Molekulare Phylogenie und freie Software

Kerstin Hoef-Emden

kerstin.hoef-emden@uni-koeln.de

Was ist „Phylogenie“?

phylon (altgriech.) = „Stamm“

genesis (altgriech.) = „Ursprung“

Phylogenie

Stammesgeschichte und Verwandtschaften der Organismen

Darstellung in Form von **phylogenetischen Bäumen**

Phylogenetik

Wissenschaft von der Stammesgeschichte der Organismen und ihrer Verwandtschaften

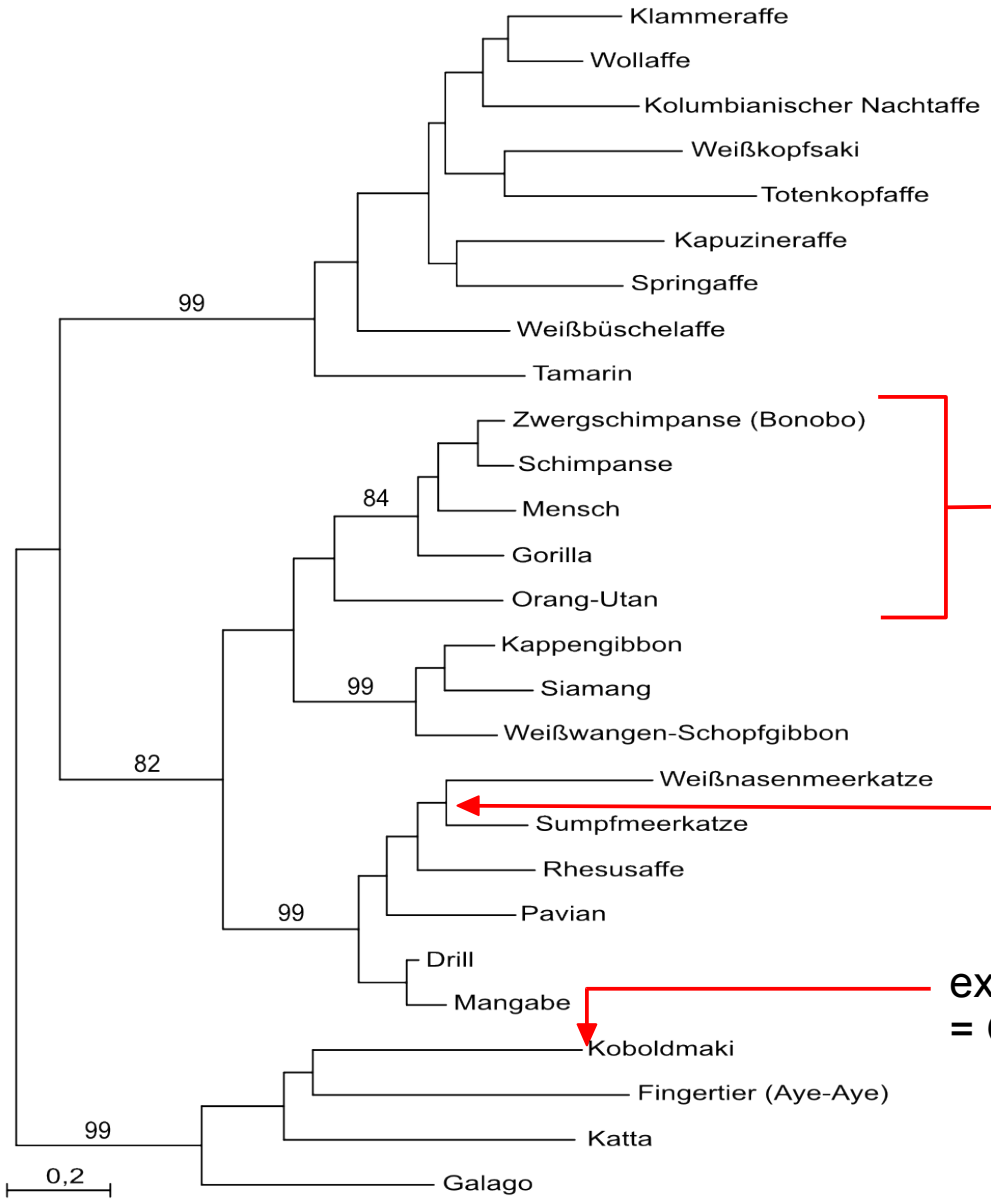
Systematik

Klassifizierung von Organismen anhand ihrer Stammes- und Evolutionsgeschichte

Taxonomie

Klassifizierung und Benennung von Organismen nach den Regeln der Nomenklatur-Codes (ICBN, ICZN, INCB)

Interpretation von phylogenetischen Bäumen



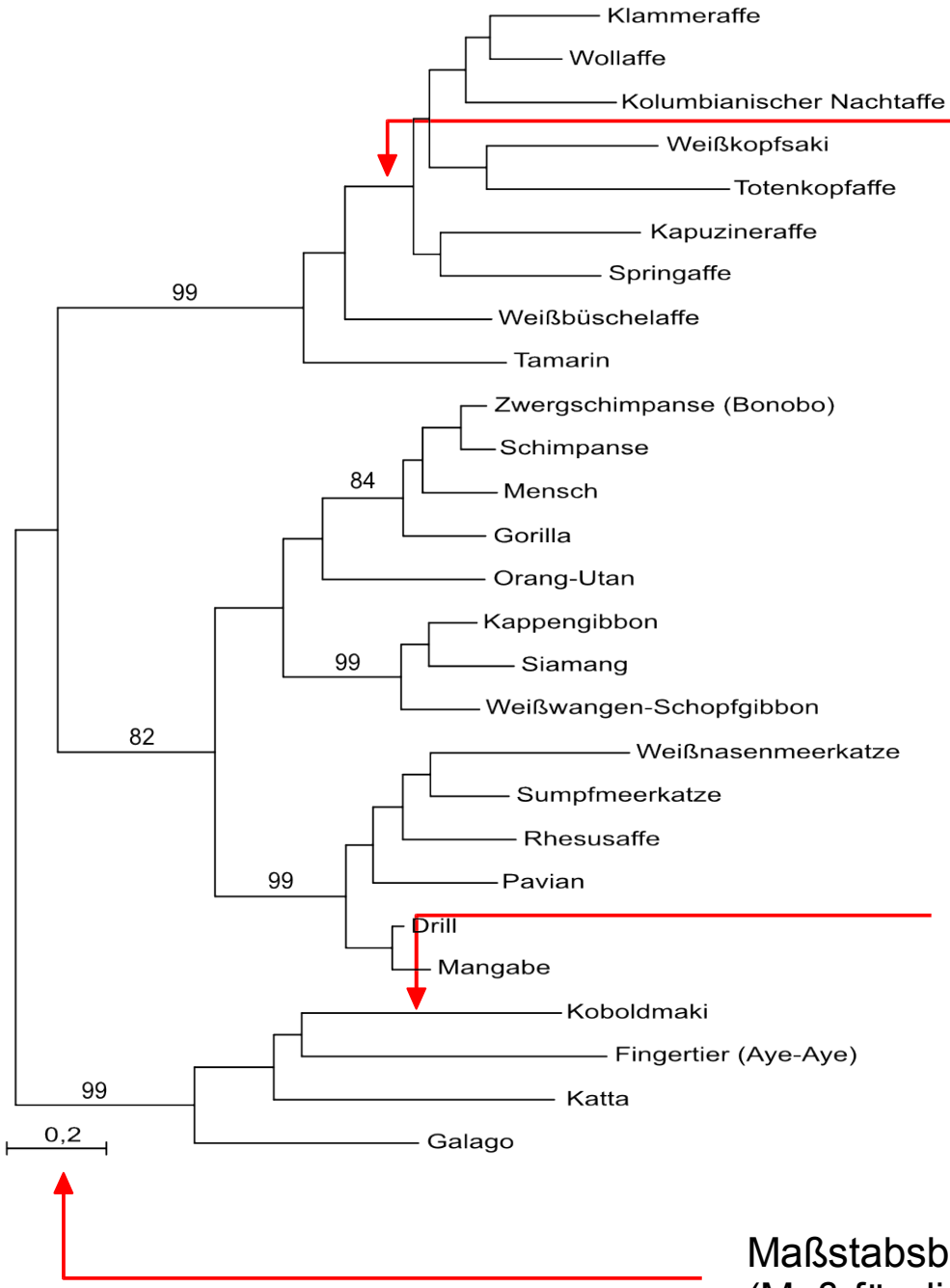
Klados = Gruppe

interner Knoten = hypothetischer Vorfahr (Vergangenheit)

externer Knoten = heute lebende Art (Gegenwart) = OTU (*operational taxonomic unit*)

Zurück in die Vergangenheit

Interpretation von phylogenetischen Bäumen



interner Ast = Entwicklungsweg zwischen hypothetischen Vorfahren

externer Ast = Entwicklungsweg vom letzten gemeinsamen Vorfahren bis heute (hier von Koboldmaki und Fingertier)

Maßstabsbalken = Angabe in Mutationen pro Position (Maß für die Evolutions-/Mutationsrate)

Was ist „Phylogenie“?

phylon (altgriech.) = „Stamm“

genesis (altgriech.) = „Ursprung“

Phylogenie

Stammesgeschichte und Verwandtschaften der Organismen

Darstellung in Form von **phylogenetischen Bäumen**

Phylogenetik

Wissenschaft von der Stammesgeschichte der Organismen und ihrer Verwandtschaften

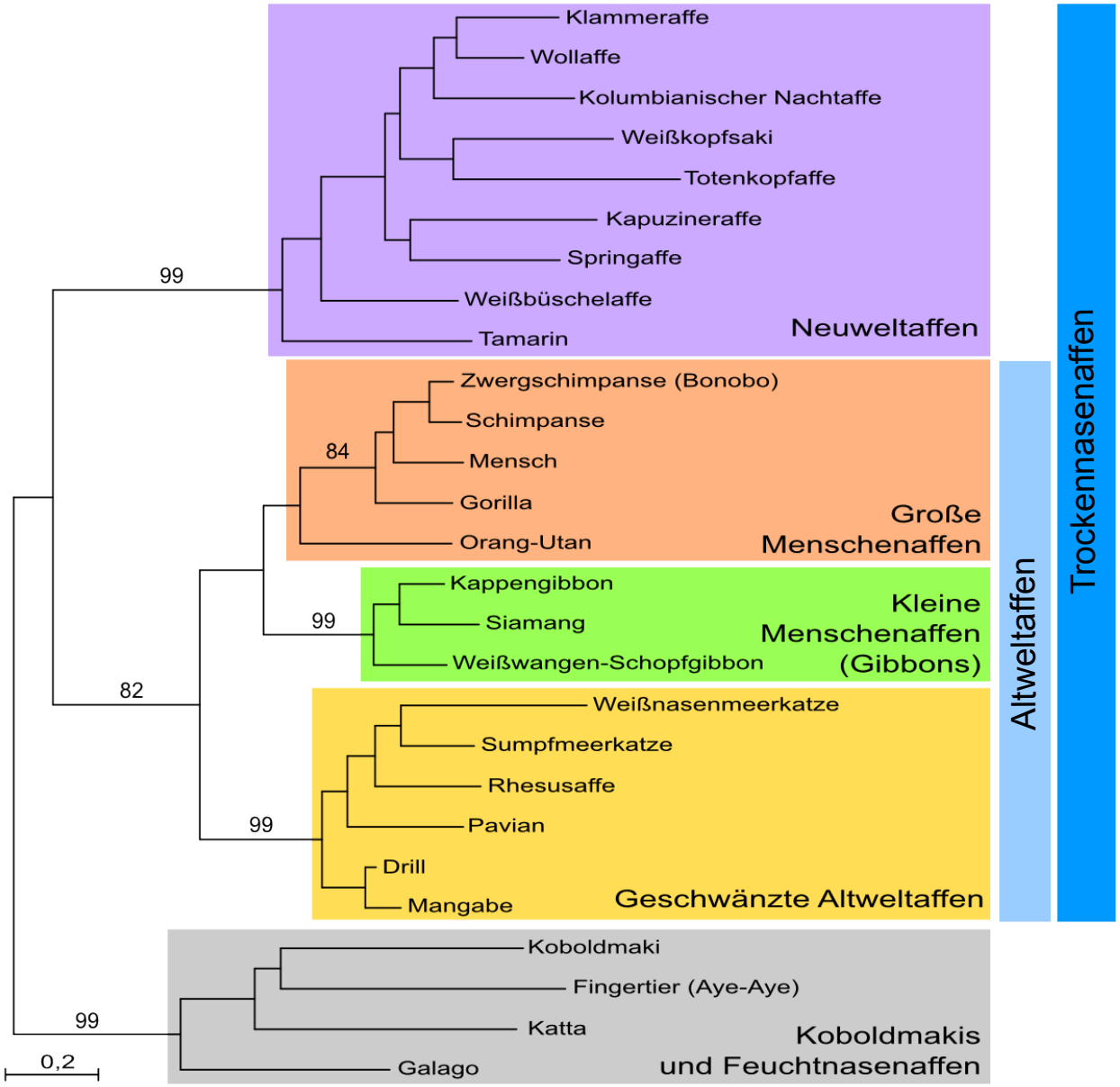
Systematik

Klassifizierung von Organismen anhand ihrer Stammes- und Evolutionsgeschichte

Taxonomie

Klassifizierung und Benennung von Organismen nach den Regeln der Nomenklatur-Codes (ICBN, ICZN, INCB)

Interpretation von phylogenetischen Bäumen



Systematik =

Klassifizierung der Organismen aufgrund ihrer Verwandtschaften und Evolutionsgeschichte

Taxonomie =

Berücksichtigung der Regeln der Nomenklaturcodes bei der Klassifizierung (ansonsten ist die Klassifizierung ungültig)

Die ältesten Datenmatrizen: Morphologische Daten

(Schädel-, Gebissformen, Blütenfarbe, Wuchsformen, Blütenstruktur etc.)

Heute fast nur noch bei Fossilien angewandt (z. B. Dinosaurier).

Nachteile:

- Meist nur wenige morphologische Merkmale verfügbar (wenige 100), daher keine Bäume mit hohen Auflösungen.
- Keine tiefen Phylogenien über den gesamten Baum des Lebens hinweg (Phylogenie unter Einbeziehung von Pflanzen **und** Tieren **und** Einzellern?)
- Äußeres Erscheinungsbild von vielen Genen abhängig, daher keine Berechnung von Wahrscheinlichkeiten für Mutationen berechenbar.

Morphologische Merkmale: Datenmatrix

OTU	Mundwerkzeuge	Flügel	Antennen	...	Merkmal n
Biene	1	1	1		
Fliege	3	2	2		
Mücke	2	2	3	Zeile = Merkmalssequenz einer OTU	
Wespe	4	1	1		
Maikäfer	4	3	4		

Mundwerkzeuge: 1 = saugend, 2 = stechend-saugend, 3 = leckend-saugend, 4 = beißend

Flügel: 1 = 2 Paar Flügel, 2 = Flügel + Schwingkölbchen, 3 = Deckflügel + häutige Flügel

Antennen: 1 = abgeknickt, 2 = kurze Stummel, 3 = gerade, 4 = gefächert

Morphologische Merkmale: Datenmatrix

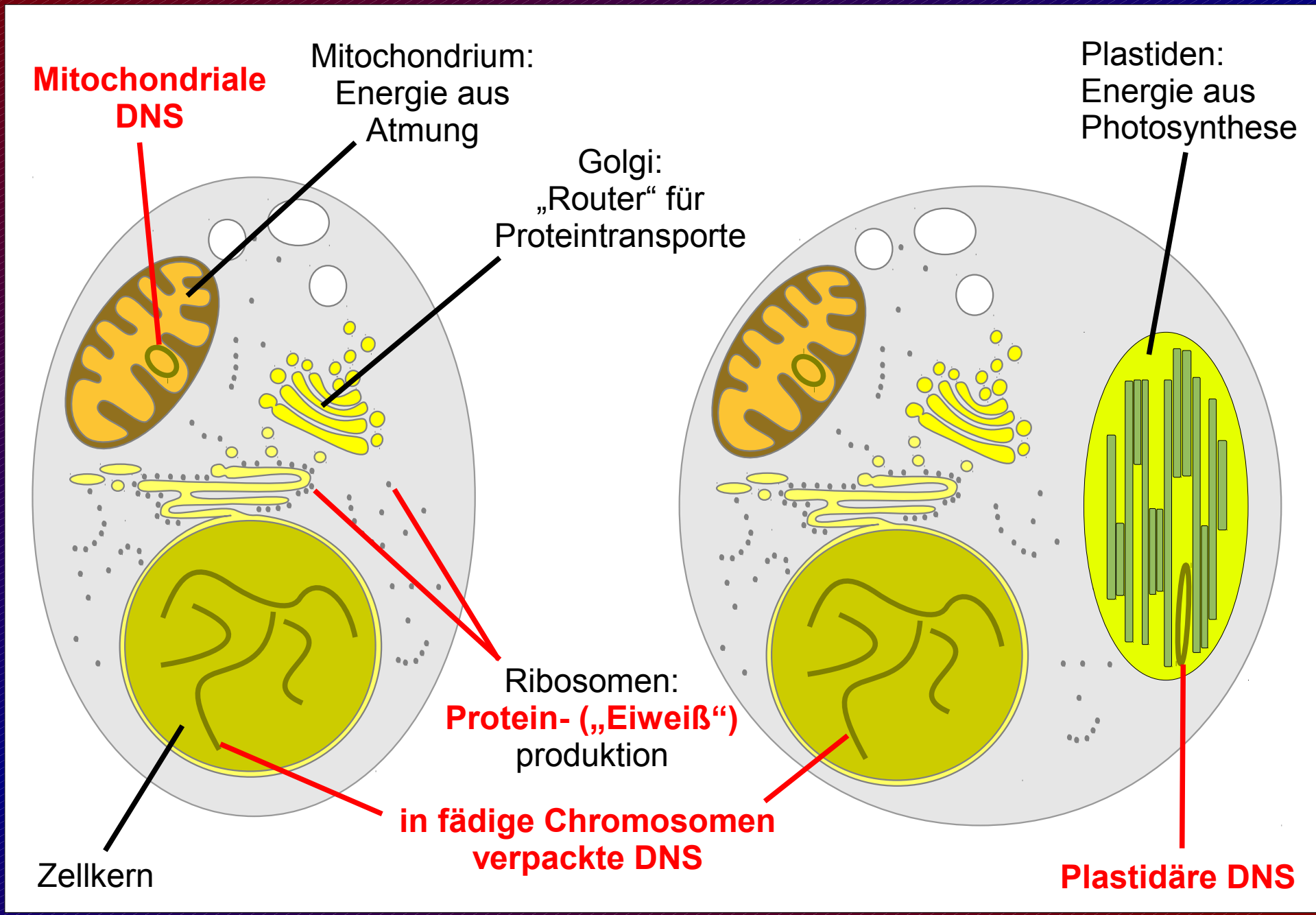
OTU	Mundwerkzeuge	Flügel	Antennen	...	Merkmal n
Biene	1	1	1		
Fliege	3	2	2		
Mücke	2	2	3		
Wespe	4	1	1		
Maikäfer	4	3	4		
	Jede Spalte = Merkmalszustände eines Merkmals				

Mundwerkzeuge: 1 = saugend, 2 = stechend-saugend, 3 = leckend-saugend, 4 = beißend

Flügel: 1 = 2 Paar Flügel, 2 = Flügel + Schwingkölbchen, 3 = Deckflügel + häutige Flügel

Antennen: 1 = abgeknickt, 2 = kurze Stummel, 3 = gerade, 4 = gefächert

Welche Daten werden heute für Stammbäume verwendet?



Welche Daten werden für Stammbäume verwendet?

Proteingene | **RNS-Gene**

DNS

C A T G T T G A T
G T A C A A C T A

Transkription
(= Kopieren eines Stranges)

Boten-RNS
(mRNA)

C A U G U U G A U

Translation
(= Übersetzung
des Triplett-Codes
in Aminosäuren)

His-Val-Asp-Glu-Pro-Ala-...

Faltung

Protein

Bindegewebsfasern
(Kollagen ...)

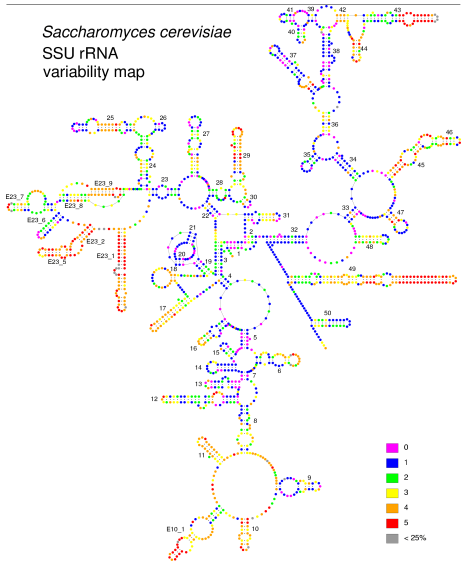
Sekretion
(Schweiß, Hormone ...)

Zellproteine
(Ionenkanäle, Rezeptoren ...)

Enzyme
(Beschleunigung/Steuerung
von chemischen Reaktionen:
Produktion von Kohlenhydraten,
Fetten ...)

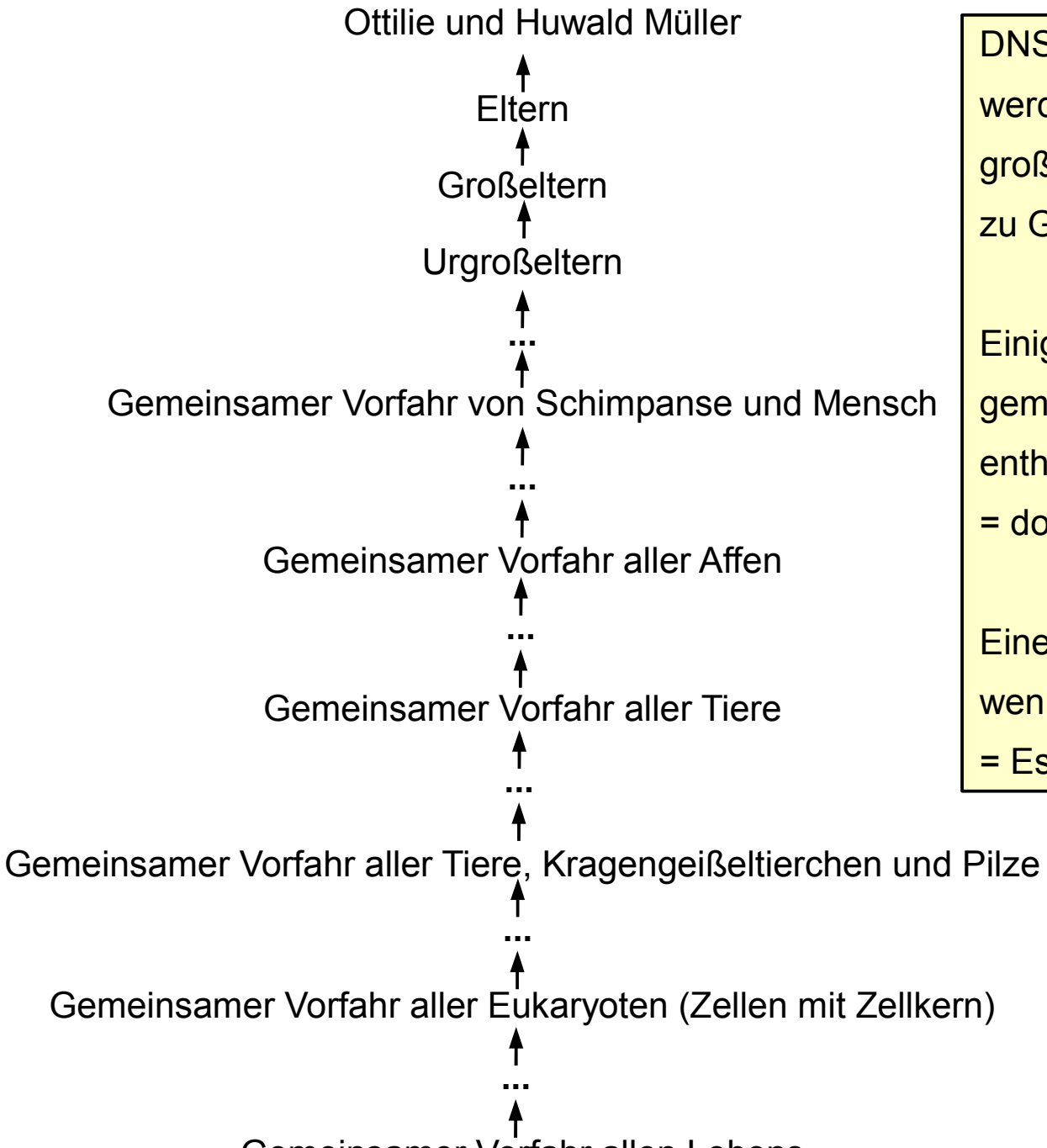
rRNS

Faltung



http://bioinformatics.psb.ugent.be/webtools/rRNA/

Warum DNS und Proteine?

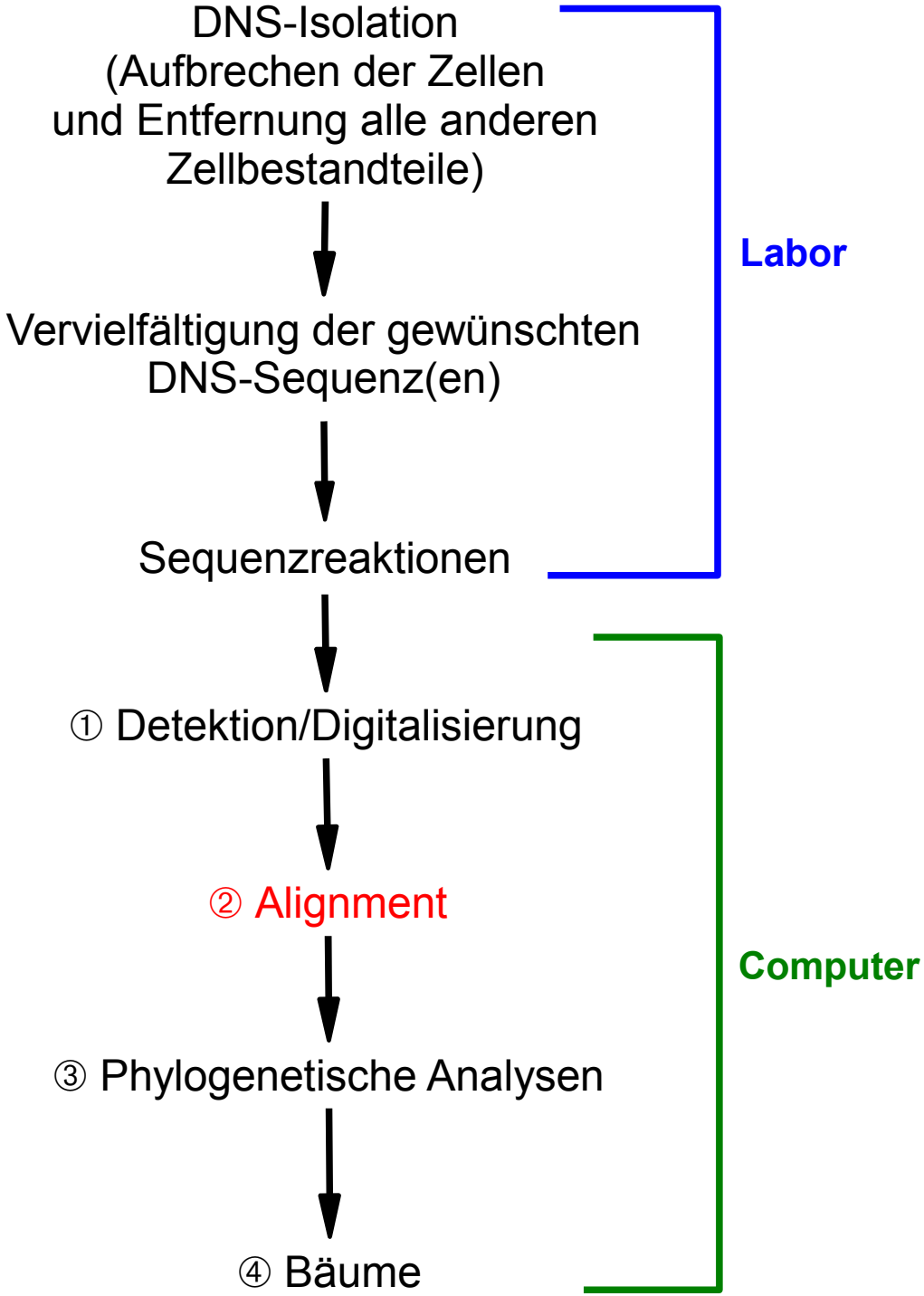


DNS-Sequenzen (und indirekt Proteine) werden unter stetigen kleinen oder großen Veränderungen von Generation zu Generation weiter vererbt.

Einige Gene waren bereits im gemeinsamen Vorfahren allen Lebens enthalten.
= dokumentieren > 3 Mrd. Jahre Evolution!

Eine Vererbung der Gene ist nur möglich, wenn die Linie nicht unterbrochen wird.
= Es gibt keine *Missing Links*!

Ablauf eines Phylogenieprojektes



Pipeline verschiedener Programme:

① Detektion/Digitalisierung:
proprietär

② Alignment:
Public Domain und freie Software

③ Phylogenetische Analyse-programme für DNA- oder Proteinsequenzen:
meist freie Software

④ Graphische Darstellung:
Public Domain oder freie Software

Proteingene: der genetische Code

Signal/ Aminosäure	Kürzel	Codons (DNS)
Start		ATG
Methionin	Met, M	ATG
Tryptophan	Trp, W	TGG
Tyrosin	Tyr, Y	TAT TAC
Phenylalanin	Phe, F	TTT TTC
Cystein	Cys, C	TGT TGC
Asparagin	Asn, N	AAT AAC
Aspartat	Asp, D	GAT GAC
Glutamin	Gln, Q	CAA CAG
Glutamat	Glu, E	GAA GAG
Histidin	His, H	CAT CAC
Lysin	Lys, L	AAA AAG
Isoleucin	Ile, I	ATT ATC ATA
Glycin	Gly, G	GGT GGC GGA GGG
Alanin	Ala, A	GCT GCC GCA GCG
Valin	Val, V	GTT GTC GTA GTG
Threonin	Thr, T	ACT ACC ACA ACG
Prolin	Pro, P	CCT CCC CCA CCG
Leucin	Leu, L	CTT CTC CTA CTG TTA TTG
Serin	Ser, S	TCT TCC TCA TCG AGT AGC
Arginin	Arg, R	CGT CGC CGA CGG AGA AGG
Stop		TAA TAG TGA

Alignment = molekulare Datenmatrix

Proteingene:

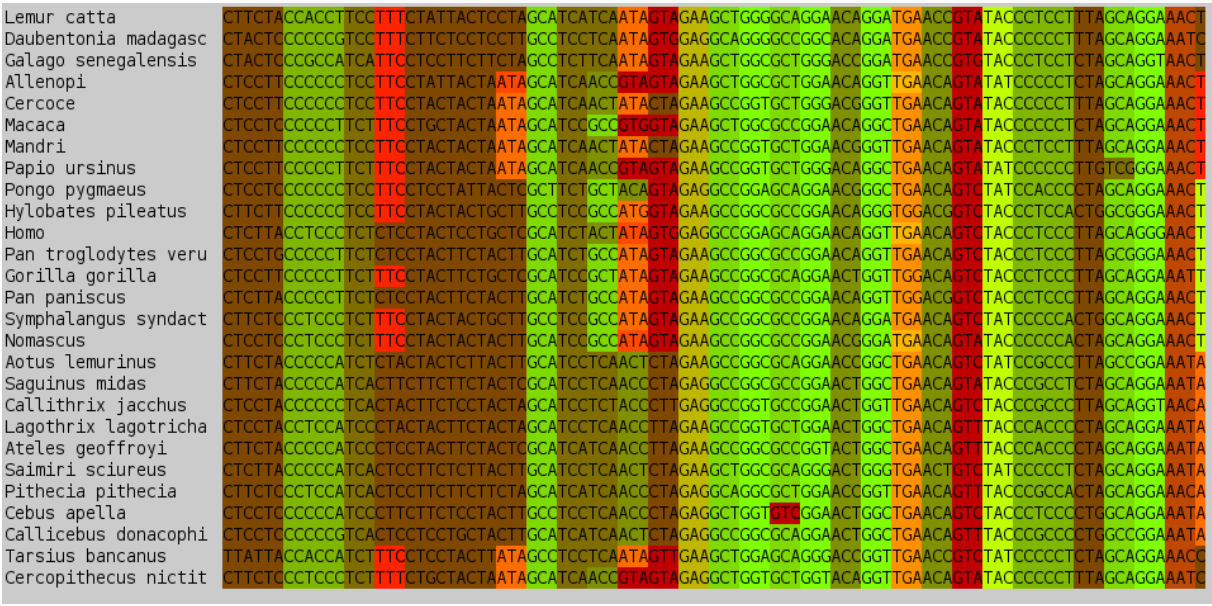
DNS-Sequenzen

mit Triplett-Code;

4 Nucleotide (C, A, T, G);

Leseraster muss

eingehalten werden

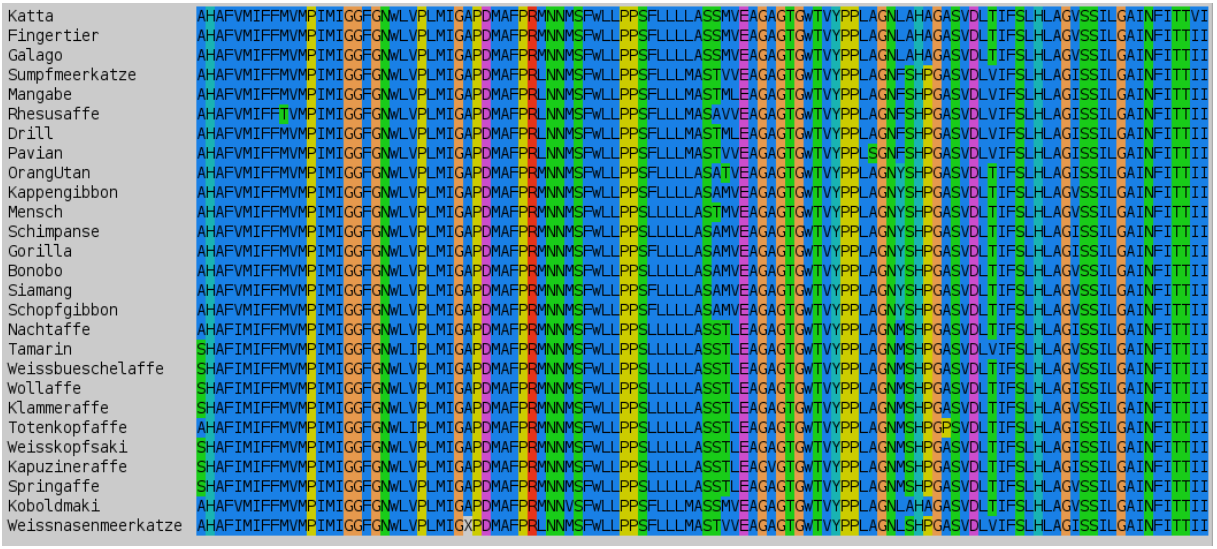


Proteine:

Genprodukt der DNS

= Tripletts übersetzt in

20 Aminosäuren



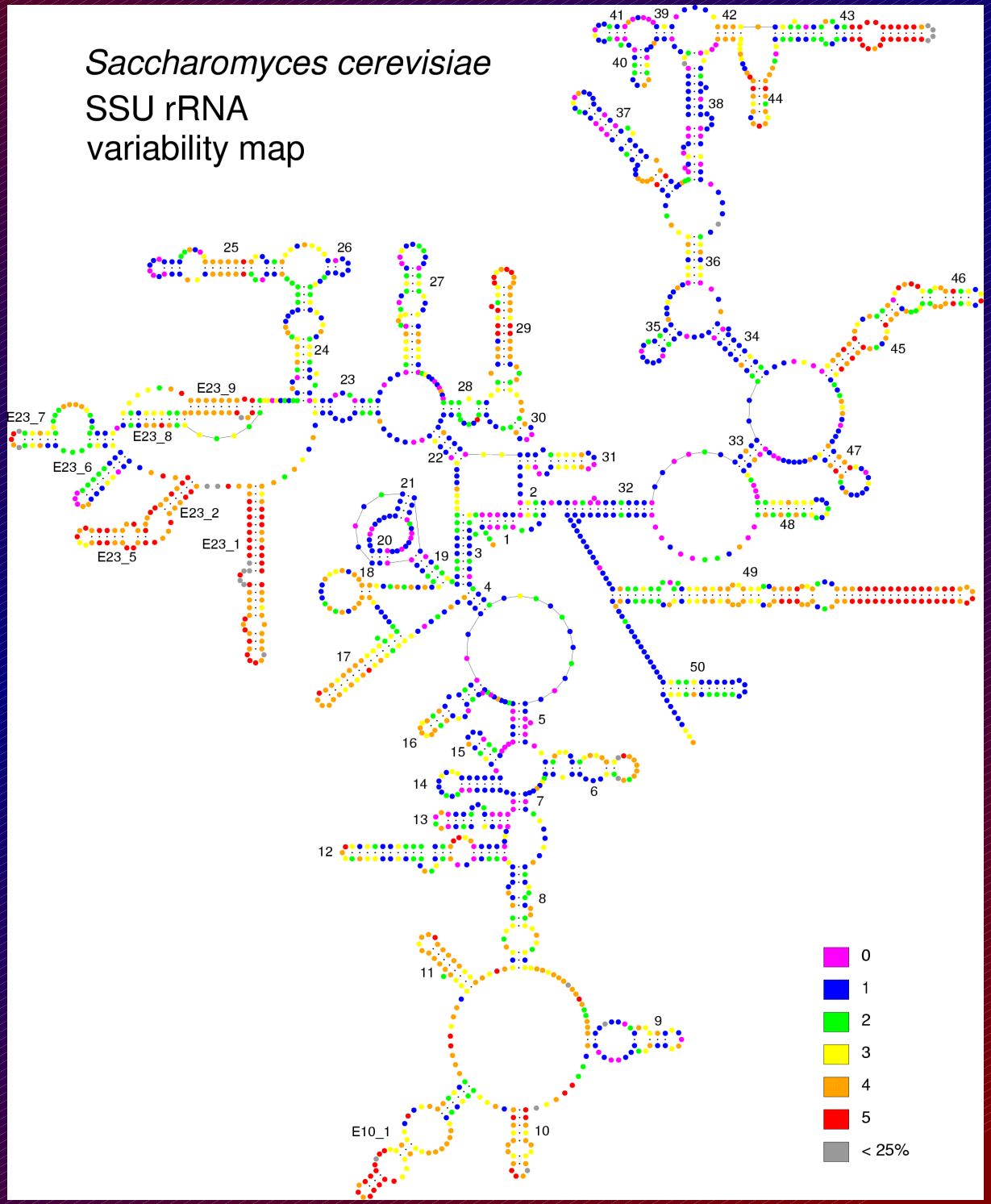
rRNS-Gene

rRNS-Gene:

DNS-Sequenzen mit vielen Einfügungen und Lücken

Ursache:

Hochvariable Bereiche sind funktionell weniger wichtig und mutieren daher schneller.



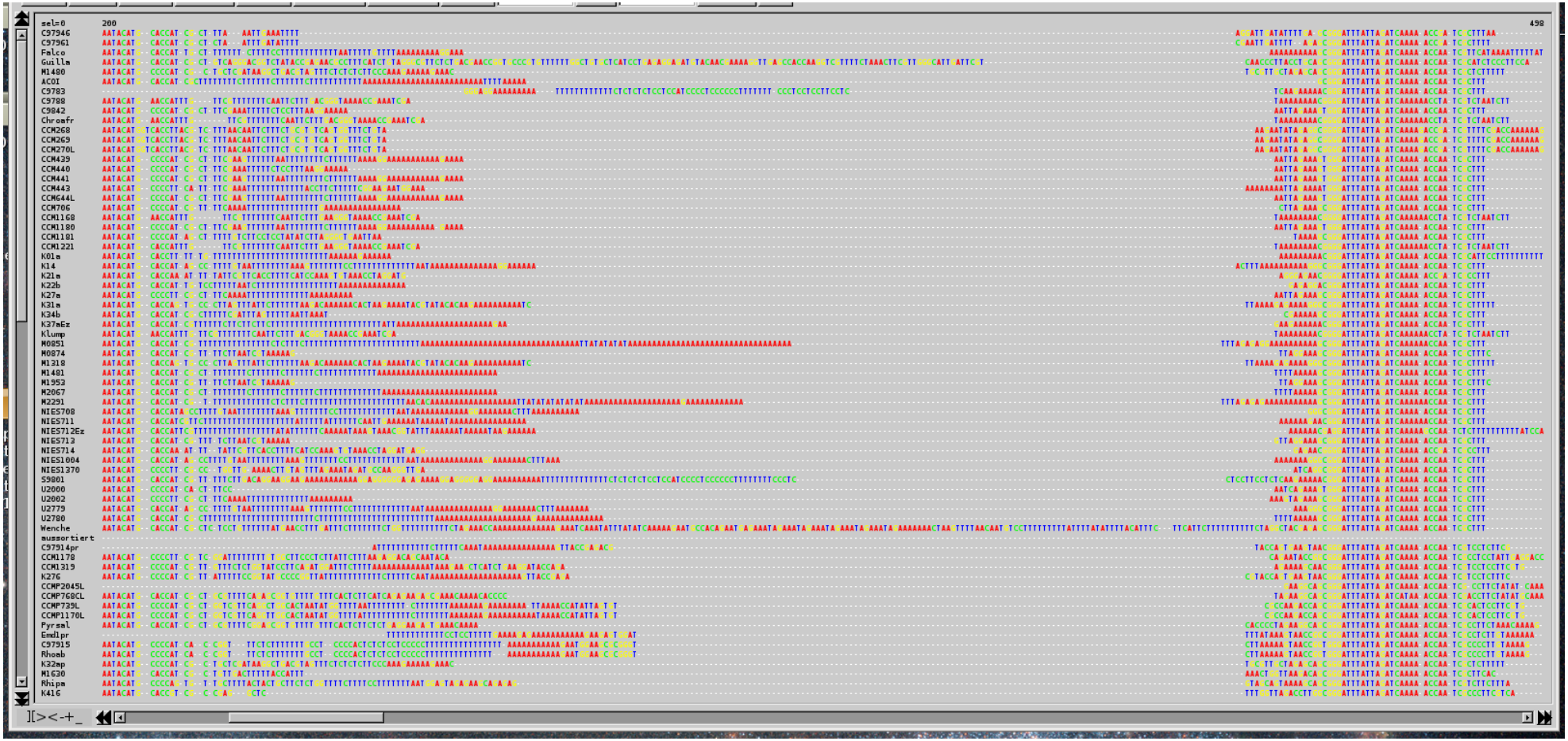
http://bioinformatics.psb.ugent.be/webtools/rRNA/

Alignment = molekulare Datenmatrix

rRNS-Gene:

DNS-Sequenzen mit vielen Insertionen und Lücken

Ursache: Sekundärstruktur



Interaktive Sequenzalignment-Editoren

- Import von DNS- oder Proteinsequenzen (Download aus Datenbanken, Dateiimport)
- Export von Dateien im Nexus- und Phylipformat für phylogenetische Analysen
- Selektion von Positionen, die in Analysen verwendet werden sollen

  (= Entfernung von nicht-alinierbaren Regionen)
- Schnittstelle zu automatischen Alignmentprogrammen für ein erstes Alignment
- Interaktives Editieren der Daten, um Alignmentfehler manuell zu korrigieren
- Löschsperre (verhindert versehentliches Löschen von Nucleotiden oder Aminosäuren)
- Suchen von Motiven

Interaktive Sequenzalignment-Editoren

BioEdit (http://www.mbio.ncsu.edu/bioedit/bioedit.html) - *Closed Source*, nur Windows

Se-Al (http://tree.bio.ed.ac.uk/software/seal/) - *Closed Source*, nur MacOS

STRAP (http://www.bioinformatics.org/strap/) - Java-Programm, nur Proteinsequenzen,

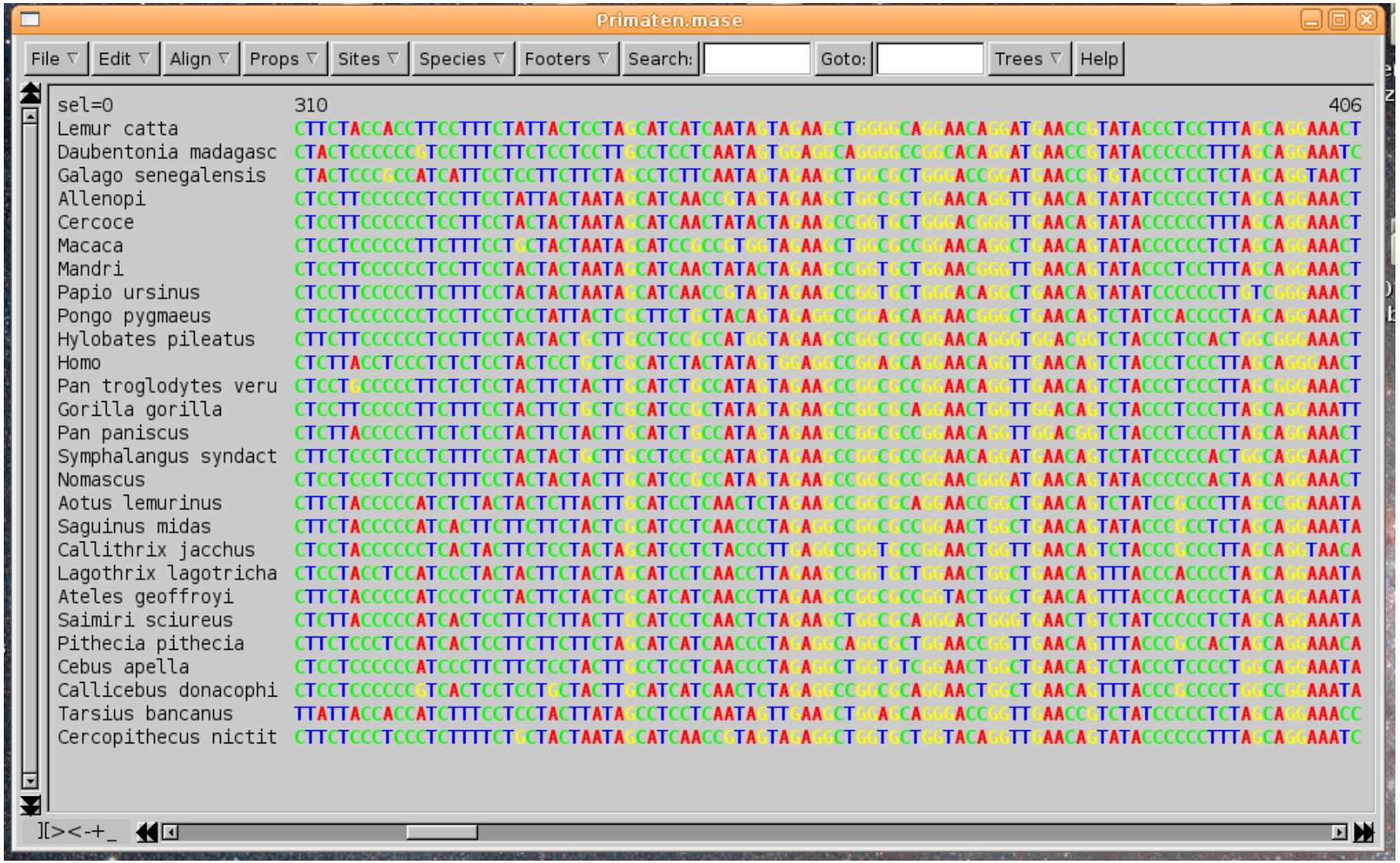
Quellen angeblich zugänglich im *advanced user mode*

ARB (http://www.arb-home.de/) - Quellen, nur RNS-Gene, Pakete für Linux-Distros

SeaView (http://pbil.univ-lyon1.fr/software/seaview.html) - Quellen frei, keine Lizenz erwähnt

Jalview (http://www.jalview.org/source/source.html) - GPL3

Der Sequenzeditor SeaView



If you use SeaView in a published work, please cite the following reference:

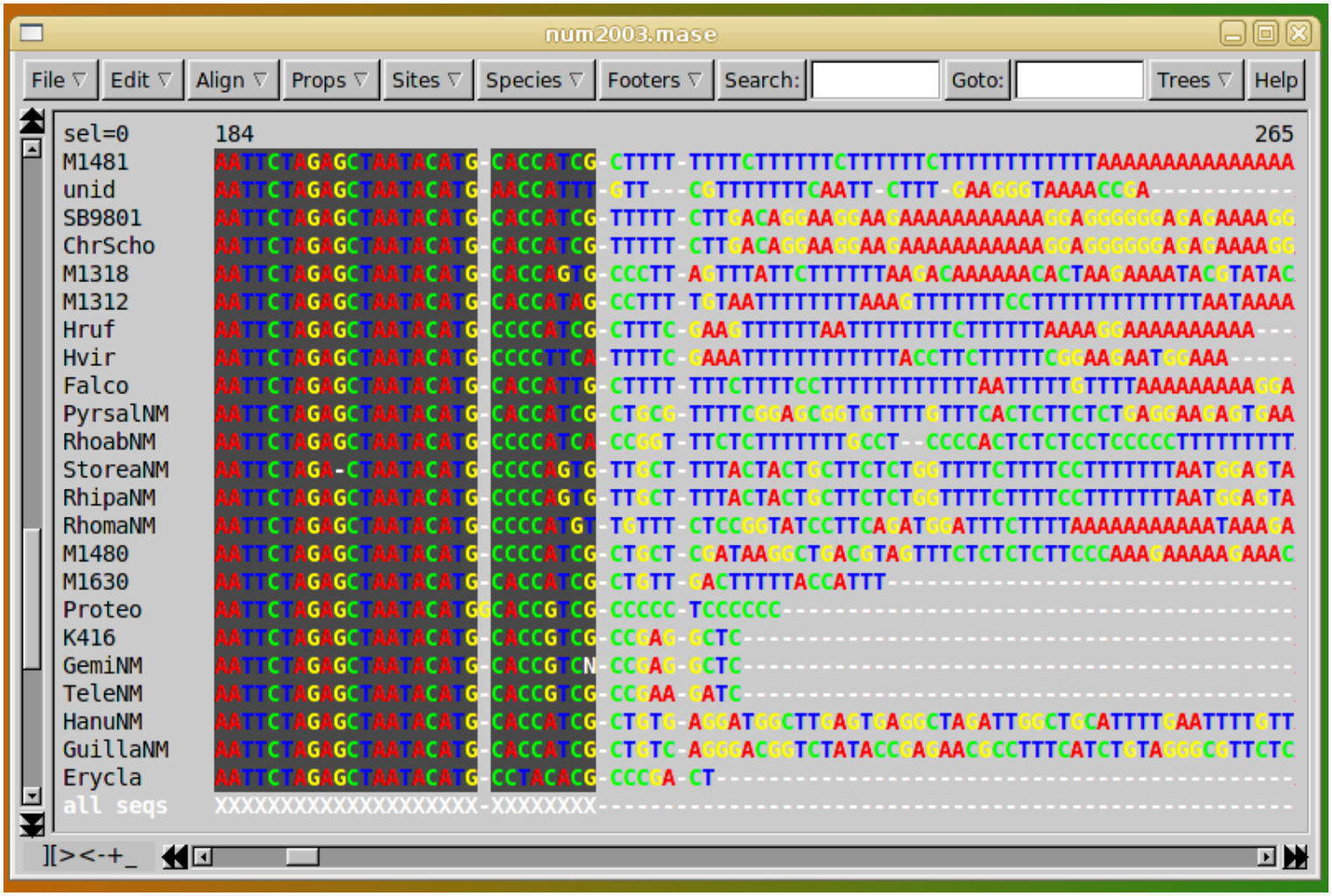
Gouy M., Guindon S. & Gascuel O. (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular Biology and Evolution 27(2):221-224.

Demo

Vorbereitung eines Analysedatensatzes

Ausschluss von nicht-alinierbaren Regionen aus dem Alignment.

Speicherung des Analyse-Datensatzes in einem passenden Dateiformat.



Phylogenetische Analysemethoden

Gängigste Methoden:

Distanzen

Maximum Parsimony

Maximum Likelihood

Bayesische Analysen

Distanzanalysen

- Paarweiser Vergleich aller Sequenzen
- Im einfachsten Fall:

  Zählen der Unterschiede und Umrechnung in Prozent für jedes Sequenzpaar

```
Sumpfmeerkatze   CCTGGTAATC TACTAGGTAG TGACCATCTT TATAACGTCA TCGTAACAGC CCATGCATTC
Mensch           CCAGGCAACC TTCTAGGTAA CGACCACATC TACAACGTTA TCGTCACAGC CCATGCATTT
                   ^  ^  ^   ^       ^ ^     ^^ ^   ^      ^       ^                ^


Sumpfmeerkatze   CCTGGTAATC TACTAGGTAG TGACCATCTT TATAACGTCA TCGTAACAGC CCATGCATTC
Schopfgibbon     CCTGGCAACC TCCTGGGCAA CGACCATATT TATAATGTCA TCGTGACAGC CCACGCATTC
                      ^  ^   ^  ^  ^ ^ ^       ^         ^          ^          ^


                                        usw.
```

Beisp:

13 Unterschiede auf 60 Nucleotide zwischen Sumpfmeerkatze und Mensch = 22 % Divergenz

11 Unterschiede auf 60 Nucleotide zwischen Sumpfmeerkatze und Schopfgibbon = 18 % Divergenz

Distanzanalysen

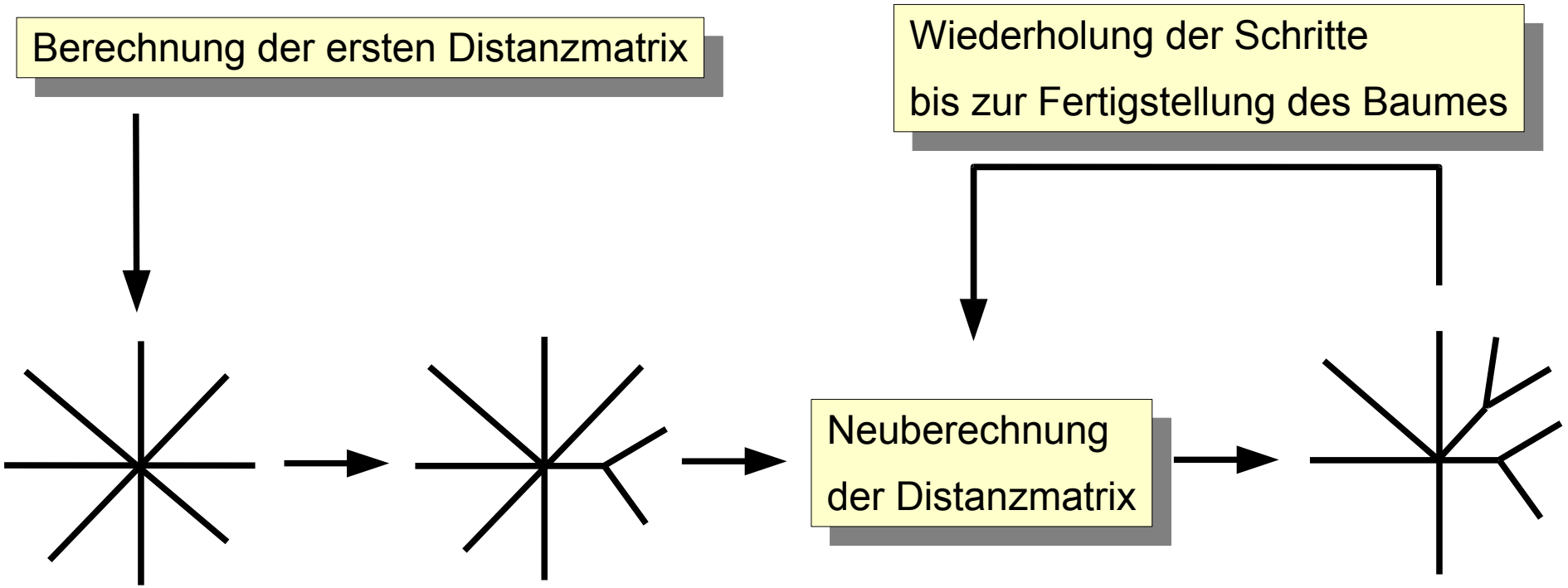
OTU 1	OTU 2	Genetische Distanz
Fingertier	Katta	0.19009902
Galago	Katta	0.19801980
Galago	Fingertier	0.19801980
Sumpfmeerkatze	Katta	0.20594059
Sumpfmeerkatze	Fingertier	0.21980198
Sumpfmeerkatze	Galago	0.19801980
Mangabe	Katta	0.21584159
Mangabe	Fingertier	0.22772278
Mangabe	Galago	0.22376238
Mangabe	Sumpfmeerkatze	0.12673268
Rhesusaffe	Katta	0.22178218
Rhesusaffe	Fingertier	0.22376238
Rhesusaffe	Galago	0.20198020
Rhesusaffe	Sumpfmeerkatze	0.12277228
Rhesusaffe	Mangabe	0.14455445
Drill	Katta	0.20594059

Ausgabe jedoch meist als Distanzmatrix

Baumkonstruktion: *Neighbor-Joining*

Berechnungsschritte:

1) Berechnung einer Distanzmatrix
2) (Virtueller) sternförmiger Startbaum
3) Gruppierung der beiden OTUs mit geringster Distanz auf einem Ast
4) Neuberechnung der Distanzmatrix; gruppierte OTUs werden wie eine OTU behandelt.
5) Wiederholung der Schritte 3 und 4, bis Baum vollständig aufgelöst ist.



Maximum Parsimony und *Maximum Likelihood*

Unterschiede zu Distanz/*Neighbor-Joining-A*nalysen:

1.) Sequenzen werden **nicht** in genetische Distanzen umgerechnet.

Methoden arbeiten direkt mit den Nucleotiden/Aminosäuren pro Position.

2.) Baum wird **nicht** direkt aus dem Datensatz berechnet.

Stattdessen:

Bäume werden vorgeschlagen und mittels *Parsimony* oder *Likelihood* bewertet.

Ergebnis:

Der oder die best-bewerteten Bäume nach dem gewählten Kriterium

(= der Baum der den Datensatz am besten erklärt)

Maximum Parsimony

„*parsimony*“ = Geiz

Kann auch auf morphologische Daten angewandt werden.

Heute fast nur noch für Fossilien eingesetzt (z. B. Dinosaurier)

Ein Baum wird zur Bewertung vorgelegt.

Wieviele Mutationen sind notwendig, um anhand des Baumes die Evolutionsgeschichte der Nucleotide/Aminosäuren des Analysedatensatzes zu erklären?

Jede Mutation = ein Strafpunkt (= „*tree scores*“ sind immer ganzzahlig.)

Bester Baum → geringste Anzahl an Mutationen

(= je niedriger der „*tree score*“, umso besser).

Meist mehrere gleichwertige Bäume als Ergebnis

Edwards AWF, Cavalli-Sforza LL (1963). The reconstruction of evolution. Ann Human Genet 27: 105-106

Maximum Likelihood

Probability = absolute Wahrscheinlichkeit

Likelihood = bedingte Wahrscheinlichkeit

Suche nach dem Baum mit der höchsten Wahrscheinlichkeit

(= *maximum likelihood*) für den Datensatz.

Wahrscheinlichkeitswerte sind abhängig vom Evolutionsmodell

Evolutionsmodelle

Anteile der Nucleotide: A, C, G, T ≠ 25 %

Mutationsraten: (A ↔ C) ≠ (A ↔ G) ≠ (A ↔ T) ≠ (C ↔ G) ≠ (C ↔ T) ≠ (T ↔ G)

Positionsabhängige Mutationsraten: 1., 2., 3. Position eines Tripletts?

Helices oder Endschlaufen einer rRNS?

Anteile der Nucleotide, Mutationsraten für Punktmutationen,

positionsabhängige Mutationsraten

= *likelihood estimators* oder „*nuisance parameters*“

Die *Likelihood*-Formel für phylogenetische Analysen

$$L = P(Data/Tree) = \prod_{i=1}^{m} P(Data^{(i)}/Tree)$$

P (Data/Tree) = Bedingte Wahrscheinlichkeit für einen Datensatz bei gegebenem Baum ist das Produkt aus allen Wahrscheinlichkeiten für jede einzelne Position in einem Alignment.

$$P(Data^{(i)}/Tree) = \sum_{x} \sum_{y} P(Seq1, Seq2, Seq3, Seq4, x, y/Tree)$$

P (Data$^{(i)}$/Tree) = Bedingte Wahrscheinlichkeit für eine Position des Alignments bei gegebenem Baum ergibt sich aus den Summen für die Wahrscheinlichkeiten für alle vier Nucleotide an allen (bekannten) terminalen und allen internen Knoten des gegebenen Baumes (Beisp. gilt für einen DNS-Datensatz mit vier OTUs)

Seq1 – Seq4 = Nucleotide der vier Sequenzen an Position i; x, y = interne Knoten

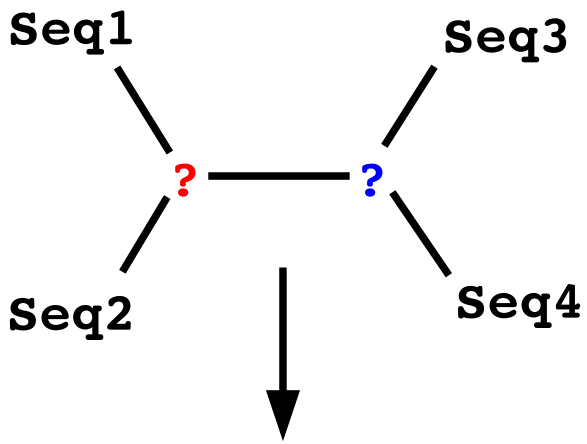
Felsenstein J (1981). Evolutionary trees from DNA sequences – a maximum likelihood approach. J Mol Evol 17: 368-376

Berechnung des *Likelihoods*

Beispiel: Datensatz mit vier OTUs und 1500 Nucleotidpositionen

```
Alignment

        1234...

Seq 1   ATTA...
Seq 2   ACTA...
Seq 3   CCTA...
Seq 4   GGTG...
```

$$P(Data^{(1)}/Tree)=\sum_{x}\sum_{y}P(A,A,C,G,x,y/Tree)$$



Baum mit vier OTUs =

16 mögliche Kombinationen

von ursprünglichen

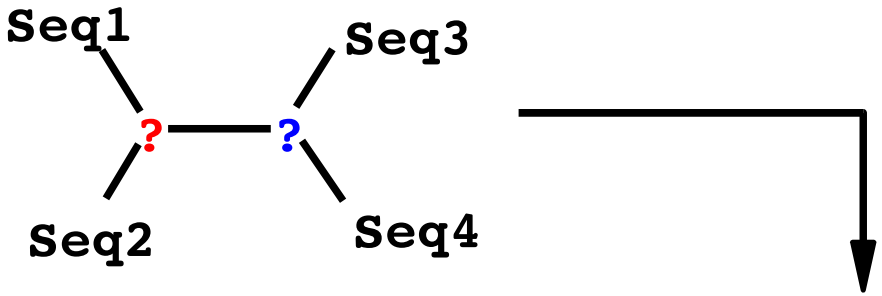
Merkmalszuständen an

internen Knoten

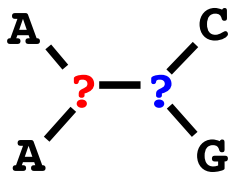
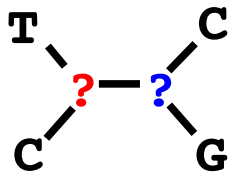
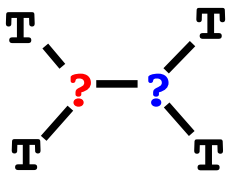
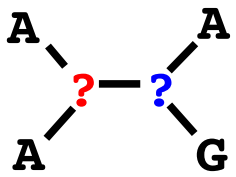
```
?-? =

A-A   C-A   G-A   T-A
A-C   C-C   G-C   T-C
A-G   C-G   G-G   T-G
A-T   C-T   G-T   T-T
```

Berechnung des *Likelihoods*

Zu bewertender Baum:



Berechnung der Wahrscheinlichkeiten für jede der 1500 Positionen im Alignment („*site-wise log likelihoods*“)

Pos. 1	Pos. 2	Pos. 3	Pos. 4	usw. bis Pos. 1500
				...
A-A C-A G-A T-A A-C C-C G-C T-C A-G C-G G-G T-G A-T C-T G-T T-T	A-A C-A G-A T-A A-C C-C G-C T-C A-G C-G G-G T-G A-T C-T G-T T-T	A-A C-A G-A T-A A-C C-C G-C T-C A-G C-G G-G T-G A-T C-T G-T T-T	A-A C-A G-A T-A A-C C-C G-C T-C A-G C-G G-G T-G A-T C-T G-T T-T	...
Summe aus 16 Wahrschein-lichkeiten	Summe aus 16 Wahrschein-lichkeiten	Summe aus 16 Wahrschein-lichkeiten	Summe aus 16 Wahrschein-lichkeiten	1500 positions-spezifische Wahrscheinlich-keiten

Multiplikation aller 1500 „*site-wise log likelihoods*“ = Gesamt-Wahrscheinlichkeit für den Datensatz als negativer natürlicher Logarithmus (-lnL)

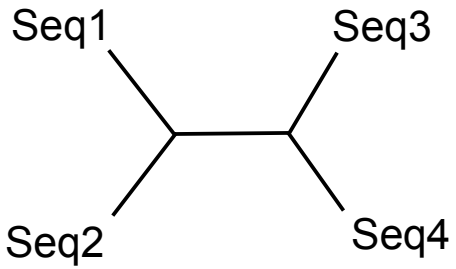
$$L = P(Data/Tree) = \prod_{i=1}^{m} P(Data^{(i)}/Tree)$$

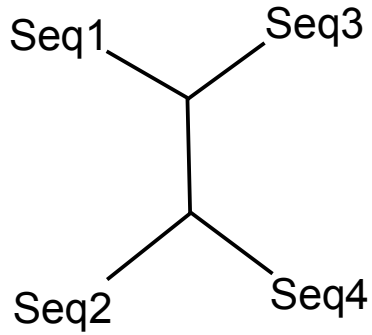
Woher kommen die zu bewertenden Bäume?

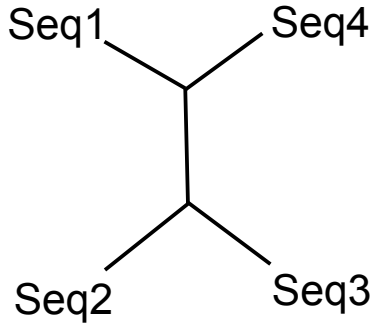
Beste, weil gründlichste Methode:

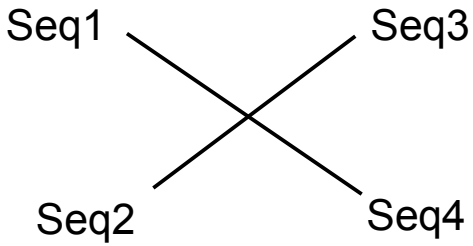
Berechnung **aller** potentiell möglichen Verzweigungsmuster

= erschöpfende Suche (*exhaustive search*)









4 OTUs: 4 Bäume sind potentiell möglich:

Alle 4 Bäume werden berechnet und der beste Baum wird gewählt.

Parsimony: Baum mit der geringsten Anzahl an Mutationen

Likelihood: Baum mit der höchsten Wahrscheinlichkeit

Dimensionen von Analysedatensätzen

Pröschold et al. (2001):	156 DNS-Sequenzen (= OTUs), 1642 Positionen
Saldarriaga et al. (2003):	78 DNS-Sequenzen (= OTUs), 1488 Positionen
Marin et al. (2003):	112 DNS-Sequenzen (= OTUs), 1588 Positionen
Murphy et al. (2001):	64 OTUs, 18 Gene, 9779 Positionen
Patron et al. (2007):	34 OTUs, 102 Proteine, 16459 Positionen
Rodríguez-Ezpeleta et al. (2007):	64 OTUs, 143 Proteine, 31604 Positionen
Shalchian-Tabrizi et al. (2007):	175 DNS-Sequenzen (= OTUs), 1159 Positionen
Hoef-Emden (2008):	96 DNS-Sequenzen (= OTUs), 1556 Positionen
	34 OTUs, 3 Gene, 4083 Positionen
Shalchian-Tabrizi et al. (2008):	134 DNS-Sequenzen (= OTUs), 1582 Positionen

Das Problem

Anzahl aller möglichen gewurzelten dichotom gegabelten Bäume (aus Felsenstein, *Inferring Phylogenies,* 2004):

6 OTUs = 945

10 OTUs = 34.459.425

15 OTUs = 213.458.046.676.875 Bäume

50 OTUs = $2{,}75292 \times 10^{76}$

Die Rechenzeit von phylogenetischen Analysen wächst faktoriell (= schneller als exponentiell) mit der Anzahl der OTUs: mathematisch ein sogenanntes NP-hartes Problem.

Suchalgorithmen

Die Menge aller möglichen Bäume wird mit einer Landschaft aus Tälern und Hügeln verglichen (*tree space*).

Die optimalen Bäume mit der besten Bewertung befinden auf den Spitzen der Hügeln, die schlechtesten Bäume in den Tälern.

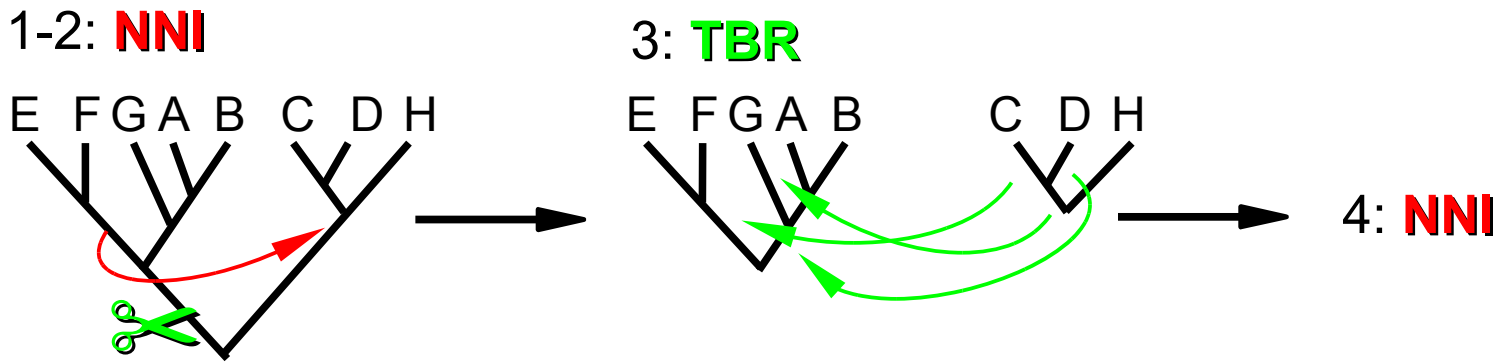
Absolutes Maximum = bester Baum

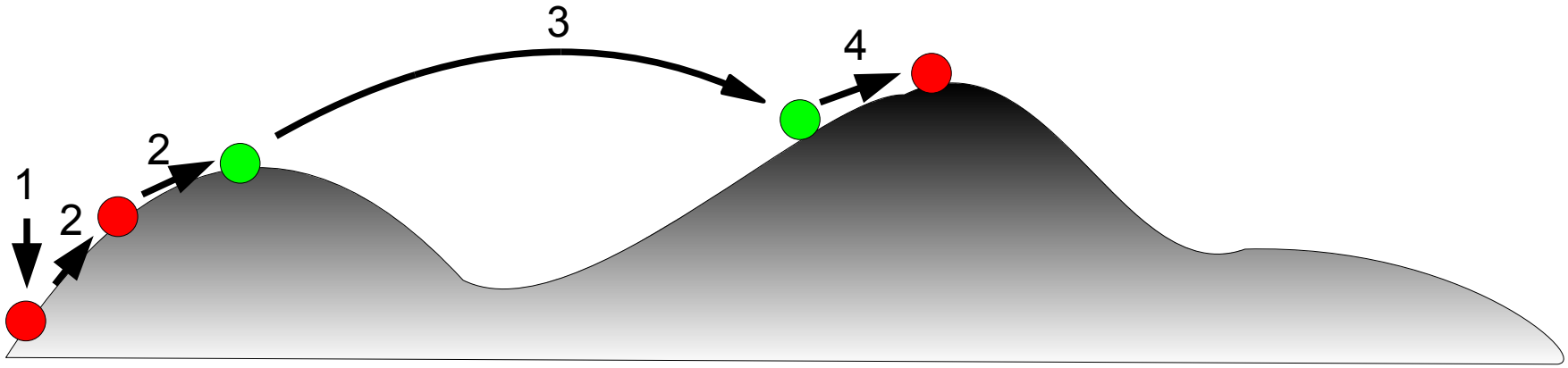
Lokale Maxima = suboptimale Bäume

Wie kann man schnellstmöglich den besten Baum finden, ohne alle Bäume im „*tree space*“ bewerten zu müssen?

Heuristische Suche

1.) Ein Zufalls-Startbaum wird erstellt.
Bewertung des Starbaumes nach dem *Parsimony*- oder *Likelihood*-Kriterium.
2.) Startbaum wird unter ständiger Neubewertung umarrangiert:
kleinere Umbauten an endständigen Verzweigungen (NNI)
3.) globales Umarrangieren (Auseinanderschneiden des Baumes und
Einfügen der Teilstücke an verschiedenen Positionen;TBR)
4.) kleinere Umbauten an endständigen Verwzeigungen (NNI)
5.) Wiederholung von (1) 2-3, bis kein besserer Baum gefunden werden kann.





Software-Anfänge: PAUP und PHYLIP

PAUP* 4b10

(**P**hylogenetic **A**nalyses **U**sing **P**arsimony - * and other methods)

von David Swofford

- ***Closed Source***, Löhnware
- kompilierte *Binaries* für MacOS9, MacOSX, Windows, Linux, versch. Unixe
- *Parsimony*- (auch Proteine), Distanz- und *Likelihood*-Analysen nur von DNS-Sequenzen
- „Klicki-Bunti“ nur für MacOS9, ansonsten Kommandozeile

PHYLIP 3.69 (2010: 30 Jahre alt)

(**Phyl**ogenetic **I**nference **P**ackage)

von Joseph Felsenstein

- ***Open Source***, vorkompilierte *Binaries* für Windows und MacOS
- *Parsimony*-, Distanz-, *Likelihood*-Analysen von DNS- und Proteinsequenzen
- Textmenüs (ncurses-ähnlich)

Beide „Veteranen“-Programme: nicht parallelisierbar, langsame Baumsuchalgorithmen

Felsenstein J (1989). PHYLIP – Phylogeny Inference Package (Version 3.2). Cladistics 5: 164-166

Swofford DL (1993). PAUP – A computer program for phylogenetic inference using maximum parsimony. J gen Physiol 102: A9-A9 (Meeting Abstract)

Die neue Generation: *Open Source*

Alle Programme: parallelisierbar (MPI oder Pthreads), Protein- und DNS-Sequenzen, partitionierte Datensätze (Alignments aus verschiedenen Genen/Proteinen, für jedes Gen wird das Evolutionsmodell separat berechnet), kein Checkpointing

MrBayes: Bayesische Analysen mit Markov-Ketten-Monte-Carlo-Simulation
von John Huelsenbeck, Bret Larget, Paul van der Mark, Fredrik Ronquist
http://mrbayes.csit.fsu.edu/

PhyML: Maximum Likelihood-Analysen von Proteinen und DNS mit heuristischer Suche
von Olivier Gascuel und Stéphane Guindon (Quellcode nur auf Anfrage)
http://www.atgc-montpellier.fr/phyml/binaries.php

RAxML: Maximum Likelihood-Analysen von Proteinen und DNS mit heuristischer Suche
von Alexandros Stamatakis
http://wwwkramer.in.tum.de/exelixis/software.html

Bayesische Analysen

Das Bayes-Theorem in der Phylogenetik:

$$P(Tree/Data) = \frac{P(Data/Tree) \times P(Tree)}{P(Data)}$$

Data = Analysedatensatz

Tree = vorgeschlagener Baum

P (Tree/Data) = *a posteriori*-Wahrscheinlichkeit für einen Baum bei vorgegebenem Alignment

P (Tree) = *a priori*-Wahrscheinlichkeit für einen Baum

(Standardeinstellung: alle Bäume gelten als gleich wahrscheinlich)

P (Data) = *a priori*-Wahrscheinlichkeit für das beobachtete Alignment

***P (Data/Tree)* = Likelihood** für das Alignment bei gegebenem Baum

MrBayes 3.1.2

MCMCMC oder MC^3= **M**etropolis-**C**oupled **M**arkov **C**hain **M**onte **C**arlo

Metropolis-Hastings-Algorithmus entscheidet, ob eine Baumtopologie mit ihren Evolutionsmodellparametern behalten und modifiziert oder verworfen wird (entspricht einer Generation).

Die Markov-Kette wird mit einem Parsimony-Startbaum und *a priori*-Wahrscheinlichkeiten für alle Evolutionsparameter initialisiert. Durch den MH-Algorithmus entsteht eine Kette aus hintereinander geschalteten Generationen von getesteten Bäumen, die die Häufigkeitsverteilung der verknüpften *a posteriori*-Wahrscheinlichkeiten durchwandert.

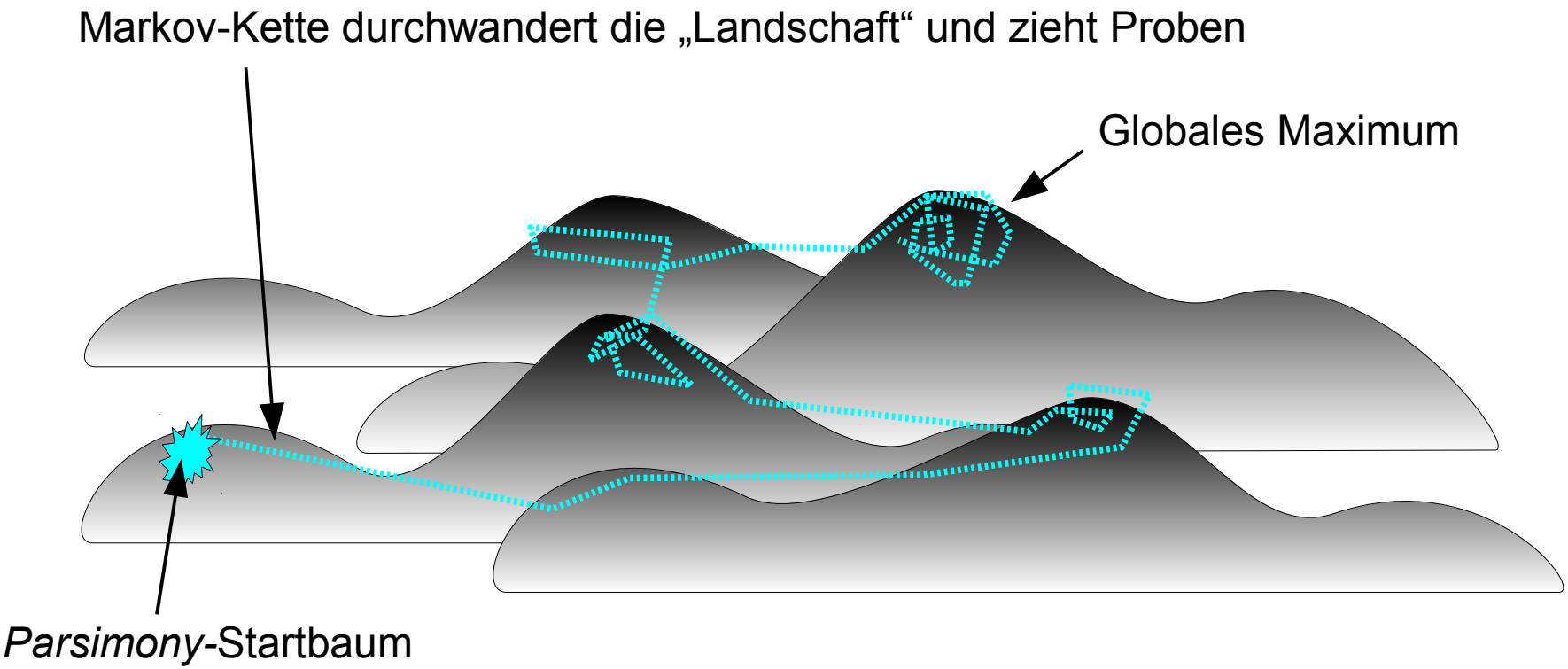
Alle 1000 Generationen werden Baumstruktur und Evolutionsmodellparameter abgespeichert.

Ronquist F, Huelsenbeck JP (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572-1574

MrBayes 3.1.2

Häufigkeitsverteilung der verknüpften *a posteriori*-Wahrscheinlichkeiten

= Landschaft ähnlich dem „*Tree Space*“



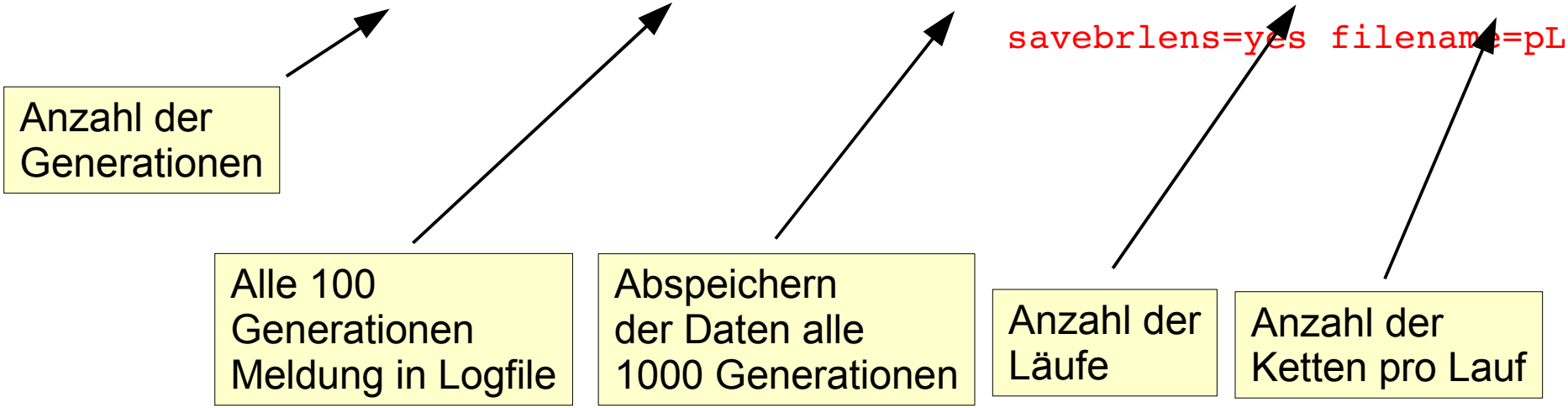
Bayesische Analysen: MrBayes

Einstellen des Evolutionsmodells:

```
MrBayes> lset nst=6 rates=invgamma ngammacat=4 covarion=yes
```

Einstellen der Parameter für die Markov-Ketten:

```
MrBayes> mcmcp ngen=5000000 samplefreq=100 printfreq=1000 nruns=2 nchains=4
                                                savebrlens=yes filename=pLSU
```



Initialisieren der Markov-Ketten:

```
MrBayes> mcmc
```

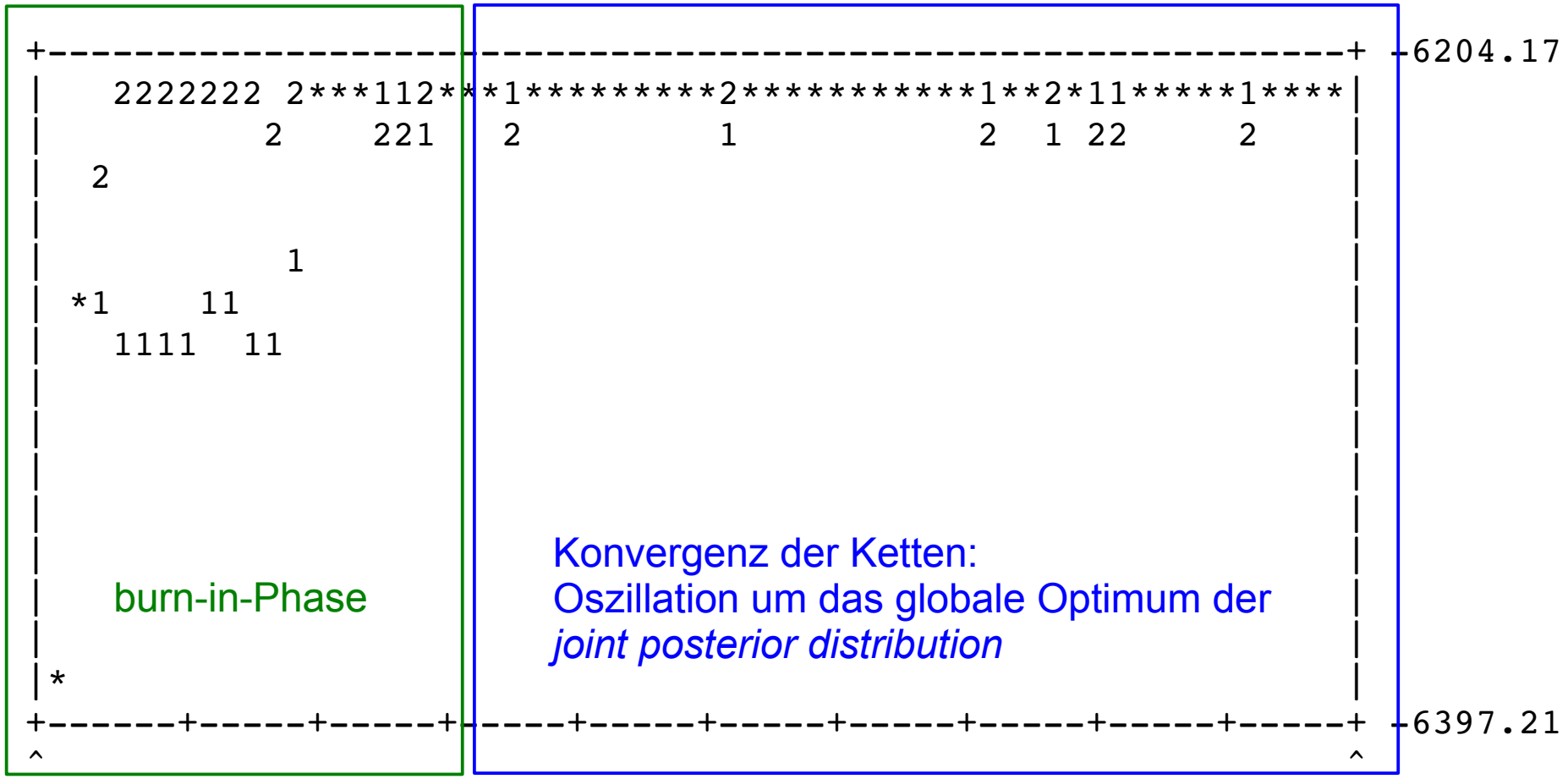
Beenden des Programms nach Ablauf der Analyse:

```
MrBayes> quit
```

MrBayes: Auswertung

Überprüfung von burn-in-Phase und Konvergenz der kalten Ketten aus beiden Läufen:

```
MrBayes> sump burnin=10000 filename=pLSU
                                        ...
 Overlay plot for both runs:
    (1 = Run number 1; 2 = Run number 2; * = Both runs)
```



RAxML 7.2.6

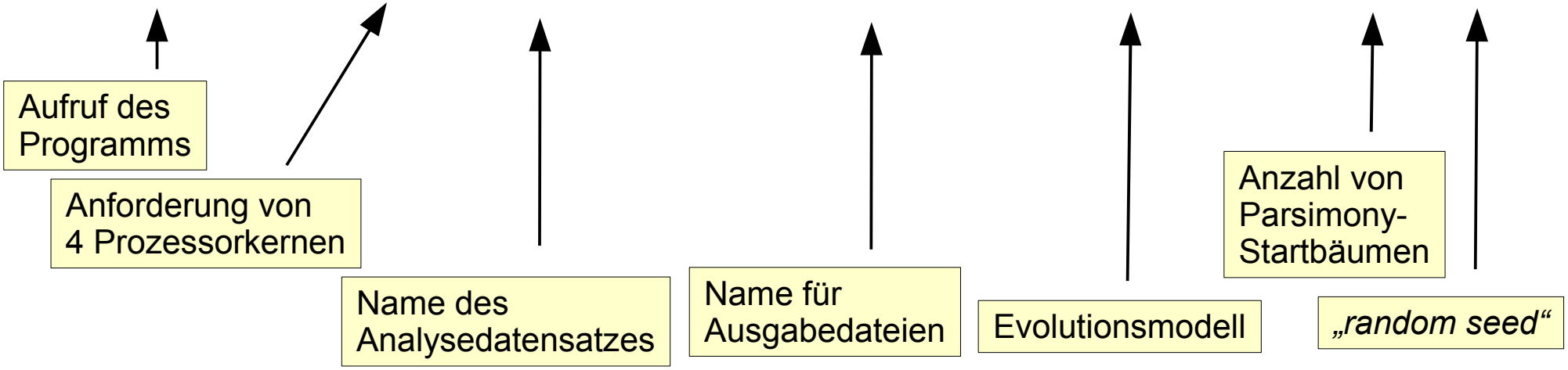
Kommandozeilenbeispiele

Rechner mit nur einem Prozessorkern/Thread:

```
raxmlHPC -s Dateiname.phy -n Ausgabedatei -m PROTGAMMAIWAG -N20 -p 87
```

Prozessoren mit mehreren Kernen (Dualcore, Quadcore, Hexacore):

```
raxmlHPC-PTHREADS -T4  -s Dateiname.phy -n Ausgabedatei -m PROTGAMMAIWAG -N20 -p 87
```



Stamatakis A (2006). RaxML-VI-HPC (2006). Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models”, Bioinformatics 22::2688–2690

Batch Queue Processing

Die Analysen werden auf einem ***H**igh **P**erformance **C**omputing Server* durch einen *Job Scheduler* gestartet und laufen nicht-interaktiv und ohne offenes Terminal ab.

Dies bedeutet:

Das Programm darf während der Analyse nicht in den Eingabemodus wechseln.

Alle notwendigen Kommandos müssen zum Zeitpunkt des Starts vorhanden sein.

CHEOPS

Erst in diesem Jahr (2010) offiziell in Betrieb genommen.

1. Ausbaustufe:

256 Knoten mit je 2 Intel Nehalem Quadcore-Prozessoren (2,66 GHz)

jeder Knoten mit 24 GB RAM (1333 Mhz)

Batch Queue Processing mit MrBayes

MrBayes-Kommandoblock wird an das Ende des Datensatzes eingefügt:

```
#NEXUS
[saved by seaview on Tue Nov  9 23:28:14 2010]
BEGIN DATA;
  DIMENSIONS NTAX=27 NCHAR=505;
  FORMAT DATATYPE=DNA
  GAP=-
  ;
MATRIX
[1] Katta
cctgggtctctgttaggagacgaccagatctacaacgttgttgtaacagctcatgctttc
gtcataatctttttcatagttatacctattataattggaggcttcggaaactgattagtt
Ccattaataattggagctcctgatatagcattcccccgaataaacaatataagcttctga

[…]

End;

Begin mrbayes;
        set autoclose=yes nowarn=yes;
        log start filename=sum.log;
        lset  nst=6 nucmodel=4by4 rates=invgamma ngammacat=4 covarion=yes;
        mcmcp ngen=3000000 samplefreq=100 printfreq=1000 nruns=2 nchains=4 savebrlens=yes filename=Prim;
        mcmc;
        quit;
End;
```

Batch Queue Processing mit MrBayes

Shell-Script für MrBayes:

PBS = **P**ortable **B**atch **S**ystem

```
#!/bin/bash -l
#PBS -q default
#PBS -l nodes=1:ppn=8
#PBS -l mem=23G
#PBS -l walltime=30:00:00
#PBS -j oe
#PBS -r n
#PBS -A UniKoeln

module load mrbayes

TESTDIR=$HOME/Primaten/MB
INPUT=$TESTDIR/Primaten.bnex
OUTPUT=$TESTDIR/Prim.log

cd $TESTDIR

time mpirun -np 8 mb $INPUT > $OUTPUT
sleep 1
```

Anfordern der CPUs

Maximale Rechenzeit

Arbeitsverzeichnis

Analysedatensatz

Log-Datei

Start von MPI

Anzahl der CPUs

Start von MrBayes

Start der Analyse:

```
[aeb25@cheops1 ~]$ qsub shell-script.sh
```

Batch Queue Processing mit RAxML

Shell-Script für RAxML:

```
#!/bin/bash -l
#PBS -q default
#PBS -l nodes=1:ppn=8
#PBS -l mem=23GB
#PBS -l walltime=24:00:00
#PBS -j oe
#PBS -r n
#PBS -A UniKoeln

module load raxml

RAXML=raxmlHPC-PTHREADS
TESTDIR=$HOME/Primaten/
INPUT=$TESTDIR/Primaten.phy
OUTPUT1=heur
OUTPUT2=boot
OUTPUT3=supp

cd $TESTDIR

time $RAXML -s $INPUT -T8 -n $OUTPUT1 -m GTRGAMMAI -N 20 -p 977

time $RAXML -s $INPUT -T8 -n $OUTPUT2 -m GTRGAMMAI -N 1000 -b 987

time $RAXML -f b -T 8 -z RAxML_bootstrap.boot -t RAxML_bestTree.heur -m GTRGAMMAI -s $INPUT -n $OUTPUT3
```

Start der Analyse:

```
[aeb25@cheops1 ~]$ qsub shell-script.sh
```

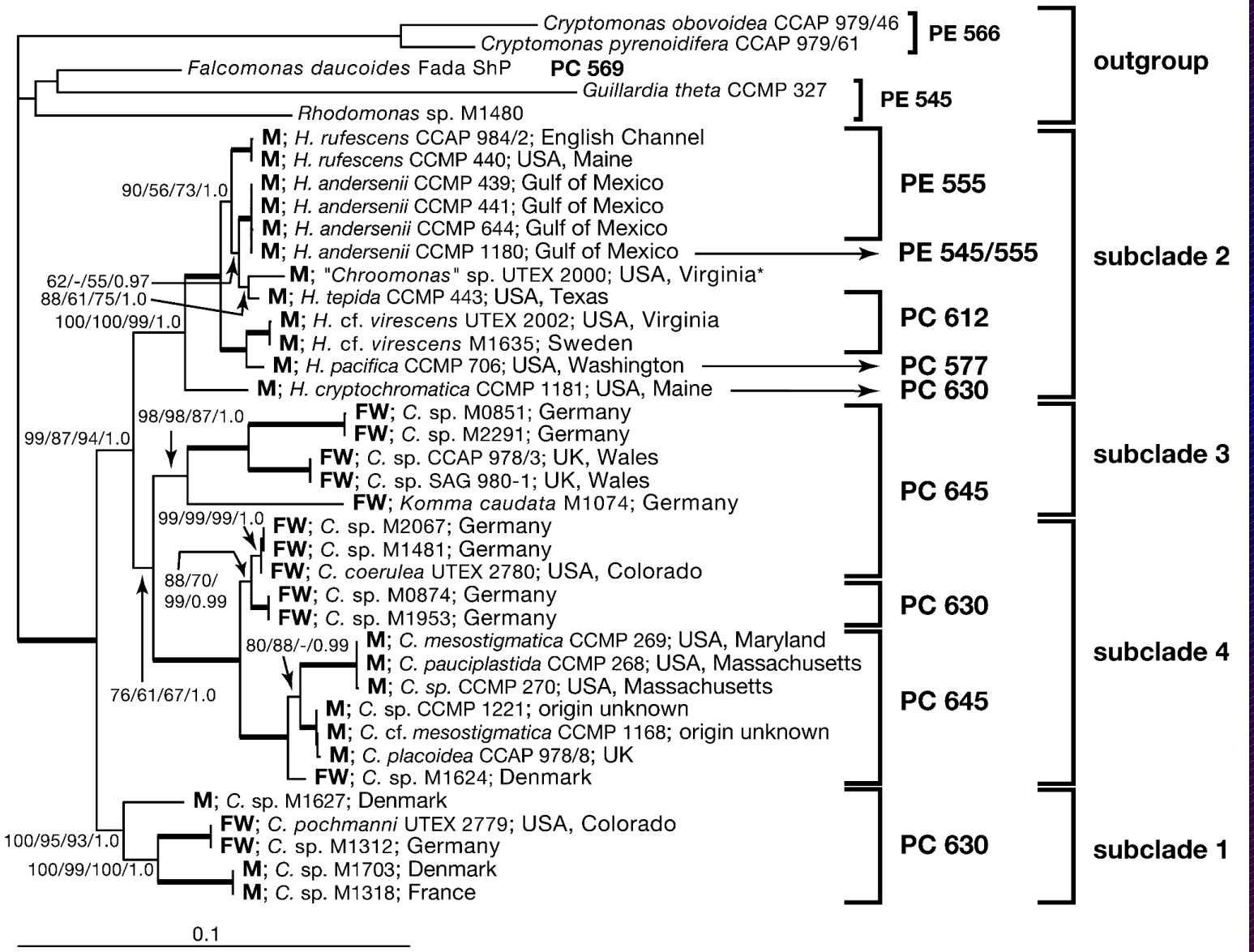


FIG. 2. Rooted maximum-likelihood (ML) tree of concatenated nuclear SSU rDNA, partial nuclear LSU rDNA, and nucleomorph SSU rDNA sequences. The tree shows the distribution of biliproteins across the PC-containing and related cryptophytes and their geographical distribution and habitats. Evolutionary model: GTR + I + Γ (Rodríguez et al. 1990). Names of *Hemiselmis* species and type of PC in *Hemiselmis cryptochromatica* according to taxonomic revision of Lane and Archibald (2008). The positions of strains M1635 and UTEX 2002 corresponded to the position of *Hemiselmis virescens* in the nucleomorph SSU rDNA phylogeny of Lane and Archibald, but their genetic identity with the *Hemiselmis virescens* strains could not yet be confirmed. The species name *Hemiselmis rufescens* was assigned to the strains CCAP 984/2 and CCMP 440 because CCAP 984/2 was a duplicate of strain PCC 14 of the Plymouth Culture Collection according to the CCAP strain data. Lane and Archibald examined PCC 14. *Strain UTEX 2000 displayed a blue–green color but died before its type of PC could be determined. Support values from left to right: ML bootstrap/neighbor-joining (NJ) bootstrap/maximum-parsimony (MP) bootstrap/posterior probabilities. Bold branches, 100% bootstrap support in all analyses and a posterior probability of 1.0. M, marine or brackish; FW, freshwater; PC, phycocyanin; scale bar, substitutions per site.

Ergebnis:

Graphisch bearbeiteter, publikations-tauglicher Baum