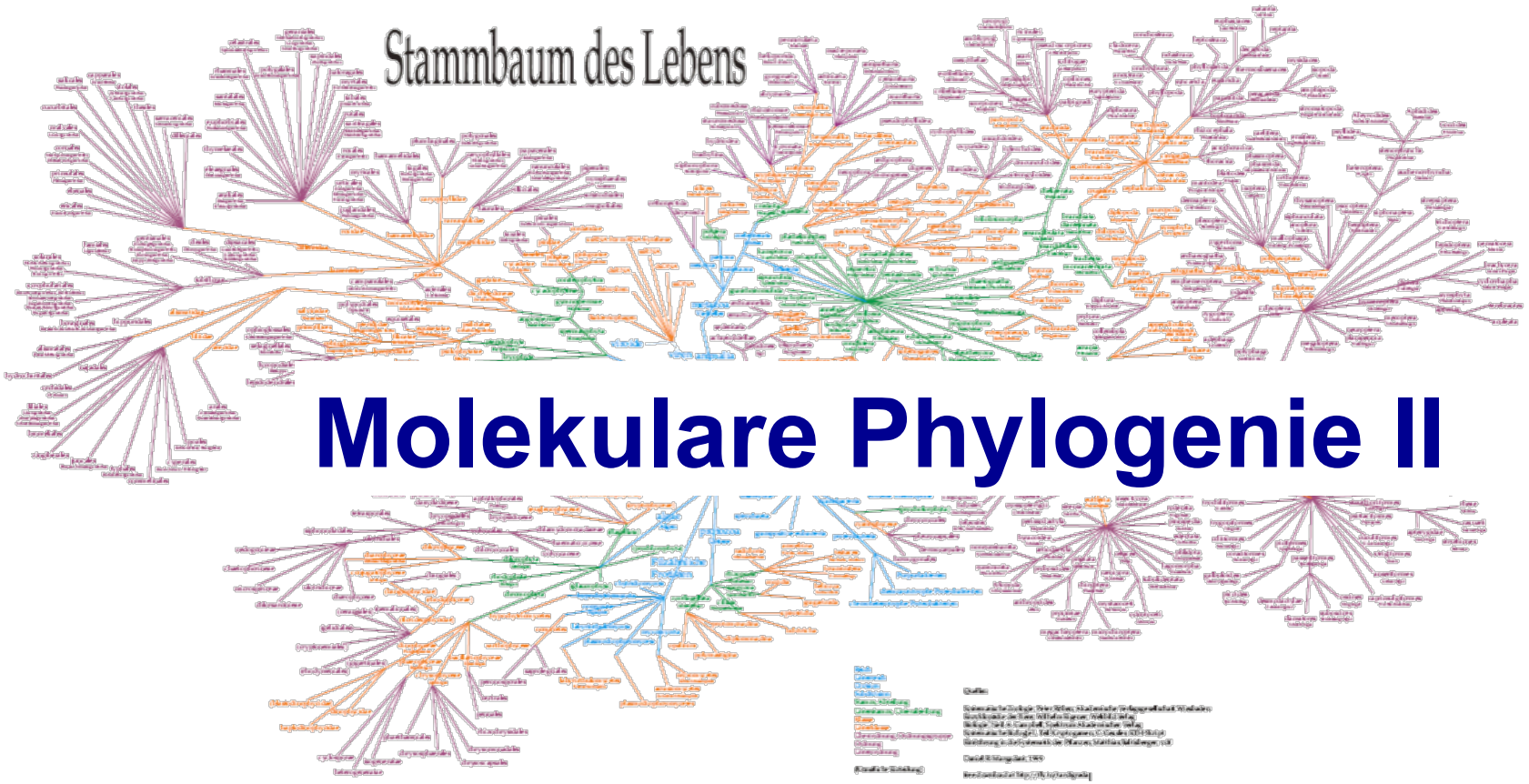
WS 2018/2019

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

Thomas Hankeln



Molekulare Phylogenie II

Stammbaumerstellung

1. Matrix-orientierte Methoden

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

=> Sequenzen in Distanzmatrix konvertiert

2. Charakter-orientierte Methoden

- **Parsimony**
- **Maximum Likelihood, Bayes etc.**

=> jede Position als informative Einheit

Datentypen

Distanzen

Sequenz 1	0,000	0,236	0,621	0,702	1,510
Sequenz 2		0,000	0,599	0,672	1,482
Sequenz 3			0,000	0,112	1,561
Sequenz 4				0,000	1,425
Sequenz 5					0,000

Charaktere

Sequenz 1	TATAAGCATGACTAGTAAGCTTAGCAAT
Sequenz 2	TAT---CATGACTGGTAACCTCAACAAT
Sequenz 3	TAT---CATGACTAGCAGGCTTAACATT
Sequenz 4	TGTTGCCACGATTAGCTACCATAGCGAT
Sequenz 5	CGTAGCTATGACCAACGGGCACAGCGAT

Wo stehen wir?

Sequenzen

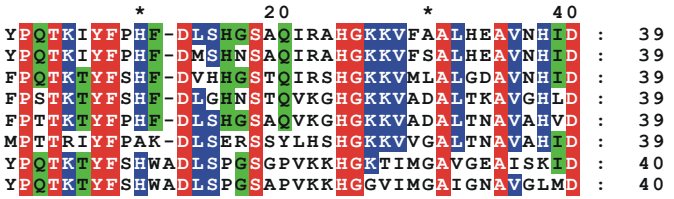
Multiples Alignment

Auswahl der Methode

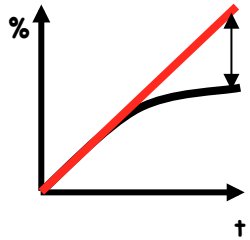
Auswahl EvoModell

Stammbaumberechnung

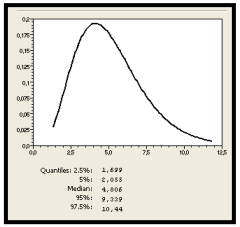
Ergebnisüberprüfung



Distanz ⟺ Charakter



Evolutionsmodell und gamma shape „Modeltest“ -> hLRT



WIE machen wir mit der korrigierten Distanzmatrix nun den **Baum**?

Distanzmatrix-Methoden

Zwei Schritte:

1. Berechnen der korrigierten paarweisen Abstände zwischen den Sequenzen
   => Evolutionsmodelle!
   DNA: JC, K2P ...
   Protein: PAM, BLOSUM...

2. Erstellen eines Stammbaums anhand dieser Abstandsdaten

Distanzmatrix

Berechnen des paarweisen Abstands

Sequenz 1	0,000	0,236	0,621	0,702	1,510
Sequenz 2		0,000	0,599	0,672	1,482
Sequenz 3			0,000	0,112	1,561
Sequenz 4				0,000	1,425
Sequenz 5					0,000

- Ausgedrückt i.d.R. als Mutationen pro Position
- Abstand kann > 1 werden!

Bsp. Jukes-Cantor: $K = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right)$

$p = 0.6 \Rightarrow K = 1.21$

Vorgehensweise

- Algorithmus berechnet aus den Distanzen den „besten“ Stammbaum

- Sequenzen selbst werden nicht mehr berücksichtigt

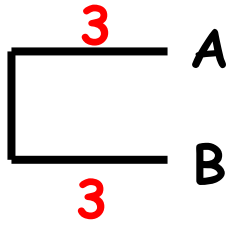
UPGMA

Unweighted Pair-Group Method with Arithmetric Means

1.

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

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

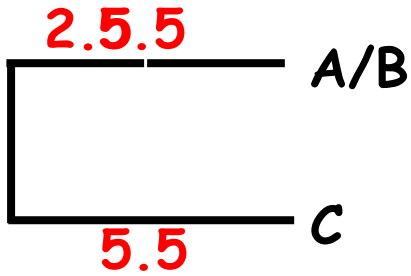


$\frac{d\,AC + d\,BC}{2}$ $\frac{d\,AD + d\,BD}{2}$

2.

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

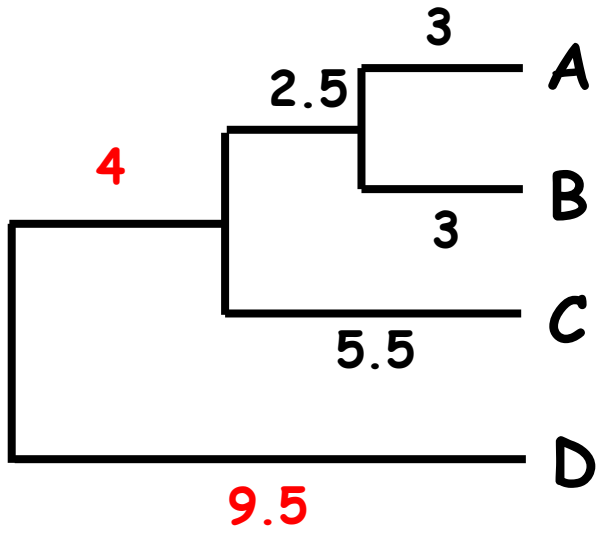
$\frac{d_{(AB)C}}{2} = 5{,}5$



UPGMA

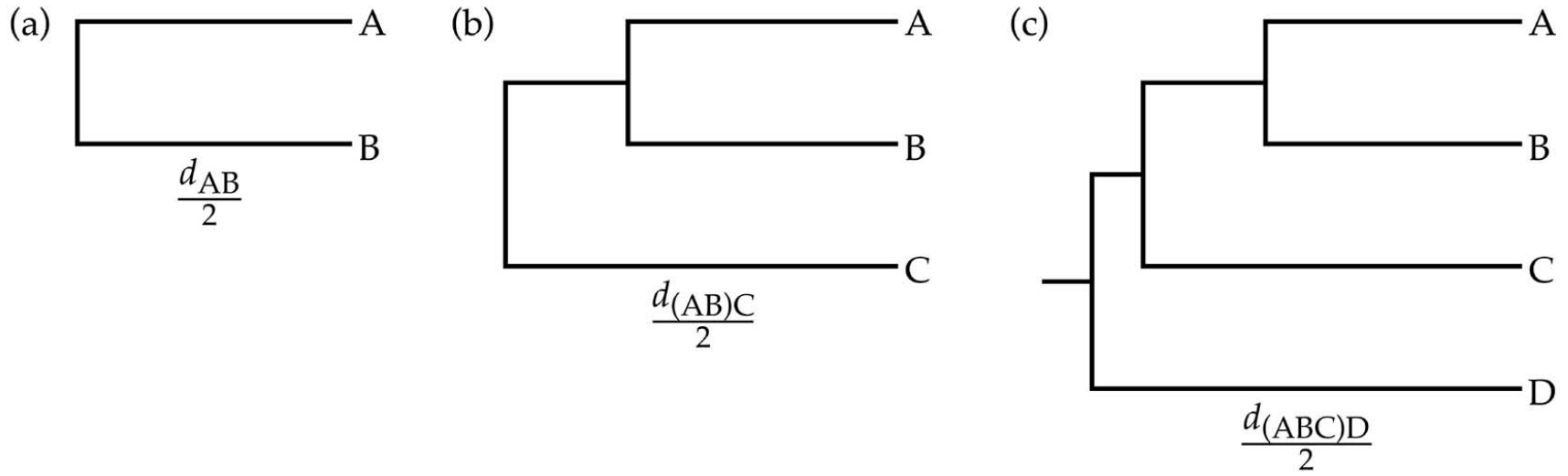
3.

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



- nimmt **konstante Evolutionsraten** auf allen Ästen der Phylogenie an (= „molecular clock“)
- Außengruppe wird „automatisch“ bestimmt

UPGMA

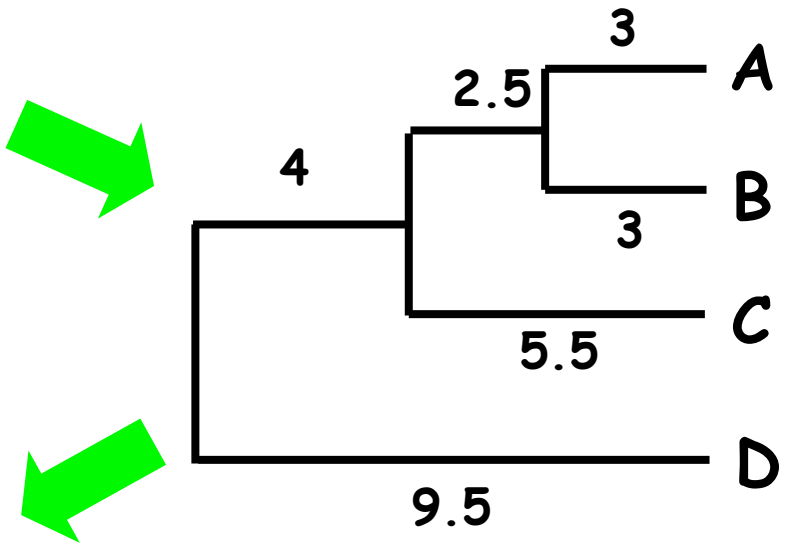


- UPGMA ist eine typische ***"Clustering"-Methode“***: OTUs werden durch sequenzielles Clustern nach absteigender Ähnlichkeit gruppiert.

UPGMA-Problem

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

....passt nicht überein!

UPGMA-Problem!

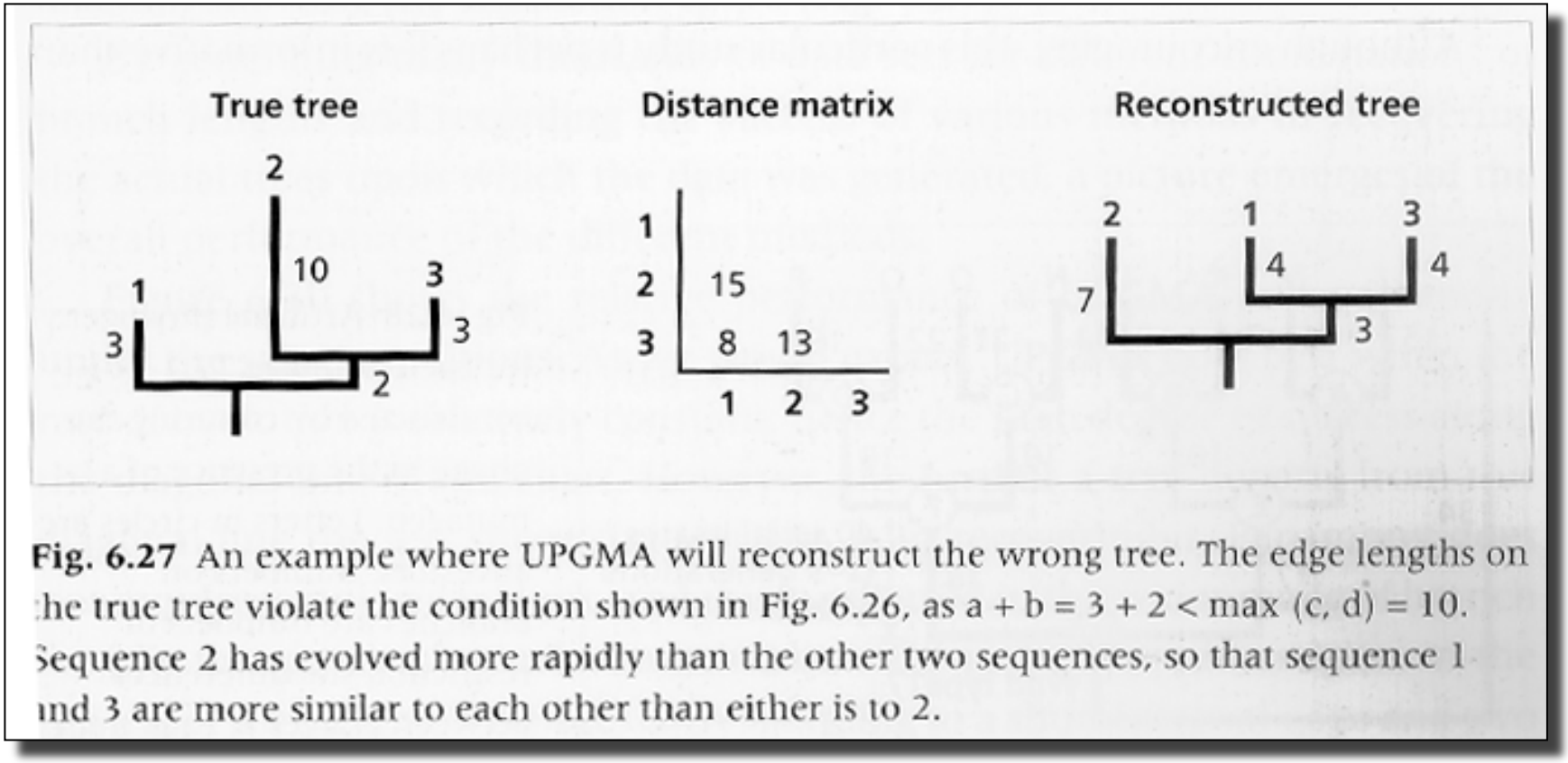
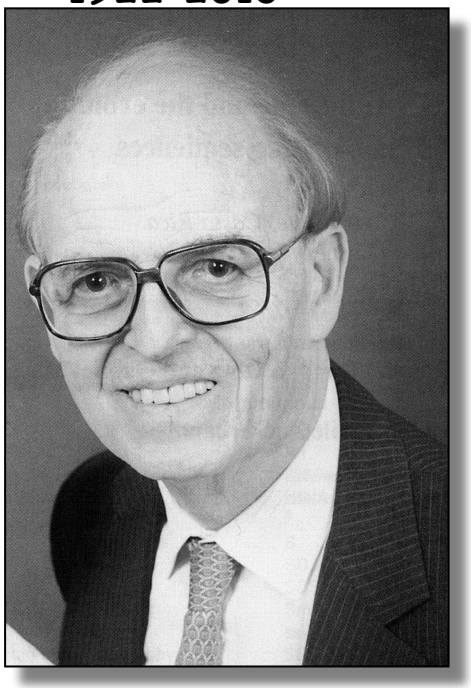


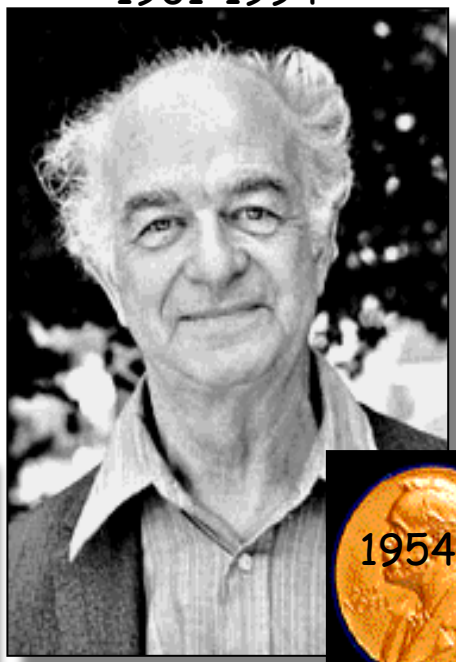
Fig. 6.27 An example where UPGMA will reconstruct the wrong tree. The edge lengths on the true tree violate the condition shown in Fig. 6.26, as a + b = 3 + 2 < max (c, d) = 10. Sequence 2 has evolved more rapidly than the other two sequences, so that sequence 1 and 3 are more similar to each other than either is to 2.

UPGMA liefert falsche Topologie bei im Stammbaum ungleich verteilten Evolutionsraten !!

1922-2013



1901-1994



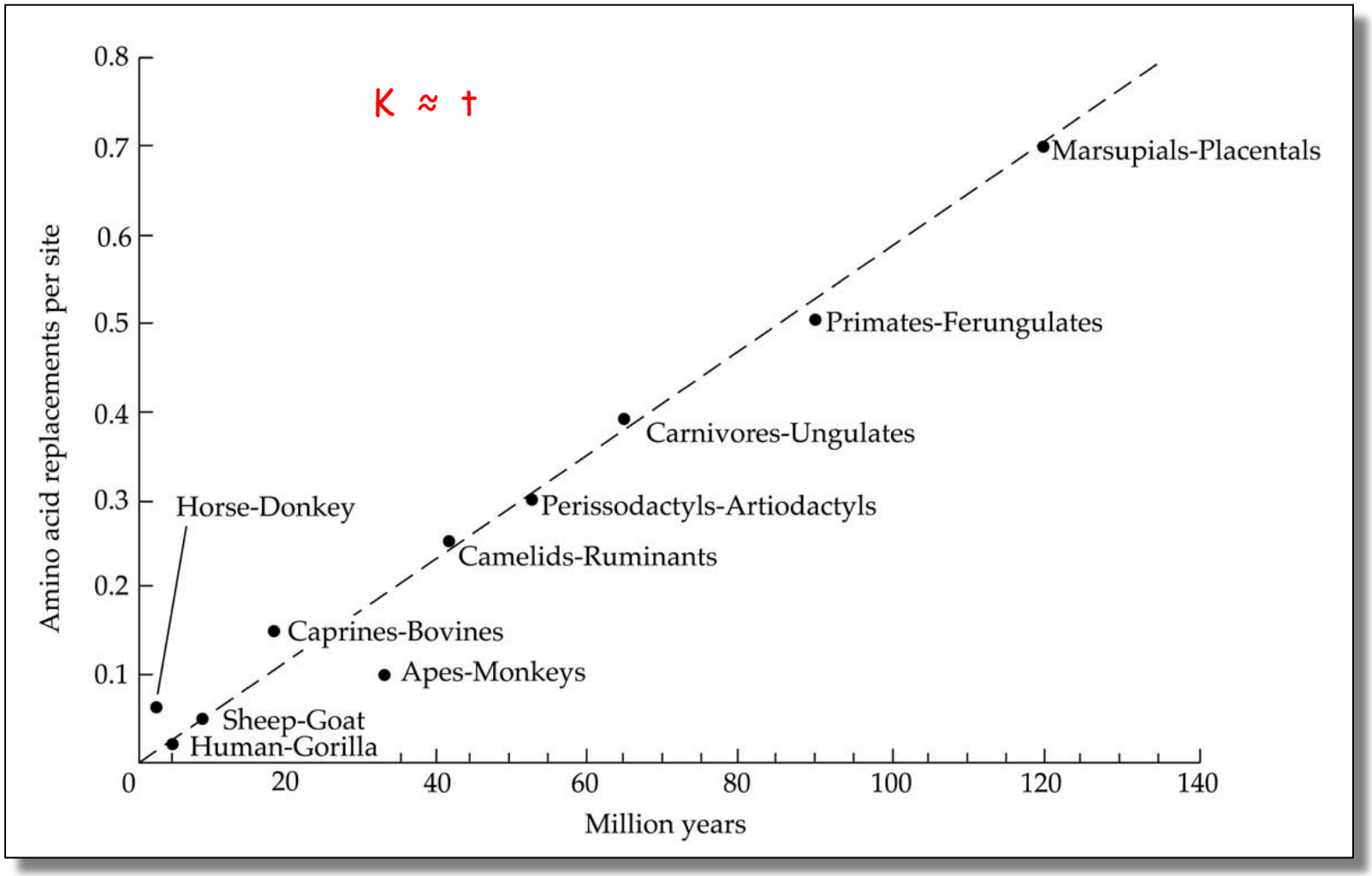
J. Theoret. Biol. (1965) **8**, 357–366

Molecules as Documents of Evolutionary History

EMILE ZUCKERKANDL AND LINUS PAULING

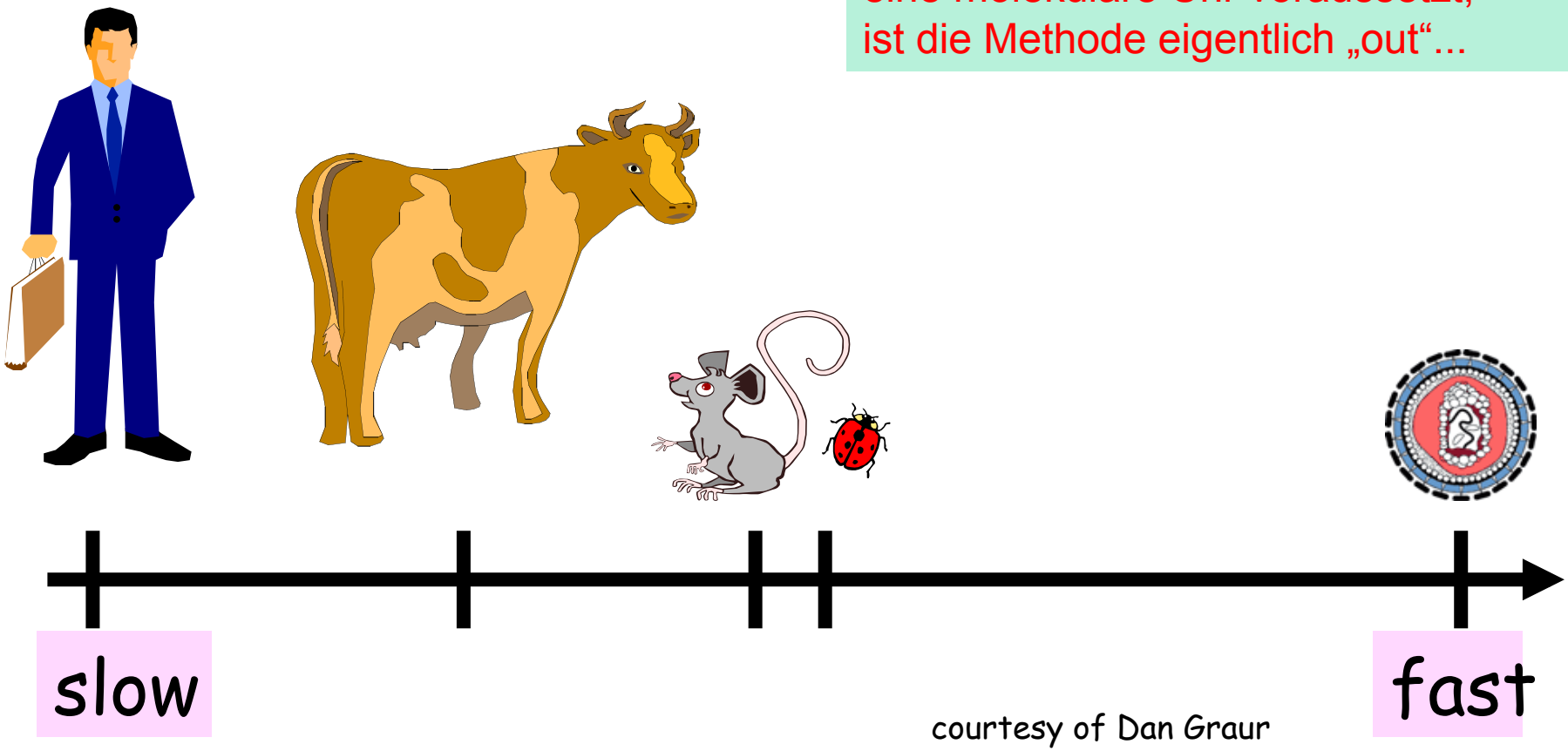
courtesy of Dan Graur

Molekulare Uhr bei Säuger-Proteinen



Es gibt keine „universelle“ Molekulare Uhr, wohl aber gut funktionierende „lokale Uhren“!

Da UPGMA aber konzeptionell bedingt eine molekulare Uhr voraussetzt, ist die Methode eigentlich „out“...

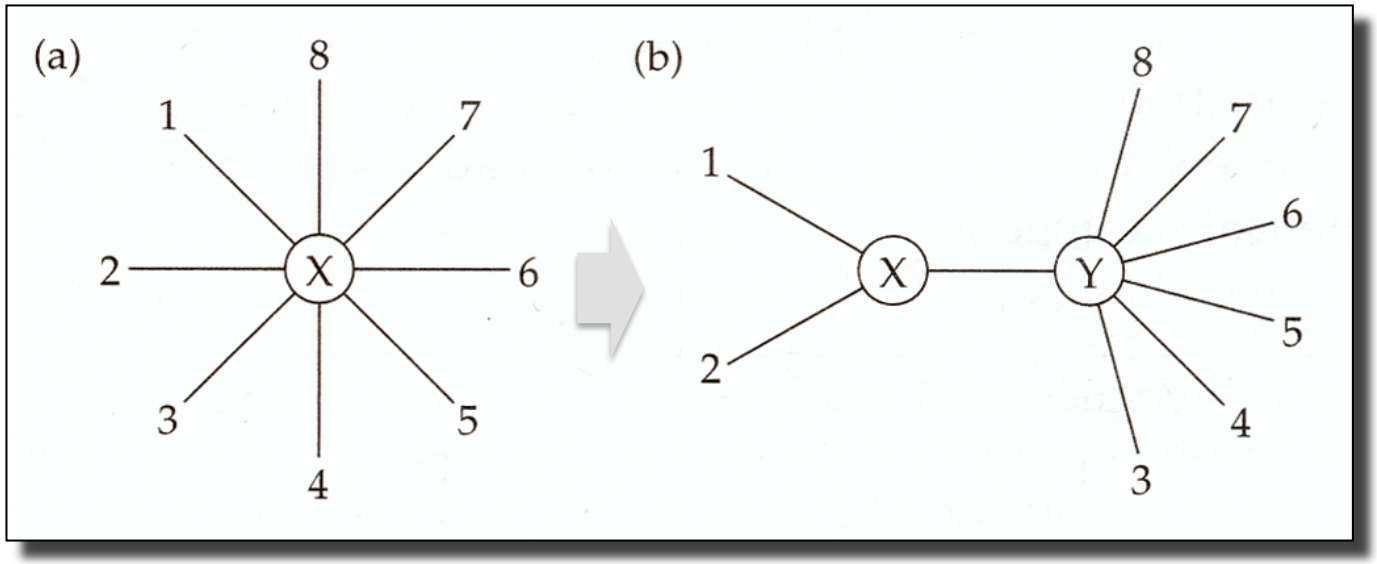


courtesy of Dan Graur

Neighbor-Joining (NJ)

Saitou, Nei 1987

- **viel besser als UPGMA: berücksichtigt unterschiedliche Evolutionsraten!**
- **Prinzip: Baum-Topologie und Astlängen werden getrennt ermittelt!**



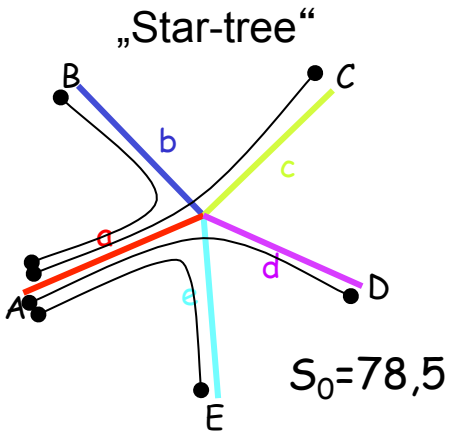
- **Prinzip:**

  Suche nach dem Baum mit der kleinsten Summe an Astlängen
  („minimum evolution tree“)

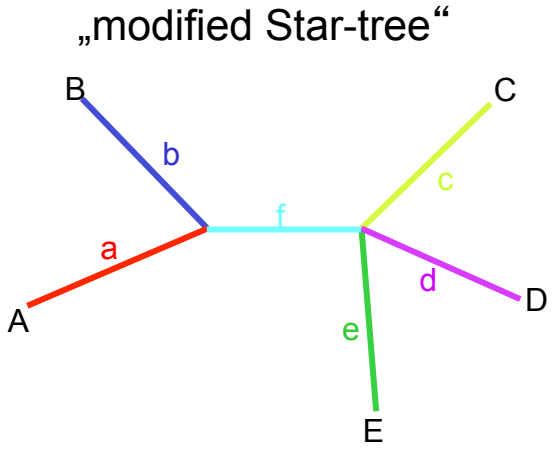
  Starte mit „star-like-tree“; identifiziere sukzessive Nachbar-Taxa
  (**NJ ist daher auch ein Clustering-Algorithmus**)

Neighbor-joining (NJ)

Ziel => Minimierung der Summe aller Astlängen



	A	B	C	D	E
OTU A	0	22	39	39	41
OTU B		0	41	41	43
OTU C			0	18	20
OTU D					10
OTU E					0



$S_0 = a+b+c+d+e$

$$S_0 = \left(\sum_{i \le j} d_{ji}\right)/N-1$$

S_0 = Summe aller Astlängen
d_{ij} = Distanzen zwischen allen OTUs
N = Anzahl der OTUs

Welche Paare müssen kombiniert werden, damit man den „kürzesten Baum“ erhält?

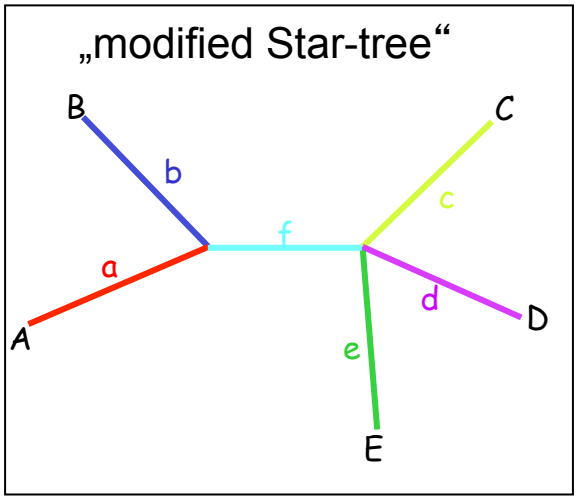
Neighbor-joining (NJ)

„Star tree“ -> Baumlänge berechnen

Grouping -> Baumlängen berechnen
-> kürzesten Baum identifizieren

Astlängen -> berechnen durch Fitch-Margoliash-Methode (FM)

Neues Taxon (AB) -> neue Matrix



Topologie
und Astlängen
separat bestimmt!

Neighbor-joining (NJ)

Erst einmal das Bestimmen der Topologie...

Errechnen der Summe aller Einzel-Distanzen, dann der durchschnittlichen Distanzen einer Gruppe (z.B. hier A+B) …

z.B. $(S_A+S_B)/N-2$

	A	B	C	D	E	Summe
OTU A	0	22	39	39	41	141
OTU B		0	41	41	43	147
OTU C			0	18	20	118
OTU D				0	10	108
OTU E					0	114

…und zuletzt Errechnen der „Distanzunterschiede“ („rate corrected distance“)

z.B. $D_{AB}=d_{AB} - (S_A+S_B)/N-2$

	A	B	C	D	E	Summe
OTU A	0	22	39	39	41	141
OTU B	-74	0	41	41	43	147
OTU C	-47,3	-47	0	18	20	118
OTU D	-46	-44	-57,3	0	10	108
OTU E	-44	-44	-57,3	-60,6	0	114

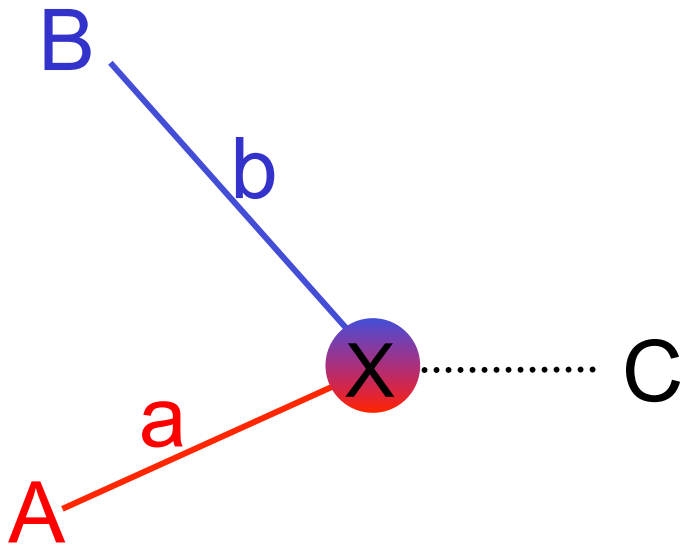
Grouping (A mit B)

Neighbor-joining (NJ)

Der nächste Schritt:

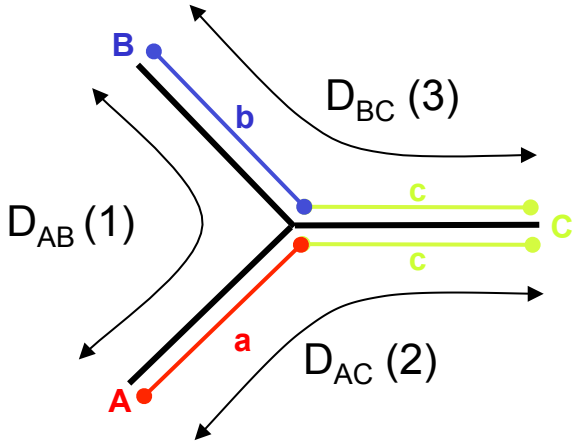
Errechnen der Astlängen nach Fitch-Margoliash (FM)

W. Fitch
1929-2011



Fitch-Margoliash-Methode

3 Taxa



	A	B	C
OTU A	**0**	**22**	**39**
OTU B		**0**	**41**
OTU C			**0**
OTU D			

Einzelabstände

(1) D_{AB} = a+b = 22
(2) D_{AC} = a+c = 39
(3) D_{BC} = b+c = 41

(2) - (3)

Astlängen nicht bekannt, **Distanzen** aber ja!

=> 2 Unbekannte (a, b); 1 “Konstante” (c)
=> z.B. aus der Differenz von (3) und (2) ist Unterschied der Äste errechenbar
=> Auflösen nach b
=> Einsetzen in (1)

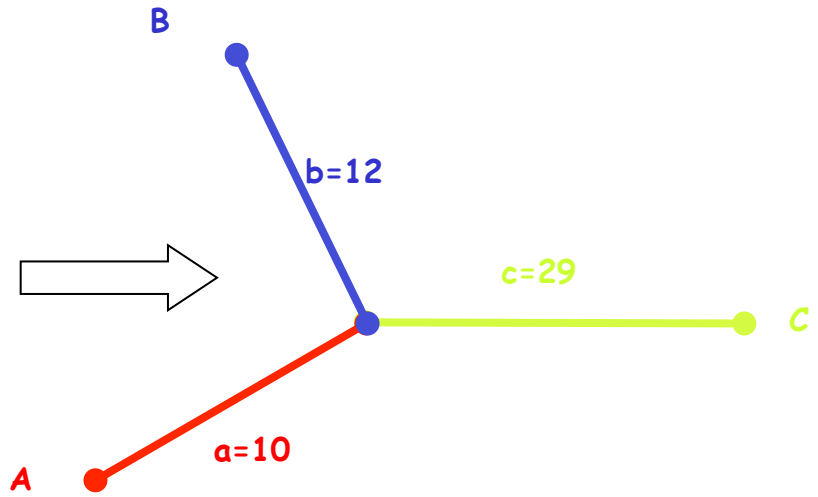
a-b = 39 – 41 = -2
-b = -2-a
b = 2+a
a+a+2 = 22
2a = 22-2
a = 10

Fitch-Margoliash-Methode

3 Taxa

	A	B	C
OTU A	0	22	39
OTU B		0	41
OTU C			0
OTU D			

(a): 10
(b): 12
(c): 29



Neighbor-joining (NJ)

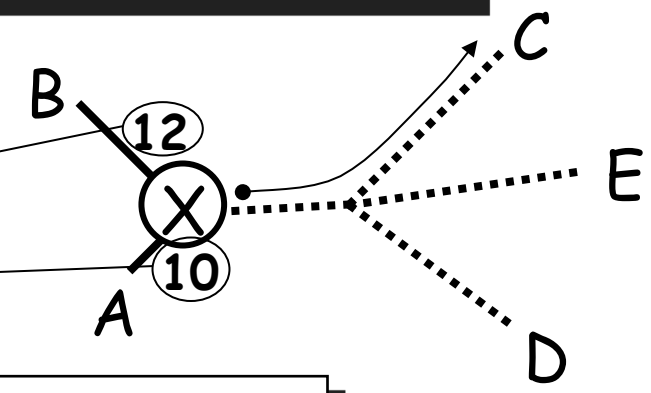
Ausgangsmatrix

	A	B	C	D	E	Summe
OTU A	0	22	39	39	41	141
OTU B		0	41	41	43	147
OTU C			0	18	20	118
OTU D				0	10	108
OTU E					0	114

Erstellen einer reduzierten Datenmatrix mit AB als composite taxon

$d_{XC} = (d_{AC} - d_{AX} + d_{BC} - d_{BX})/2$

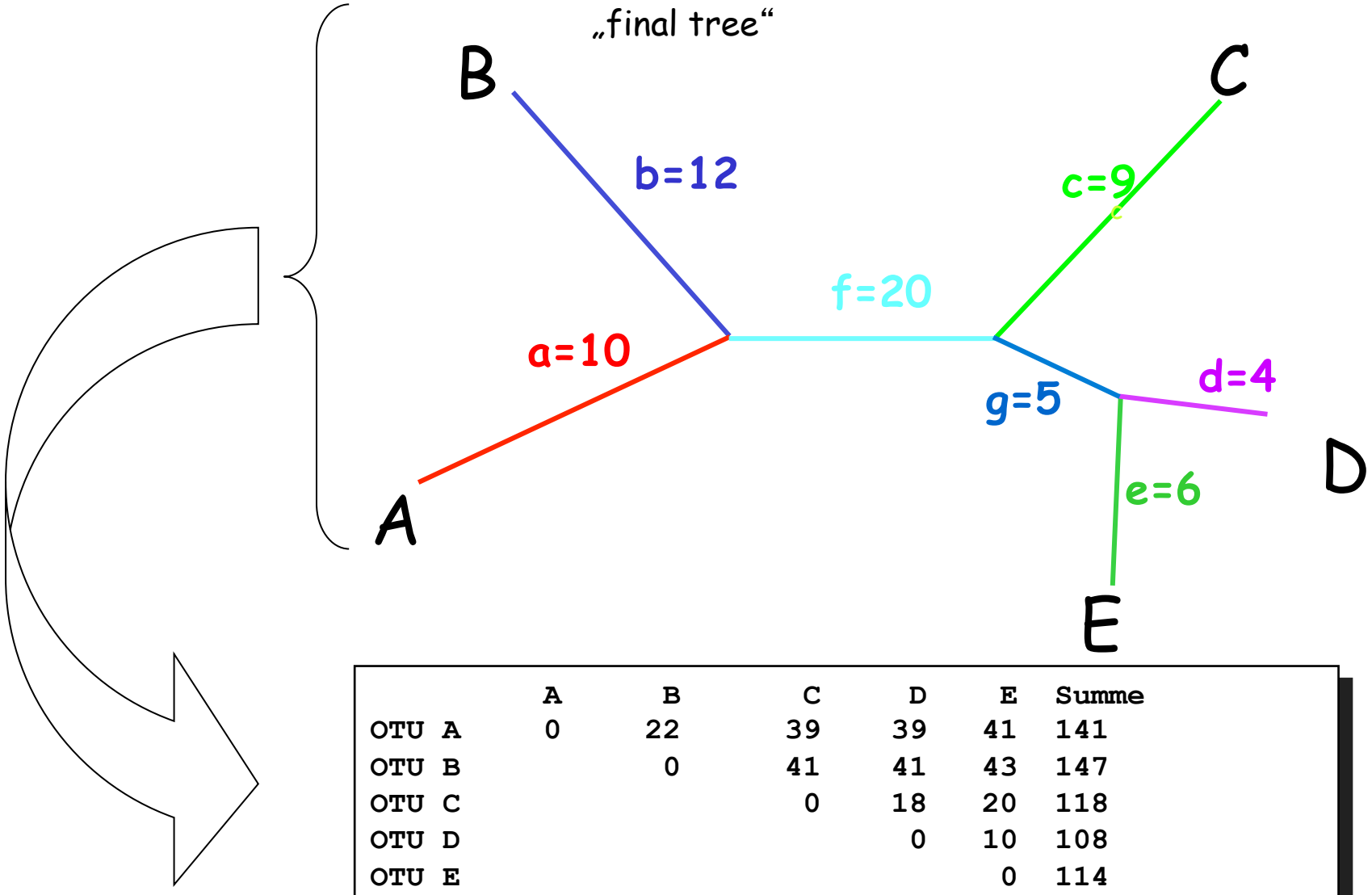
<=> (39 - 10 - 41 - 12)/2 = 29



	AB	C	D	E	Summe
OTU AB	0	29	29	31	89
OTU C	-49	0	18	20	67
OTU D	-44	-44	0	10	57
OTU E	-44	-44	-49	0	61

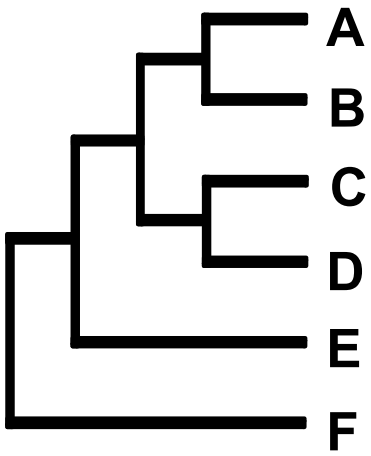
Berechnen der transformierten Matrix, Identifizierung der nächsten Gruppierung, Berechnung der Astlängen nach FM, usw...

Neighbor-joining (NJ)

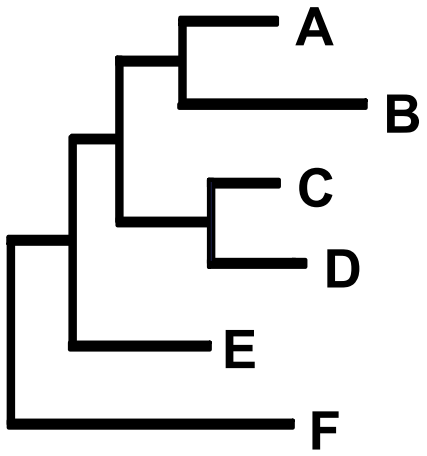


	A	B	C	D	E	Summe
OTU A	0	22	39	39	41	141
OTU B		0	41	41	43	147
OTU C			0	18	20	118
OTU D				0	10	108
OTU E					0	114

Distanzmethoden: UPGMA vs. NJ



UPGMA



NJ

Außengruppe festgelegt	⟷	Außengruppe wählbar
konstante Evolutionsrate	⟷	unterschiedliche Evolutionsraten
Verlust der realen Astlängen	⟷	Kein Astlängenverlust
Keine Matrixrekonstruktion möglich	⟷	Matrixrekonstruktion möglich

Weitere Distanzmethoden

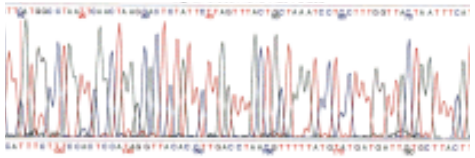
- **Least-squares-Methode**
  - Fehler (Abweichung) mit der *n* Sequenzen auf einen Baum gepasst werden
  - K_{ij} korrigierte Wert der Distanz (Distanzmatrixwert) zwischen *i* und *j*
  - P_{ij} Länge des Astes, der *i* und *j* verbindet

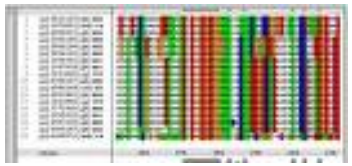
$$e = \sum_{\substack{i,j \\ 1 \le i < j < n}} (Kij - Pij)^2$$

- **Minimum Evolution**
  - Baum aus *n* Sequenzen besitzt *2n-3* Zweige
  - Jeder Zweig *z* hat Länge *l*
  - Summe dieser Zweiglängen ist die Länge des Baumes = minimal
  - Nach der LS-Formel wird dann die Abweichung der Astlängen von den Distanzen minimiert

$$L = \sum_{z=1}^{2n-3} l_z$$

Was bisher geschah...





Daten

Clustal

MSA

Evolutionsmodelle

(JC, K2P ...)

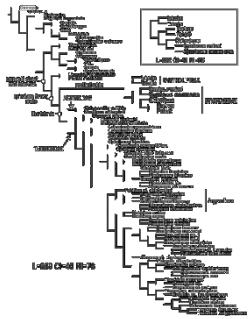
$$r_i = \frac{1}{N-2}\sum_{k=1}^{N} d_{i,k}$$

	A	B	C	D	E	Summe
OTU A	0	22	39	39	41	141
OTU B		0	41	41	43	147
OTU C			0	18	20	118
OTU D				0	10	108
OTU E					0	114

Distanzmatrix

Clustering-Algorithmus

z.B UPGMA, NJ…



Stammbaum

Stammbaum-Rekonstruktion

1. Matrix-orientierte Methoden

2. Charakter-orientierte Methoden

Maximum Parsimony (MP)

Maximum Likelihood (ML)

Bayes

Charakter-orientierte Methoden

- Arbeiten direkt mit dem Alignment
- Extrahieren mehr Information als Matrix-orientierte Methoden
- Arbeiten nicht mit Clustering, sondern durchsuchen den „tree space“ nach dem optimalen Baum

Was sind Charaktere?

- kontinuierliche oder diskontinuierliche Eigenschaften.

  1,2,3,4.... = kontinuierliche Charaktere

  A,T,G,C = diskontinuierliche Charaktere

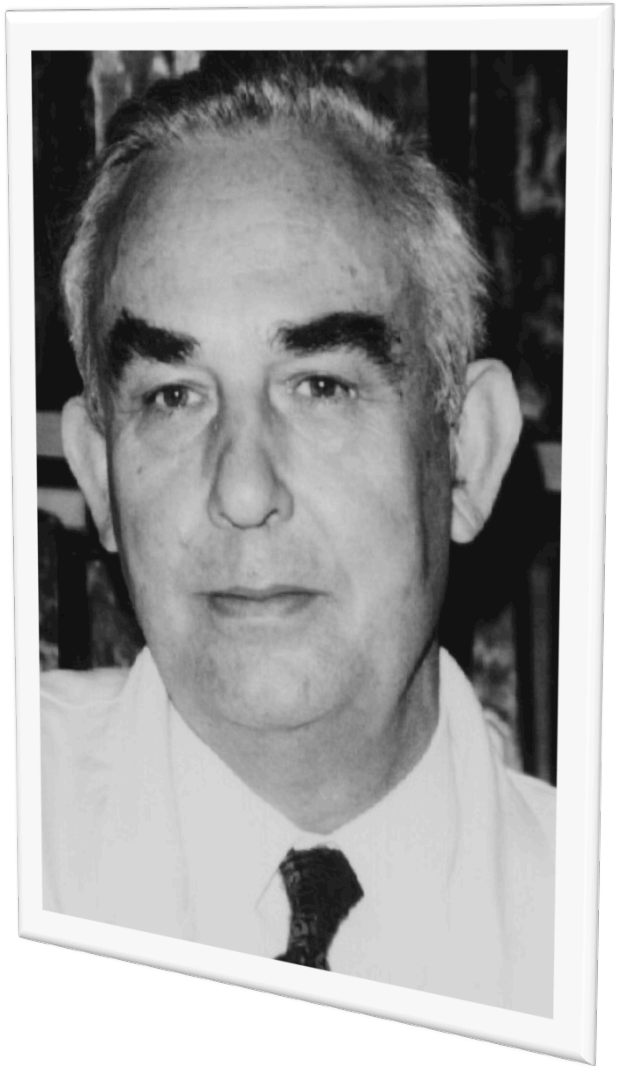
- Nukleotide und Aminosäuren können als diskrete, diskontinuierliche Charaktere behandelt werden.

- Der phylogenetische Stammbaum wird anhand des Musters der Änderungen der Charaktere berechnet

Maximum Parsimony (MP)

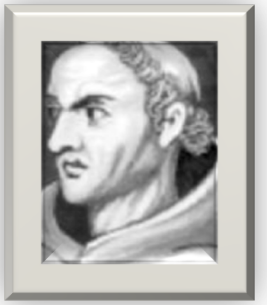
- Methode des "maximalen Geizes" bzw. der "maximalen Sparsamkeit“
- Entwickelt für morphologische Charaktere

1950 *„Grundzüge einer Theorie der phylogenetischen Systematik“,*



Willi Hennig
1913-1976

Maximum Parsimony



William of Ockham (1285-1349)

- *“Ockham's razor”* : "*Pluralitas non est ponenda sine neccesitate*" ("*Ohne Notwendigkeit soll keine Vielfältigkeit hinzugefügt werden*")
- Annahme: Evolution ging den kürzesten Weg (“Ökonomie-Prinzip”)
- kürzester Stammbaum wird berechnet, d.h. der die wenigsten evolutiven Schritten benötigt
- „Schritte“ = Änderungen von Charakteren

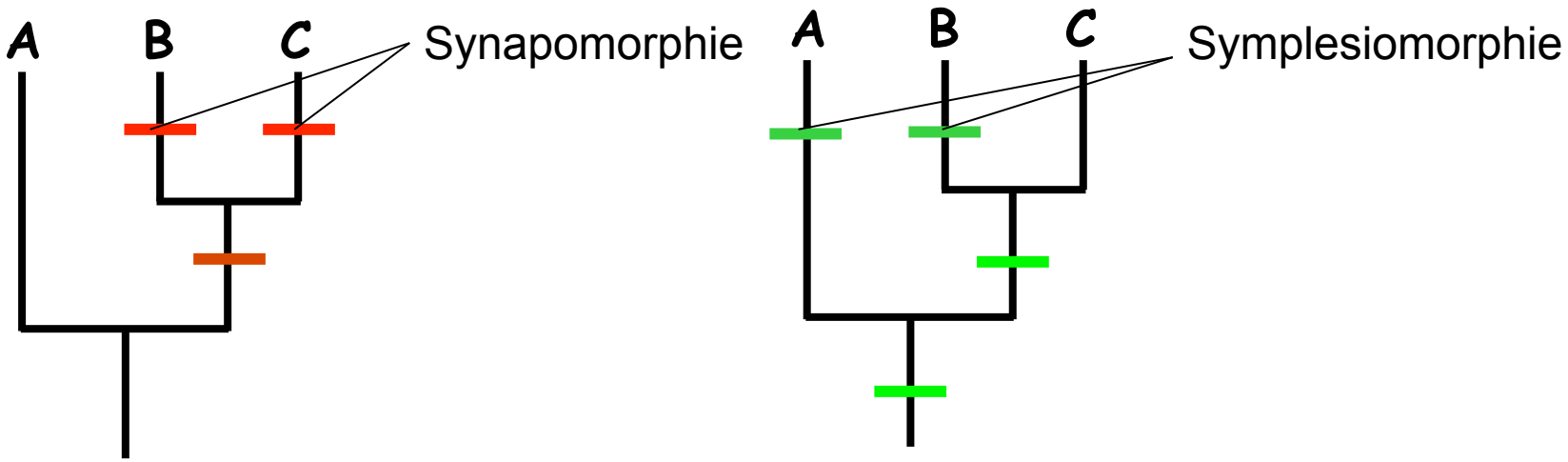
Maximum Parsimony

- Erklärung mit morphologischen Charakteren möglich

- Gleiche Prinzipien sind für Sequenzen (Basenpaare, Aminosäuren) gültig

Maximum Parsimony

Apomorphie: Abgeleiteter Charakter.
Synapomorpie: Abgeleiteter Charakter, welcher mehreren Taxa gemeinsam ist.
Plesiomorphie: Primitiver Charakter.
Symplesiomorphie: Primitiver Charakter, welcher mehreren Taxa gemeinsam ist.

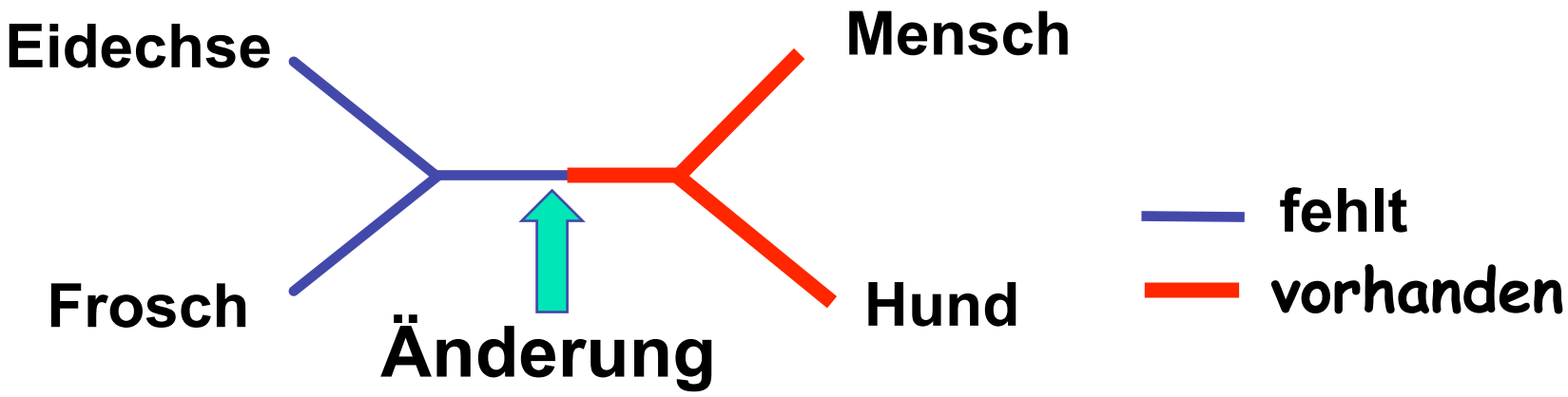


Nur Synapomorphien sind in MP zu verwerten!

Synapomorphie

- Beispiel Haare:

  Haare sind in der Evolution nur einmal entstanden.
  D.h., der Besitz von Haaren ist ein synapomorphes Merkmal der Säugetiere.



Synapomorphie = "richtige" Information

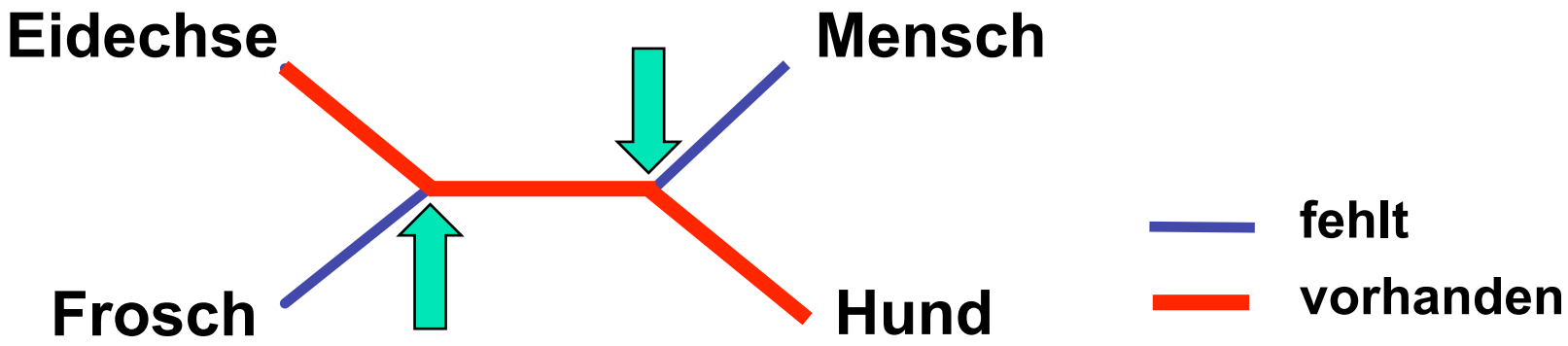
Homoplasie

- **Homoplasie** ist Übereinstimmung ohne Homologie (d.h., keine gemeinsame Abstammung)
- **Homoplasie** resultiert aus unabhängiger Evolution (**Konvergenz**, Reversion)
- **Homoplasie** ist „falsche“ Information, die zu falschen Stammbäumen führen kann
- MP ist anfällig für Homoplasie

Homoplasie-Konvergenz

- Beispiel Schwanz:

  Schwanz ging unabhängig in den Fröschen und beim Menschen verloren.



Anwendung auf Sequenzen

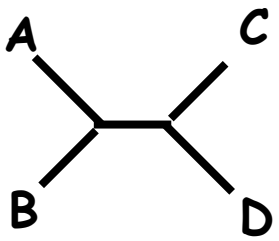
- Nukleotide und Aminosäuren sind diskrete, diskontinuierliche Charaktere
- 4 (Nukleotide) bzw. 20 (Aminosäuren) Charaktere
- Lücken ("gaps") können als 5. bzw. 21. Charakter behandelt werden

Maximum Parsimony

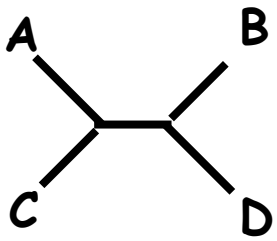
Beispiel:

	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

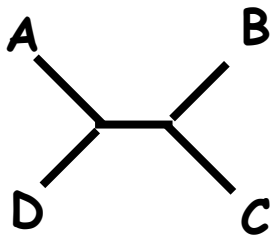
3 mögliche Stammbäume



((A,B),(C,D))



((A,C),(B,D))



((A,D),(B,C))

Maximum Parsimony

Welche Positionen sind **informativ**, bevorzugen also eine bestimmte Topologie?

	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

3 Positionen invariabel => nicht informativ

Maximum Parsimony

	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

6 Positionen sind variabel

=> aber auch informativ?

Maximum Parsimony

	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

3 Positionen sind zwar variabel,

aber nicht informativ

Maximum Parsimony

Welche Positionen sind aber nun informativ?

	Position										
Sequenz	1	2	3	4	5	6	7	8	9	10	11
A	A	A	G	A	G	T	G	C	A	-	A
B	A	G	C	C	G	T	G	C	G	-	G
C	A	G	A	T	A	T	C	C	A	C	G
D	A	G	A	G	A	T	C	C	G	C	G
					*		*		*	*	

=> nur 3 von 9 Positionen sind informativ, d.h., favorisieren eine best. Topologie.

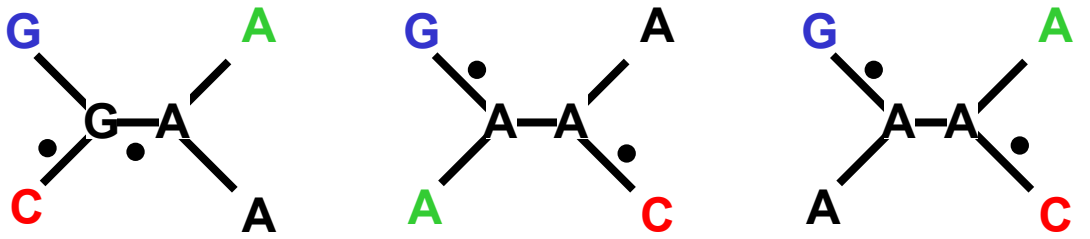
=> Indels sind Charaktere!

Maximum Parsimony

Sequenz	Position 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
			-		+				+

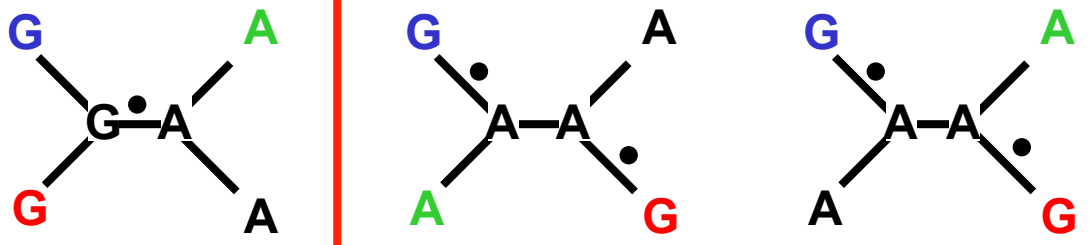
((A,B),(C,D)) ((A,C),(B,D)) ((A,D),(B,C))

Position 3:

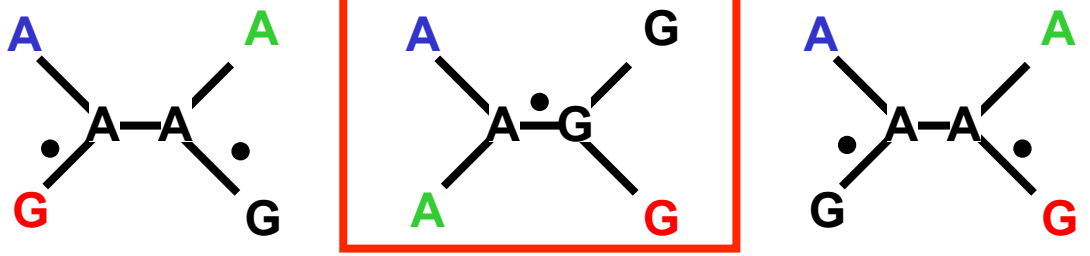


nicht informativ

Position 5:



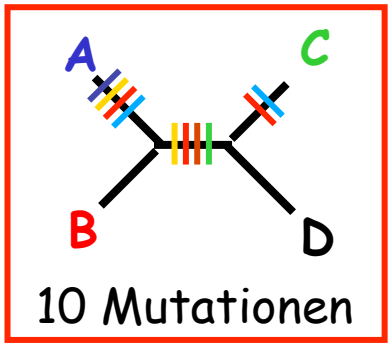
Position 9:



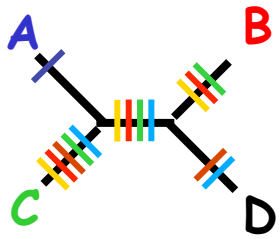
Maximum Parsimony

	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
					*		*		*

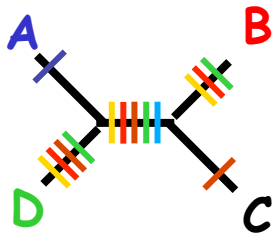
3 mögliche Stammbäume



10 Mutationen



15 Mutationen



14 Mutationen

Maximum Parsimony...

... durchsucht den „tree space“!

Exhaustive = Alle Stammbäume werden untersucht, der beste Stammbaum wird erhalten (garantiert).

Branch-and-Bound = Einige Stammbäume werden berechnet, bester Stammbaum garantiert.

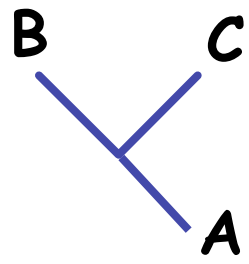
Heuristic = Einige Stammbäume werden berechnet, bester Stammbaum nicht garantiert.

MP Exhaustive Search

MP Exhaustive Search

Start: 3 beliebige Taxa

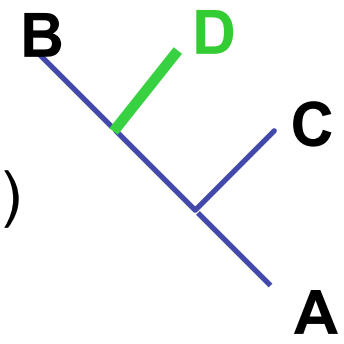
(1)



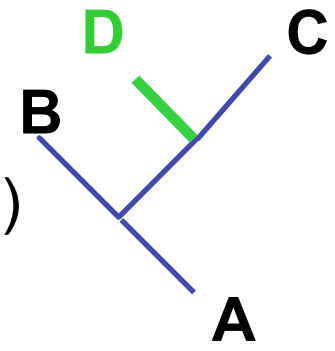
"Branch addition"

+ 4. Taxon (D) in jeder möglichen Position -> 3 Bäume

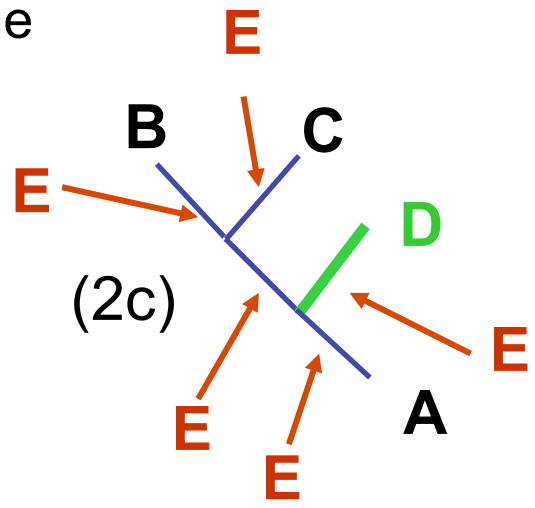
(2a)



(2b)



(2c)



+ 5. Taxon (E) in jeder der fünf möglichen Positionen
=> 15 Stammbäume etc.

MP Exhaustive Search

Problem: Anzahl der möglichen Stammbäume

Number of OTUs	Number of rooted trees	Number of unrooted trees
2	1	1
3	3	1
4	15	3
5	105	15
6	954	105
7	10 395	954
8	135 135	10 395
9	2 027 025	135 135
10	34 459 425	2 027 025

=> bei > ~10 Sequenzen ausführliche Suche aller Stammbäume *de facto* unmöglich

MP tree search

1. Lösung "Branch and bound":

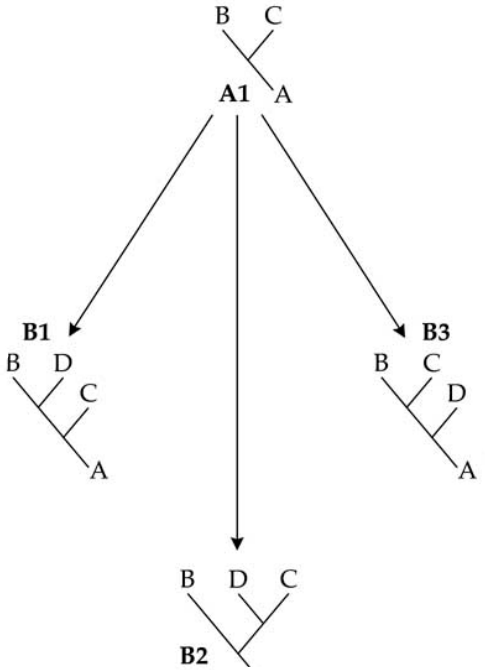
- Erster Stammbaum wird mit schneller Methode (z.B. NJ) berechnet > die Anzahl der notwendigen Schritte (L) wird berechnet.

- => verwirft Gruppen von Bäumen, die nicht kürzer werden können als L.

- Kann für Problemlösungen mit ~ 20 Taxa verwendet werden.

MP branch & bound

„verzweigen und beenden“



MP tree search

2. Lösung: Heuristische Verfahren

- „**stepwise addition**“ drei Taxa Baum – schrittweise Addition auf allen nächsten Ebenen (großes Problem: lokale Maxima)
- „**star decomposition**“: schrittweiser Abbau von Taxa bzw. Zusammenführung und Evaluation (großes Problem: lokale Maxima)

Kombination mit anderen Algorithmen

- „**branch swapping**“ (Zweige vertauschen):

  Nearest neighbor interchange (**NNI**)

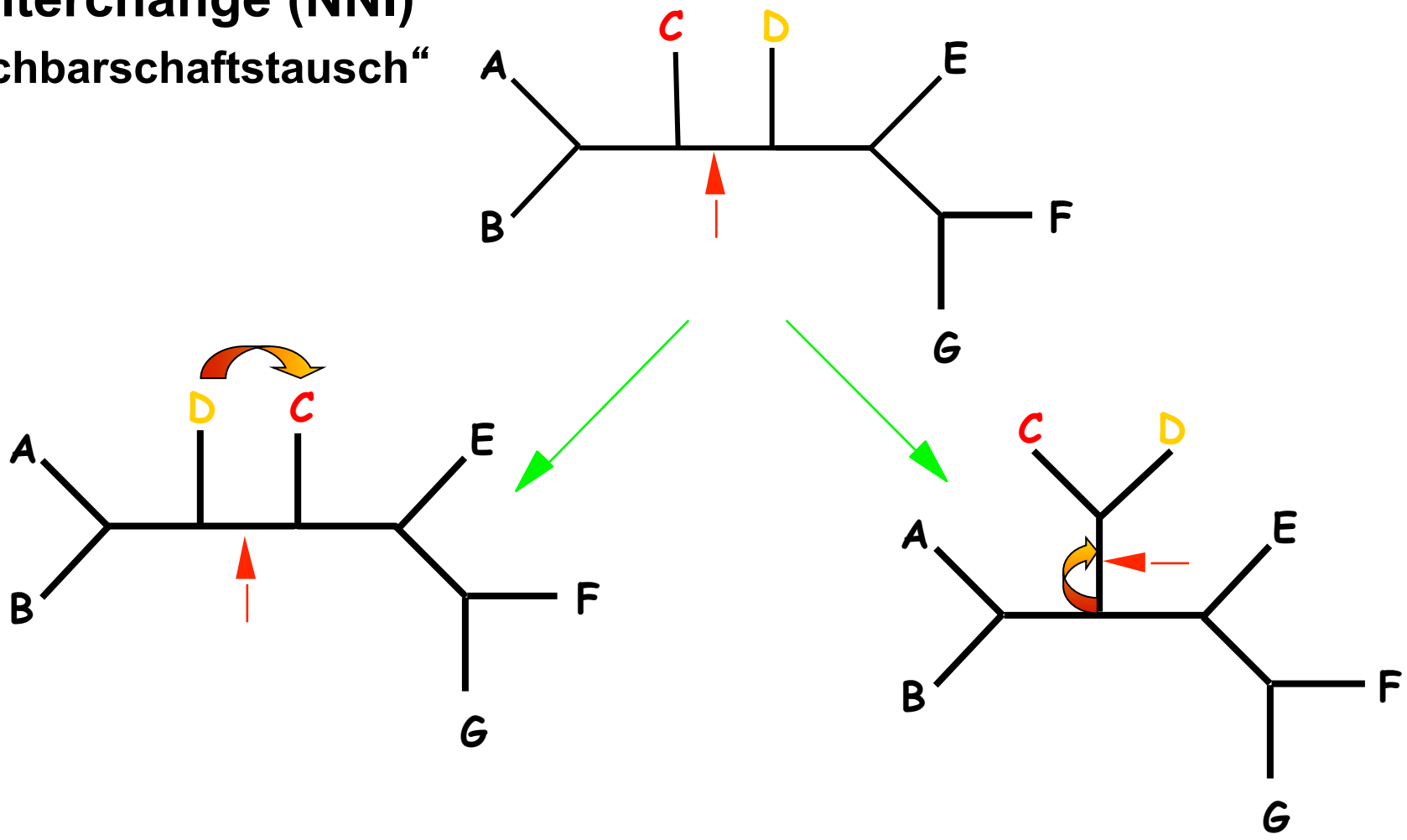
  Subtree pruning and regrafting (**SPR**)

  Tree bisection and reconnection (**TBR**)

MP heuristic tree search

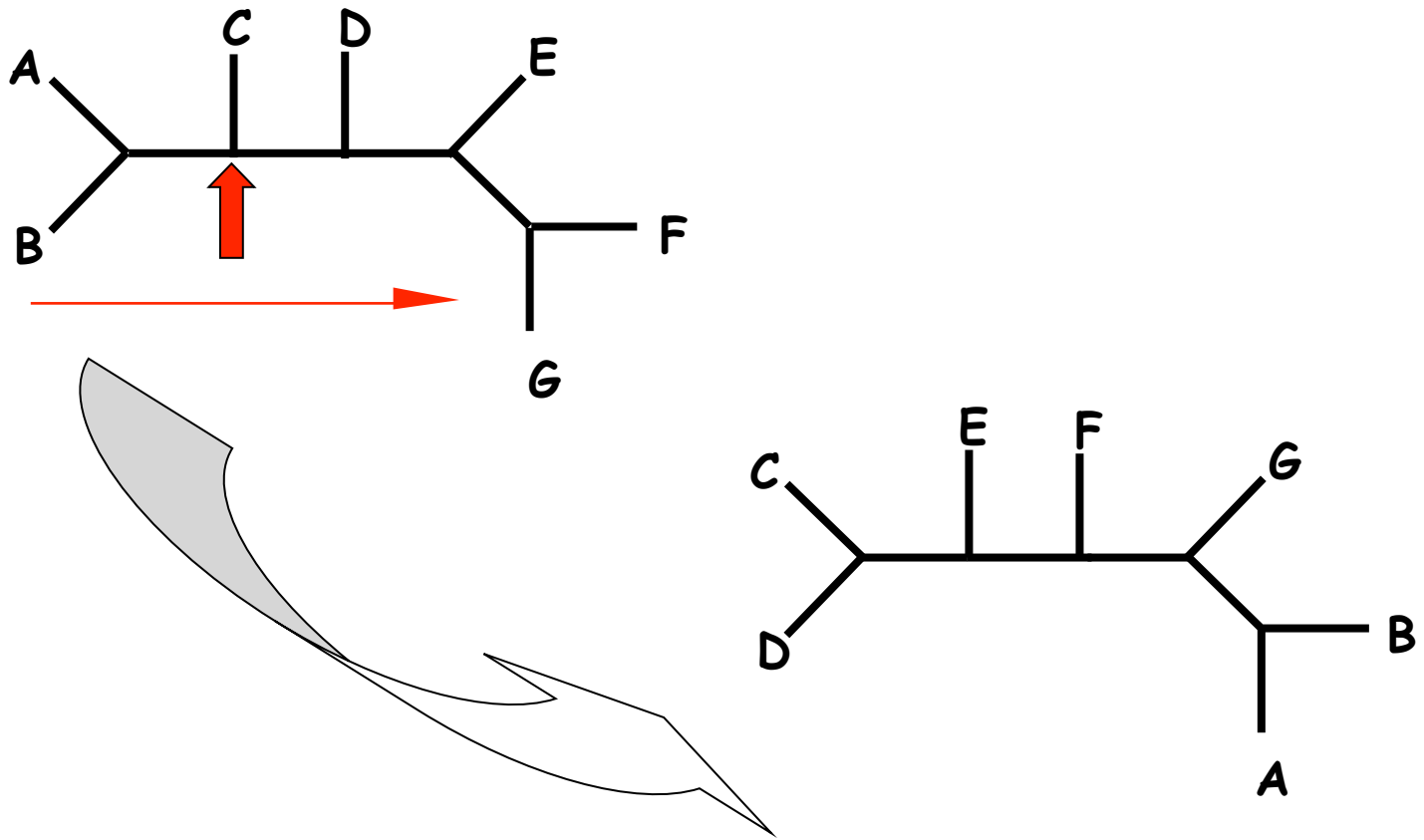
Nearest neighbor interchange (NNI)

„Nachbarschaftstausch“



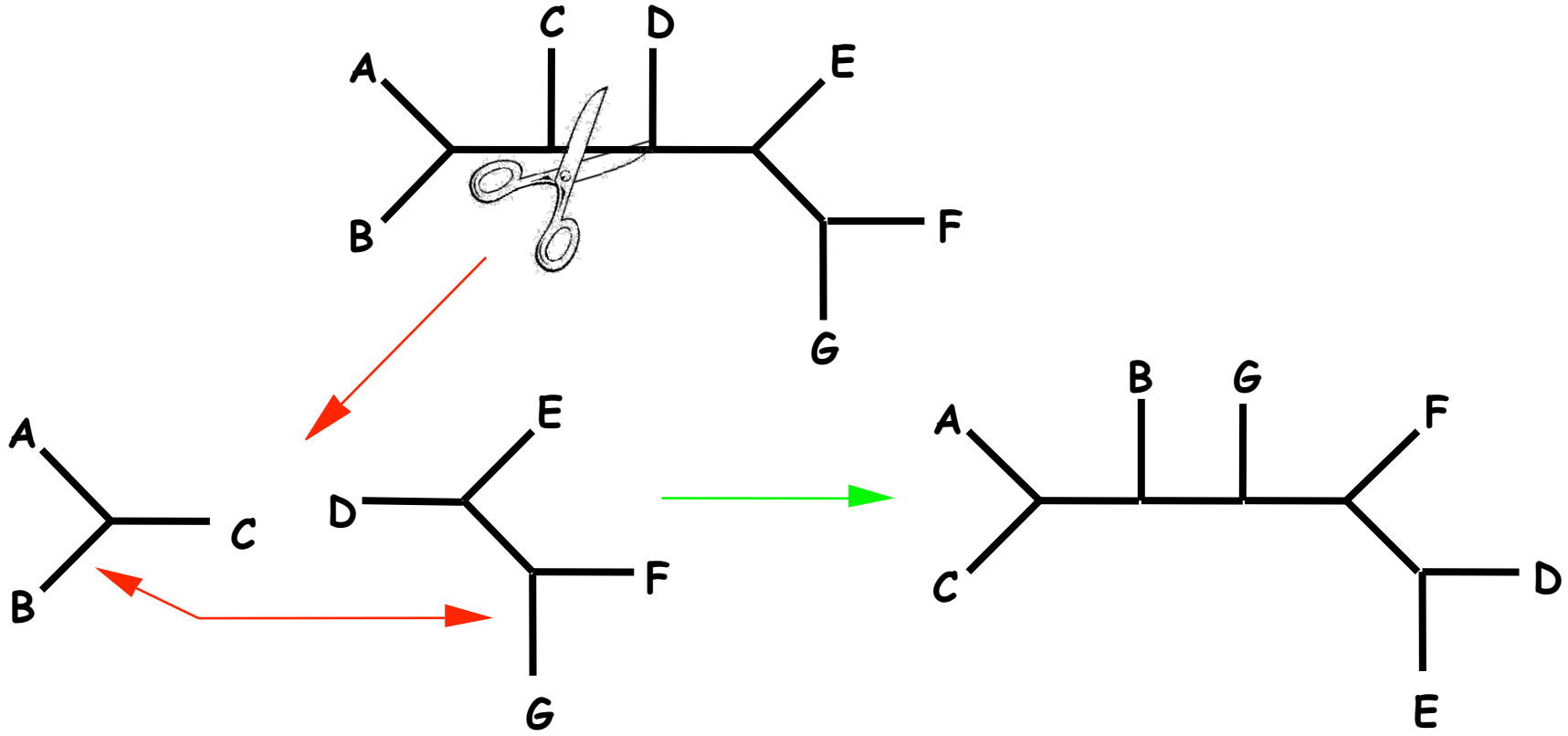
MP heuristic tree search

Subtree pruning and regrafting (SPR) „Astverpflanzung“



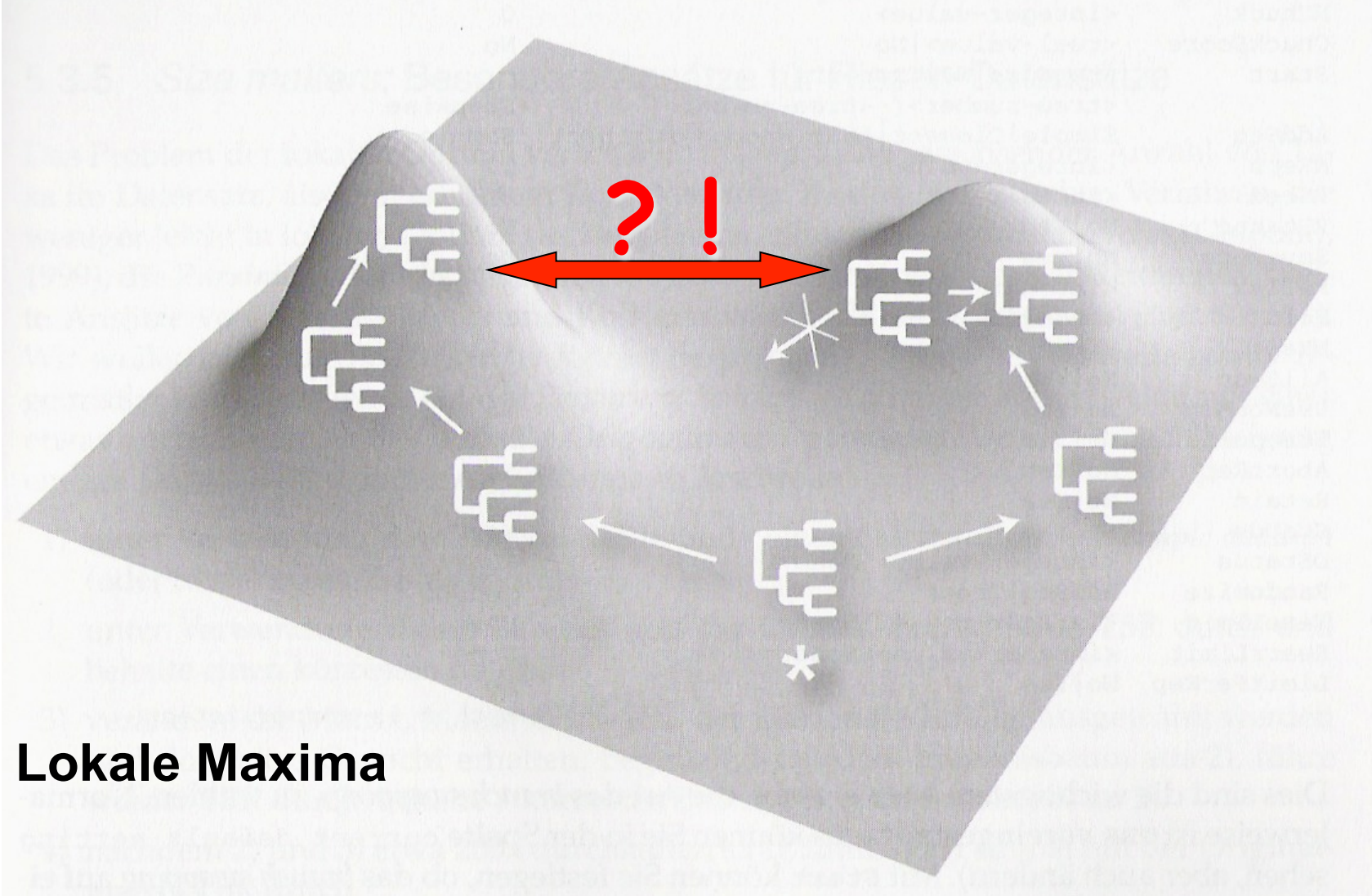
MP heuristic tree search

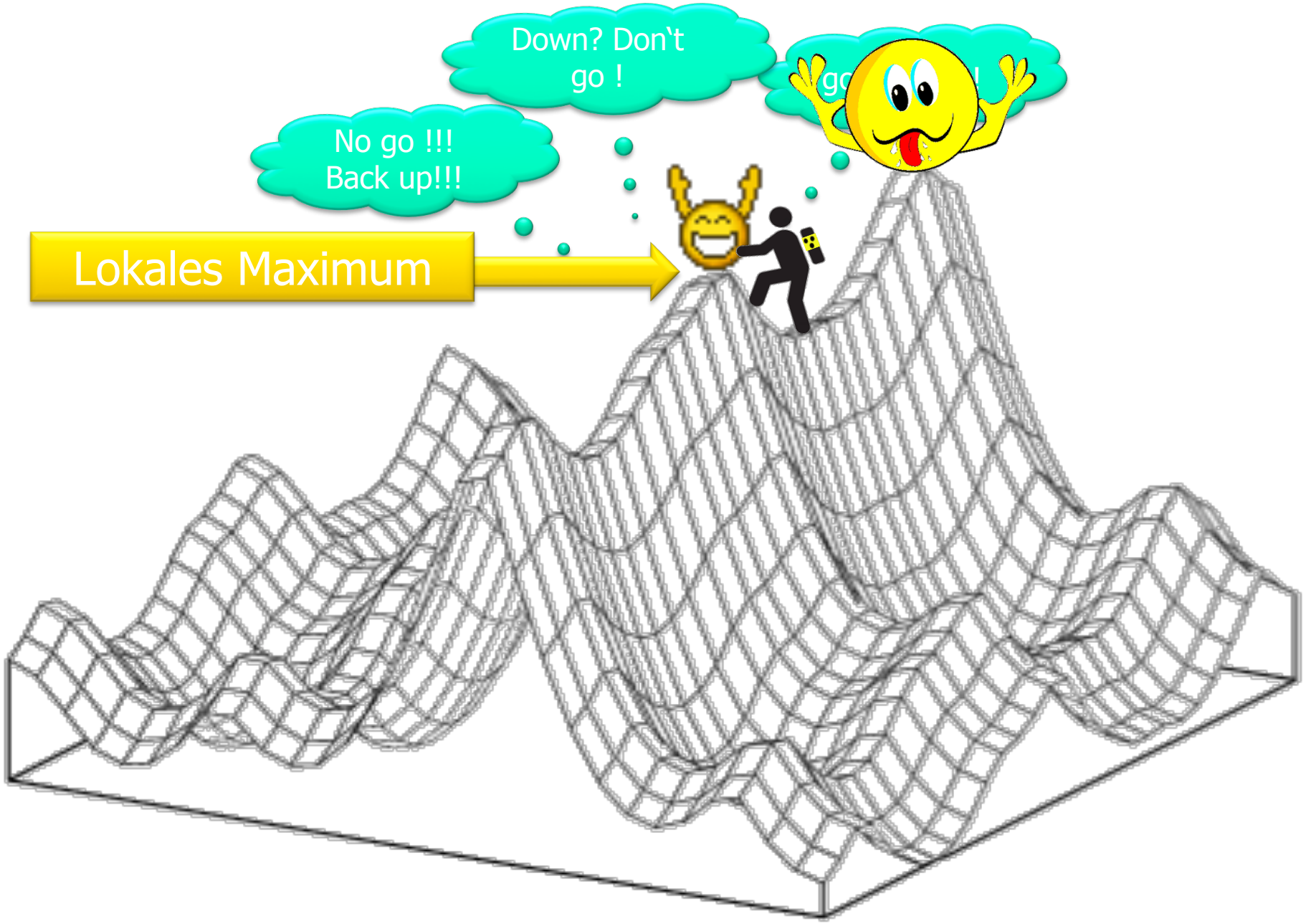
Tree bisection and reconnection (TBR)
„Baumschnittwiederverknüpfung“ (effektiv)



Gutes Durchmischen, aber CPU-aufwändig

Das Problem des blinden Bergsteigers...




Down? Don't
go !
No go !!!
Back up!!!
Lokales Maximum

„long branch attraction“

- OTUs mit hoher Evolutionsrate und vielen Veränderungen („long branches“) enthalten notwendigerweise zahlreiche Homoplasien/Konvergenzen

- diese Homoplasien führen dazu, daß MP die „long branch“-OTUs im Baum fälschlicherweise zueinandergruppiert

> u. U. Taxa mit long branches entfernen!

Größtes Problem bei MP:

„long branch attraction“

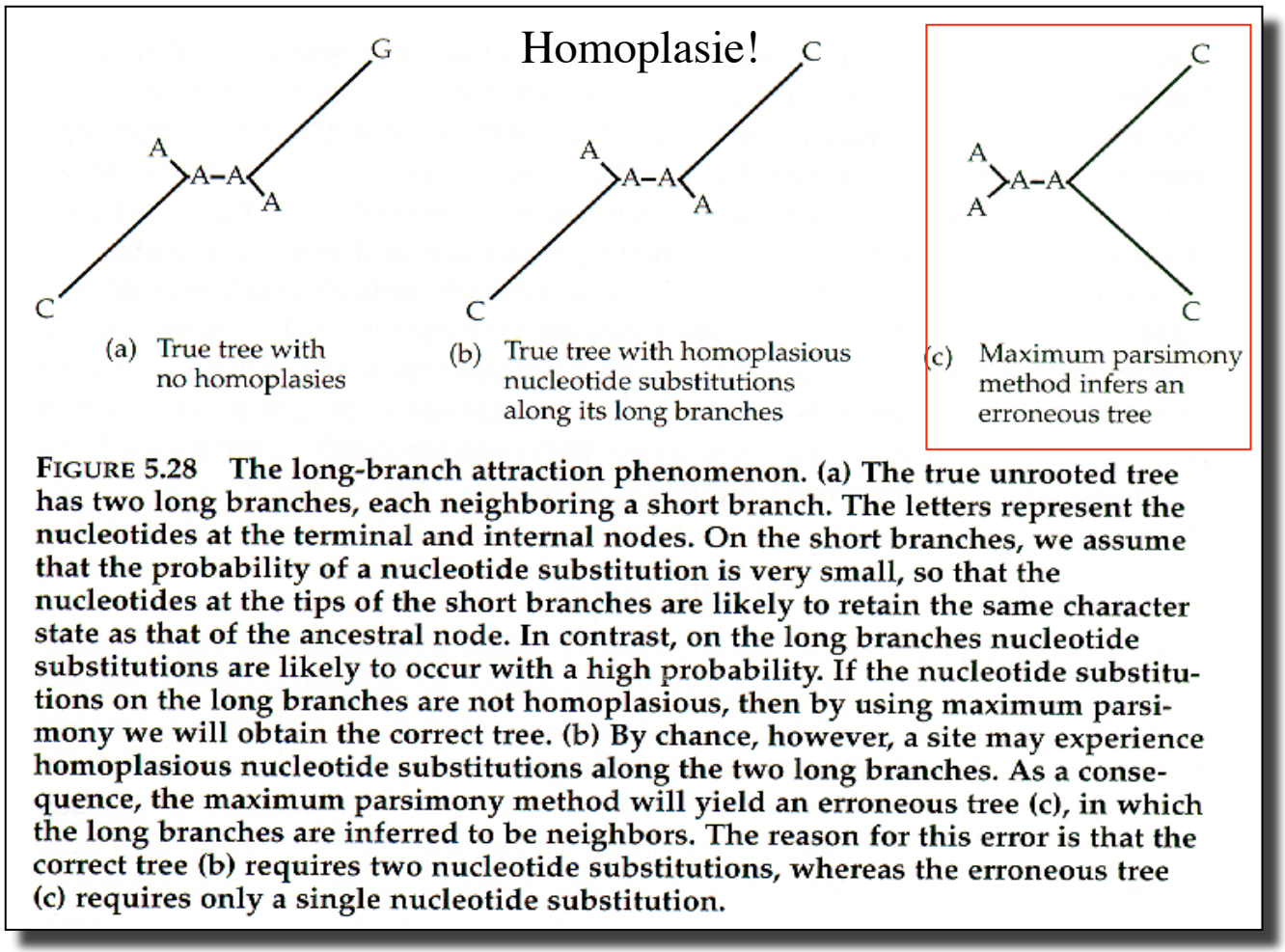


FIGURE 5.28 The long-branch attraction phenomenon. (a) The true unrooted tree has two long branches, each neighboring a short branch. The letters represent the nucleotides at the terminal and internal nodes. On the short branches, we assume that the probability of a nucleotide substitution is very small, so that the nucleotides at the tips of the short branches are likely to retain the same character state as that of the ancestral node. In contrast, on the long branches nucleotide substitutions are likely to occur with a high probability. If the nucleotide substitutions on the long branches are not homoplasious, then by using maximum parsimony we will obtain the correct tree. (b) By chance, however, a site may experience homoplasious nucleotide substitutions along the two long branches. As a consequence, the maximum parsimony method will yield an erroneous tree (c), in which the long branches are inferred to be neighbors. The reason for this error is that the correct tree (b) requires two nucleotide substitutions, whereas the erroneous tree (c) requires only a single nucleotide substitution.

Falsche Topologie!

„LBA“ oder
„Felsenstein zone“

Maximum Parsimony

Vorteile:

- einfach
- „ohne“ konkretes Evolutionsmodell
- Errechnung anzestraler Positionen
- funktioniert gut mit konsistenten Datensätzen

Nachteile:

- empfindlich gegen Homoplasien (Konvergenz)
- empfindlich gegen "Long Branch Attraction"
- Astlängen werden unterschätzt
- kein Evolutionsmodell möglich!

Methoden-Übersicht

Rekonstruktionsmethode

	Datentyp: Distanzen	Datentyp: Character
Clustering-Algorithmus	UPGMA Neighbor joining	
Such-Strategie	Minimum Evolution	Maximum Parsimony Maximum Likelihood Bayes