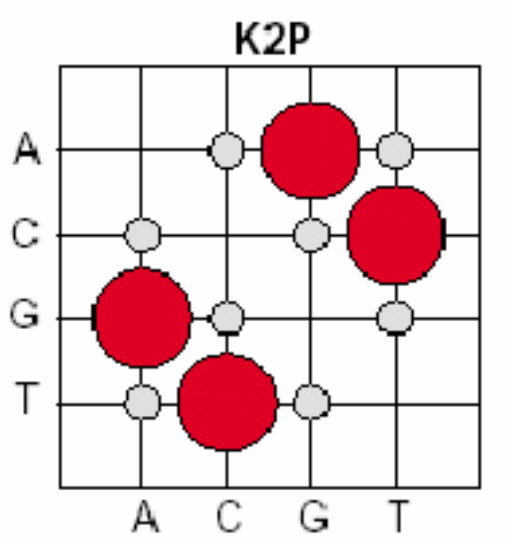


Evolutionsmodelle (DNA)



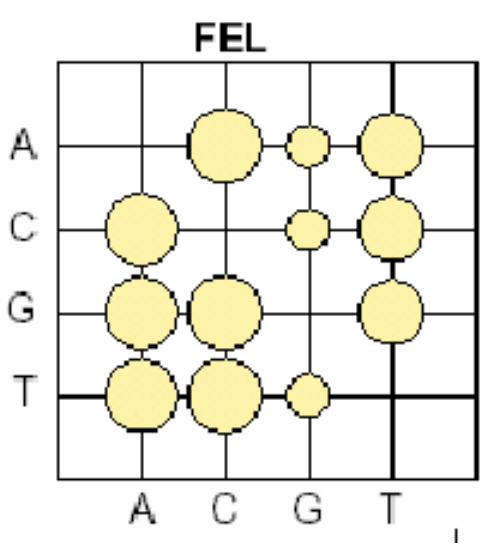
- alle Substitutionen gleich häufig

- erwartete Nukleotid-zusammensetzung identitsch



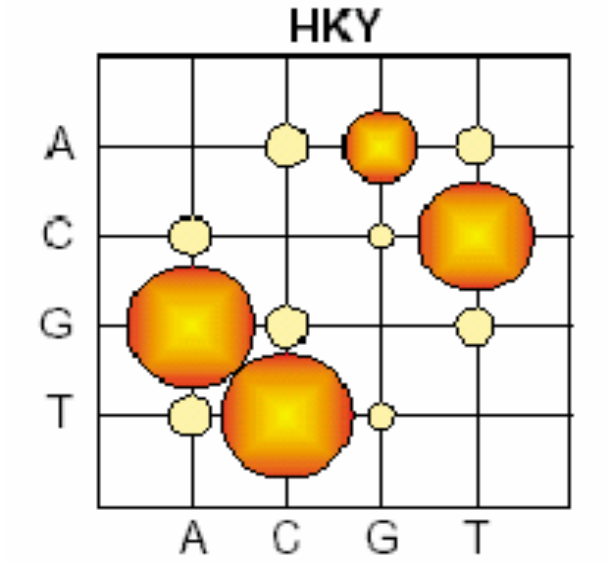
-Transitionen und Transversionen unterschiedlich häufig

- erwartete Nukleotidzusammen-setzung identitsch

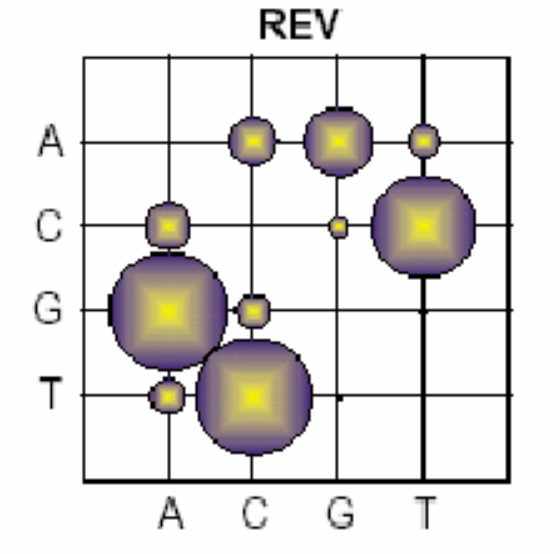


- erwartete Nukleotid-zusammensetzung unterschiedlich

Evolutionsmodelle (DNA)

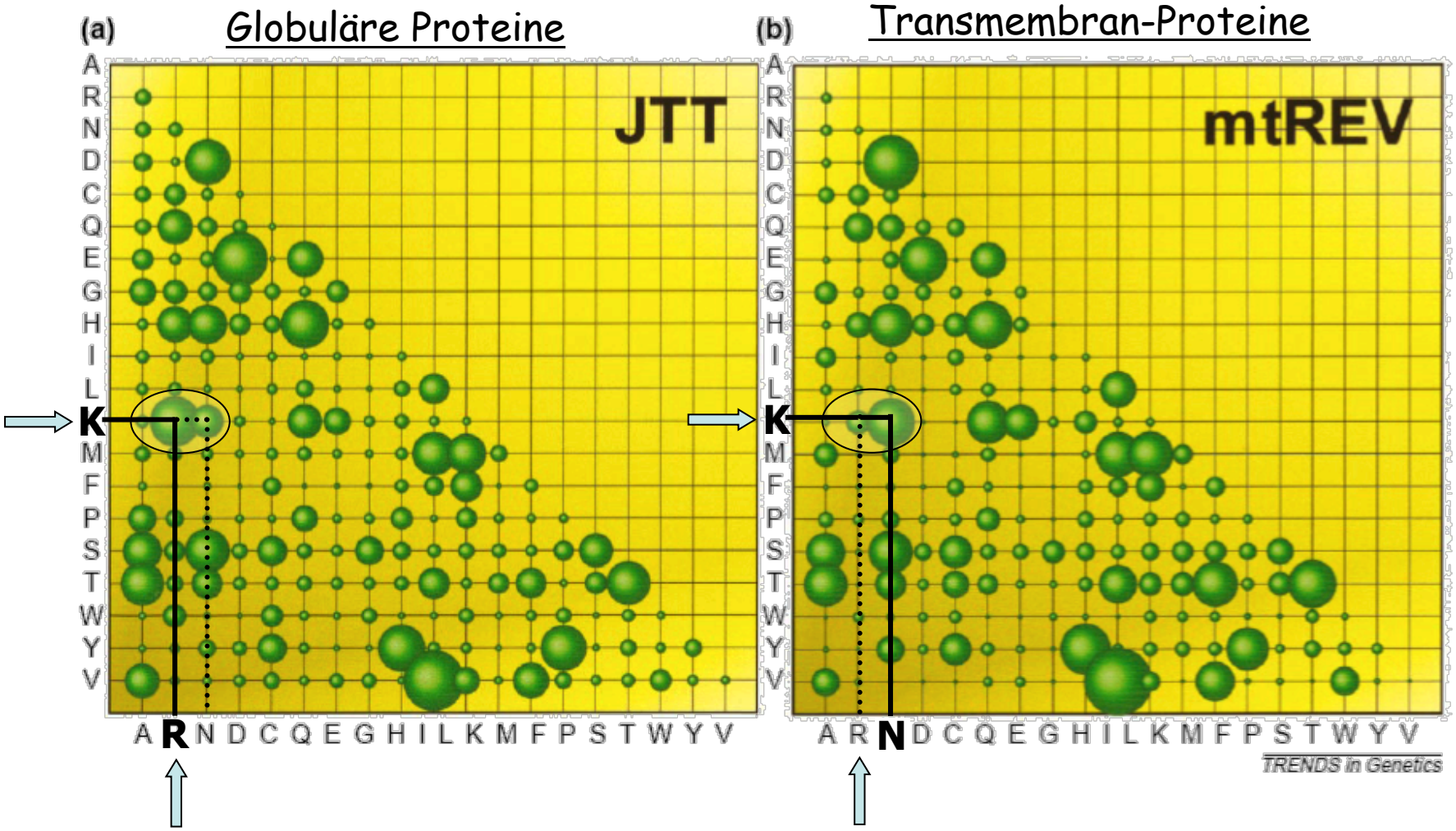


Transitionen und Transversionen und Nukleotidzusammensetzung sind unterschiedlich häufig



Alle Parameter (Austausche und Austauschrichtungen) und Nukleotidzusammensetzung dürfen variieren

Spezialisierte Evolutionsmodelle für bestimmte Protein-Typen



Variation der Substitutionsraten

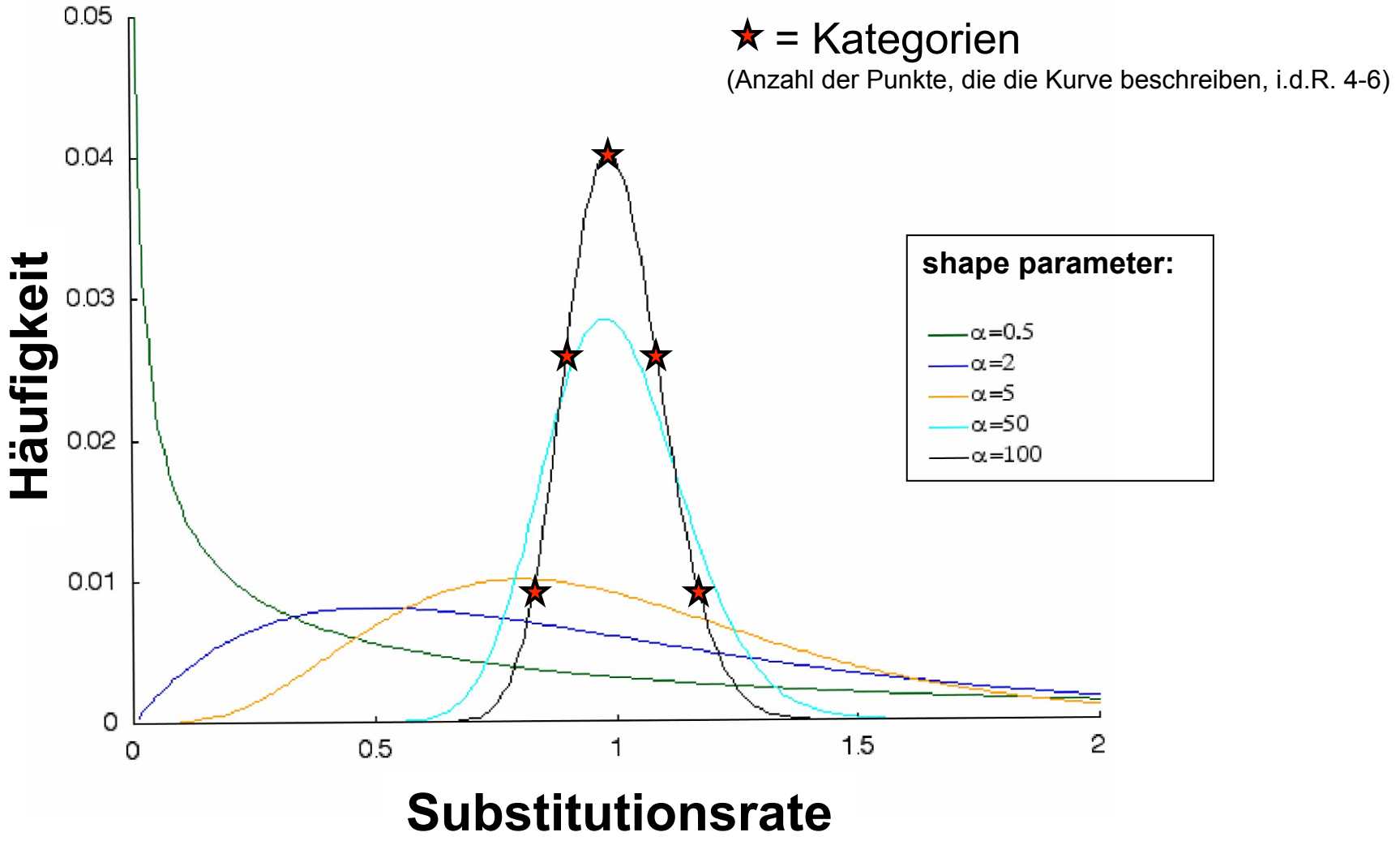
- Jede Säule im MSA hat prinzipiell eine eigene Substitutionsrate („***among site rate variation***“ → ∞)

- Annahme: Die Variation der Rate lassen sich über eine ***Gamma-*** {Γ}***Verteilung*** beschreiben

- der ***"shape"-parameter α*** gibt die relative Verteilung der unterschiedlichen Substitutionsraten inder Gamma-Verteilung -wieder

α gross	→ geringe Streuung der Substitutionsraten
α klein	→ grosse Streuung der Substitutionsraten

Variation der Substitutionsraten



Wie komplex soll das Modell sein?

Je komplexer das Modell (also je mehr Annahmen), desto genauer und realistischer unsere Berechnung der Substitutionsraten

ABER

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

Das heißt?

→ möglichst "gute" Daten
→ möglichst einfaches Modell, das dennoch exakt ist

Bestimmung eines geeigneten Evolutionsmodells

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

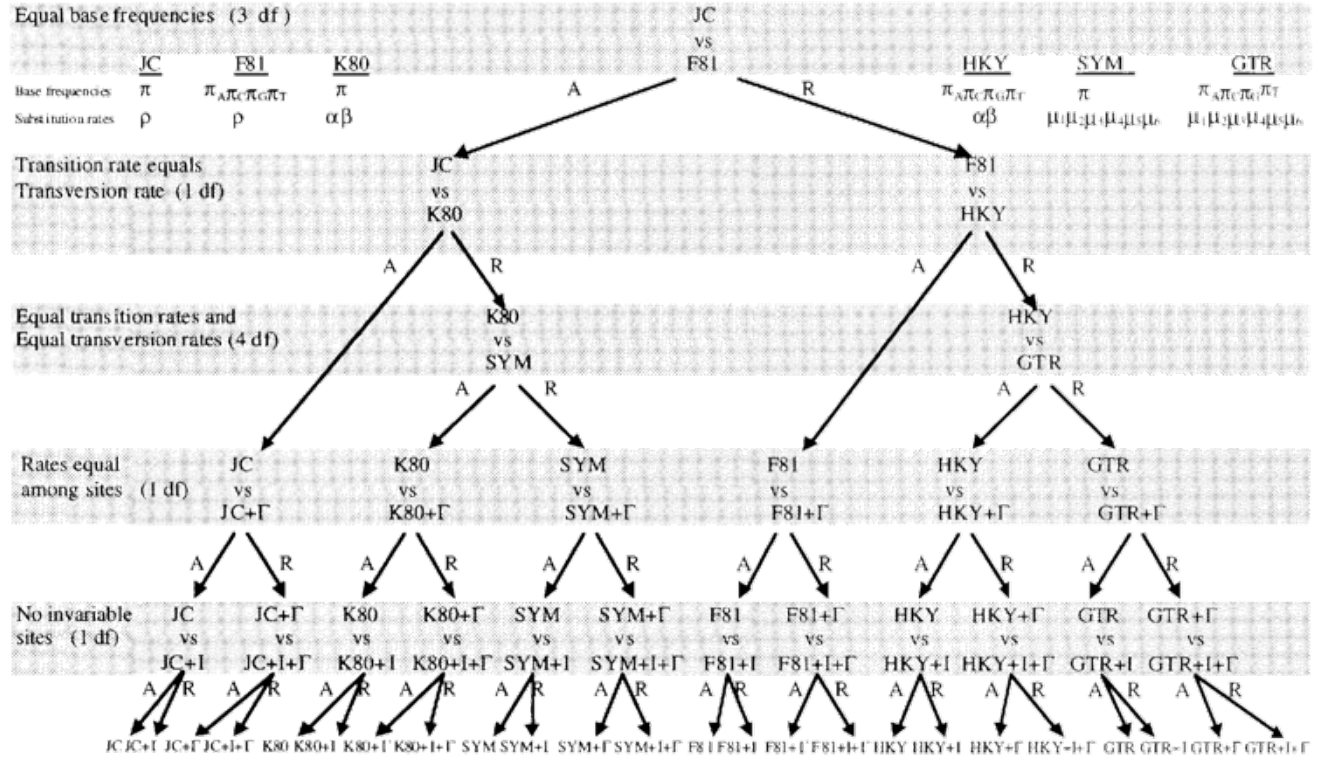


Fig. 1. Hierarchical hypothesis testing in MODELTEST. At each level the null hypothesis (upper model) is either accepted (A) or rejected (R). The models of DNA substitution are: JC (Jukes and Cantor, 1969), K80 (Kimura, 1980), SYM (Zharkikh, 1994), F81 (Felsenstein, 1981), HKY (Hasegawa *et al.*, 1985), 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.

hierarchischer Likelihood ratio test (→ hLRT)

- schrittweise kompliziertere Modelle

- signifikant besser als das vorherige Modell?

Bestimmung eines geeigneten Evolutionsmodells

Programme wie ModelTest, FindModel, ProtTest sortieren die Ergebnisse nach bestimmten Kriterien:

AIC (Akaike Information Criterion)

$$AIC = -2 \ln L + 2K$$

AICc (corrected Akaike Information Criterion)

$$AICc = AIC + 2K(K+1)/(N-K-1)$$

BIC (Bayesian Information Criterion)

$$BIC = -2\ln L + K\log N$$

L = model likelihood, K = number of estimatable parameters, N = sample size

Bestimmung eines geeigneten Evolutionsmodells

Neben dem Substitutionsmodell an sich werden mitunter noch weitere Einstellungen empfohlen:

Variation der Substitutionsraten

+G Gamma-Verteilung der Substitutionsraten

+I Anteil an invariablen Stellen

Aminosäure/Nukleotid-Frequenzen

+F die Gleichgewichts-Frequenzen werden aus dem Datensatz abgeschätzt

Von der Sequenz zum Baum

Sequenzen — Orthologie!

Multiples Sequenz Alignment — gap penalties, …

Auswahl eines Evolutionsmodells — hLRT, gamma-shape, …

Auswahl von Methode & Algorithmus

Stammbaumberechnung
(mit/ohne statistischer Auswertung)

Methoden für die Stammbaumerstellung

1. Distanz-orientierte Methoden

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

→ **Sequenzen werden in Distanzmatrix konvertiert**

2. Charakter-orientierte Methoden

- **Maximum Parsimony**
- **Maximum Likelihood**
- **Bayes'sche Methoden**

→ **jede Position wird als informative Einheit betrachtet**

Berechnung einer Distanzmatrix

Distanz = durchschnittliche Änderung pro Position

→ Berechnen der paarweisen Abstände zwischen den Sequenzen

```
Sequenz 1   TATAAGCATGACTAGTAAGC
Sequenz 2   TATTAGCATGACTGGTAACC
Sequenz 3   TATTGGCATGACTAGCAGGC
Sequenz 4   TGTTGCCACGATTAGCTACC
Sequenz 5   CGTAGCTATGACCAACGGGC
```

Beispiel: Seq1 vs. Seq2

3 von 20 Positionen verändert

	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

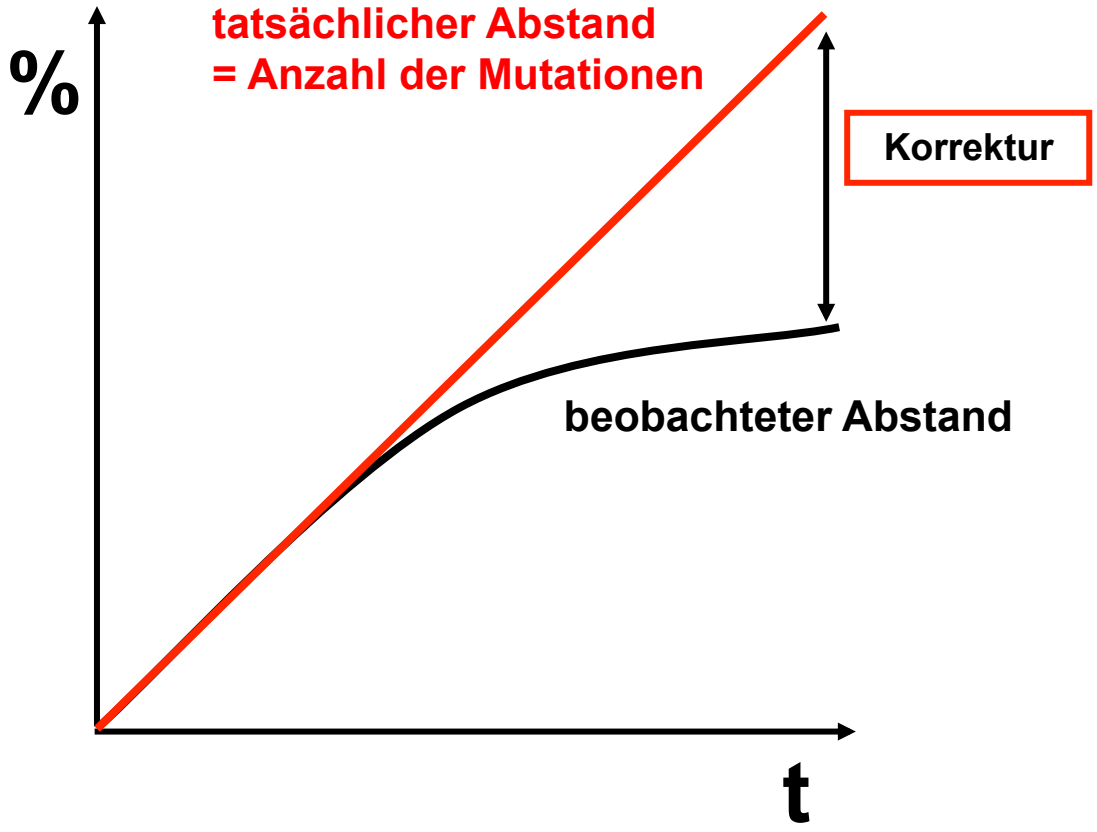
Korrektur der Distanzen

Warum?

multiple Austausche,
Rückmutationen, etc.

Wie?

Evolutionsmodelle!

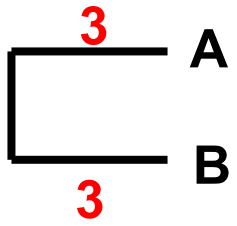


Distanzmethoden: UPGMA

UPGMA Unweighted Pair-Group Method with Arithmetric Means

	A	B	C	D
OTU A	0	6	10	18
OTU B		0	12	20
OTU C			0	19
OTU D				0

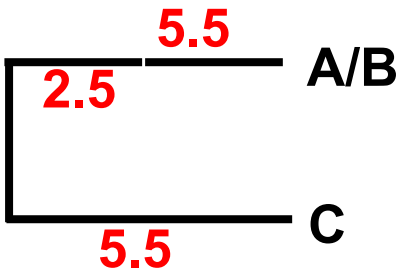
$\frac{d_{AB}}{2} = 3$



$\frac{d\,AC + d\,BC}{2}$ $\frac{d\,AD + d\,BD}{2}$

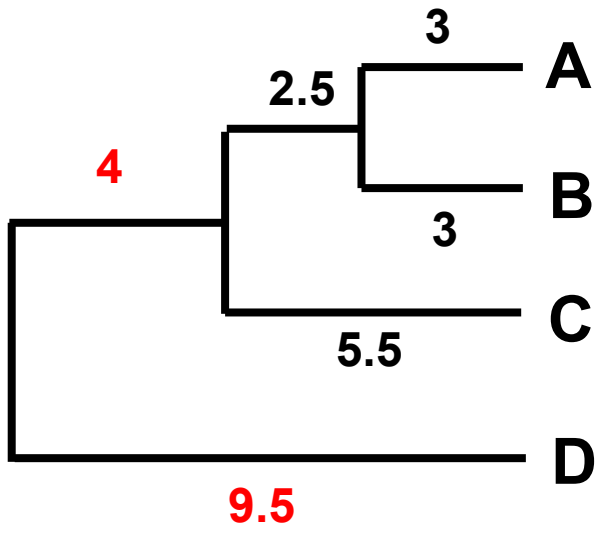
	A/B	C	D
OTU A/B	0	11	19
OTU C		0	19
OTU D			0

$\frac{d_{(AB)C}}{2} = 5{,}5$



Distanzmethoden: UPGMA

	A/B/C	D
Sequenz A/B/C	0	19
Sequenz D		0



- nimmt konstante Evolutionsraten an
- Außengruppe wird „automatisch“ bestimmt

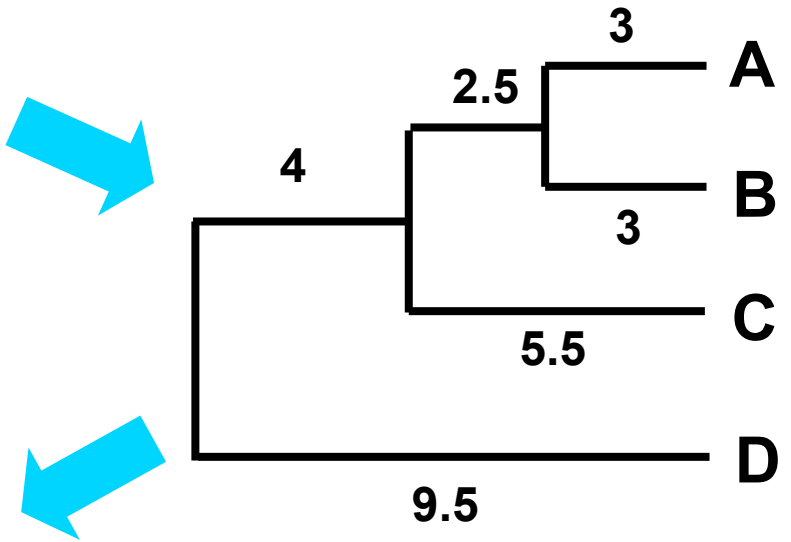
Distanzmethoden: UPGMA

Ausgangsmatrix

	A	B	C	D
OTU A	0	6	10	18
OTU B		0	12	20
OTU C			0	19
OTU D				0

rekonstruierte Matrix

	A	B	C	D
OTU A	0	6	11	19
OTU B		0	11	19
OTU C			0	19
OTU D				0



Rekonstruktion der Matrix mißlingt!

Distanzmethoden: Neighbor Joining

Grundlage: „Minimum Evolution“

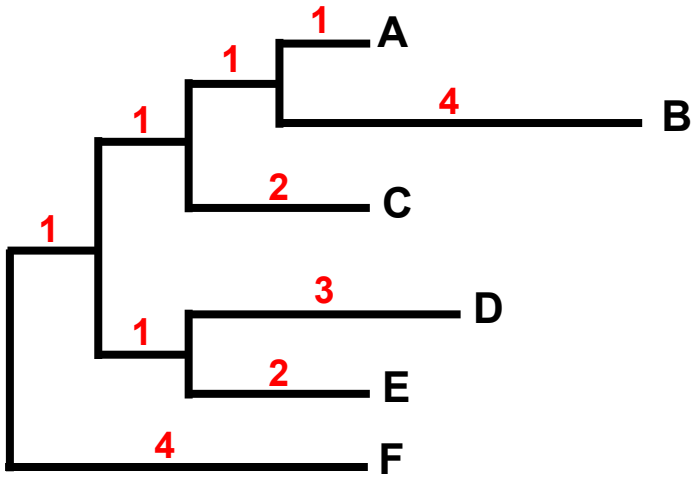
→ Minimierung der Summe aller Astlängen im Baum

→ Astlänge entspricht Distanz zwischen taxa

Um den kürzesten Baum zu finden müssten ALLE Bäume berechnet werden, dies ist mit steigender taxa-Anzahl quasi unmöglich

→ ***Neighbor Joining (NJ)*** ist ein heuristischer Clustering-Algorithmus

Distanzmethoden: Neighbor Joining



Paarweise Distanzen:

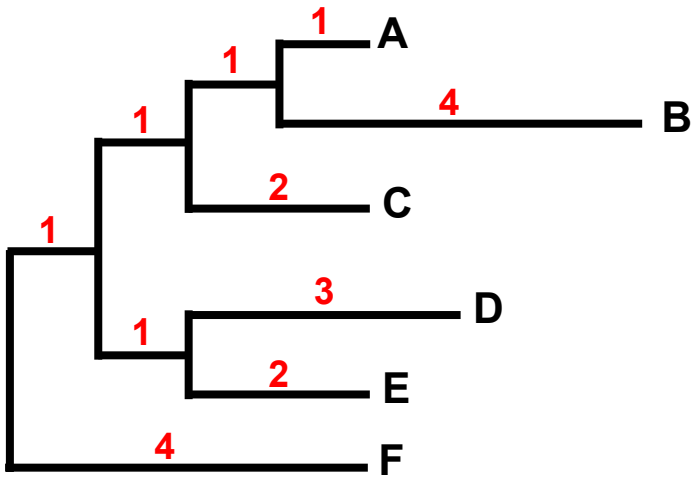
	A	B	C	D	E	F
A	0	5	4	7	6	8
B		0	7	10	9	11
C			0	7	6	8
D				0	5	9
E					0	8
F						0

→ Was würde ***UPGMA*** tun?

→ Cluster A+C

→ Was tut ***NJ*** um dies zu verhindern??

Distanzmethoden: Neighbor Joining



Paarweise Distanzen:

	A	B	C	D	E	F
A	0	5	4	7	6	8
B		0	7	10	9	11
C			0	7	6	8
D				0	5	9
E					0	8
F						0

Schritt 1: Berechne die ***„Gesamtdistanz“ r*** für jedes taxon

(= summierte Abstände eines taxons zu allen anderen taxa)

r_A = 5 + 4 + 7 + 6 + 8 = **30**

r_B = 5 + 7 + 10 + 9 + 11 = **42**

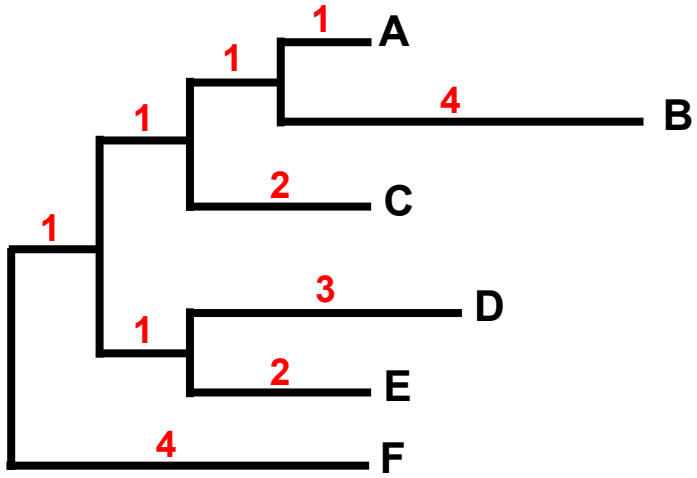
r_C = 4 + 7 + 7 + 6 + 8 = **32**

r_D = 7 + 10 + 7 + 5 + 9 = **38**

r_E = 6 + 9 + 6 + 5 + 8 = **34**

r_F = 8 + 11 + 8 + 9 + 8 = **44**

Distanzmethoden: Neighbor Joining



Paarweise Distanzen

	A	B	C	D	E	F
A	0	5	4	7	6	8
B	-13	0	7	10	9	11
C	-11,5		0	7	6	8
D				0	5	9
E					0	8
F						0

Gesamtdistanzen

$r_A = 30$ $r_D = 38$
$r_B = 42$ $r_E = 34$
$r_C = 32$ $r_F = 44$

Schritt 2: Erstelle eine „***Raten-korrigierte Distanzmatrix“*** mit Hilfe dieser Gesamtdistanzen

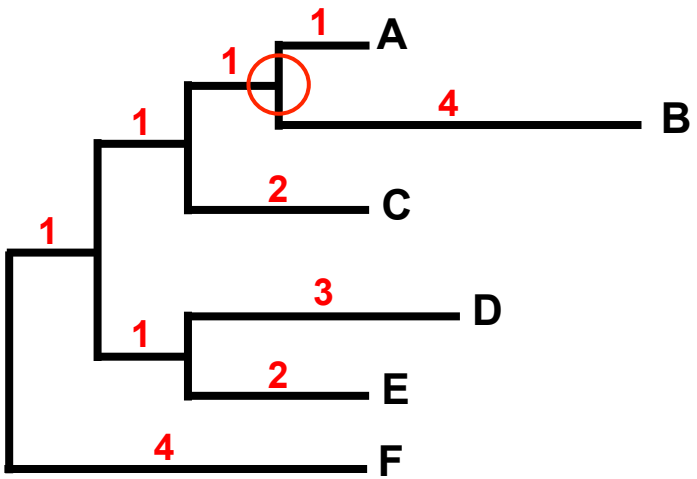
$M_{AB} = d_{AB} - (r_A + r_B) / (N-2)$

$M_{AC} = d_{AC} - (r_A + r_C) / (N-2) = 4 - (30 + 32) / 4 = \underline{-11,5}$

$M_{AD} = \ldots$

N = Anzahl der taxa

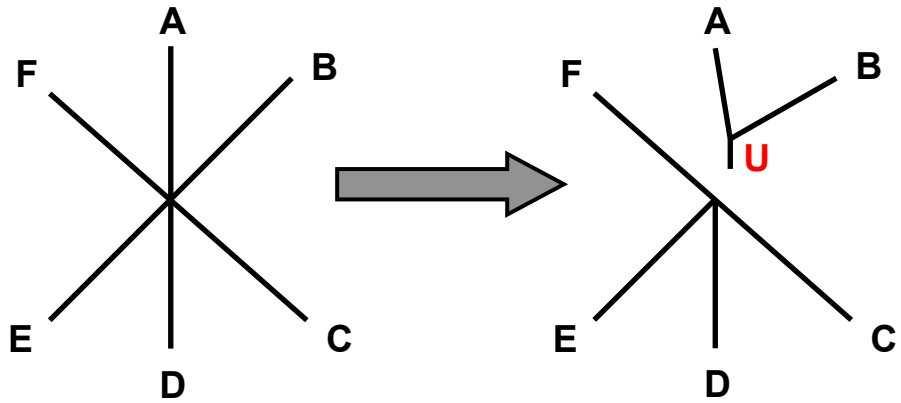
Distanzmethoden: Neighbor Joining



Ratenkorrigierte Distanzmatrix

	A	B	C	D	E	F
A	0	5	4	7	6	8
B	-13	0	7	10	9	11
C	-11,5	-11,5	0	7	6	8
D	-10	-10	-10,5	0	5	9
E	-10	-10	-10,5	-13	0	8
F	-10,5	-10,5	-11	-11,5	-11,5	0

Schritt 3: Gruppiere 2 taxa, für die M minimal ist → ***neuer Knoten U***



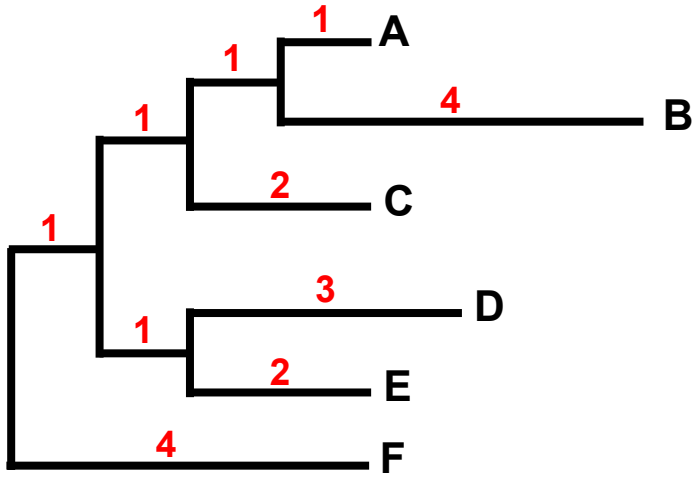
Die Astlängen gehen nicht verloren!

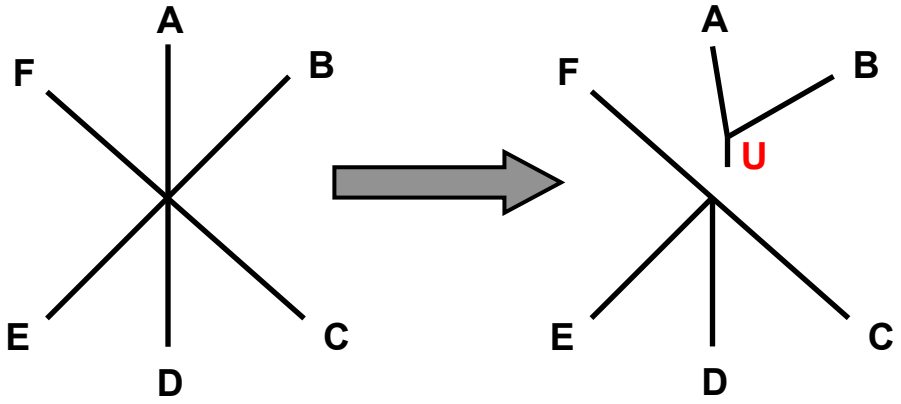
$S_{AU} = d_{AB}/2 + (r_A - r_B)/2*(N-2) = 1$

$S_{BU} = d_{AB}/2 + (r_B - r_A)/2*(N-2) = 4$

$[\ S_{BU} = d_{AB} - S_{AU} = 4\ ]$

Distanzmethoden: Neighbor Joining





Schritt 4: Berechne eine neue Distanzmatrix mit dem ***neuen Knoten U***

$d_{UC} = (d_{AC} + d_{BC} - d_{AB})/2 = (4 + 7 - 5)/2 = 3$

$d_{UD} = (d_{AD} + d_{BD} - d_{AB})/2 = (7 + 10 - 5)/2 = 6$

$d_{UE} = \ldots = 5$

$d_{UF} = \ldots = 7$

Paarweise Distanzen:

	U	C	D	E	F
U	0	3	6	5	7
C		0	7	6	8
D			0	5	9
E				0	8
F					0

Distanzmethoden: Neighbor Joining

Paarweise Distanzen

	U	C	D	E	F
U	0	3	6	5	7
C		0	7	6	8
D			0	5	9
E				0	8
F					0

Gesamtdistanzen

$r_U = 21$ $r_D = 27$

$r_E = 24$

$r_C = 24$ $r_F = 32$

Ratenkorrigierte Distanzmatrix

	U	C	D	E	F
U	0	3	6	5	7
C	-12	0	7	6	8
D	-10	-11	0	5	9
E	-10	-10	-12	0	8
F	-10,7	-10,7	-10,7	-10,7	0

→ Berechne Gesamtdistanzen r

→ Erstelle Ratenkorrigierte Distanzmatrix

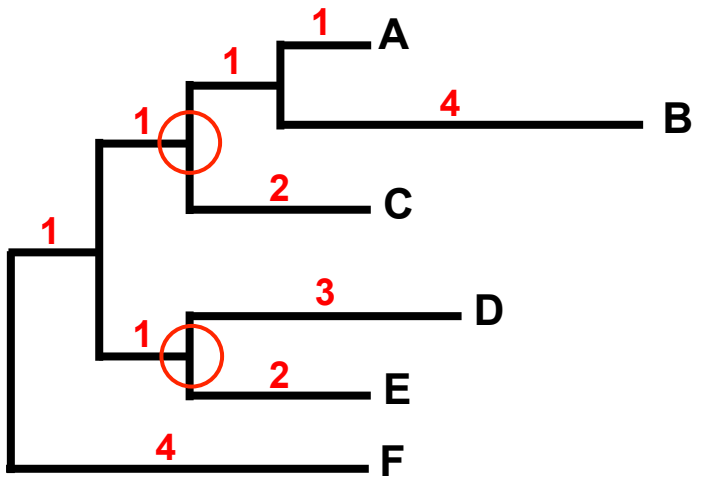
→ Gruppiere taxa, für die M minimal ist

→ Definiere neuen Knoten V

→ Speichere die Astlängen

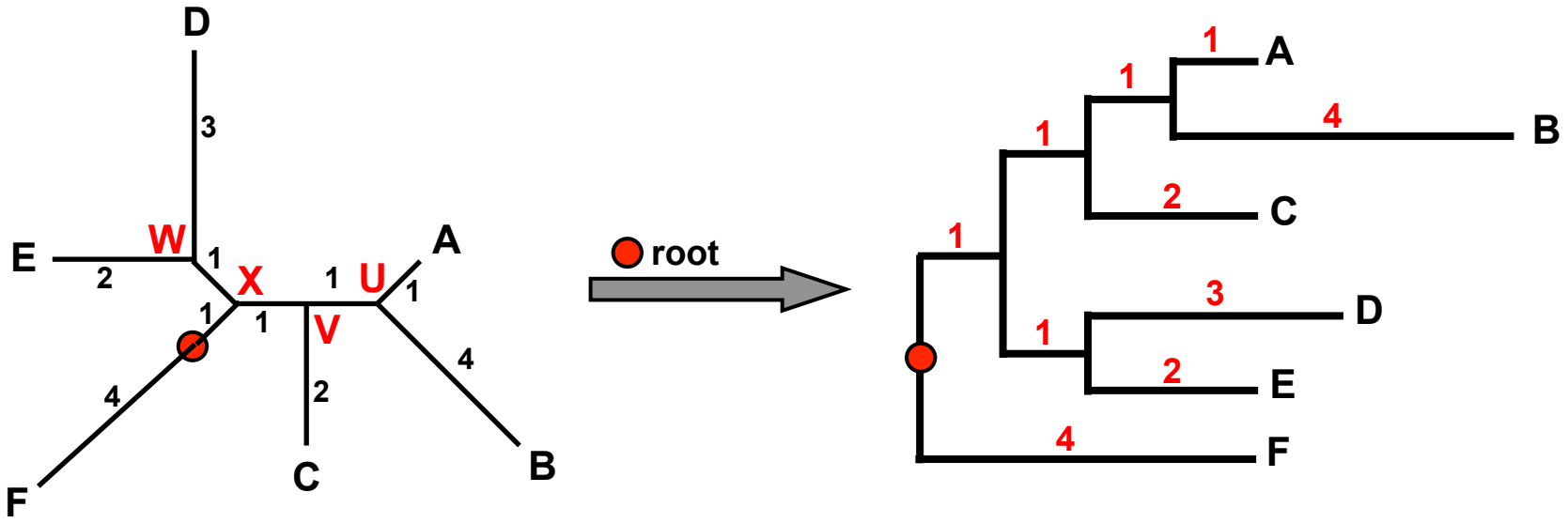
→ Berechne neue Distanzmatrix

→ …



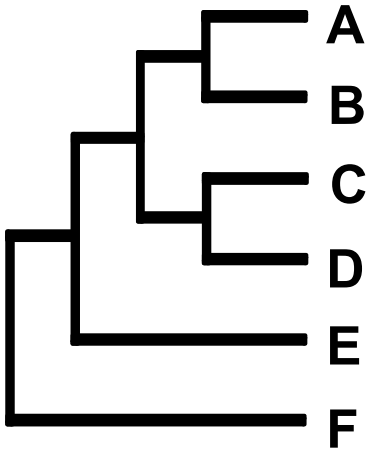
Distanzmethoden: Neighbor Joining

NJ liefert einen ***ungewurzelten*** Baum!



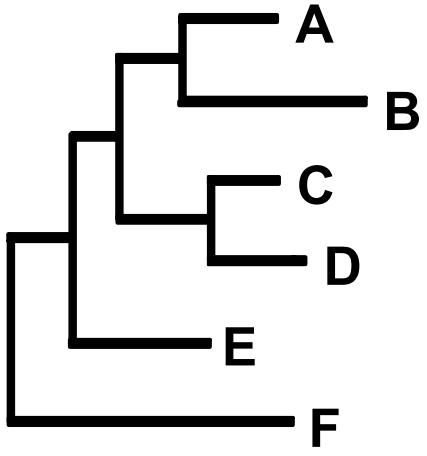
Durch Festlegen einer ***Außengruppe*** bekommt der Baum eine evolutionäre Richtung

Distanzmethoden: UPGMA vs. NJ



UPGMA

Unweighted Pair-Group Method with Arithmetric Means



NJ

Neighbor Joining

Aussengruppe festgelegt	⟷	Aussengruppe wählbar
konstante Evolutionsrate	⟷	unterschiedliche Evolutionsraten
Astlängenverlust	⟷	Kein Astlängenverlust
Keine Matrixrekonstruktion möglich	⟷	Matrixrekonstruktion möglich

Methoden für die Stammbaumerstellung

1. Distanz-orientierte Methoden

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

→ **Sequenzen werden in Distanzmatrix konvertiert**

2. Charakter-orientierte Methoden

- **Maximum Parsimony**
- **Maximum Likelihood**
- **Bayes'sche Methoden**

→ **jede Position wird als informative Einheit betrachtet**

Maximum Parsimony

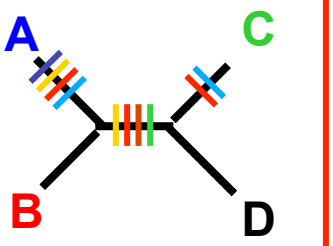
Annahme die Evolution ist sparsam

Gesucht der Baum, der die wenigsten Mutationsereignisse erfordert

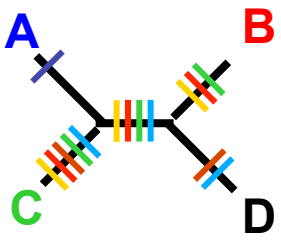
	Position								
Sequenz	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	G	C	A
B	A	G	C	C	G	T	G	C	G
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	G
					*		*		*

4 taxa

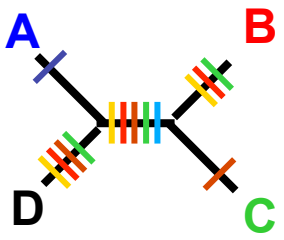
→ 3 mögliche Stammbäume



10 Mutationen



15 Mutationen

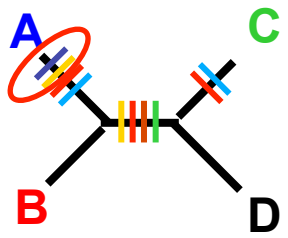


14 Mutationen

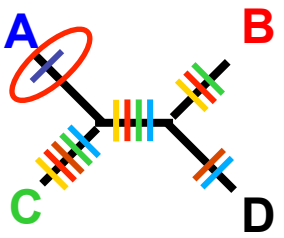
Maximum Parsimony

Achtung! Nicht jede Mutation läßt einen Schluß auf die „richtige“ Topologie zu!

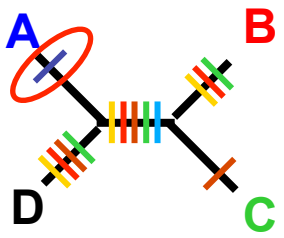
	Position								
Sequenz	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	G	C	A
B	A	G	C	C	G	T	G	C	G
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	G
					*		*		*



10 Mutationen



15 Mutationen



14 Mutationen

Parsimony-informative Positionen (*)

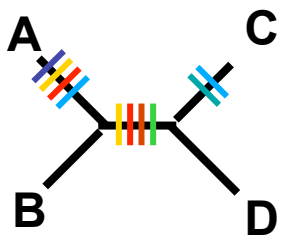
Wenigstens zwei verschiedene Charaktere an dieser Position, wobei jeder dieser Charaktere wenigstens in zwei der Sequenzen vorkommt

Maximum Parsimony

ABER Ort der Mutation nicht immer eindeutig definiert

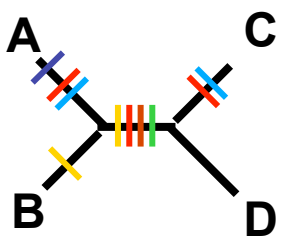
→ Parsimony berechnet zunächst **keine Astlängen**!

	Position								
Sequenz	1	2	3	4	5	6	7	8	9
A	A	A	G	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



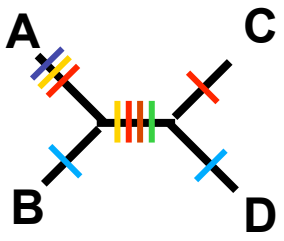
10 Mutationen

=



10 Mutationen

=



10 Mutationen

= .....

Maximum Parsimony für Proteine

1. Modell (z.B. PAUP)

- Alle Substitutionen sind gleich wahrscheinlich

- *Beispiel* Ile → Met ≡ Ile → Ala ≡ Ile → Trp ... **je 1 Schritt**

2. Modell (PROTPARS-Modell in PHYLIP)

- legt genetischen Code zugrunde
- "silent site mutations" werden ignoriert

- *Beispiel*	Ile → Met:	ATA/C/T → ATG	**ein Schritt**
	Ile → Ala:	ATA/C/T → GCN	**zwei Schritte**
	Ile → Trp:	ATA/C/T → TGG	**drei Schritte**

Maximum Parsimony

Nachteile

- empfindlich bei stark unterschiedlichen Evolutionsraten

  „Long Branch Attraction“

- Astlängenberechnung erfordert separaten Schritt

- Evolutionsmodelle sind nur eingeschränkt anwendbar

- hoher Rechenaufwand bei >20 taxa

- nur „Parsimony-informative“ Positionen können Auskunft über die Topologie geben!

Die Charakter-basierten Methoden ***Maximum Likelihood***

und ***Bayes*** werden im Rahmen des Phylogenomik-Projekts erläutert…

Distanz vs. Charaktermethoden

Distanzmethoden (UPGMA, NJ)

- Sukzessive Rekonstruktion des Stammbaums (Clustering-Algorithmus)

- UPGMA: „gemittelte“, konstante Evolutionsraten (Molekulare Uhr)

- NJ: „echte“ Evolutionsraten

Charaktermethoden (MP, ML)

- Stammbaum vorgegeben („tree space“), „hill-climbing“-Algorithmus

- Analyse aller Möglichkeiten, diesen Stammbaum mit den vorhandenen Daten zu erklären

- MP: kürzester (sparsamster) Baum

- ML: zutreffendster Baum gemäß Evo-Modell

Von der Sequenz zum Baum

Sequenzen — Orthologie!

Multiples Sequenz Alignment — gap penalties, …

Auswahl eines Evolutionsmodells — hLRT, gamma shape, …

Auswahl von Methode & Algorithmus — Distanz- vs. Charaktermethoden

Stammbaumberechnung
(mit/ohne statistischer Auswertung)

Statistische Bewertung

…oder… „Wie gut oder glaubhaft ist mein Stammbaum?“

→ Die häufigste Methode ist ***„Bootstrapping“***

- hierbei werden aus dem Alignment sogenannte Pseudosamples (i.d.R. 100 Stück) gleicher Länge erstellt
- man erhält 100 Bäume aus 100 Pseudo-Alignments…

Orginalsequenzen

	Position								
Sequence	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

	Position								
Sequence	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

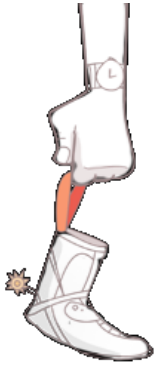
...

Statistische Bewertung

Bootstrapping

→ 100 Pseudosamples (Alignments)

→ 100 Stammbäume



→ ***Consensusbaum***

- strict consensus — nur Gruppen, die in ALLEN Bäumen monophyletisch waren, werden aufgelöst angezeigt

- majority rule consensus — Gruppen, die in DEN MEISTEN Bäumen monophyletisch waren, werden aufgelöst angezeigt

- support values — werden für die jeweilige monophyletische Gruppe angegeben

Statistische Bewertung

Jackknifing

→ „Ziehen ohne Zurücklegen“

→ entstehende Pseudosamples sind nicht so lang wie das Originalalignment (i.d.R. N/2)

→ auch hier gewöhnlich 100 Pseudosamples

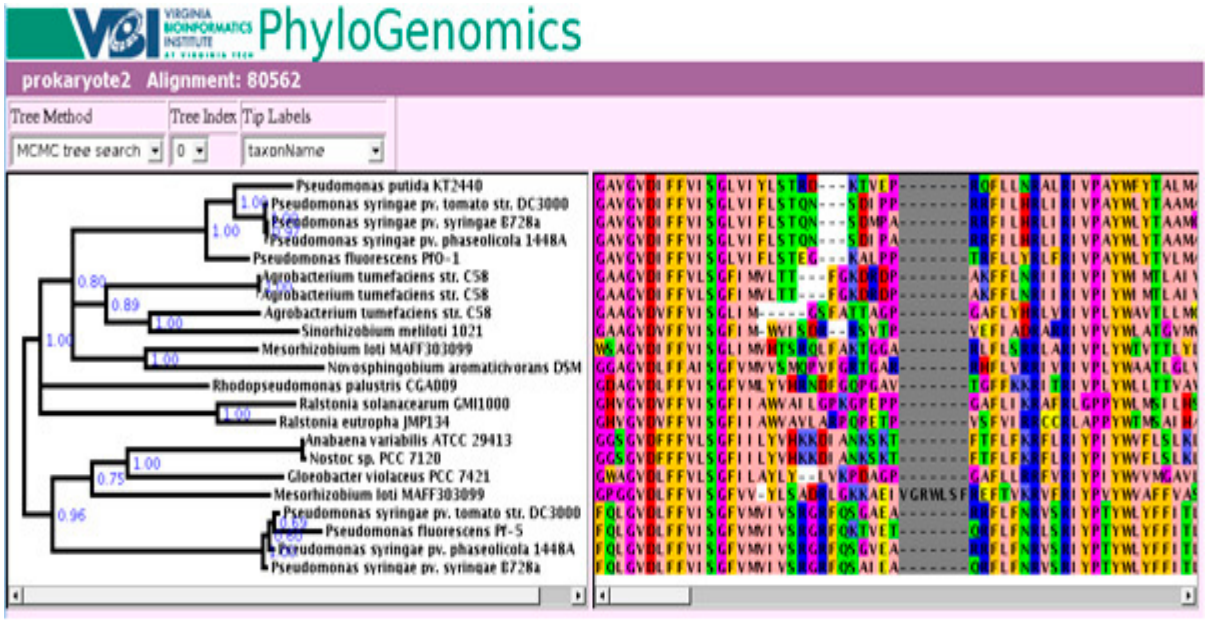
→ 100 Stammbäume

Von der Sequenz zum Baum

Sequenzen	Orthologie!
Multiples Sequenz Alignment	gap penalties, …
Auswahl eines Evolutionsmodells	hLRT, gamma shape, …
Auswahl von Methode & Algorithmus	Distanz- vs. Charaktermethoden
Stammbaumberechnung (mit/ohne statistischer Auswertung)	Bootstrapping, Jackknifing, …

…dann mal los!

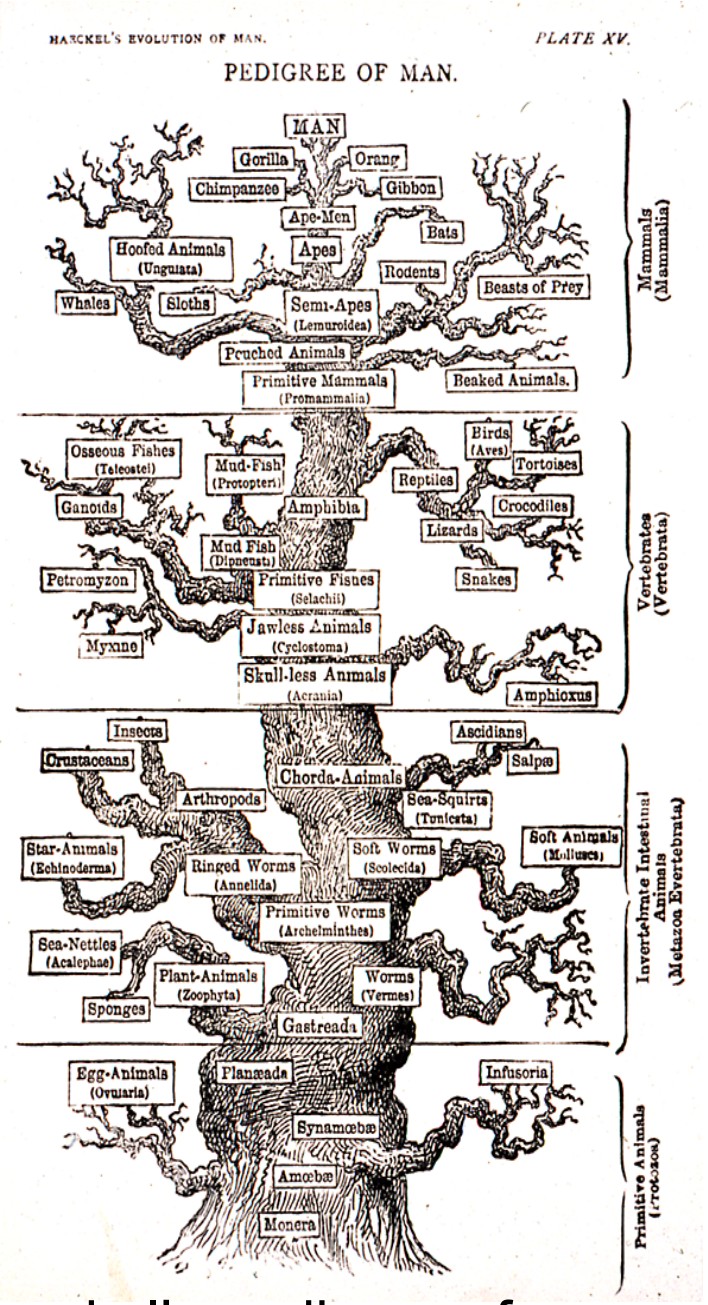
Phylogenomik: Lösen große Datensätze die Probleme?



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 trees of each great kingdom of nature"

(*genealogical* = phylogenetic*)*



Haeckel's pedigree of man

Die ‚Neue Metazoen-Phylogenie‘

Morphologie

rRNA-Sequenzen

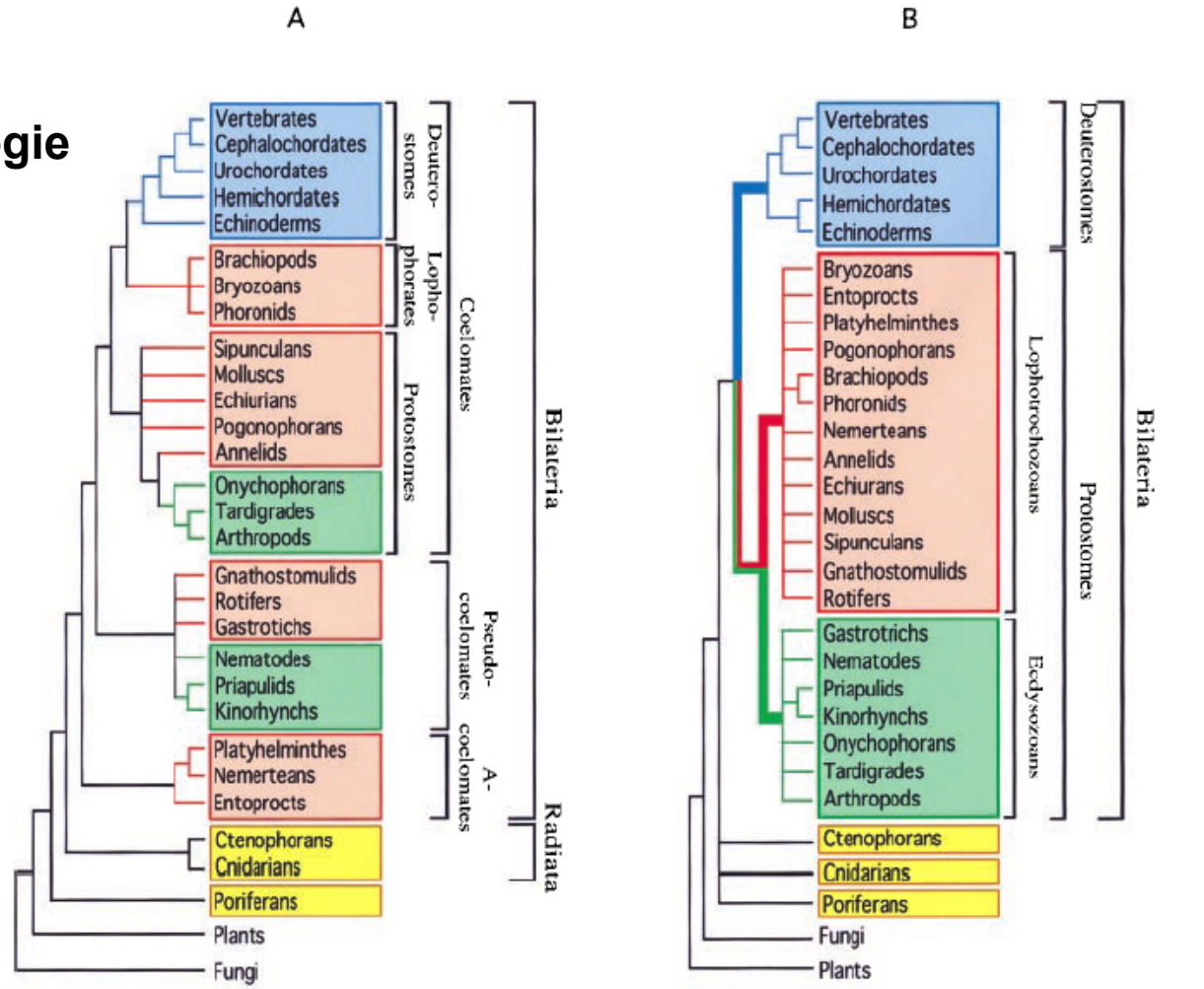
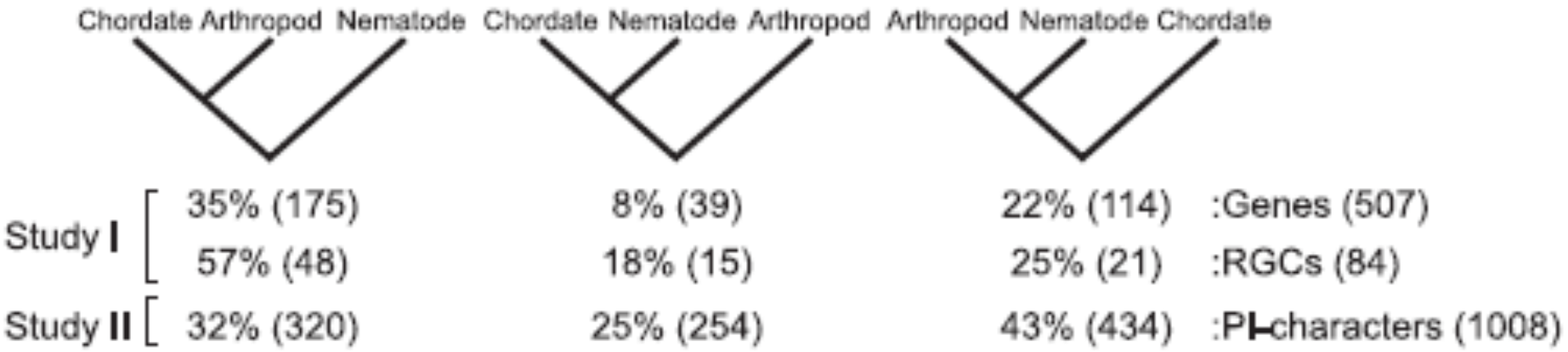


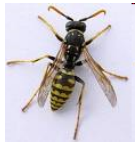
Fig. 1. Metazoan phylogenies. (*A*) The traditional phylogeny based on morphology and embryology, adapted from Hyman (11). (*B*) The new molecule-based phylogeny. A conservative approach was taken in *B*: i.e., some datasets provide resolution within some of the unresolved multifurcations displayed,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

Das Problem „Inkongruenz“

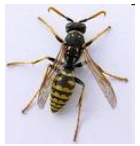
Beispiel 1 **550 Mya**



Coelomata
(Chordata + Arthropoda)

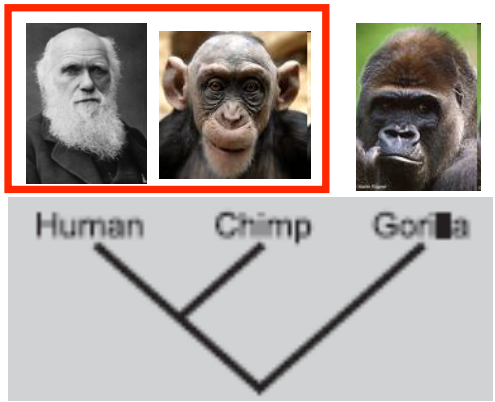


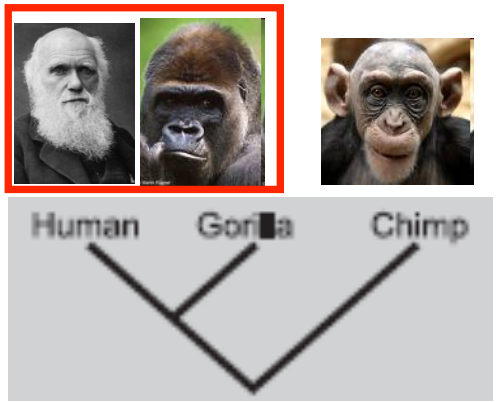
Ecdysozoa
(Nematoda + Arthropoda)

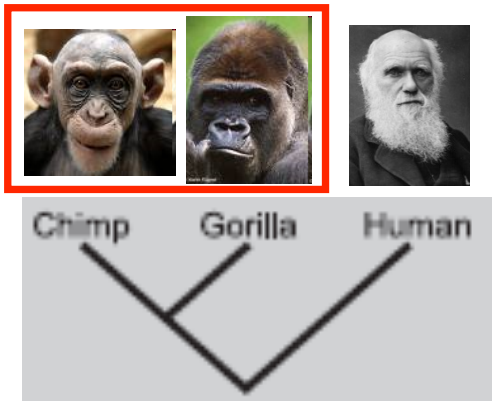


Das Problem „Inkongruenz“

Beispiel 2 **5-8 Mya**







Oben: Gene (n=98)
Mitte: Parsimony-informative Orte (n=174)
Unten: rare genomic changes (n=8)

Das Problem „Inkongruenz“

Beispiel 3

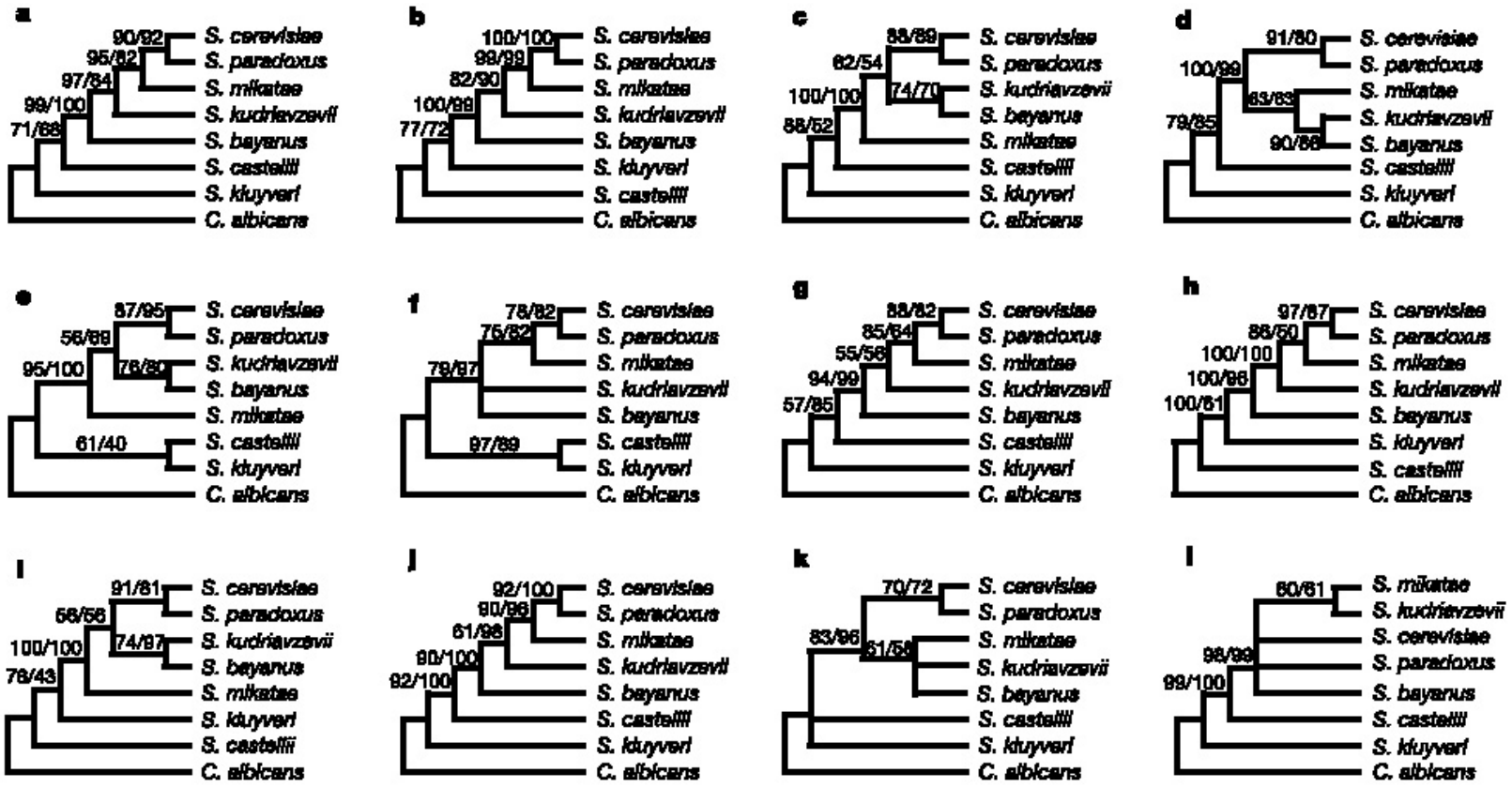


Figure 1 Single-gene data sets generate multiple, robustly supported alternative topologies. Representative alternative trees recovered from analyses of nucleotide data of 106 selected single genes and six commonly used genes are shown. The trees are the 50% majority-rule consensus trees from the genes YBL091C (**a**), YDL031W (**b**), YER005W (**c**), YGL001C (**d**), YNL155W (**e**) and YOL097C (**f**), as well as those from the commonly used genes actin (**g**), hsp70 (**h**), β-tubulin (**i**), RNA polymerase II (**j**) elongation factor 1-α (**k**) and 18S rDNA (**l**). Numbers above branches indicate bootstrap values (ML on nucleotides/MP on nucleotides).

Rokas et al., Nature 23. Okt. 2003

Darwin zum Thema

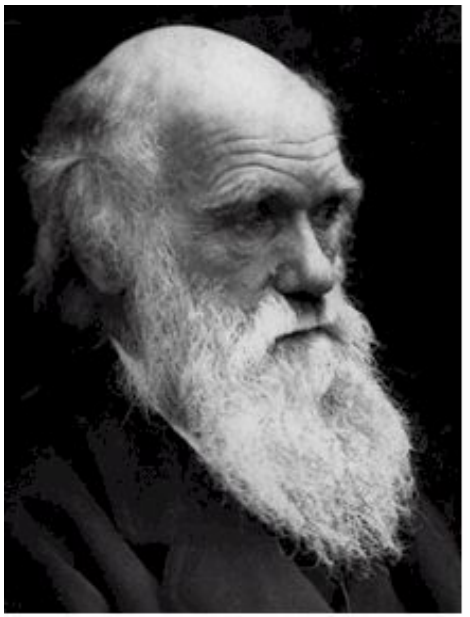
Inkongruente Bäume

The importance, for classification, of trifling characters, mainly depends on their being correlated with several other characters of more or less importance.

The value indeed of an aggregate of characters is very evident.

…a classification founded on any single character, however important that my be, has always failed.“

Charles Darwin Origin of Species Kap. 13



Gründe für Inkongruenz

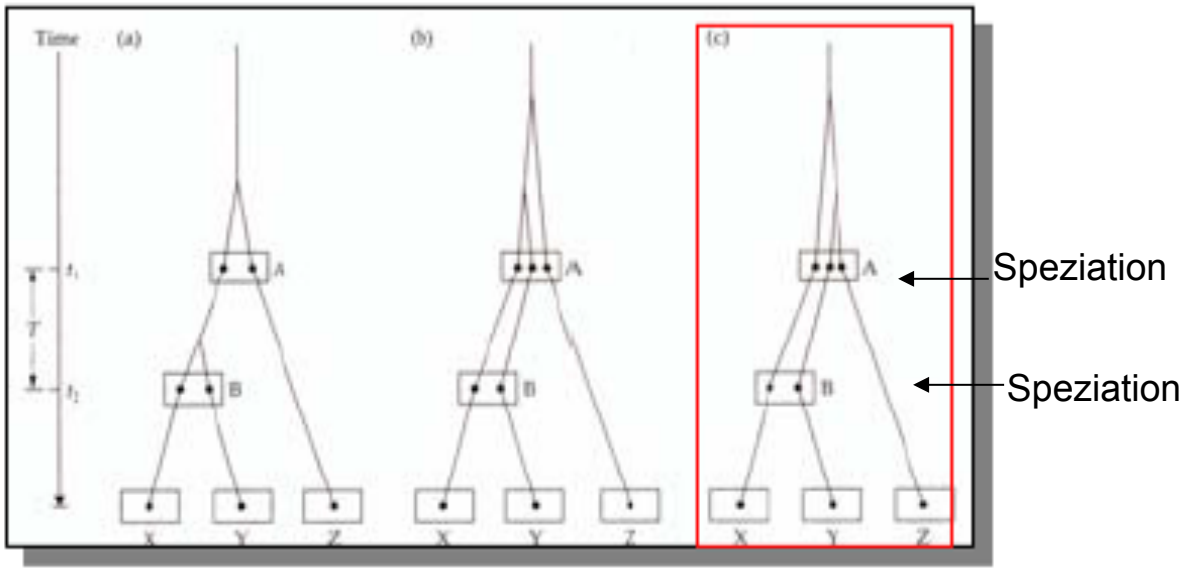
1. Stochastische Fehler in den Daten:
   - → meist bei wenigen Daten
   - → einige Positionen zeigen Homoplasie (durch multiple Austausche) und produzieren so eine falsche Baum-Topologie

2. Gen-Baum = Spezies-Baum?
   - → fälschlicherweise keine Orthologen verglichen (versteckte Paralogie)
   - → horizontaler Gentransfer
   - → multiple Allelie

3. Systematisch-methodische Fehler:
   - → Evolutionsmodelle, Rekonstruktionsmethoden

Gründe für Inkongruenz

Gen-Baum = Spezies-Baum? → Beispiel multiple Allelie



Bei Vergleich der Gene scheinen Y u. Z Schwestertaxa zu sein. In Wahrheit hatten die Taxa X u. Y einen gemeinsamen Vorläufer

Gründe für Inkongruenz

Gen-Baum = Spezies-Baum? → Beispiel multiple Allelie

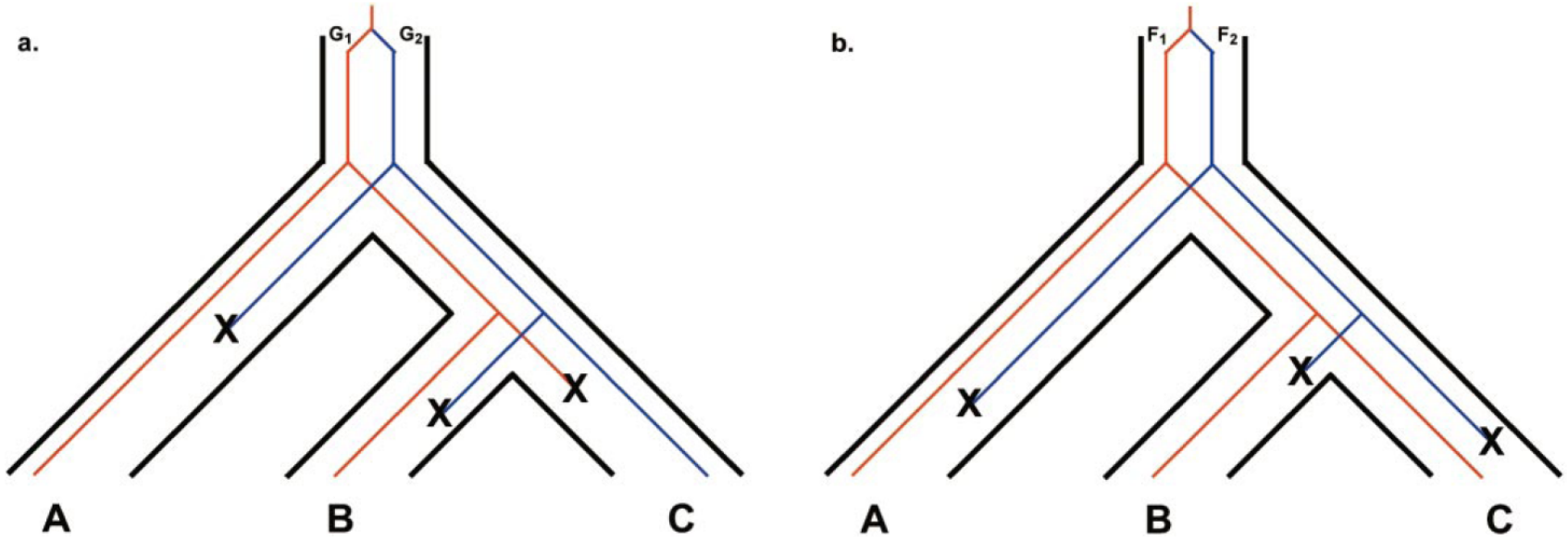


Fig. 2. An example of the gene tree/species tree problem. The species phylogeny is represented by black lines. The gene trees are represented by colored lines. (a) Prior to the root of the ABC clade, a gene (G_1, in red) is either duplicated or mutates to produce a new allele (G_2, in blue). Both versions of G are inherited at the two speciation events, but G_2 is lost in the lineages leading to species A and B, and G_1 is lost in the lineage leading to species C. The tree that would be reconstructed from the paralogous versions of G would incorrectly indicate that species A is the sister species of B. (b) A second gene (F) from the ABC clade where only F_2 is lost. Using the F_1 orthologs produces the correct set of species relationships. Note that the two genes produce incongruent trees, which indicates the possibility that B is a hybrid of A and C.

Linder & Rieseberg 2004

Inkongruenz & Phylogenomik

1. **Stochastische Fehler**
2. **Gen-Baum = Spezies-Baum?**

} Verschwinden in großen Datensätzen
→ **„Phylogenomik“!!**

3. **Systematisch-methodische Fehler**

→ Mehr Daten hätten hier gegenteiligen Effekt und würden falschen Baum sicherer erscheinen lassen („Inkonsistenz“)

Inkonsistenz

Nature, Oct 2011

Phylogenomics reveals deep molluscan relationships

Kevin M. Kocot[1], Johanna T. Cannon[1], Christiane Todt[2], Mathew R. Citarella[3], Andrea B. Kohn[3], Achim Meyer[4], Scott R. Santos[1], Christoffer Schander[2], Leonid L. Moroz[3,5], Bernhard Lieb[4] & Kenneth M. Halanych[1]

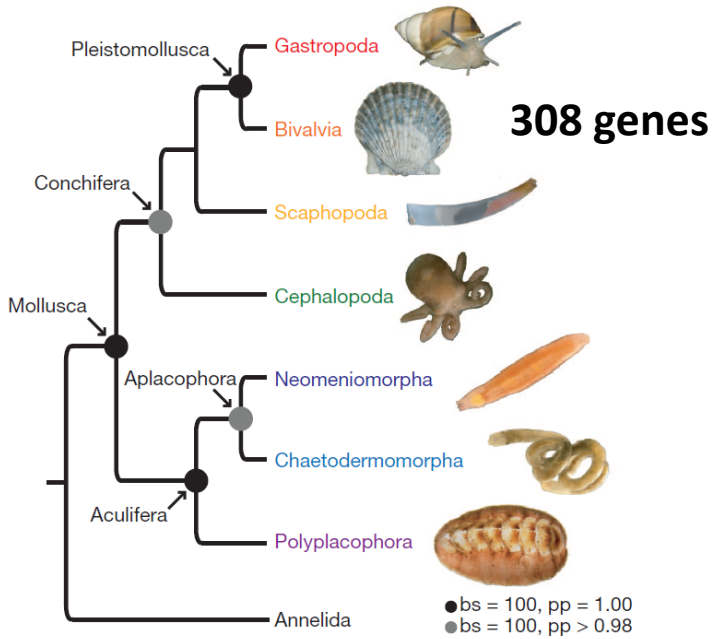


Figure 4 | Deep molluscan phylogeny as inferred in the present study. Black circles represent nodes with bs = 100 and pp = 1.00. Gray circles represent nodes with bs = 100 and pp ≥ 0.98. The actual specimens of *Polyschides* and *Hanleya* used in this study are shown. Photos are not to scale. A full-page version of this figure is presented in Supplementary Fig. 1.

Nature, Nov 2011

Resolving the evolutionary relationships of molluscs with phylogenomic tools

Stephen A. Smith[1,2], Nerida G. Wilson[3,4], Freya E. Goetz[1], Caitlin Feehery[1,4], Sónia C. S. Andrade[5], Greg W. Rouse[4], Gonzalo Giribet[5] & Casey W. Dunn[1]

1185 genes

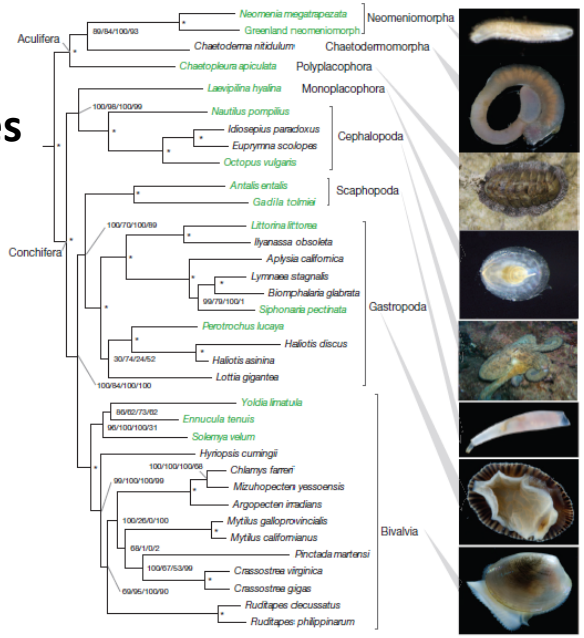


Figure 2 | Phylogram of the RAxML maximum likelihood analysis of the big matrix (216,402 amino acids) under the WAG+Γ model. Support values for the topology obtained from four analyses are listed as percentages in the order A/B/C/D. A is the bootstrap support from RAxML analysis under the WAG model for the big matrix. B is the bootstrap from RAxML analysis under the WAG model for the small matrix. C is the posterior probability from MrBayes under the WAG model for the small matrix. D is the posterior probability from PhyloBayes under the CAT model for the small matrix. Asterisks indicate 100/100/100/100 support. Taxa with new data are shown in green. Scale bar, 0.08 expected changes per site.

Systematische Fehler

1. **unterschiedlicher Nt/AS-Gehalt zwischen Taxa (compositional bias):**

   z.B. gleiches Nt in entfernten Taxa wegen ähnlichem GC-Gehalt

2. **Eingeschränkte Variabilität an Sequenzpositionen:**

   erhöhte Chance für konvergentes Auftreten eines/r bestimmten Nt/AS

3. **Unterschiedliche Mutationsrate an verschiedenen Positionen:** besonders viel Homoplasien in schnell evolvierenden Taxa

   → long-branch attraction-Phänomen

4. **Nicht-Unabhängigkeit von Sequenzpositionen durch positionelles Zusammenwirken**

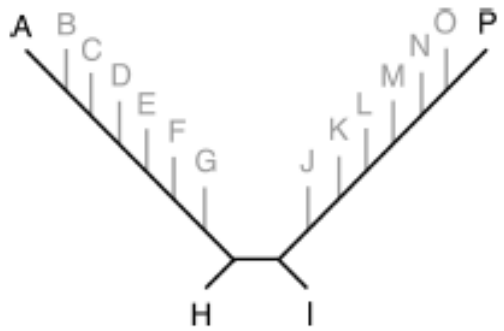
…und die mögliche Lösung

1. Geeignete Substitutionsmodelle auswählen

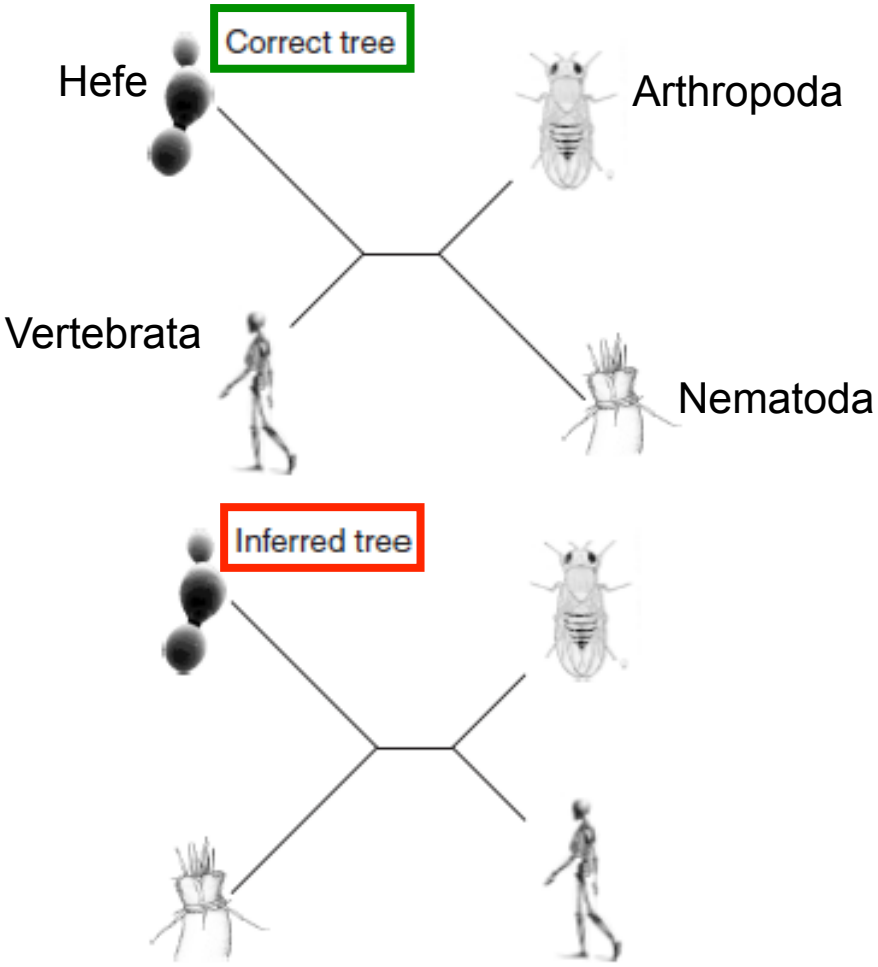
2. Schnell evolvierende Sequenzpositionen aus Alignment entfernen

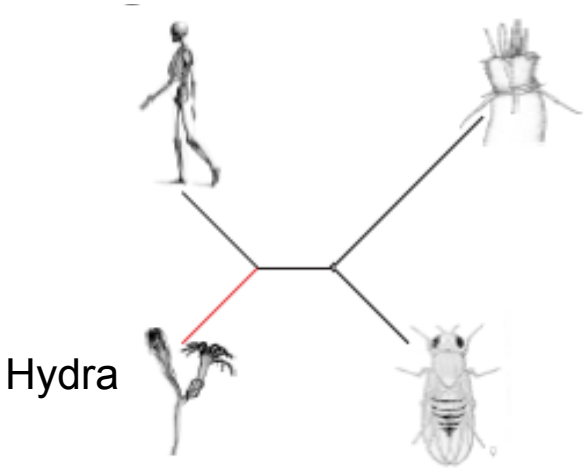
3. Schnell evolvierende Taxa u. U. komplett entfernen

4. Mehr Taxa hinzunehmen, die lange Äste „brechen“



Andere Außengruppe gegen LBA

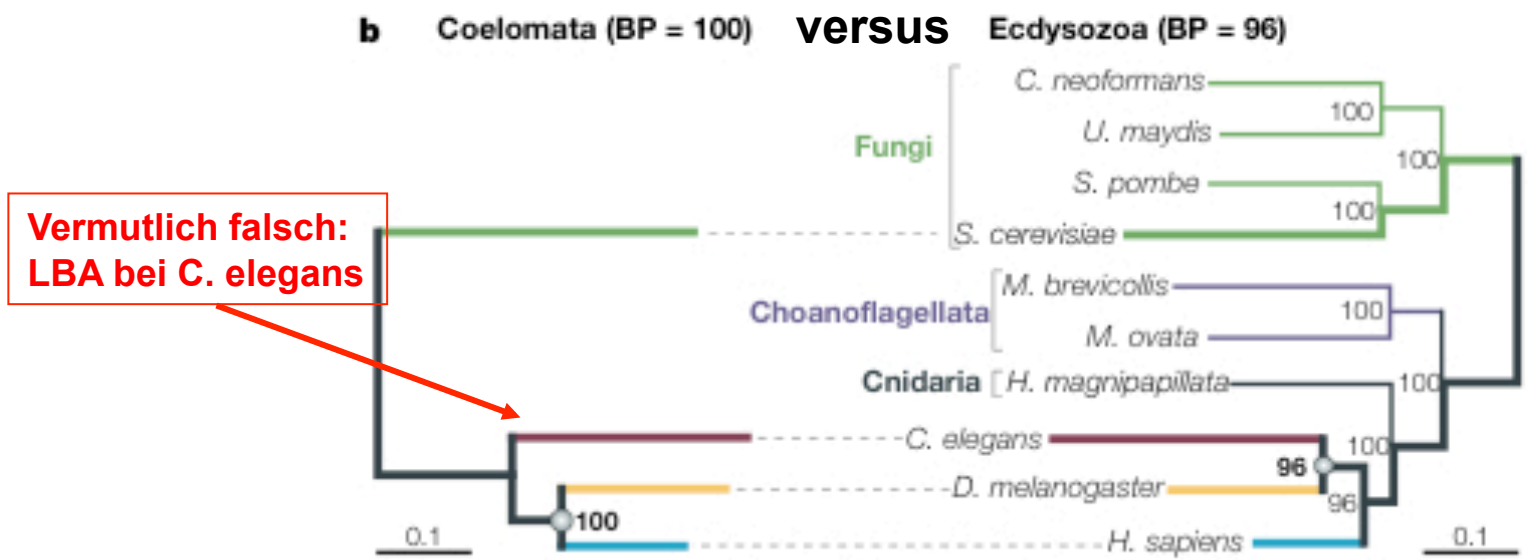




Telford & Copley 2005

Baumrekonstruktion & Phylogenomik

Systematische Fehler können auch Phylogenomik-Daten betreffen: Garbage in, garbage out!!



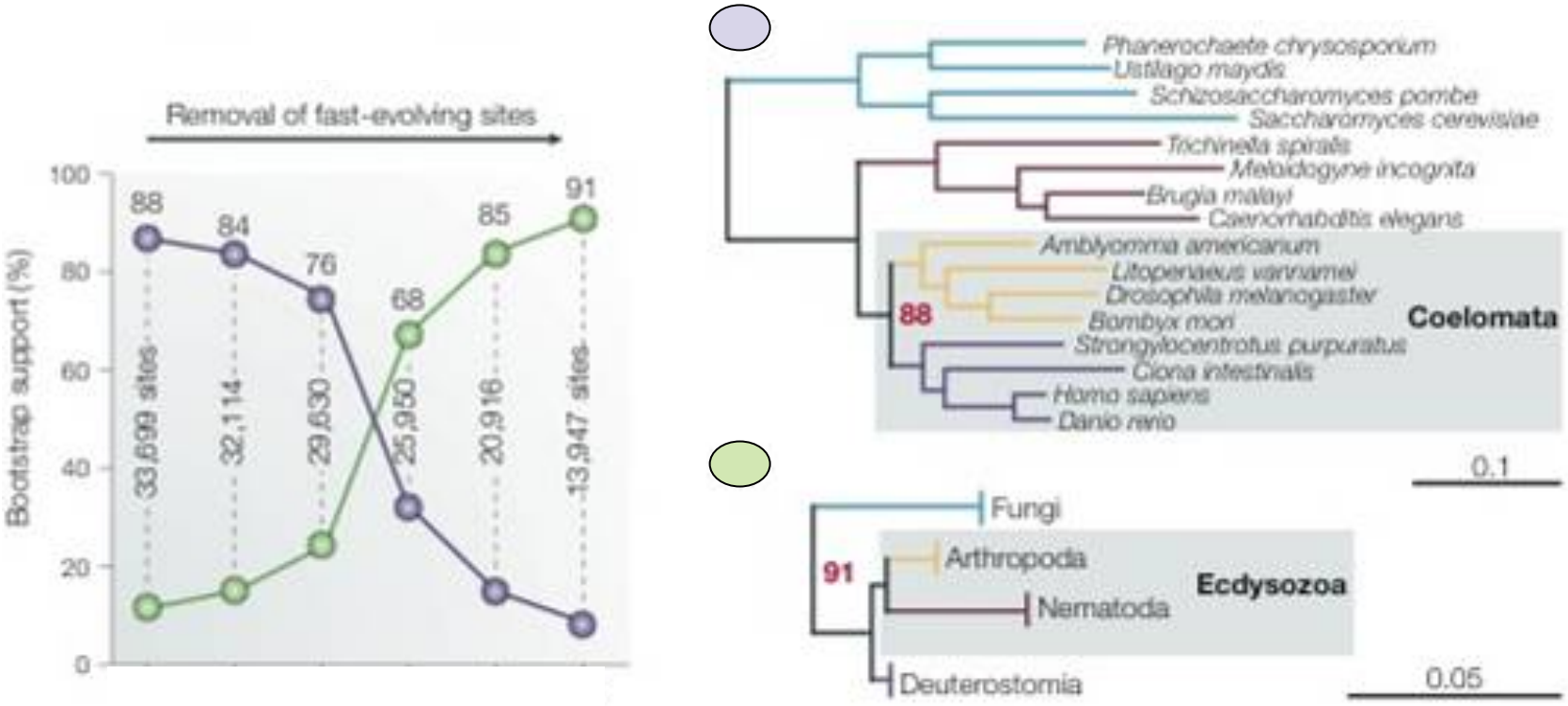
Datensatz: 35000 As (146 Kerngene), ML (Modell: JTT+Γ)
Unterschied: Taxon-Sampling

Also: mehr Taxa untersuchen!

Delsuc et al. 2005

Baumrekonstruktion & Phylogenomik

Systematische Fehler (z.B. LBA) beseitigen!



Also: schnell-evolvierende Sequenzpositionen opfern!

Delsuc et al. 2005

Zwei Methoden für die Phylogenomik

1. **Viele einzelne Gene → einzelne Bäume → kombinieren**

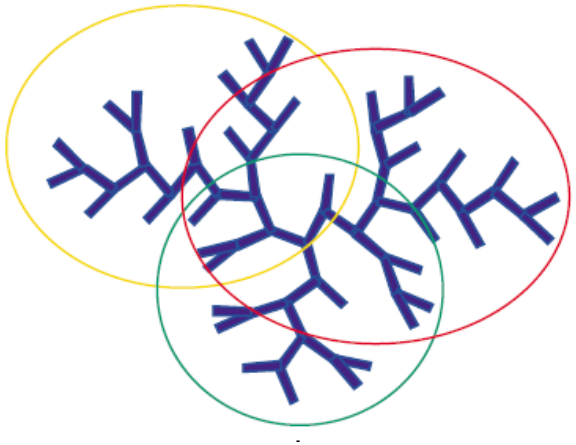
Supertree

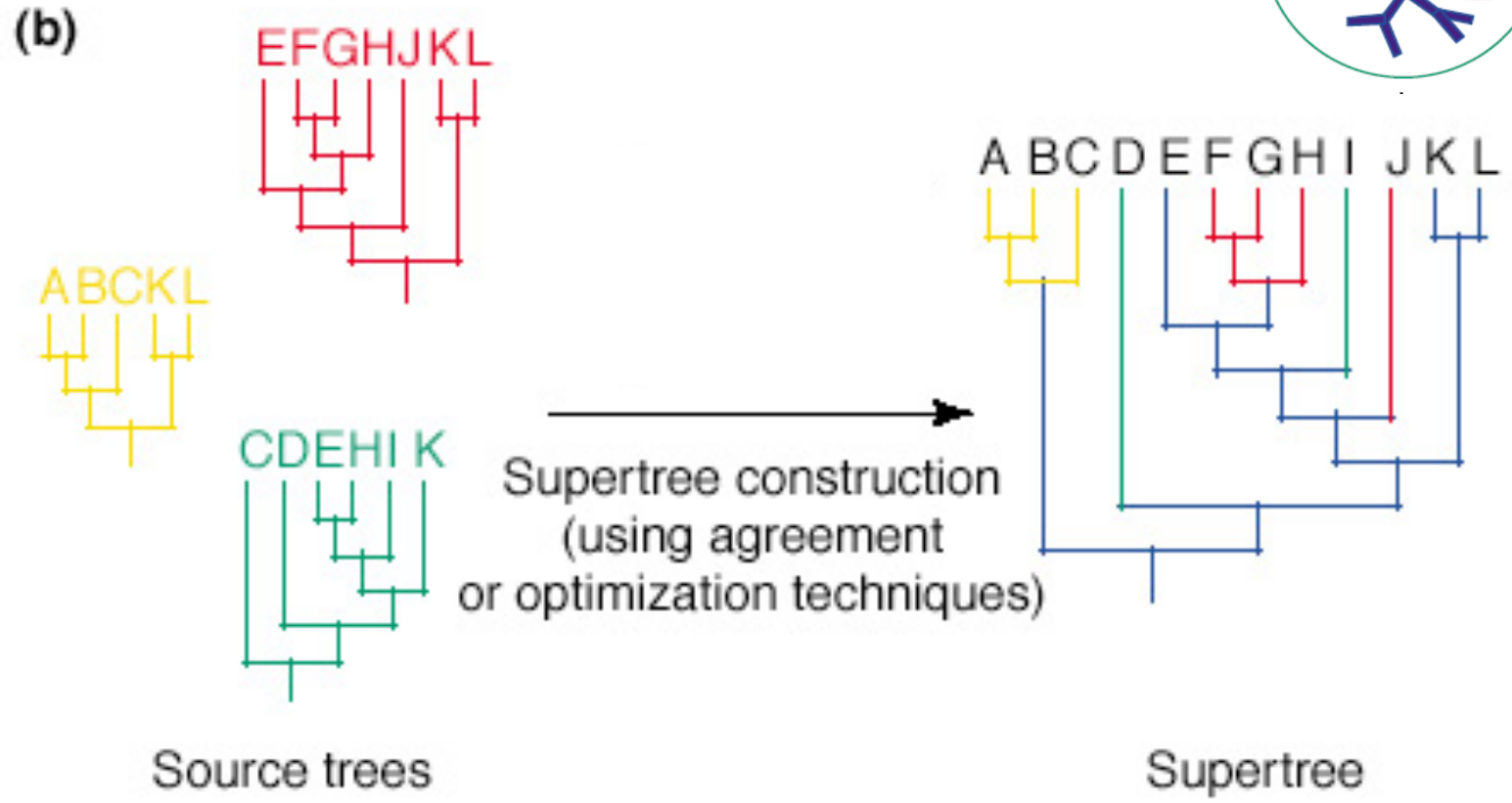
---

2. **Viele einzelne Gene → konkateniertes Alignment → Baum**

Supermatrix

Supertrees

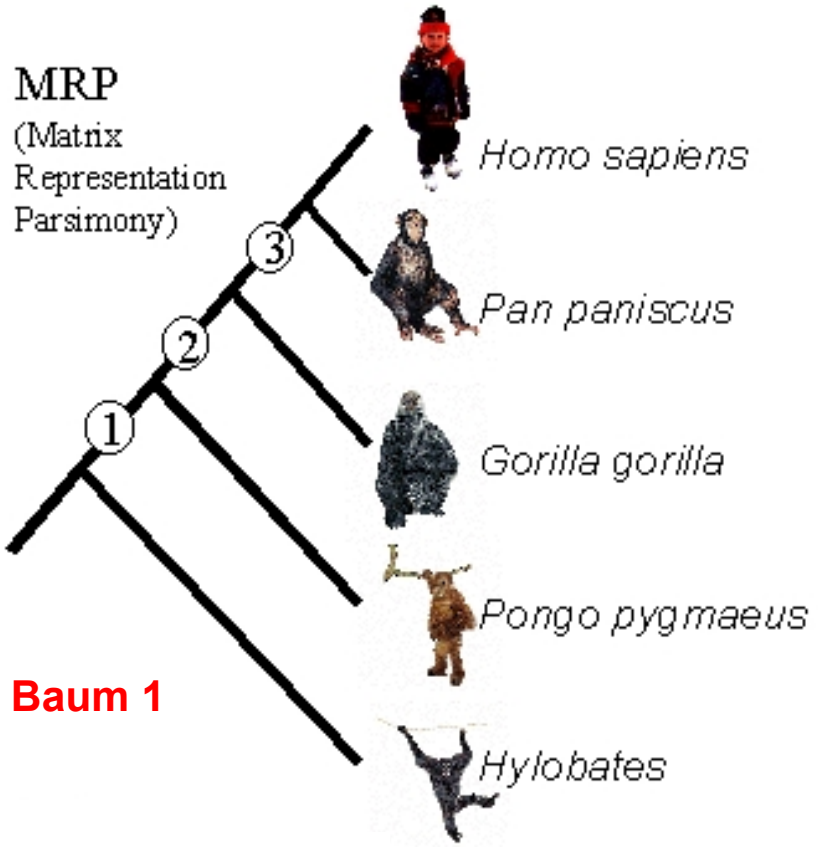




…ideal, um z.B. molekulare Bäume und morphologische Bäume zu verknüpfen!

Bininda-Emonds 2004

Supertrees: MRP



Internal branch

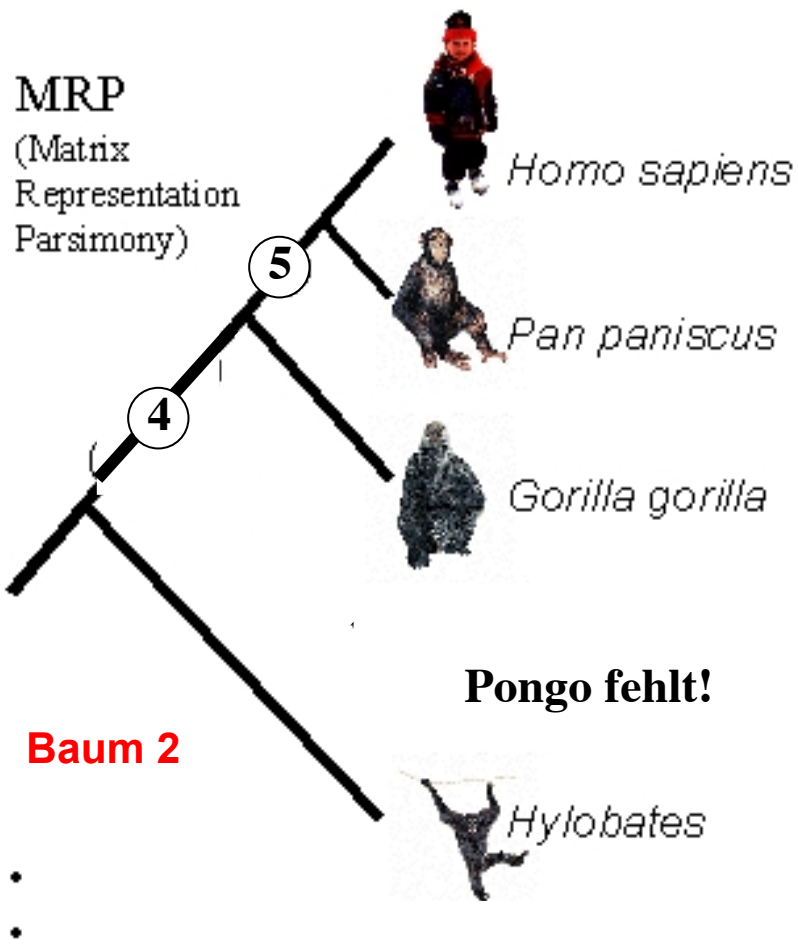
1	2	3
1	1	1
1	1	1
1	1	0
1	0	0
0	0	0

…ist wie Datenmatrix und kann daher zur Stammbaumrekonstr. z.B. per MP verwendet werden

…alle Taxa innerhalb ‚branch 1‘

courtesy of Rod Page

Supertrees: MRP

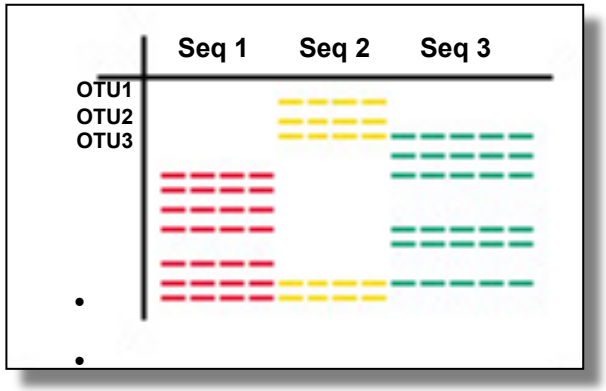


Internal branch

4	5
1	1
1	1
1	0
?	?
0	0

Supermatrix

Konkatenierte Daten



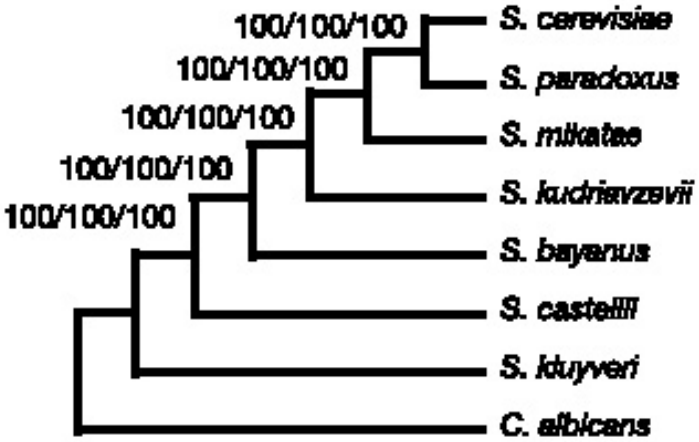
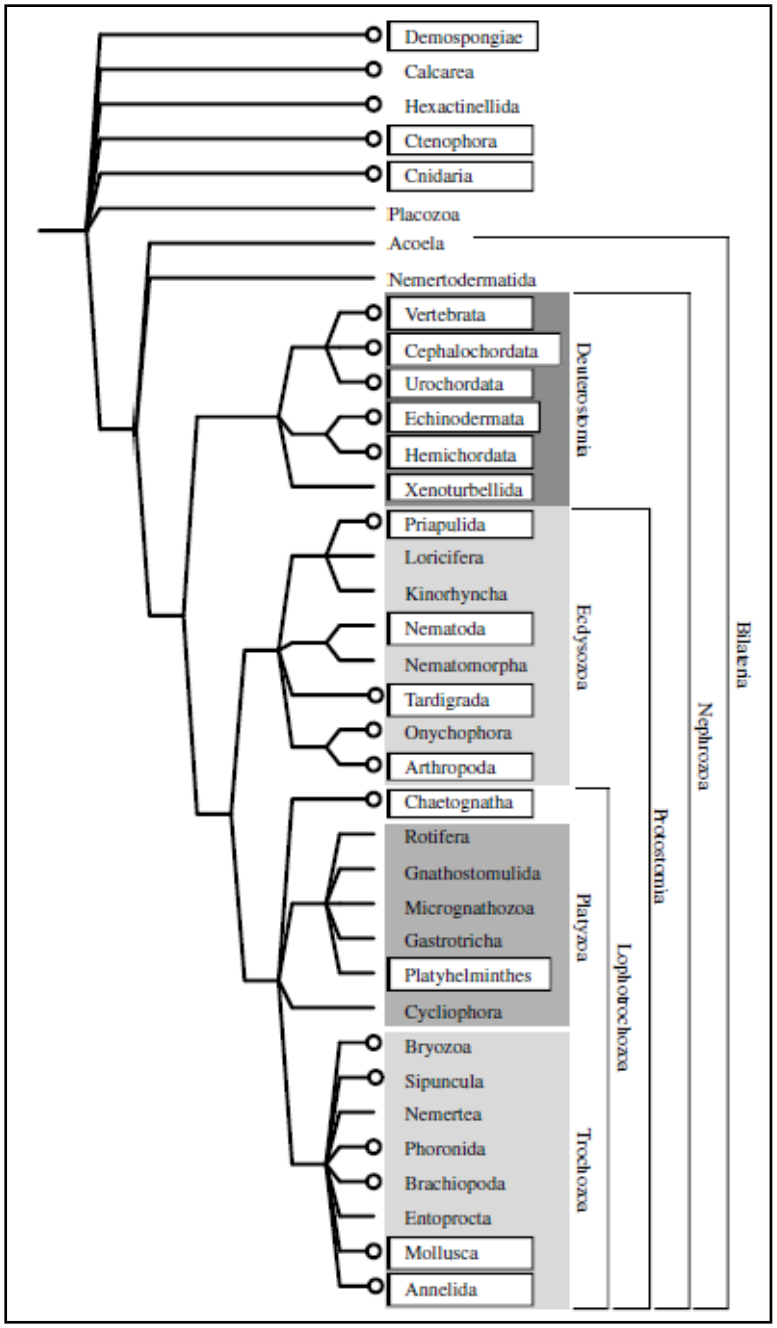


Figure 4 Phylogenetic analyses of the concatenated data set composed of 106 genes yield maximum support for a single tree, irrespective of method and type of character evaluated. Numbers above branches indicate bootstrap values (ML on nucleotides/MP on nucleotides/MP on amino acids).

20 aneinander gehängte Gene reichen aus, um stabilen, kongruenten Baum der acht Hefen zu produzieren!

Rokas et al., Nature 23. Okt. 2003



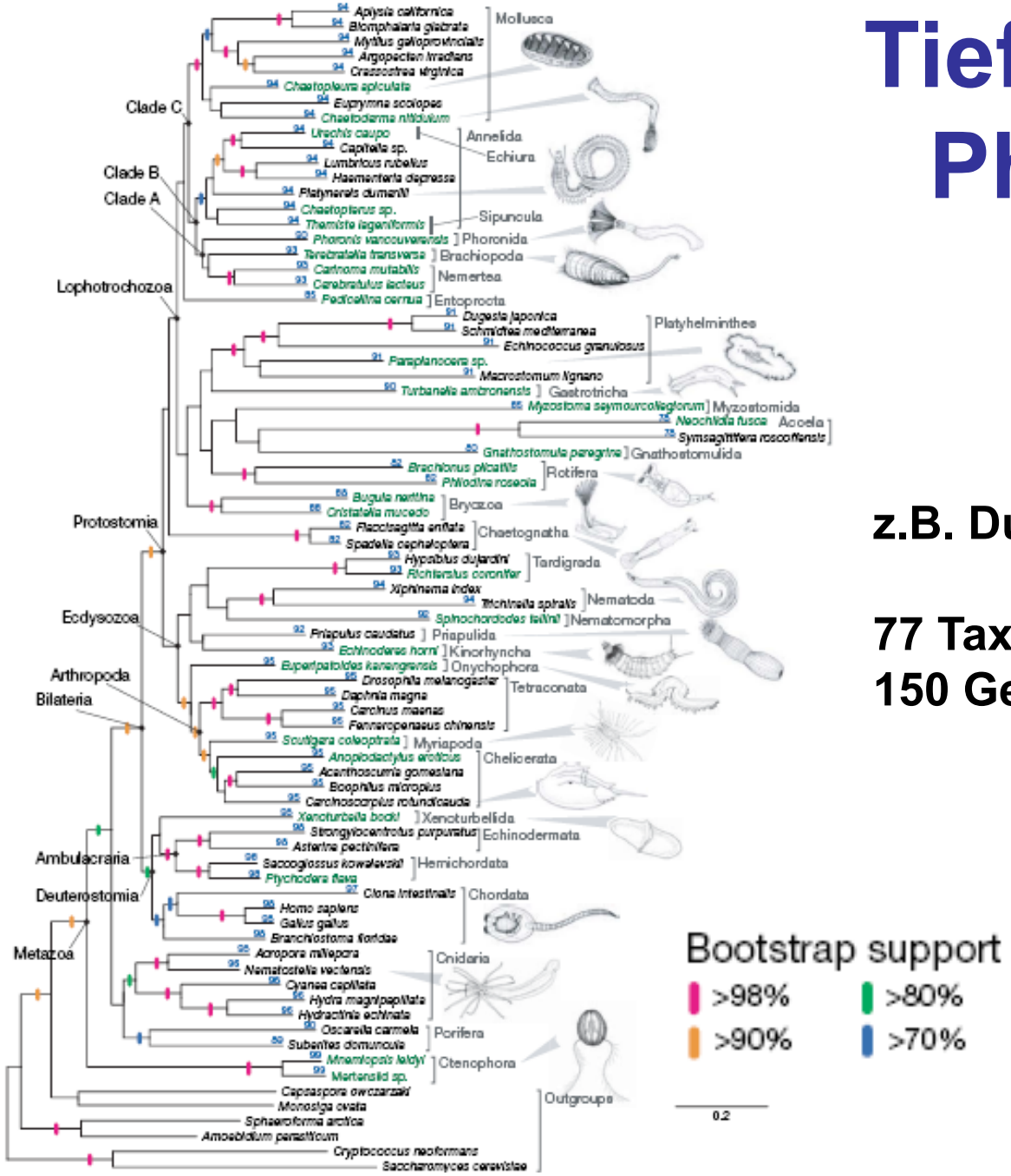
Supermatrix

…das klappt leider nicht immer so gut…

→ Tiefe Metazoen-Phylogenie ist schwierig!

Giribet 2008

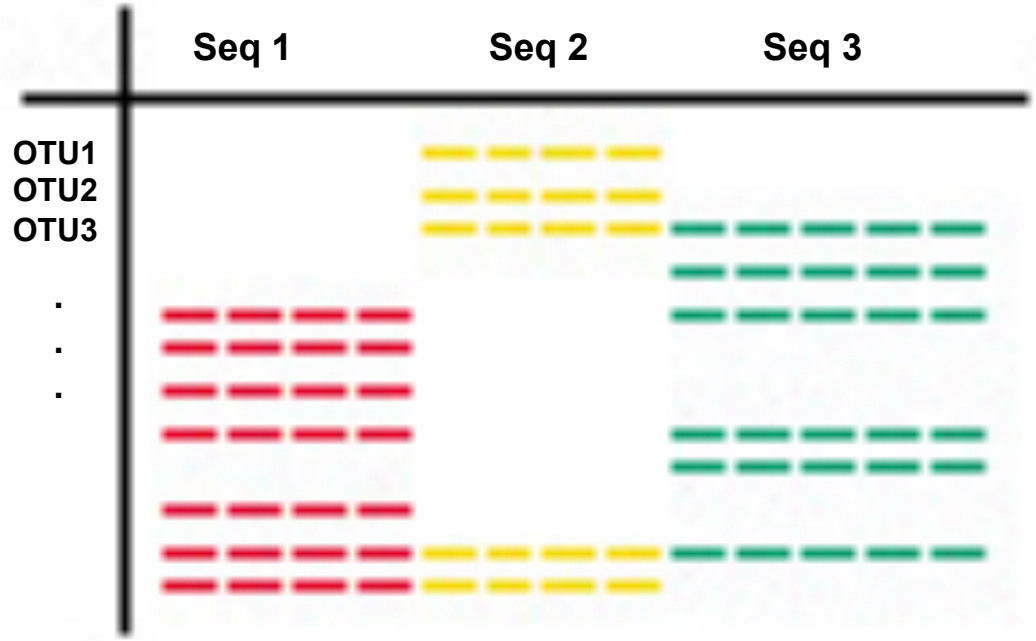
Tiefe Metazoen-Phylogenie ist schwierig



z.B. Dunn et al 2008 (Nature)

77 Taxa aus 21 Phyla
150 Gene

Supermatrix & fehlende Daten



Simulationen zeigen: Menge der vorhandenen Daten ist wichtiger als Menge der fehlenden Daten!!!

(also: es ist besser ein Taxon mehr zu haben mit vielleicht nur 50% Datenabdeckung, als dieses Taxon deswegen wegzulassen. vgl. Wiens 2005, 2006)

Woher kommen die Daten?

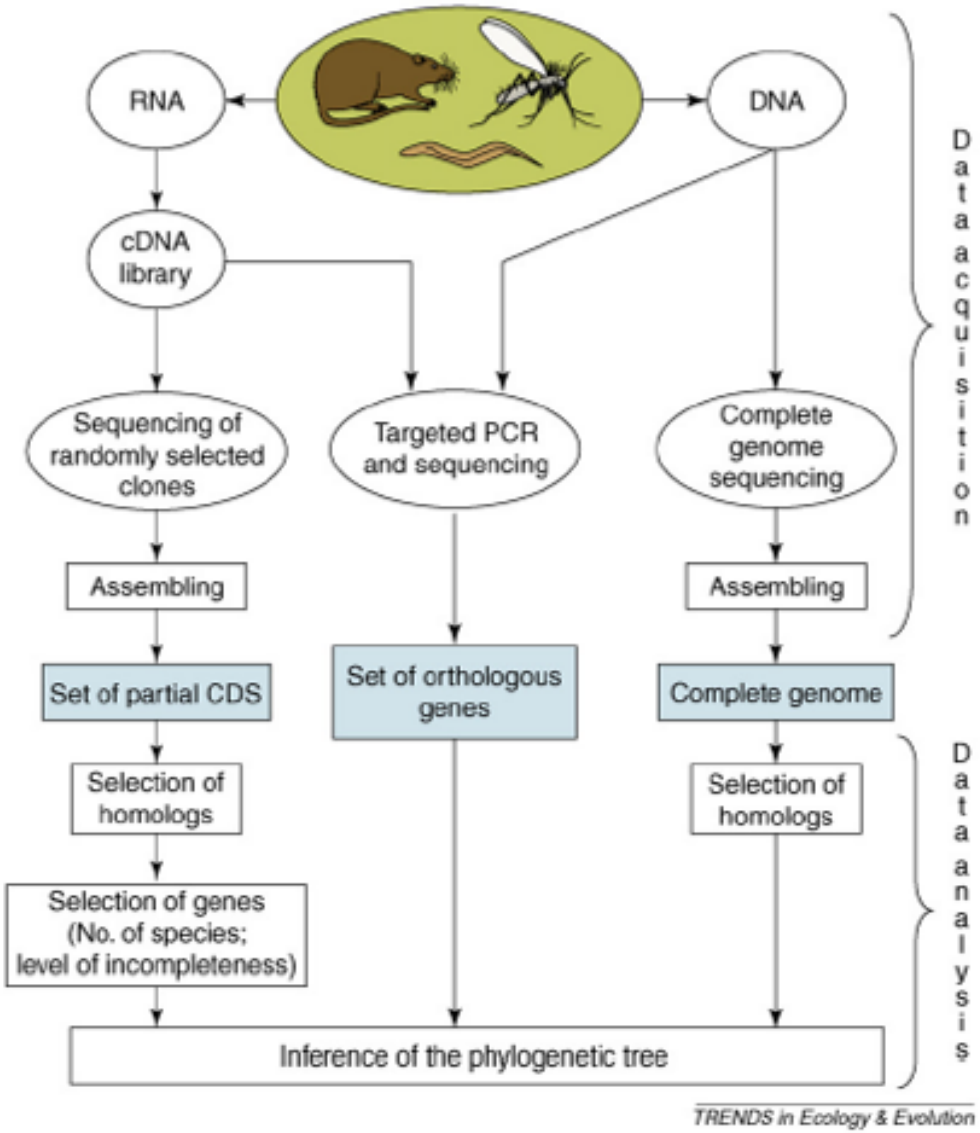
1. **Gesamt-Genomprojekte**
   - **→ Gen/Proteinsequenzen**
   - **→ ‚Rare genomic changes‘**

     Genanordnung, Gengehalt, Genfusionen, Intronpositionen, Transposonpositionen, InDels

2. **EST/RNA-SEq-Projekte (billiger - erst recht seit NGS!!!)**
   - **→ Proteinsequenzen**

3. **PCR (gezielte Isolierung von Einzelgenen)**

Woher kommen die Daten?



EST-Projekte & Phylogenomik

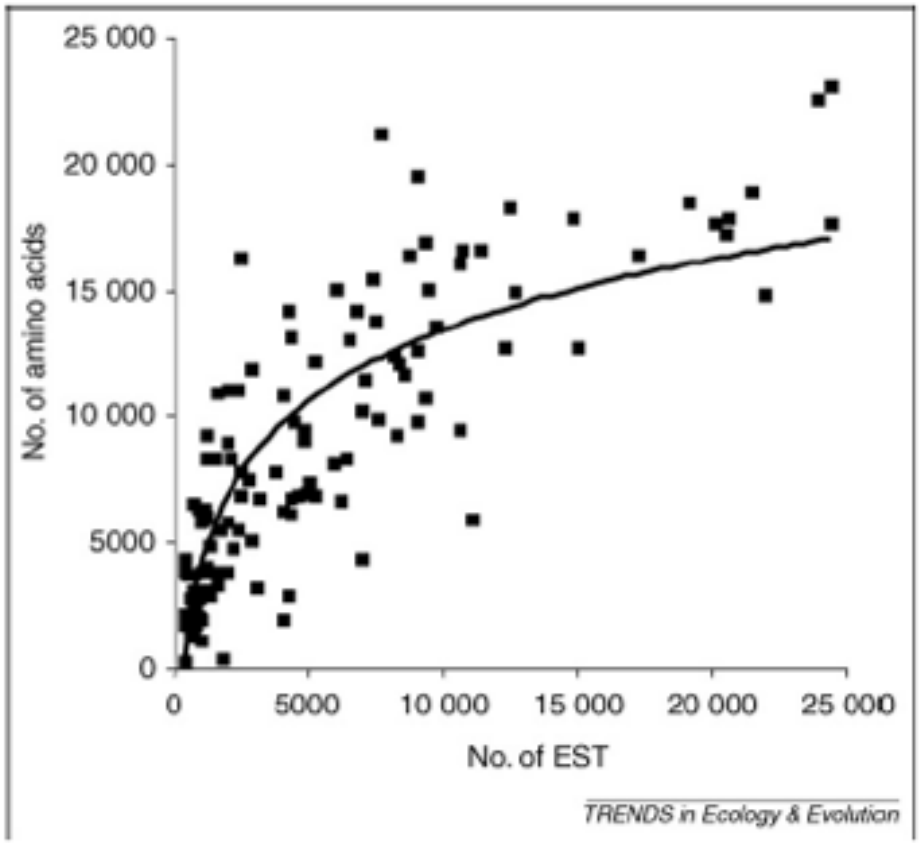


Figure 1. Estimating how many ESTs should be sequenced for inferring phylogeny. For 127 species of choanoflagellates and animals, the number of unambiguously aligned amino-acid positions found for a set of 146 evolutionarily conserved genes [19,31] are plotted against the number of ESTs that have been sequenced for each species. A logarithmic curve is fitted to the data. The sequencing of ~5000 ESTs from a single library provides ~10 000 phylogenetically informative positions, and constitutes a good compromise between cost and amount of information gained.

Schon 1000 ESTs
machen einen respektablen
Datensatz (>5000 As)

EST-Projekte & Phylogenomik

Welche Gene machen schnell ein möglichst komplettes konkateniertes Alignment?

Liste mit EST-Clustern

22	beta-thymosin [Sycon raphanus]
20	ADP/ATP carrier [Trypanosoma brucei brucei]
17	actin
16	elongation factor 1 alpha [Axinella verrucosa]
14	Syndecan binding protein (syntenin) [Xenopus tropicalis]
11	beta-tubulin [Suberites domuncula]
8	cathepsin L-like cysteine proteinase A [Rhipicephalus haemaphysaloides haemaphysaloide
8	Rab7 [Aiptasia pulchella]
8	PREDICTED: hypothetical protein XP_690365 [Danio rerio]
7	chloride intracellular channel 2 [Homo sapiens]
7	phosphoenolpyruvate carboxykinase [Schistosoma mansoni]
6	snail soma ferritin [Lymnaea stagnalis]
6	methionine adenosyltransferase II, alpha [Homo sapiens]
5	laminin receptor 1 [Danio rerio]
5	no match
→ 5	ribosomal protein L32 isoform B [Lysiphlebus testaceipes]
5	elongation factor-2 [Eurypauropus spinosus]
5	betaine-homocysteine methyltransferase [Bos taurus]
5	S-adenosylhomocysteine hydrolase [Danio rerio]
5	ribosomal protein S12 [Branchiostoma belcheri]
5	unnamed protein product [Homo sapiens]
→ 5	ribosomal protein L36 [Homo sapiens]
5	unnamed protein product [Tetraodon nigroviridis]
5	ubiquitin C [Homo sapiens]
→ 5	40S ribosomal protein S2 [Ictalurus punctatus]
4	heat shock 70kDa protein 5 (glucose-regulated protein, 78kDa) [Gallus gallus]
4	hypothetical protein SaroDRAFT_1577 [Novosphingobium aromaticivorans DSM 12444]
4	alpha-1 tubulin [Hirudo medicinalis]
4	glyceraldehyde-3-phosphate dehydrogenase [Astatotilapia burtoni]

- stark exprimierte Haushaltsgene

- kaum Paraloge

→ z.B. Gene für ribosomale Proteine (RPs)!

ABER: Vorsicht bei Verwendung spezieller Proteinklassen!

Compositional Heterogeneity and Phylogenomic Inference of Metazoan Relationships

Maximilian P. Nesnidal,[1] Martin Helmkampf,†[1] Iris Bruchhaus,[2] and Bernhard Hausdorf*[1]

Mol. Biol. Evol. 27(9):2095–2104. 2010

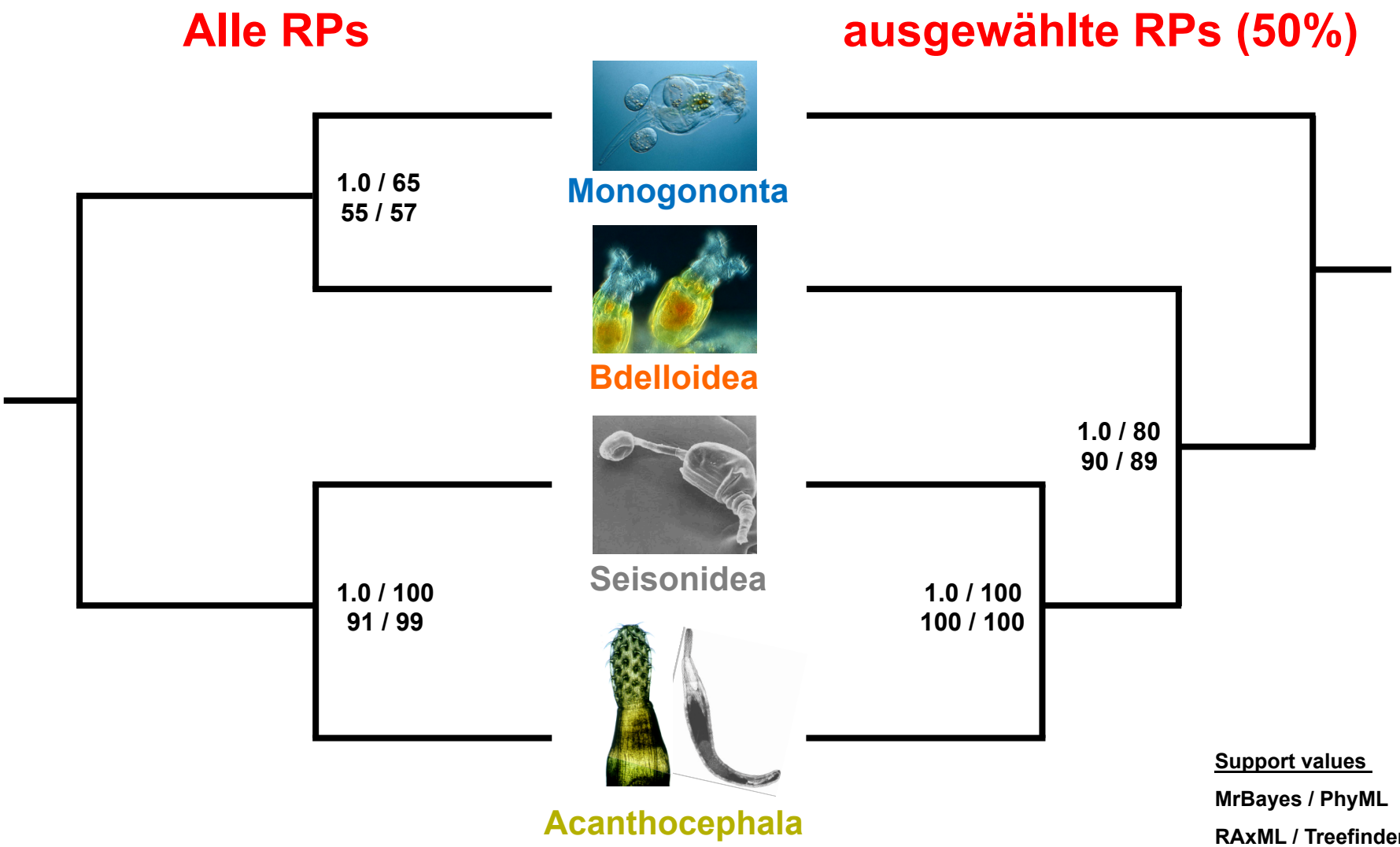
" […] to analyze a ***large ribosomal protein data set*** *representing all major metazoan taxa. Posterior predictive tests revealed that there is* ***compositional bias*** *in this data set.*

Only a few taxa with strongly deviating amino acid composition had to be excluded to reduce this bias. Thus, this is a good solution, ***if these taxa are not central to the phylogenetic question at hand****.*

Deleting individual proteins from the data matrix may be an appropriate method, if compositional heterogeneity among taxa is concentrated in a few proteins. However, ***half of the ribosomal proteins had to be excluded*** *to reduce the compositional heterogeneity […]"*

Vorsicht mit RPs!

Beispiel interne Phylogenie der Syndermata (= Rädertiere + Kratzwürmer)



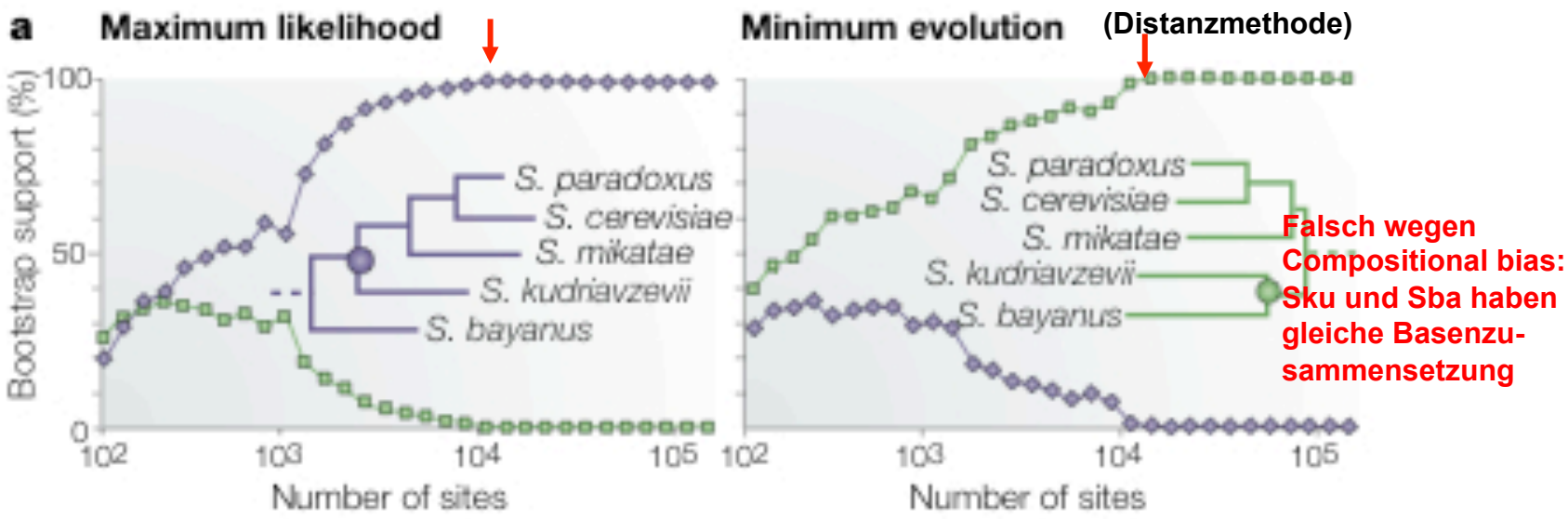
Baumrekonstruktion & Phylogenomik

Welche Methoden funktionieren mit großen Datensätzen am besten?

- **Likelihood-basierende Methoden (ML, Bayes) gelten insbesondere bei schwierigen Phylogenien als konsistenter und genauer als z.B. MP- und Distanzmethoden**

- **Likelihood-Methoden können mit ihren Substitutionsmodellen die Wirklichkeit besser abbilden**

ML versus Distanzmethode



Beide Male verwendet: Datensatz 127000 Bp, gleiches Modell (GTR+I+Γ)

Dennoch: unterschiedliche Topologie, beide Male mit Bootstrapsupport 100%!

Delsuc et al. 2005

Neue schnelle ML-Algorithmen

Übliche Heuristik: „hill climbing“

→ schrittweise Taxon-Hinzunahme & topologische Rearrangements
(z.B. „Nearest neighbour interchange“)
→ für jeden neu entstehenden Baum: Astlänge optimieren, LnL bestimmen
→ Verbesserung? Wenn ja, dann weiter verändern…
→ STOP wenn keine Verbesserung möglich

Langsam wegen getrennter Optimierung von Astlängen und Topologie

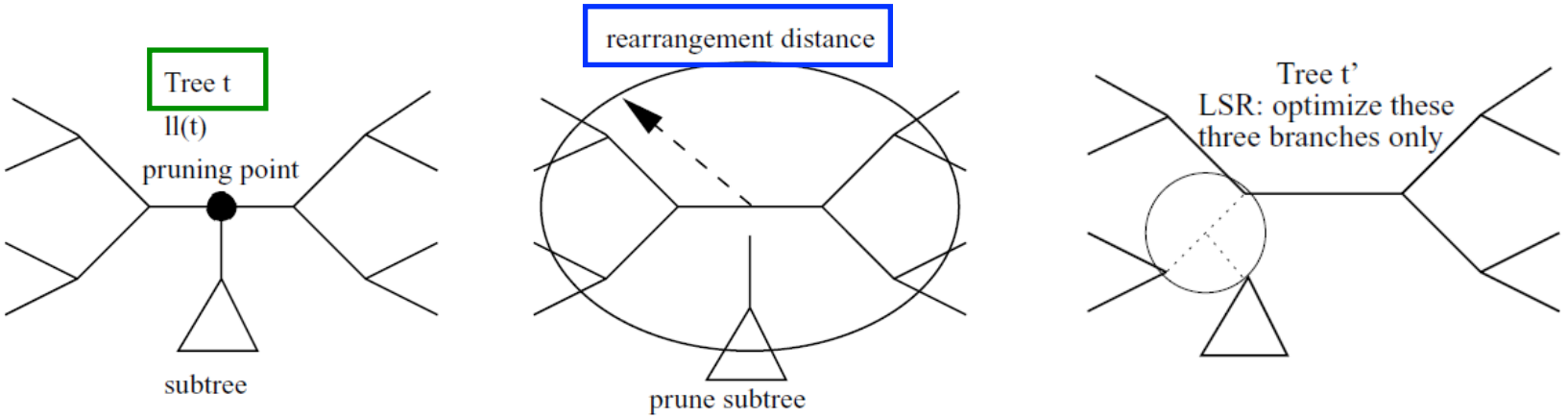
RAxML (Stamatakis et al. 2005)

→ verbesserte „search space heuristics“
→ mit „rapid bootstrap algorithmus“ (Stamatakis et al. 2008)
→ dramatische Zeitreduktion

RAxML

(Randomized A(x)ccelerated Maximum Likelihood)

Verbesserte „search space heuristics“ durch LSR
(= Lazy subtree rearrangement)

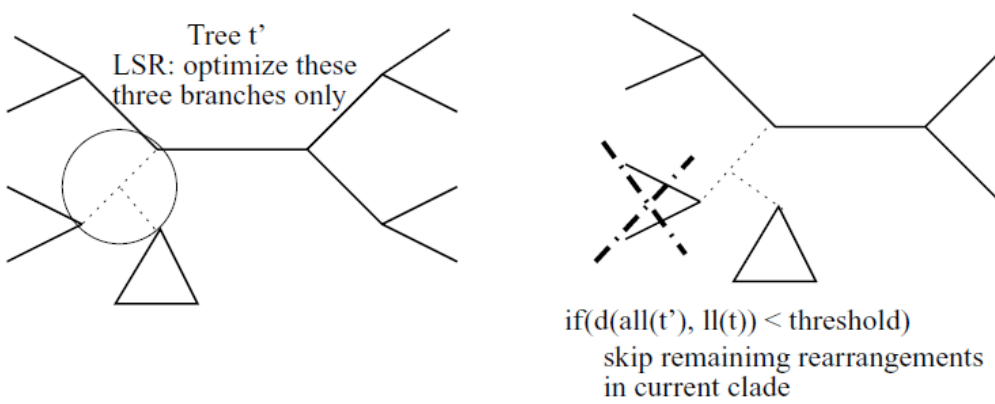


1. Ausgehend vom derzeit besten Baum *t*
→ Verpflanzung eines subtrees innerhalb eines Radius *n* (n = 5 – 15)
→ Es entsteht ein Baum *t‘*
→ Nur die Astlängen der an die Insertionstelle anschließenden Äste werden optimiert → approximate log likelihood *all(t‘)* score
→ Vorsortierung, nur die 5 besten Topologien werden gründlich optimiert (overall tree score)

RAxML

(Randomized A(x)ccelerated Maximum Likelihood)

Weitere Verbesserung: „likelihood cutoff heuristics“



Nicht alle LSR-Schritte pro subtree werden durchgeführt
Berechnung eines dynamischen likelihood-cutoff (lh_{cutoff})-Werts bei jeder Iteration

wenn δ ($all(t')$, $ll(t)$) > lh_{cutoff} wird das LSR für diesen subtree abgebrochen

→ 2,5x schneller als normales LSR!

RAxML

(Randomized A(x)ccelerated Maximum Likelihood)

Rapid bootstrap Algorithmus (RBS)

RBS ist heuristisches Verfahren („quick and dirty bootstrap“)
ergibt fast gleiche Unterstützungswerte, ist aber wesentlich schneller (vor allem bei großen Datensätzen)

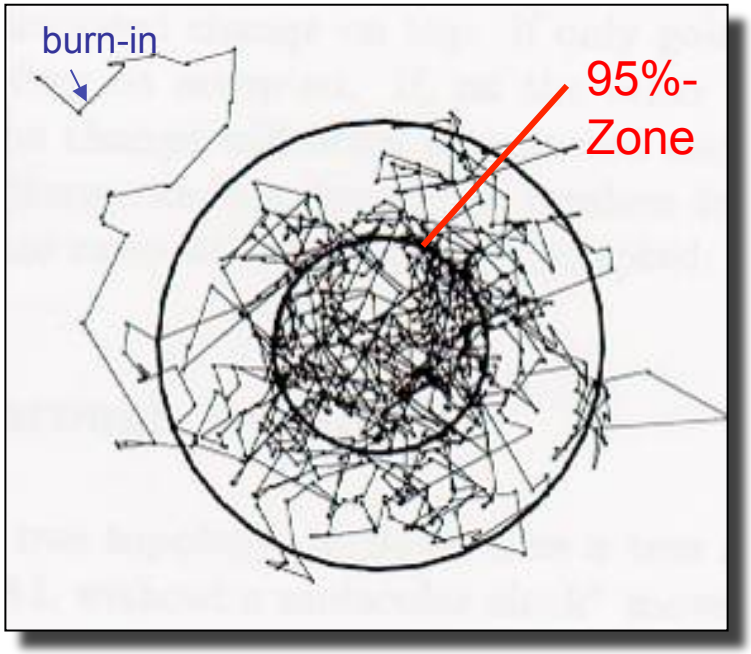
1. Random starting tree aus dem Originalalignment
2. ML-Model-Parameter und Astlängen-Optimierung für den starting tree
3. Für alle nachfolgenden RBS-Replikate keine Re-Optimierung der ML-Model Parameter mehr

→ 8 – 20 x schneller als RAxML mit Standard-bootstrap
→ 18 – 495 x schneller als PHYML

Verbesserter ML-Such-Algorithmus, der NACH dem RBS ausgeführt wird um einen ML-Baum auf das Original-Alignment zu berechnen

MCMCMC

Metropolis-coupled Markov chain Monte Carlo



- „cold chain“ sammelt
- „hot chains“ als „scouts“, um die Gipfel in der Baumlandschaft zu lokalisieren

→ der ‚MC Roboter‘ sucht nicht DEN optimalen Baum, sondern „sammelt“ die Bäume mit der höchsten posterior probability („Gipfel in der Baumlandschaft“).

→ Anhand dieser Baum-Sammlung wird ein Konsensus-Baum erstellt, dessen Verzweigungen durch die Höhe der PP-Werte gekennzeichnet und bewertet werden

ML vs. Bayes

- **ML berechnet den Baum mit der höchsten Likelihood**

  P (data | tree) → Wahrscheinlichkeit der Daten auf Grundlage eines angenommenen Baumes und eines Substitutionsmodells

  Statistischer Support erfolgt über Bootstrapping

- **Bayes ermittelt Häufigkeitsverteilung der Bäume im tree space**

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

  Statistischer Support über direkte Berechnung von PPs während der Baumsuche

Unterstützungswerte

Faustregel:

Bootstrap-Werte haben sich als eher konservativ herausgestellt
BP >80% gut, alles >50% ruhig angeben

Bayes PPs sind eher optimistisch und überschätzen Support
PP 1.0 ist ok, <1.0 eher schlecht

Achtung!

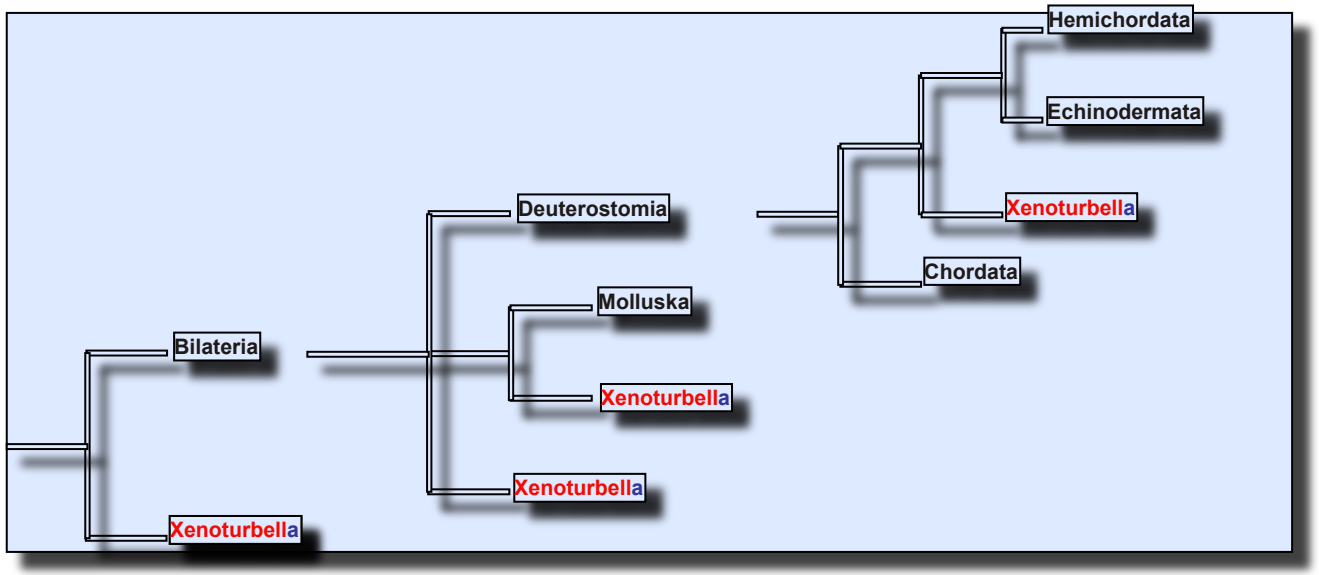
Die Monophylie der „ingroup“ wird immer zu 100% unterstützt, wenn man das Außengruppentaxon selbst wählt…

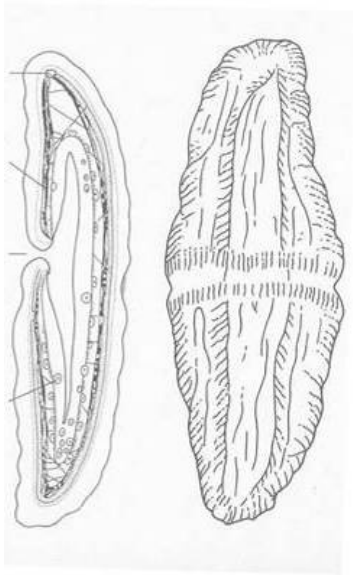
Im Kurs:

Ein Mikro-Phylogenomik-Projekt

Xenoturbella bocki

- entdeckt 1915 von Sixten Bock, erstmals beschrieben 1949
- lebt in marinen Sedimenten an der schwedischen Westküste
- bis zu ~4 cm lang
- sehr einfacher Körperbauplan
- phylogenetische Position kontrovers

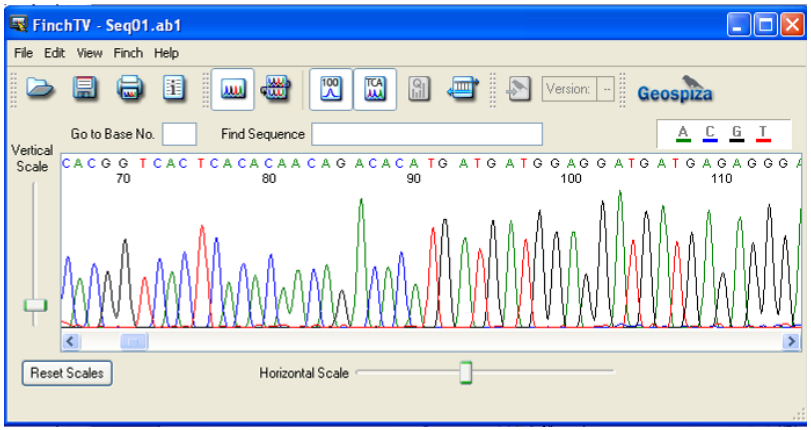


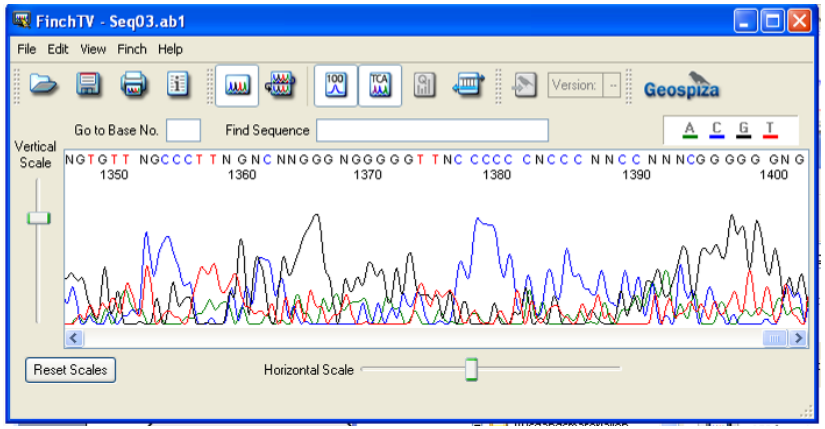


Quelle: (Bourlat et al. 2003; Ehlers and Sopott-Ehlers 1997; Israelsson 1997; Lundin 1998; Noren and Jondelius 1997; Raikova et al. 2000; Stach et al. 2005)

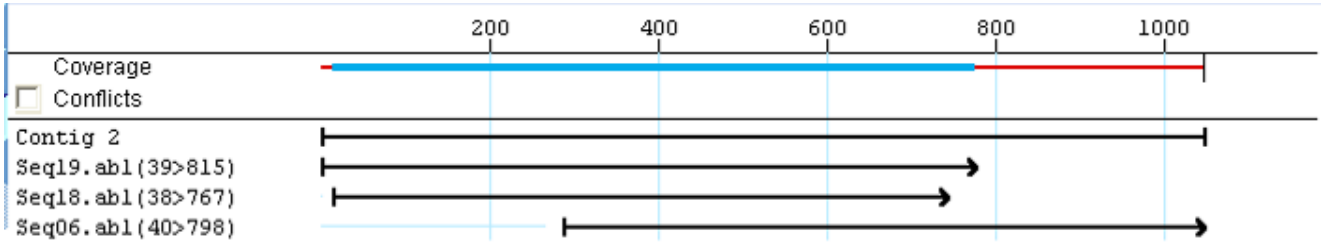
Schritt 1: zum annotierten EST

Rohdaten → **Editieren (Quality & Vectorclipping mit SeqMan)**

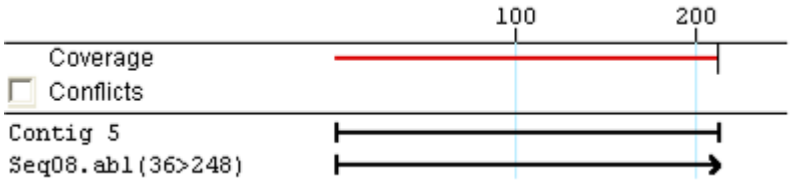




Editierte Daten → **Clustern (mit SeqMan)**

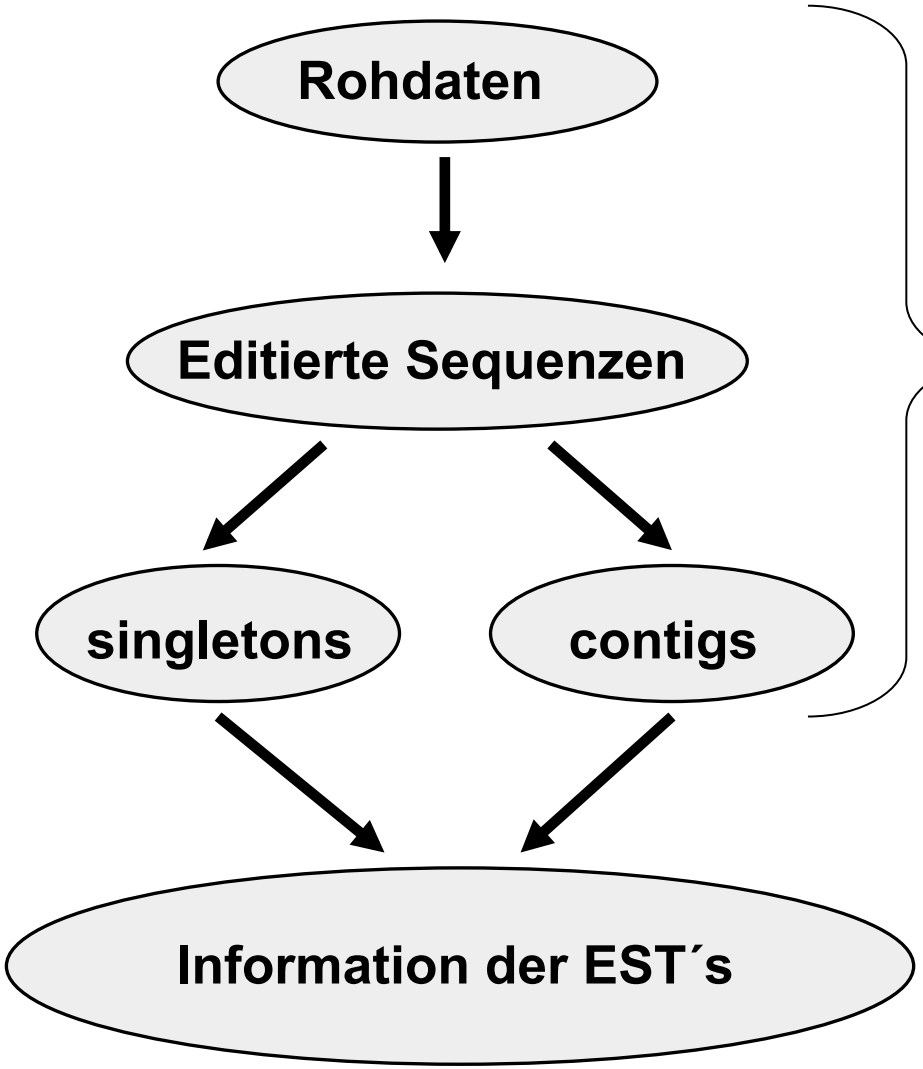


„Contig“



„Singleton“

Schritt 1: zum annotierten EST



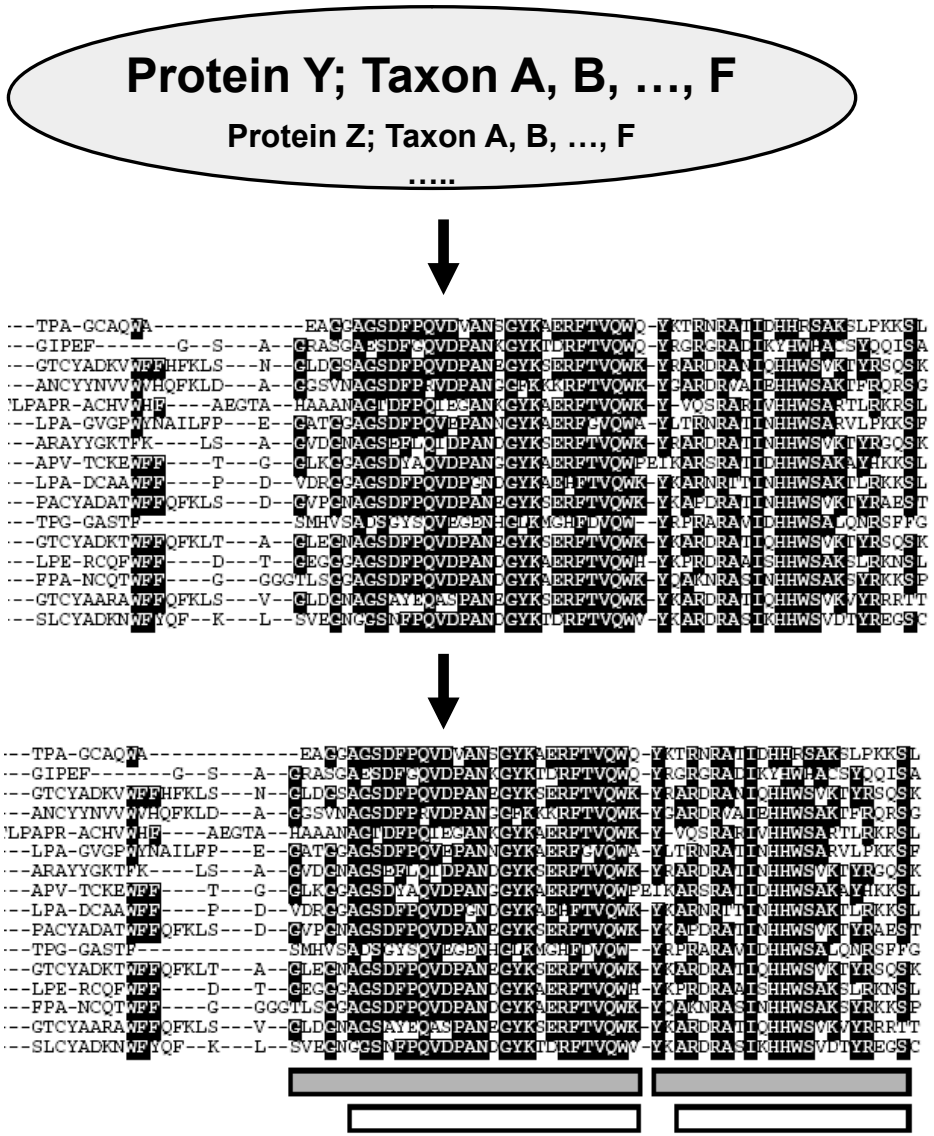
1. Editing und Clustering

- Quality clipping (SeqMan)
- Vector clipping (SeqMan)
- Clustering (SeqMan)

2. Annotation

- BlastX-Suche

Schritt 2: zum Alignment



3. DB-Suche der orthologen Proteine
- RPG
- Blastsuchen

4. Einzelproteinalignments
- MAFFT

5. Editierung
- manuell: GeneDoc
- automatisch: GBlocks

Schritt 3: zur Phylogenie

Taxon	Sequenz Y	Sequenz Z
A	aaaaaaaaaaaaaaaa	----------------
B	----------------	aaaaaaaaaaaaaaaa
C	----------------	aaaaaaaaaaaaaaaa
D	aaaaaaaaaaaaaaaa	aaaaaaaaaaaaaaaa
E	aaaaaaaaaaaaaaaa	aaaaaaaaaaaaaaaa
F	aaaaaaaaaaaaaaaa	----------------

Taxon	konkatenierte Sequenz
A	aaaaaaaaaaaaaaaa----------------
B	----------------aaaaaaaaaaaaaaaa
C	----------------aaaaaaaaaaaaaaaa
D	aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
E	aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
F	aaaaaaaaaaaaaaaa----------------

Alignment, Substitutionsmatrix S

Ein oder mehrere Bäume der Taxa $T_1,T_2,T_3,\ldots,T_X$ der Proteine $P_1,P_2,P_3,\ldots,P_X$ mit der Substitutionsmatrix S und den Methoden $M_1,M_2,M_3,\ldots,M_X$

6. Konkatenierung
- Perl-Skript

7. Substitutionsmatrix-Auswahl
- Prottest

8. Phylogenie
- Mr.Bayes
- RAxML

…dann mal los!

Phylogenie von Xenoturbella

Norén & Jondelius 1997 (Nature)

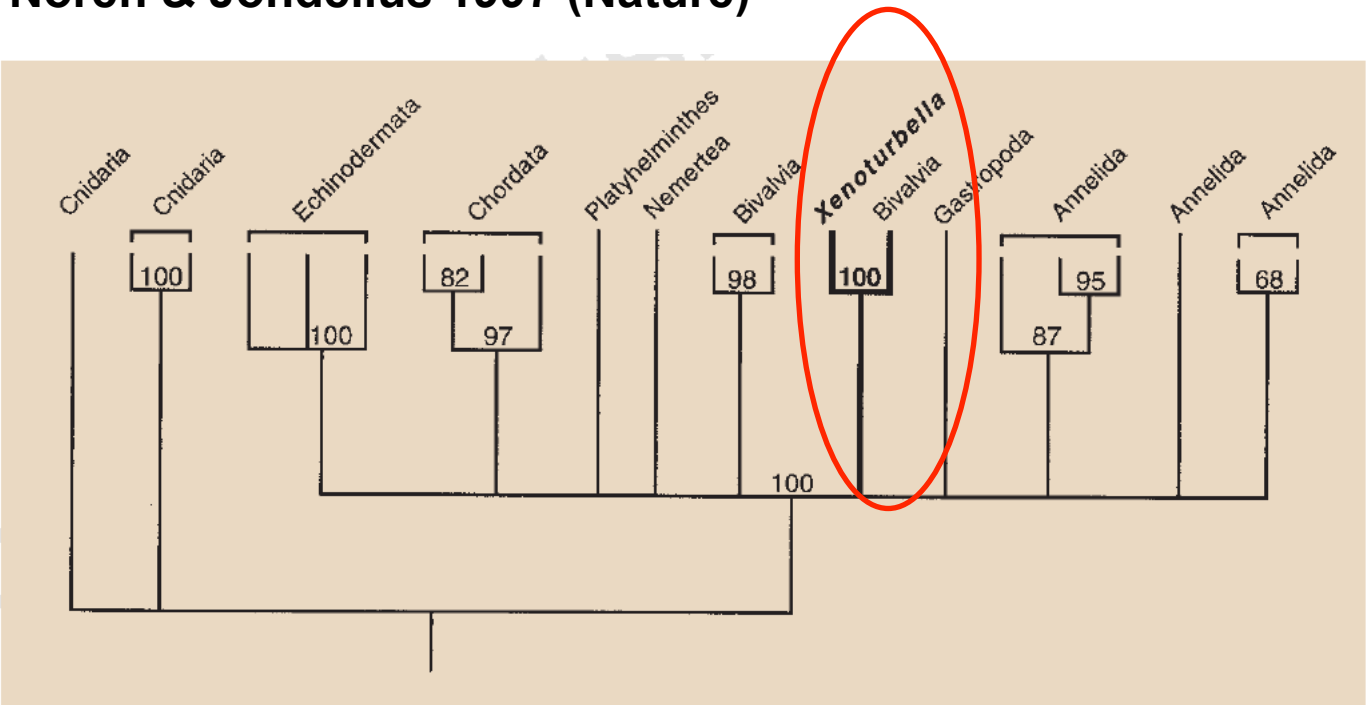


Figure 2 Consensus tree showing groups present in 60% of jack-knife replicates from analysis of COI matrix (3,000 replicates, 5 random additions and branch swapping, deletion frequency e^{-1}). Labels indicate jack-knife frequencies. Full details of tree topology and sequence alignment are available from the authors.

Mitochondriale Cytochrom-Oxidase

Phylogenie von Xenoturbella

Bourlat et al. 2003 (Nature)

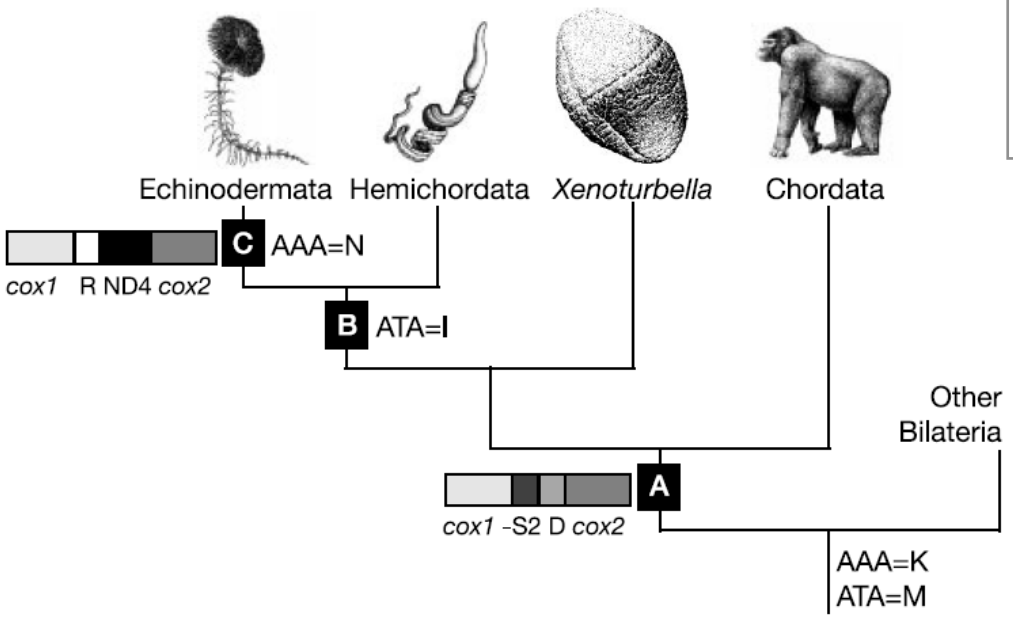


Figure 2 Position of *Xenoturbella* within the deuterostomes as suggested by our analyses of *SSU* and mitochondrial data. The distribution of synapomorphic molecular character states is indicated by a letter. A, monophyly of deuterostomes including *Xenoturbella* supported by common mitochondrial gene order; B, monophyly of Ambulacraria (hemichordates plus echinoderms) to the exclusion of *Xenoturbella* supported by one genetic code change; C, Monophyly of crown-group echinoderms supported by further genetic code change and gene order change.

***Xenoturbella* is a deuterostome that eats molluscs**

Sarah J. Bourlat[1], Claus Nielsen[2], Anne E. Lockyer[3], D. Timothy J. Littlewood[3] & Maximilian J. Telford[1]

→ Xenoturbella Schwestertaxon zu Ambulacraria

Xenoturbella EST Sequenzierung

1186 ESTs

Qualitäts-prüfung

954 informative reads

clustering

495 Singletons and 116 Contigs

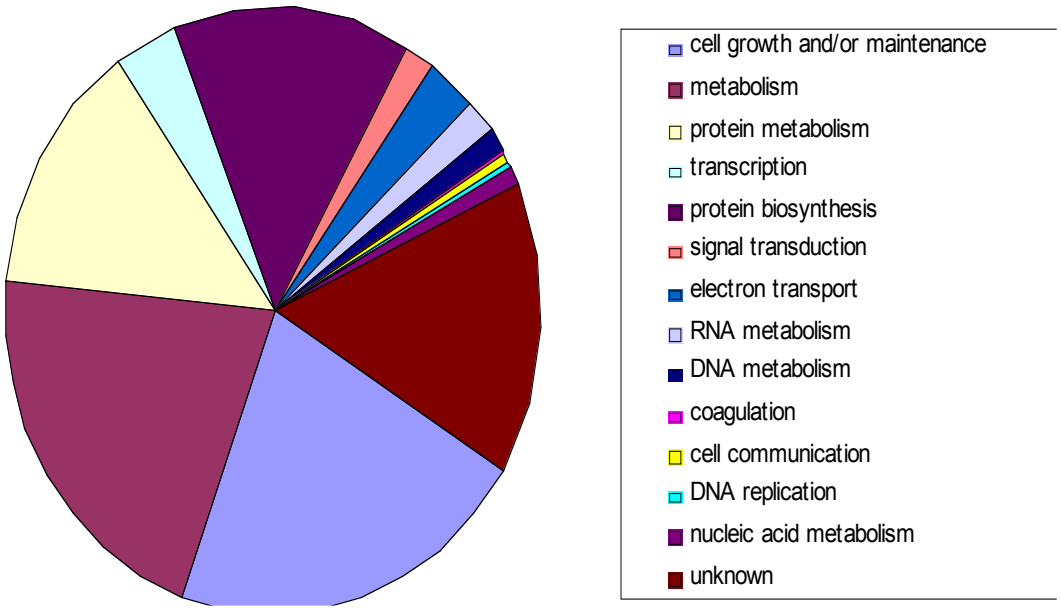
246
$< e^{-10}$

Blast-Suche

365 UniGenes

(265 $> e^{-30}$)

GO Klassifikation



RP Phylogenie

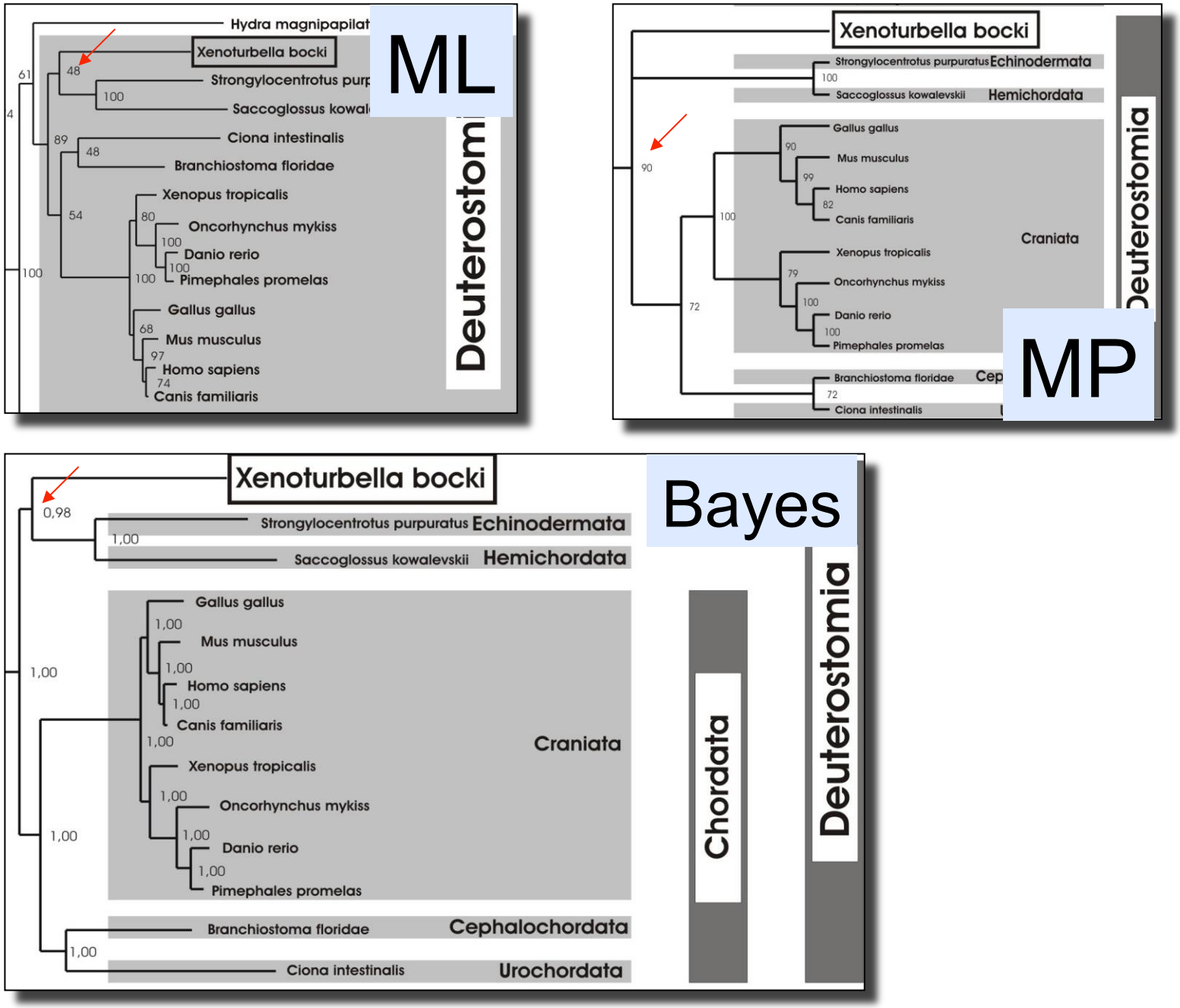
- **30 ribosomale Proteine** identifiziert
- konkatenierter Datensatz: **4757 AS**
- 20 von 28 taxa: Abdeckung >90% relativ zu Mensch

Prottest → **PHYML**, **MrBAYES**

Taxon	ribosomale Proteine														
	L3	L5	L7	L10	L11	L13	L15	L17	L23	L27	L32	L35	L36	L39	SA
Xenoturbella bocki	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Homo sapiens	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Mus musculus	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Canis familiaris	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Gallus gallus	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Danio rerio	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Pimephales promelas	x	x	x	x	x	x	x	x	x	x	x	x	x	x	0
Oncorhynchus mykiss	x	x	x	x	x	x	x	x	x	x	x	x	x	x	0
Branchiostoma floridae	x	x	x	x	x	x	x	x	x	x	x	x	x	x	0
Xenopus tropicalis	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Ciona intestinalis	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Strongylocentrotus purpuratus	x	0	x	x	x	x	x	x	x	x	x	x	x	x	x
Lumbricus rubellus	x	x	x	x	x	x	x	x	x	x	x	x	x	x	0
Ascaris suum	x	x	x	x	x	x	x	x	x	x	x	x	x	x	0
Caenorhabditis elegans	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Drosophila melanogaster	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Anopheles gambiae	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Bombyx mori	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Hydra magnipapillata	x	x	x	x	x	x	x	x	x	x	x	x	x	x	0
Saccharomyces cerevisiae	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Argopecten irradians	x	x	x	x	x	x	0	x	x	x	x	x	x	0	0
Pecten maximus	0	x	x	0	0	0	x	x	x	0	0	x	x	x	0
Crassostrea virginica	x	x	x	x	x	x	x	x	x	0	0	0	x	x	0
Crassostrea gigas	x	x	x	x	x	x	x	x	x	x	x	x	x	0	0
Biomphalaria glabrata	0	x	x	x	x	x	x	x	x	0	x	x	x	x	0
Mytilus galloprovincialis	x	x	x	x	x	x	x	x	x	x	x	x	x	x	0
Arabidopsis thaliana	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Saccoglossus kowalevskii	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Taxon	ribosomale Proteine														
	S2	S3	S4	S5	S8	S11	S12	S16	S19	S20	S23	S25	S26	S7	P1
Xenoturbella bocki	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Homo sapiens	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Mus musculus	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Canis familiaris	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Gallus gallus	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Danio rerio	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Pimephales promelas	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Oncorhynchus mykiss	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Branchiostoma floridae	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Xenopus tropicalis	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Ciona intestinalis	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Strongylocentrotus purpuratus	x	x	x	x	x	x	x	x	x	x	x	0	x	x	x
Lumbricus rubellus	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Ascaris suum	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Caenorhabditis elegans	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Drosophila melanogaster	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Anopheles gambiae	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Bombyx mori	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Hydra magnipapillata	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Saccharomyces cerevisiae	x	x	x	x	x	x	x	x	x	x	x	0	0	x	x
Argopecten irradians	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Pecten maximus	x	0	x	0	x	x	x	0	x	x	x	x	0	0	x
Crassostrea virginica	x	x	x	x	x	x	x	x	x	x	x	x	x	0	x
Crassostrea gigas	x	x	x	x	x	x	x	x	0	x	x	x	x	0	x
Biomphalaria glabrata	x	x	x	x	0	x	x	x	x	x	x	0	0	x	x
Mytilus galloprovincialis	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Arabidopsis thaliana	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Saccoglossus kowalevskii	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Datenquellen: RiboProt, dbEST, TRACE ARCHIVE


ML
Hydra magnipapilat
Xenoturbella bocki
Strongylocentrotus pur
Saccoglossus kowal
Ciona intestinalis
Branchiostoma floridae
Xenopus tropicalis
Oncorhynchus mykiss
Danio rerio
Pimephales promelas
Gallus gallus
Mus musculus
Homo sapiens
Canis familiaris
Deuterostom
MP
Xenoturbella bocki
Strongylocentrotus purpuratus Echinodermata
Saccoglossus kowalevskii Hemichordata
Gallus gallus
Mus musculus
Homo sapiens
Canis familiaris
Craniata
Xenopus tropicalis
Oncorhynchus mykiss
Danio rerio
Pimephales promelas
Branchiostoma floridae
Ciona intestinalis
Deuterostomia
Bayes
Xenoturbella bocki
Strongylocentrotus purpuratus Echinodermata
Saccoglossus kowalevskii Hemichordata
Gallus gallus
Mus musculus
Homo sapiens
Canis familiaris
Craniata
Xenopus tropicalis
Oncorhynchus mykiss
Danio rerio
Pimephales promelas
Branchiostoma floridae Cephalochordata
Ciona intestinalis Urochordata
Chordata
Deuterostomia

Xenoturbella bocki

Oct. 2006

Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida

Sarah J. Bourlat[1], Thorhildur Juliusdottir[2], Christopher J. Lowe[3], Robert Freeman[4], Jochanan Aronowicz[3], Mark Kirschner[5], Eric S. Lander[4,6], Michael Thorndyke[7], Hiroaki Nakano[7], Andrea B. Kohn[8], Andreas Heyland[8], Leonid L. Moroz[8], Richard R. Copley[2] & Maximilian J. Telford[1]

Deuterostomes comprise vertebrates, the related invertebrate chordates (tunicates and cephalochordates) and three other invertebrate taxa: hemichordates, echinoderms and *Xenoturbella*[1]. The relationships between invertebrate and vertebrate deuterostomes are clearly important for understanding our own distant origins. Recent phylogenetic studies of chordate classes and a sea urchin have indicated that urochordates might be the closest invertebrate sister group of vertebrates, rather than cephalochordates, as traditionally believed[2–5]. More remarkable is the suggestion that cephalochordates are closer to echinoderms than to vertebrates and urochordates, meaning that chordates are paraphyletic[2]. To study the relationships among all deuterostome groups, we have assembled an alignment of more than 35,000 homologous amino acids, including new data from a hemichordate, starfish and *Xenoturbella*. We have also sequenced the mitochondrial genome of *Xenoturbella*. We support the clades Olfactores (urochordates and vertebrates) and Ambulacraria (hemichordates and echinoderms[6]). Analyses using our new data, however, do not support a cephalochordate and echinoderm grouping and we conclude that chordates are monophyletic. Finally, nuclear and mitochondrial data place *Xenoturbella* as the sister group of the two ambulacrarian phyla[1]. As such, *Xenoturbella* is shown to be an independent phylum, Xenoturbellida, bringing the number of living deuterostome phyla to four.

(the urochordate plus vertebrate clade[10]) seem to be credible components of the deuterostomes, two further aspects of the phylogeny of the group remain contentious.

The first stems from a remarkable result in a phylogenomic analysis in which cephalochordates were associated with echinoderms rather than with other chordates[2]. This result, if correct, has profound implications for our understanding of early deuterostome evolution and the origins of the chordate body plan, because it suggests that a number of conserved chordate features were present in the common ancestor of all deuterostomes and were secondarily lost in echinoderms and hemichordates. Although this is clearly a controversial finding, it deserves serious consideration. The initially surprising relocation of the hemichordates from chordate sister group to the Ambulacraria has already shown that certain characteristics that were thought to be specific to the chordate linage are likely to derive from the deuterostome common ancestor; the homology of gill slits and endostyle in hemichordates and chordates is now supported by studies both of morphology and of shared gene-expression patterns[9,11,12].

The second open question concerns the position of the worm *Xenoturbella*. Consideration of its small subunit ribosomal RNA gene (SSU) indicated that *Xenoturbella* might be a deuterostome, related to the Ambulacraria but not included in either the echinoderms or the hemichordates[1]. This result is supported by the lack in

Phylogenie von Xenoturbella

Bourlat et al. 2006 (Nature)

Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida

Sarah J. Bourlat[1], Thorhildur Juliusdottir[2], Christopher J. Lowe[3], Robert Freeman[4], Jochanan Aronowicz[3], Mark Kirschner[5], Eric S. Lander[4,6], Michael Thorndyke[7], Hiroaki Nakano[7], Andrea B. Kohn[8], Andreas Heyland[8], Leonid L. Moroz[8], Richard R. Copley[2] & Maximilian J. Telford[1]

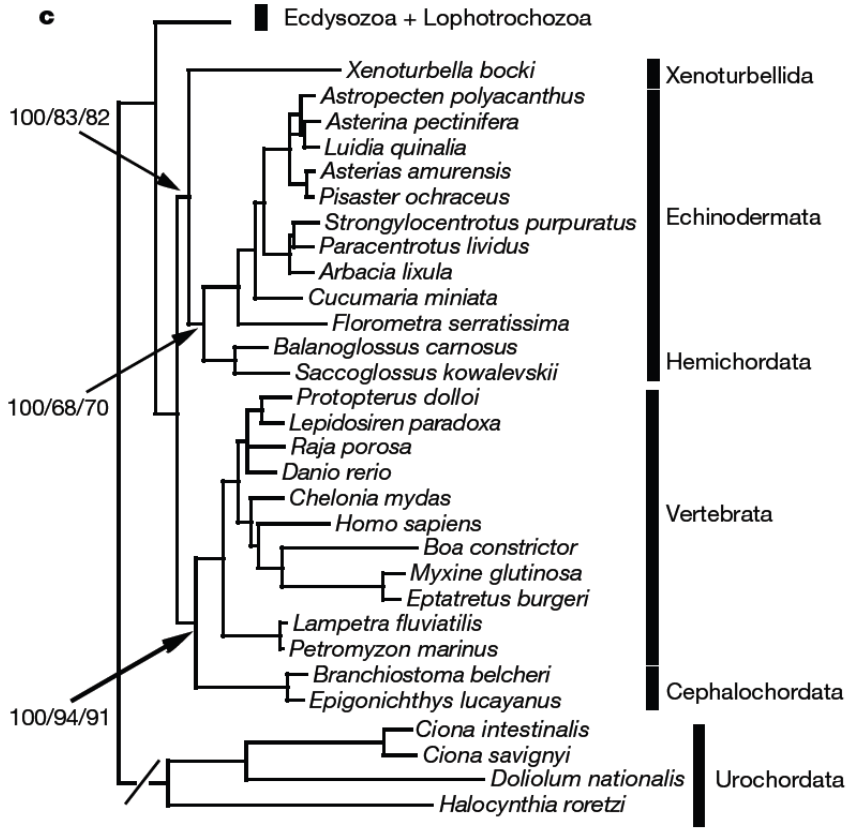


Figure 1 | Phylogenetic analyses of 170 nuclear proteins and 13 mitochondrial proteins support a monophyletic chordate clade and an independent deuterostome phylum of Xenoturbellida.

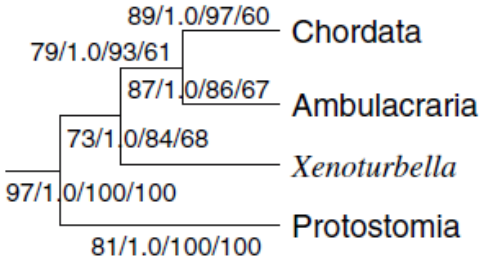
→ Xenoturbella Schwestertaxon zu Ambulacraria

Phylogenie von Xenoturbella

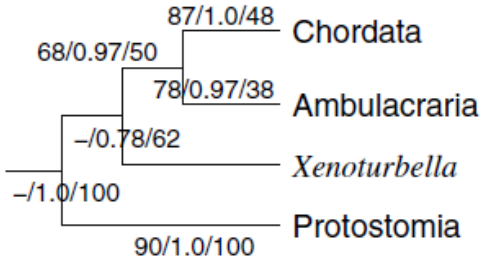
Perseke et al. 2007 (Theory Biosci.)

The mitochondrial DNA of *Xenoturbella bocki*: genomic architecture and phylogenetic analysis

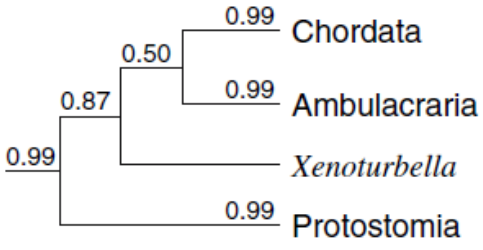
Marleen Perseke · Thomas Hankeln · Bettina Weich · Guido Fritzsch · Peter F. Stadler · Olle Israelsson · Detlef Bernhard · Martin Schlegel



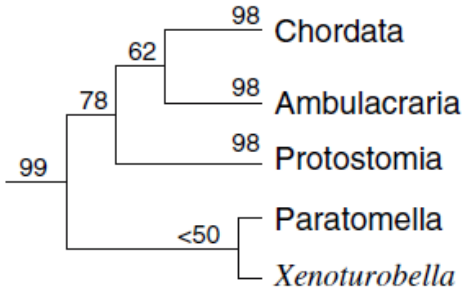
Set A



Set A (MINK)



Set B



Set C

→ Xenoturbella basaler Deuterostomia

Phylogenie von Xenoturbella

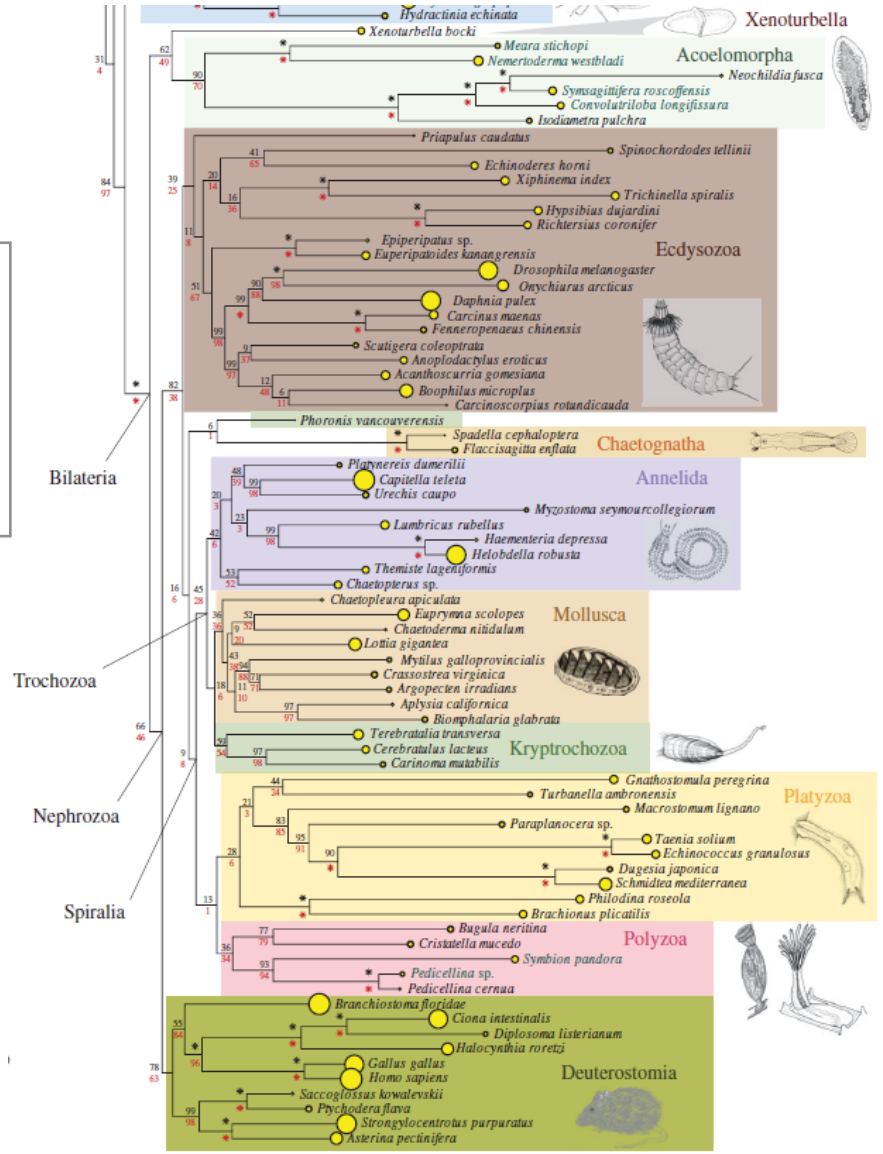
Hejnol et al. 2009 (Proc.R.Soc.B.)

Assessing the root of bilaterian animals with scalable phylogenomic methods

Andreas Hejnol[1,*], Matthias Obst[2], Alexandros Stamatakis[3], Michael Ott[3], Greg W. Rouse[4], Gregory D. Edgecombe[5], Pedro Martinez[6], Jaume Baguñà[6], Xavier Bailly[7], Ulf Jondelius[8], Matthias Wiens[9], Werner E. G. Müller[9], Elaine Seaver[1], Ward C. Wheeler[10], Mark Q. Martindale[1], Gonzalo Giribet[11] and Casey W. Dunn[12,*]

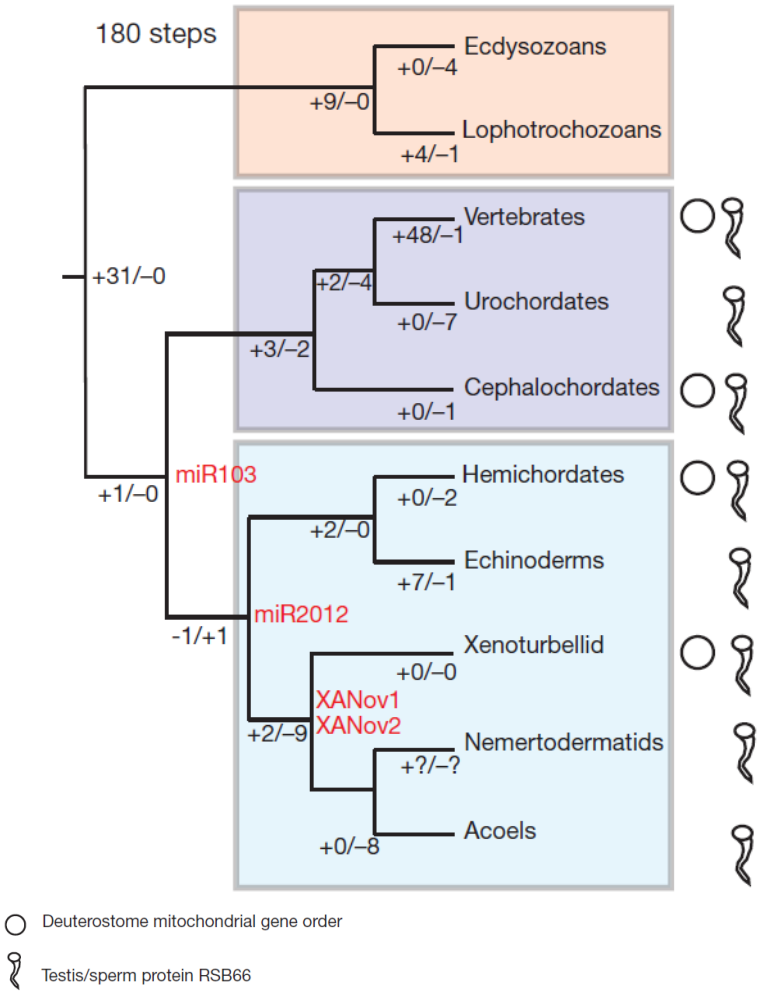
> 1000 Gene (!!)

→ Xenoturbella + Acoelomorpha basale Bilateria



Phylogenie von Xenoturbella

Phillippe et al. 2011 (Nature)



Acoelomorph flatworms are deuterostomes related to *Xenoturbella*

Hervé Philippe[1], Henner Brinkmann[1], Richard R. Copley[2], Leonid L. Moroz[3], Hiroaki Nakano[4]†, Albert J. Poustka[5], Andreas Wallberg[6], Kevin J. Peterson[7] & Maximilian J. Telford[8]

"We propose that the basal emergence of Xenoturbella plus Acoelomorpha observed by Hejnol et al. is the result of an LBA artefact stemming from the use of a sub-optimal site-homogeneous model."

Fast 200 Gene
+ miRNA-Analyse
+ mtDNA Analyse

→ Xenoturbella + Acoelomorpha Schwestertaxon zu Ambulacraria

Phylogenie von Xenoturbella

Lowe & Pani 2011 (Curr.Biol.)

Animal Evolution: A Soap Opera of Unremarkable Worms

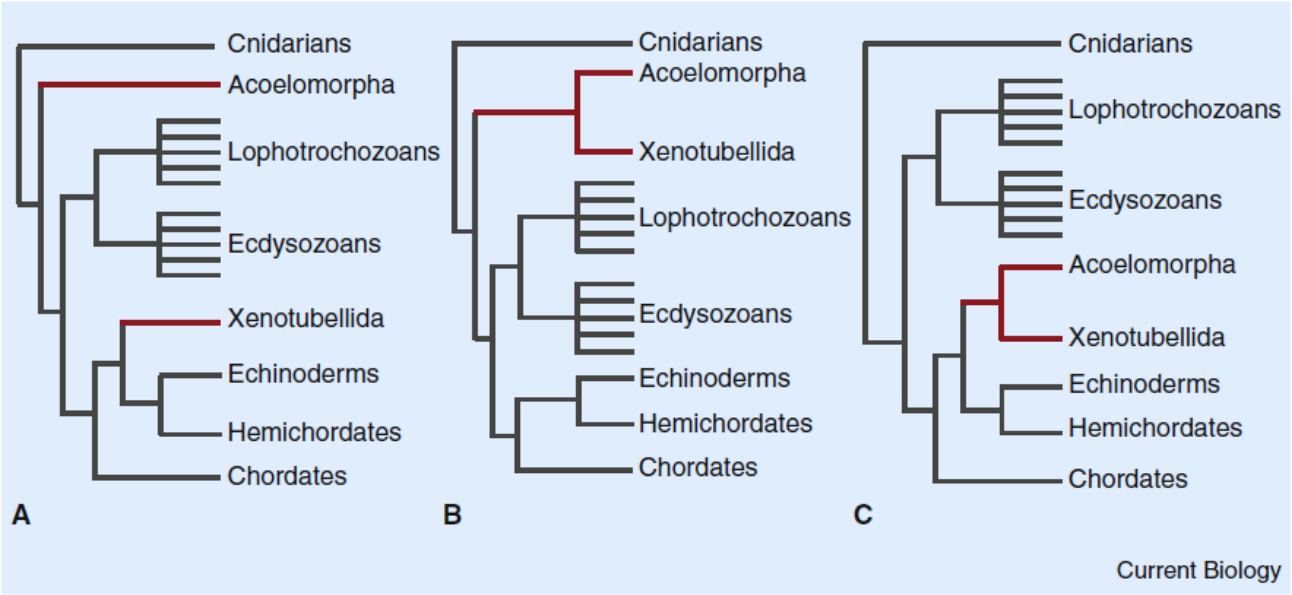


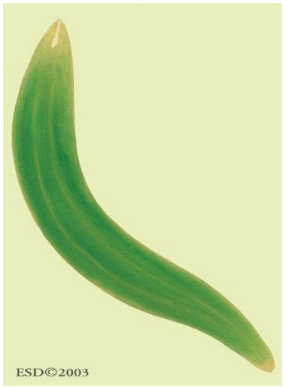
Figure 2. Alternative hypotheses for the phylogenetic position of acoelomorphs and Xenoturbellida within metazoans.

(A) Basal position of Acoelomorpha within bilaterians, but *Xenoturbella* located within the deuterostomes [10]. (B) Grouping of Acoelomorpha with *Xenoturbella* at the base of the bilaterians [6]. (C) New hypothesis of Xenacoelomorpha as sister group to echinoderms and hemichordates within the deuterostomes [2].

„Unremarkable worms“

- Kein Verdauungstrakt
- Marin
- Kein Kreislaufsystem
- Keine Atmungsorgane
- Kein Exkretionssystem
- Neuronennetzwerk
- Hermaphroditen
- Keine Gonaden

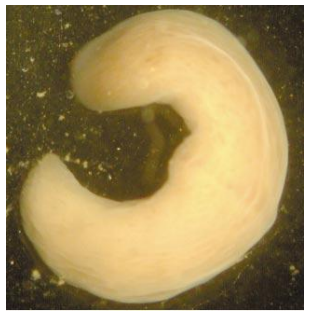
→ Sekundäre Vereinfachungen?



Symsagittifera roscoffensis
(Acoela)

Meara stichopi
(Nemertodermatida)

Nemertoderma westbladi
(Nematodermatida)



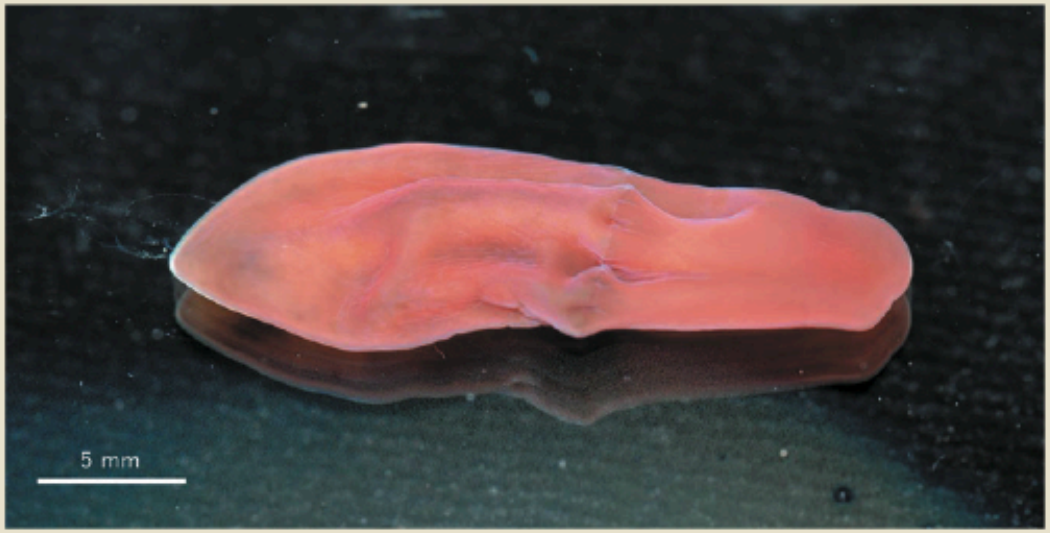
Xenoturbella bocki
(Xenoturbellida)

2016...Die Geschichte geht weiter!

PHYLOGENY

A home for *Xenoturbella*

Sometimes it is the most unassuming animals that cause the most consternation. *Xenoturbella* (pictured) are simple marine flatworms with no brain, anus, gonads, excretory system or through gut, so one would expect them to find a home among the acoels — similarly simple animals thought to lie at the base of the evolutionary tree of Bilateria, bilaterally symmetrical animals. Yet *Xenoturbella* have caused puzzlement since they were first described in 1949, because quibbles about their ultrastructure and mitochondrial DNA sequences have meant that the worms have never sat entirely happily in their assumed station.



Analysis of nuclear DNA sequences underlined the oddity: *Xenoturbella* were even thought to be highly degenerate molluscs until the revelation that molluscs are what *Xenoturbella* eat. Even stranger was the proposal that *Xenoturbella* and other acoels were most closely related to hemichordates (animals known as acorn worms and pterobranchs) and echinoderms (radially symmetrical marine animals such as sea urchins and starfish). This cast into question the timing of the evolution of several advanced characteristics, such as gill slits, that are shared by members of the deuterostome branch of Bilateria (to which hemichordates and echinoderms belong), but that are lacking in *Xenoturbella*. It even raised questions about the last common ancestor of Bilateria — perhaps *Xenoturbella* were not as simple as they looked, but had degenerated from a structurally more complex ancestor.

These questions are all but resolved by two studies in this week's issue. Cannon *et al.* (page 89)[1] present a robust phylogenetic analysis based on the gene-transcript profiles of eleven species of *Xenoturbella* and other acoels. This shows that the combined group, known as Xenacoelomorpha, indeed lies at the very base of the bilaterian radiation. Rouse *et al.* (page 94)[2] add four new species of *Xenoturbella* from the eastern Pacific Ocean to the one already known from the waters of Scotland and Scandinavia. The authors' anatomical and phylogenetic studies on these new forms add weight to the idea that these worms were the earliest to branch from other bilaterians. Zoologists can exhale, and their shy charges can resume their diet of molluscs in peace. **Henry Gee**

1. Cannon, J. T. *et al. Nature* **530,** 89–93 (2016).
2. Rouse, G. W., Wilson, N. G., Carvajal, J. I. & Vrijenhoek, R. C. *Nature* **530,** 94–97 (2016).

2016...Die Geschichte geht weiter!

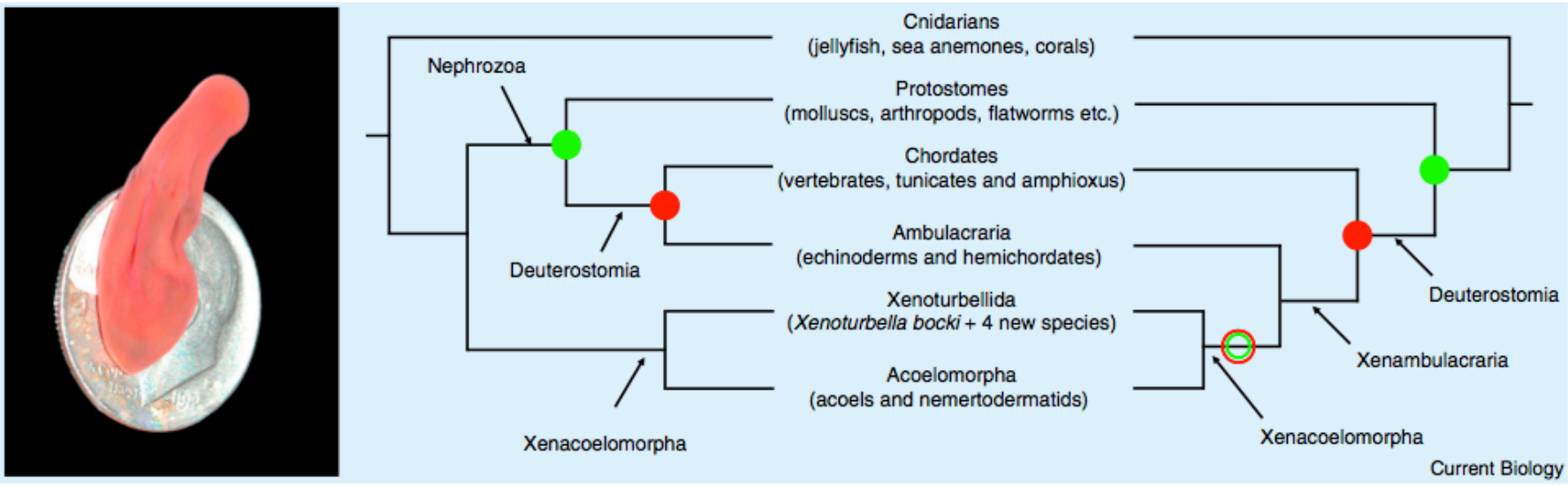


Figure 1. A new *Xenoturbella* species and possible relationships between Xenacoelomorpha and other major groups of animals.
(far left) The new species *Xenoturbella hollandorum*, approximately 2.5 cm long (Photo: Greg Rouse). On the left, the phylogenetic tree supported by Rouse *et al.* [3] and Cannon *et al.* [4] in which xenacoelomorphs branch before the common ancestor of protostomes and deuterostomes (Nephrozoa). Absences of characters result from Xenacoelomoprha diverging before the origin (red and green filled circles) of these characters. On the right, the tree supported by previous analyses of large data matrices [8]. Xenacoelomorphs are deuterostomes most closely related to Ambulacraria. Some characters common to protostomes and deuterostomes (filled green circle) or present in the deuterostome common ancestor (filled red circle) are absent from xenacoelomorphs through loss (empty red and green circles).