

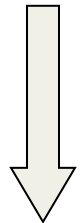
# Das Szenario ...ein neues tödliches Virus!

- Labor: Isolierung der Erbsubstanz und Sequenzierung



- Computer: Erkennen der Virusgene (*de novo* Genvorhersage)

Ähnlichkeit zu bekannten Genen? (Datenbanksuchen)



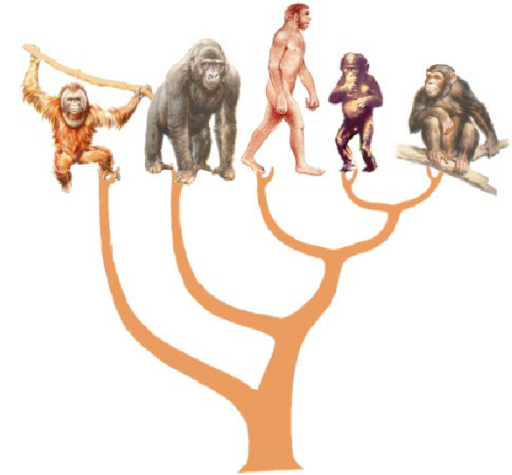
Verwandtschaft? Ausbreitung? Herkunft?  
(Phylogenetische Rekonstruktion)

Struktur der Proteine?  
(Struktur-Vorhersage, Modellierung)

Wirkstoff-Design

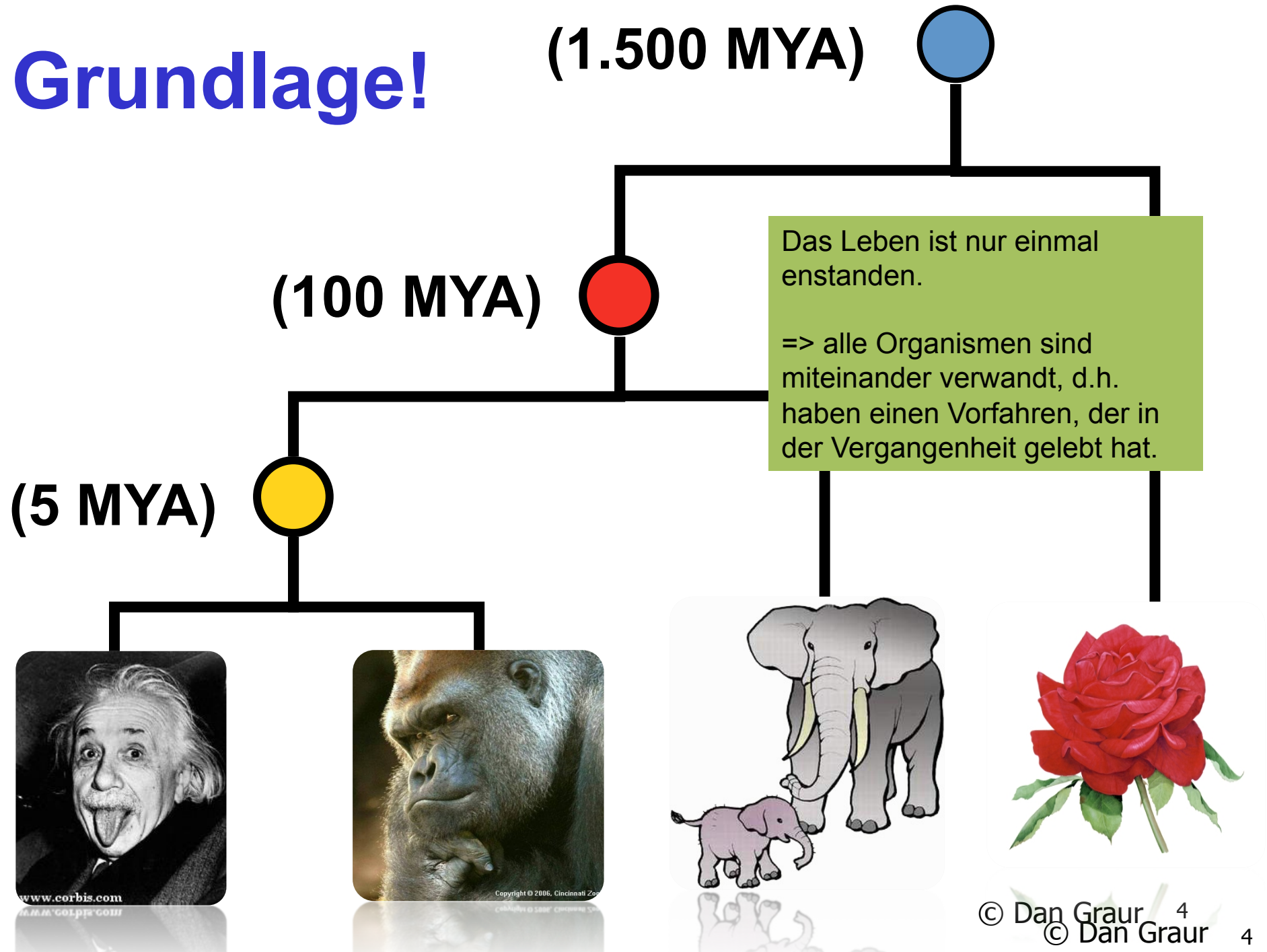
- Labor: Wirkstoff-Test

# Molekulare Phylogenie



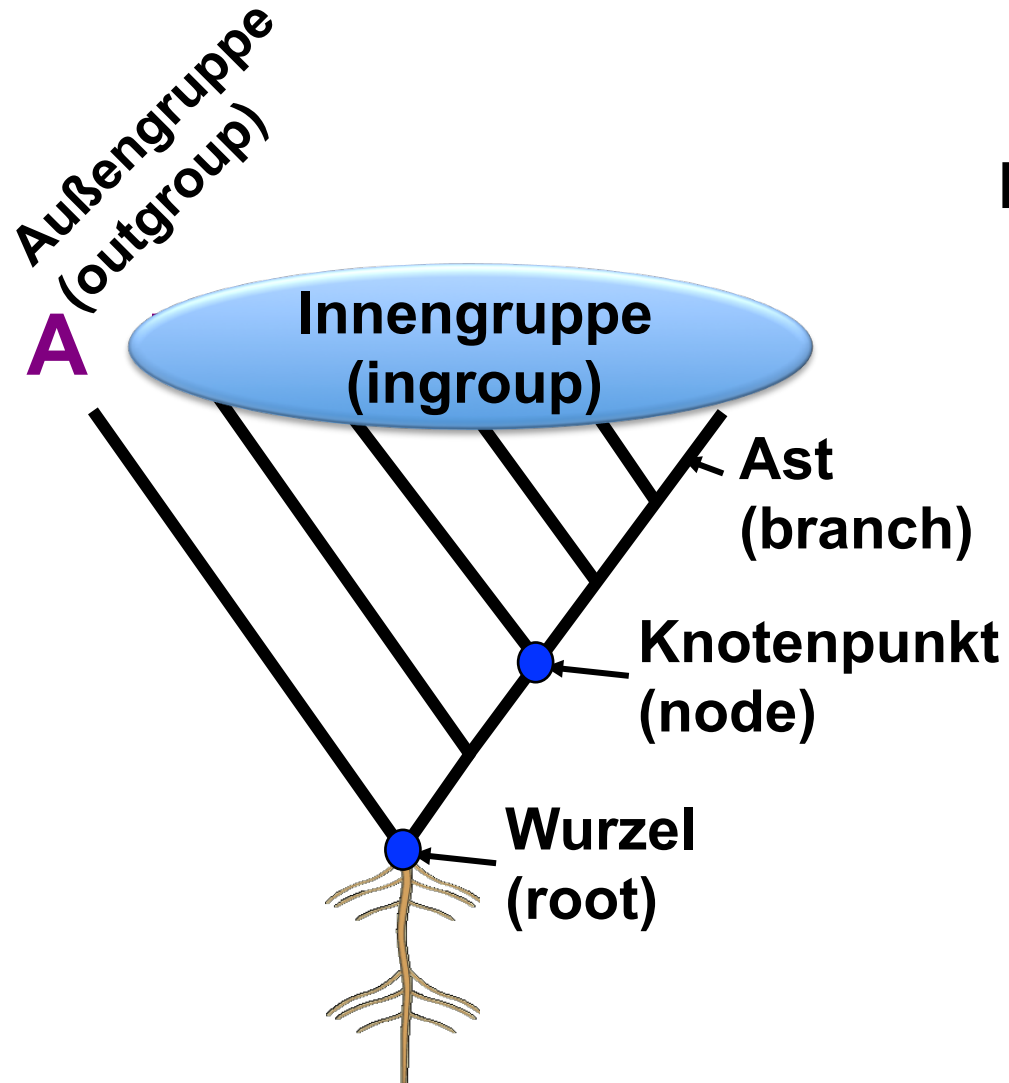
- **Verwandschaft von Organismen**  
(molekulare Systematik, Forensik)
- **Verwandschaft zwischen Genen/Proteinen**  
(Genomevolution, Gen/Proteinfunktion)
- **Ausbreitung von Lebewesen**  
(Anthropologie, Ökologie, Epidemiologie)

# Grundlage!

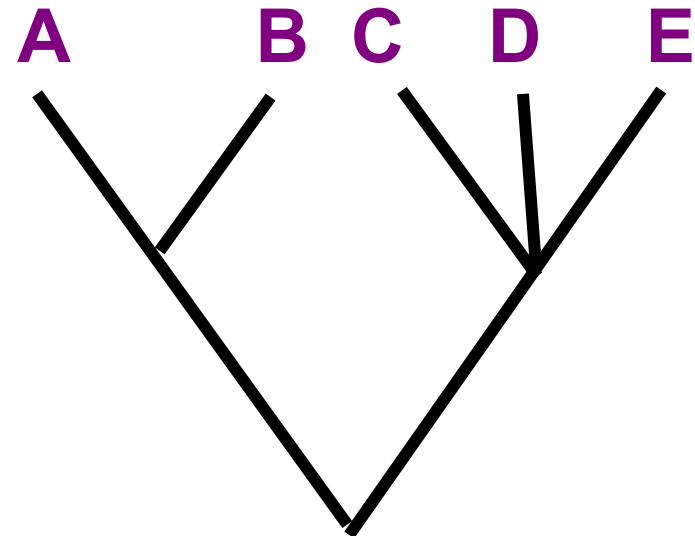




# Grundbegriffe



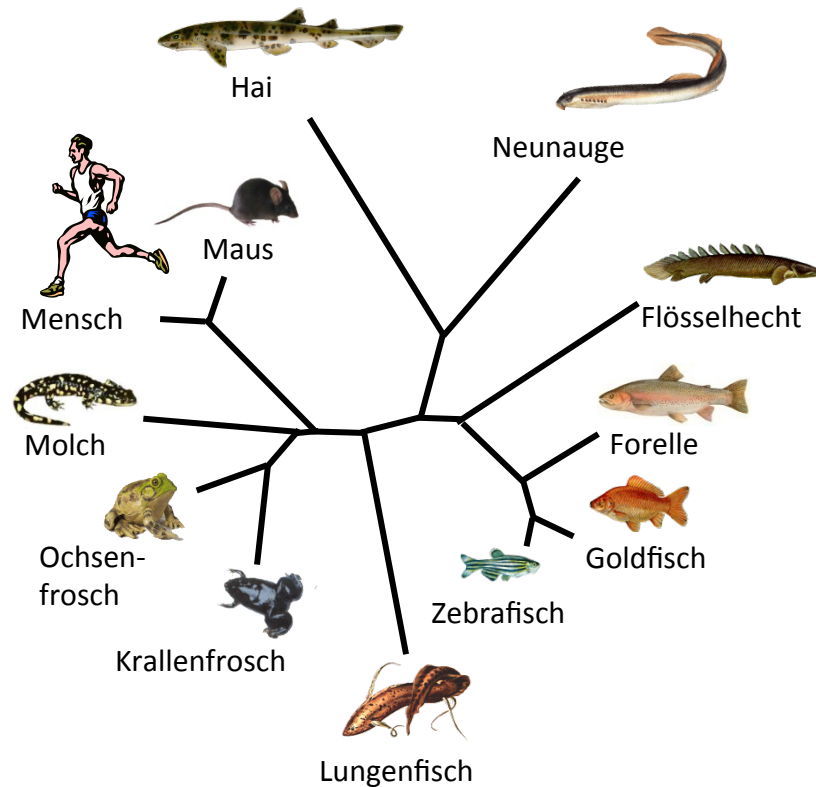
Dichotomie   Polytomie



# Stammbaum-Typen

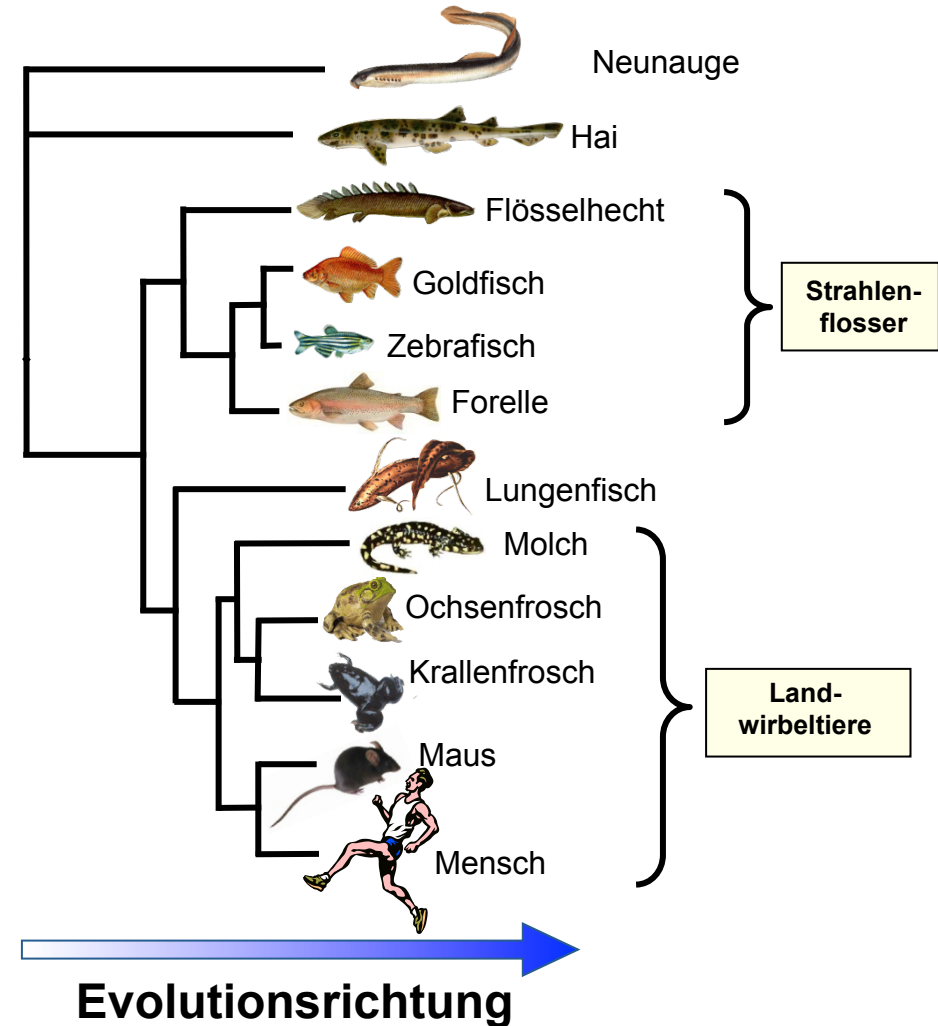


## Ohne Außengruppe:

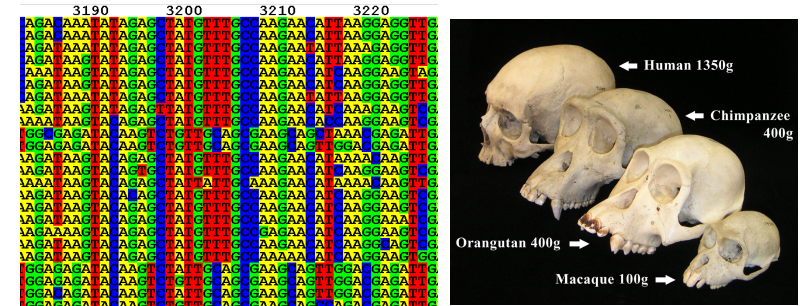


Evolutionsrichtung?

## Mit Außengruppe:



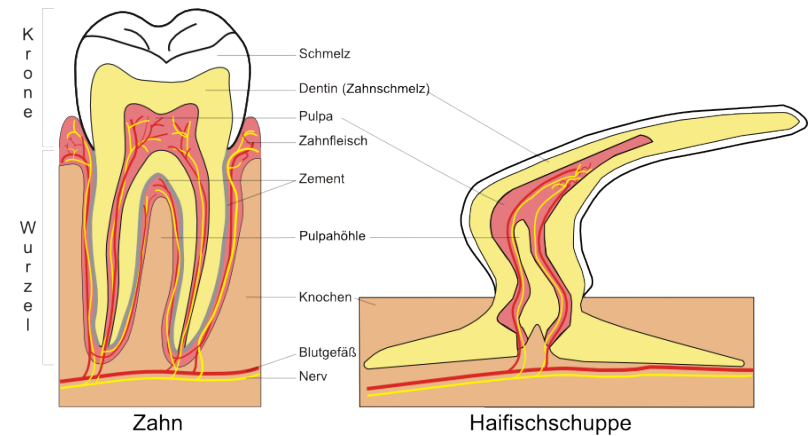
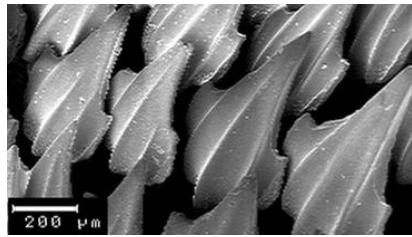
# Molekulare vs. klassische Phylogenie



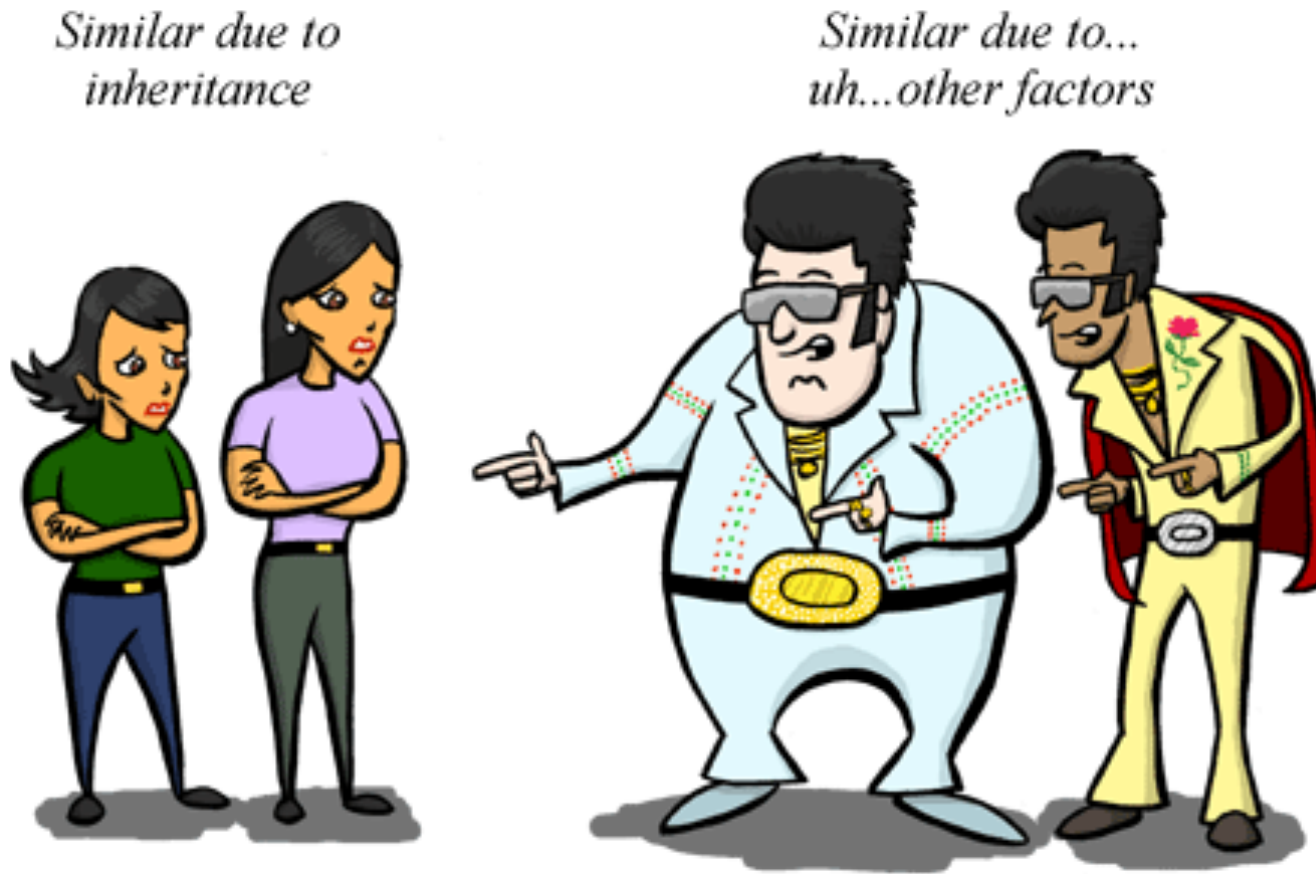
- Sequenzen sind direkt vererbt; **keine Umwelteinflüsse**
- Sequenzdaten sind in großer Menge relativ **kostengünstig** und schnell zu erhalten (Dank sei der PCR!!!)
- weitgehend **frei von Interpretationseinflüssen** („reduziert“, „etwas abgeflacht“ etc.)
- Sequenzen erlauben **Vergleiche über große Distanzen** (Tiere, Pilze, Pflanzen)

Dennoch: auch molekulare Daten können zu falschen Stammbäumen führen ☹

# Welche Vergleiche mache Sinn?



# Welche Vergleiche mache Sinn?



[http://evolution.berkeley.edu/evolibrary/article/similarity\\_hs\\_01](http://evolution.berkeley.edu/evolibrary/article/similarity_hs_01)

HOMOLOGIE heißt das Zauberwort für sinnvolle Vergleiche!

# Welche Vergleiche mache Sinn?

YWREDFGINS SHHWHHLLVYPI  
YWREDYGINVHHWHHLLIYPE  
YFREDIGINLHHWHHLLVYPE  
YFREDIGVNSHHWHHLLVYPT  
YFREDLGINLHHWHHLLVYPE  
YFREDLGVNLHHWHHLLVYPI  
YFREDIGVNAHHWHHLLVYPS  
YFREDIGINS SHHWHHLLVYPT  
YFREDIGANAHHWHHLLIYPI  
YYREDVGINAHHWHHLLVYPS  
YFGEDIGLNTHHVTWHMEFPE  
YFGEDVGMNTHHVLWHMEFPE  
YFGEDIGMNIHHVTWHMDFPE

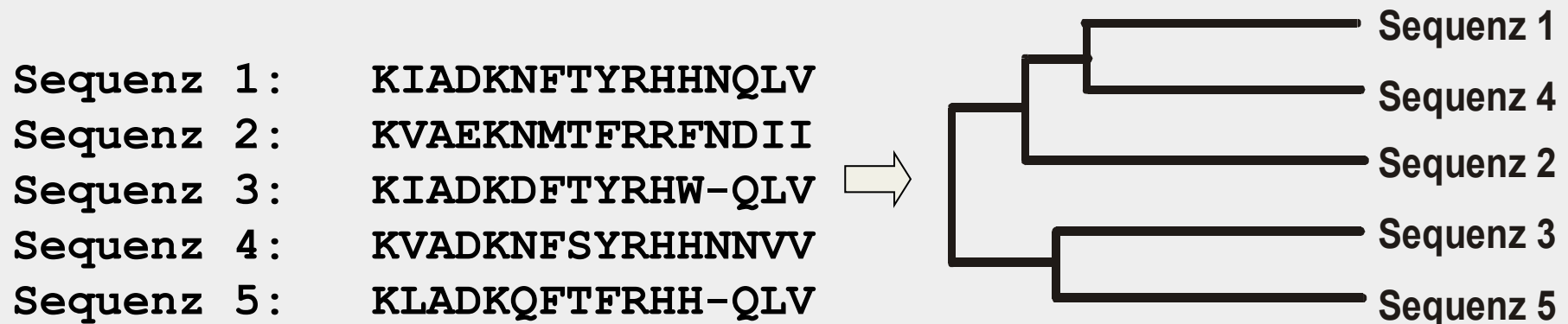
IEM-----NVN  
PAM-----GFD  
FDAADRA-IVN  
TTGPTE--VVN  
FEASDRS-IVA  
IEAPDRS-IVD  
STYDPAFFGKV  
AFYDADIFGKI  
PTWDASVMSKV  
STWNPKYFGKK  
FWWNDAYG-HH  
FWWEDSSG-RH  
FWWEDSYG-YH

VNLSRVEKLE  
LGLPKVEKLD  
NNLSRVRRYN  
NNLKKVQPLN  
NNLARVLPFN  
NHMARVQPFN  
NGLNRMIPFH  
VGLQRMIPFQ  
TGLRRMIPFH  
NGMHRMLPFN  
NYLDPVGELQ  
NHLDPVEELS  
NWLDPVDELH

?

PlePPO	YWREDFGINS SHHWHHLLVYPI	IEM-----NVN	DRKGELFYMHQQMVARYDWERLSVNLSRVEKLE	61
PmoPPO	YWREDYGINVHHWHHLLIYPE	PAM-----GFD	DRKGELFYMHQQVIARYDIERLCLGLPKVEKLD	61
BmoPPO1	YFREDIGINLHHWHHLLVYPT	FDAADRA-IVN	KDRKGELFYMHQQIARYNVERMCNNLSRVRRYN	65
DmePPOA1	YFREDIGVNSHHWHHLLVYPT	TTGPTE--VVN	KDRKGELFYMHQQILARYNVERFCNNLKKVQPLN	64
DmePPO2	YFREDLGINLHHWHHLLVYPT	FEASDRS-IVA	KDRKGELFYMHQQVIARYNAERFSNNLARVLPFN	65
DmePPO3	YFREDLGVNLHHWHHLLVYPI	IEAPDRS-IVD	KDRKGELFYMHQQIARYNAERLSNHMARVQPFN	65
EcaHcA	YFREDIGVNAHHWHHLLVYPT	STYDPAFFGKV	KDRKGELFYMHQQMCARYDCERLSNGLNRMIPFH	66
EcaHcD	YFREDIGINS SHHWHHLLVYPT	AFYDADIFGKI	KDRKGELFYMHQQMCARYDCERLSVGLQRMIPFQ	66
EcaHcF	YFREDIGANAHHWHHLLIYPI	PTWDASVMSKV	KDRKGELFYMHQQMCARYDCDRLSTGLRRMIPFH	66
LpoHc2	YYREDVGINAHHWHHLLVYPT	STWNPKYFGKK	KDRKGELFYMHQQMCARYDCERLSNGLNRMIPFH	66
PvaHc	YFGEDIGLNTHHVTWHMEFPE	FWWNDAYG-HH	DRKGELFYMHQQMCARYDCERLSNYLDPVGELQ	65
PirHcC	YFGEDVGMNTHHVLWHMEFPE	FWWEDSSG-RH	DRKGESFFWVHHQLTVRYDAERLSNHLDPVEELS	65
PirHcA	YFGEDIGMNIHHVTWHMDFPE	FWWEDSYG-YH	DRKGELFYMHQQLTARFDFERLSNWLDPVDELH	65

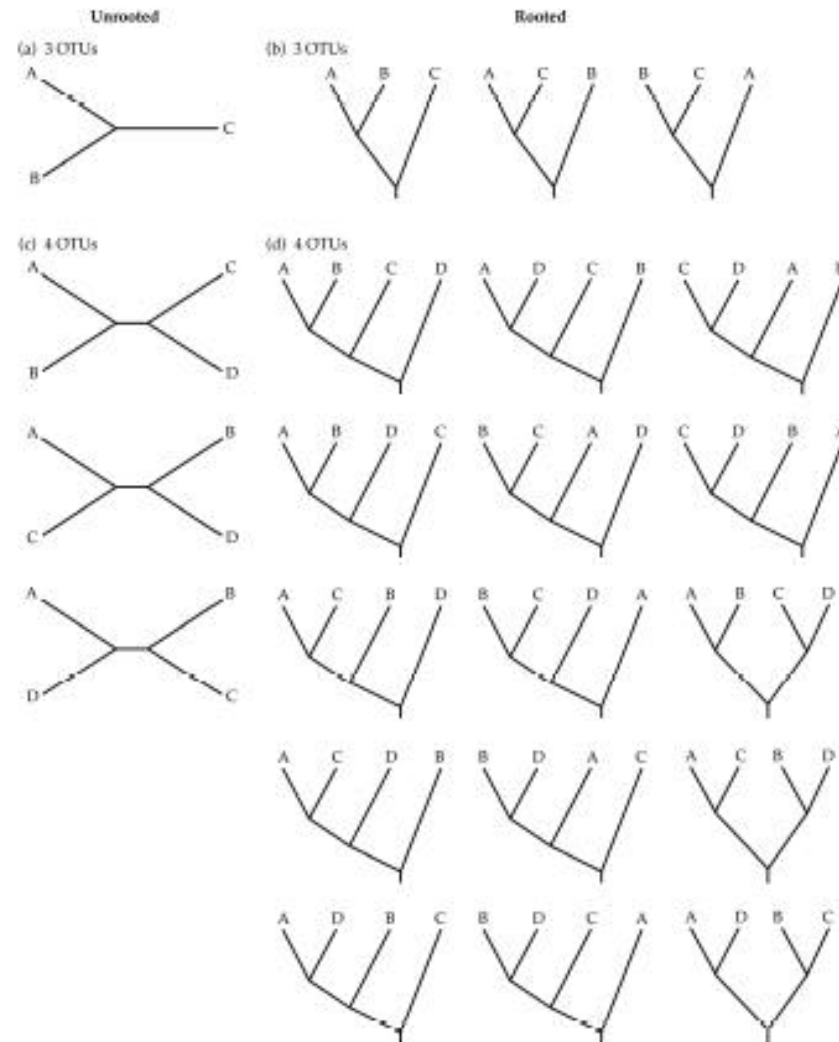
# Allgemeine Vorgehensweise...



- ➡ Multiples Sequenzalignment erstellen (**DNA oder Protein**)
- ➡ Sequenzen vergleichen > Ähnlichkeit bestimmen
- ➡ Aus Ähnlichkeitsmaß die Verwandtschaft rekonstruieren (Baum)



# Phylogenie-Rekonstruktion ist kein triviales Problem



- mit 3 Taxa kann man 1 unverwurzelten Baum erstellen, aber 3 alternative Bäume mit Wurzel

- es ist viel leichter und sicherer, einen „unrooted“ Baum zu erstellen:

**d. h. nur dann „rooten“, wenn die „Outgroup“ klar definiert ist**



# Phylogenie-Rekonstruktion ist kein triviales Problem

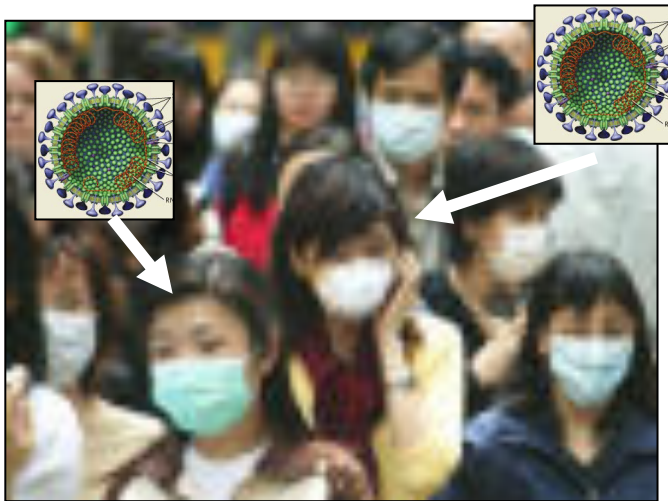
**TABLE 5.1** Possible numbers of rooted and unrooted trees up to 10 OTUs

<i>Number of OTUs</i>	<i>Number of rooted trees</i>	<i>Number of unrooted trees</i>
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

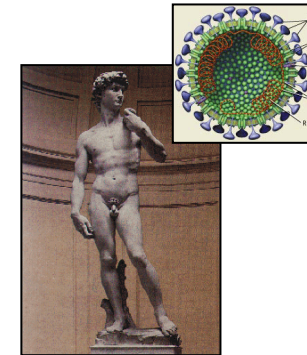
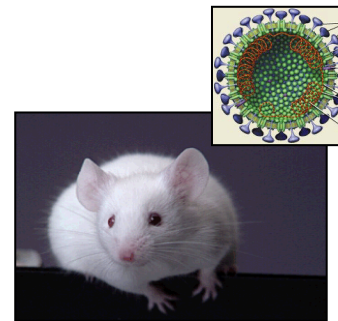
From Felsenstein (1978).

Nochmal die Frage...

# Wann DNA? Wann Protein?

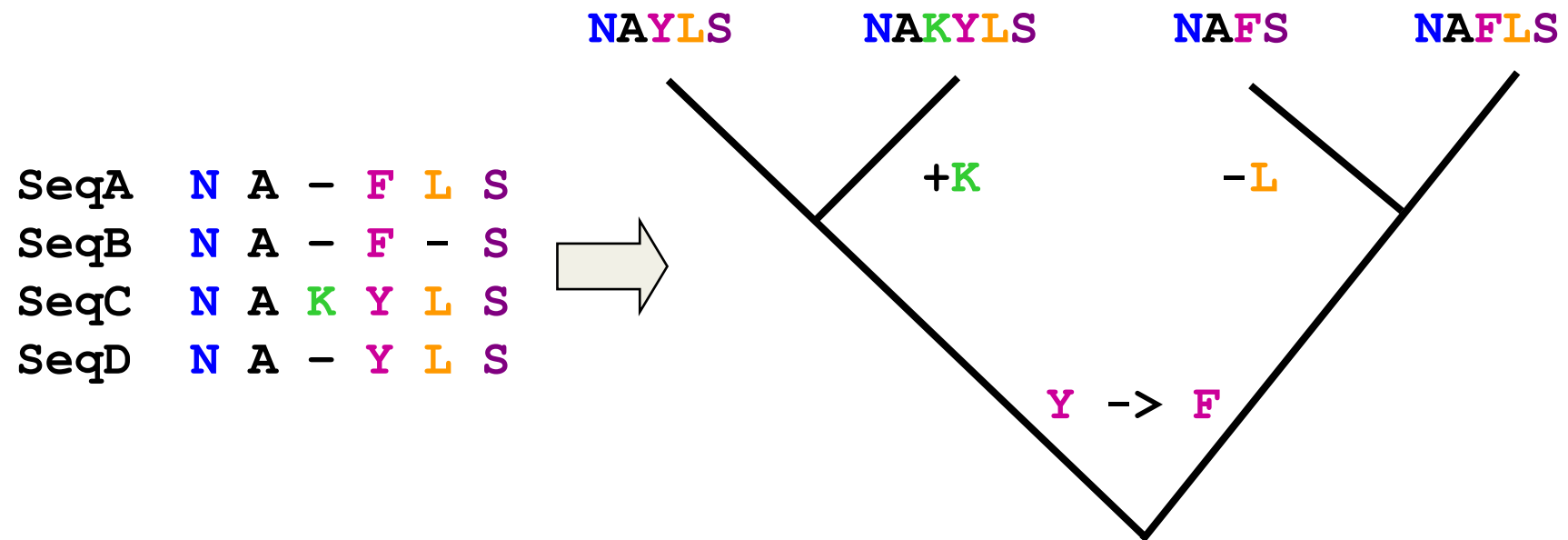


**Eng verwandte SARS-Varianten  
in der menschlichen Population**



**Corona-Virus-Gruppen  
aus verschiedenen Spezies**

# Ein Alignment ist immer eine Hypothese zur Sequenzevolution!



Wir waren ja leider nicht direkt dabei....☹

# Warum ist es nicht einfach, das „beste“ Alignment zu konstruieren?

- 2 Sequenzen à 300 Bp  
=  $10^{88}$  mögliche Alignments!!!
- Computer-Algorithmen erforderlich, die ohne Ausprobieren aller Möglichkeiten auskommen.
- „**Regelwerk**“ **notwendig**, um bestmögliches Alignment zu erkennen

# Warum ist es nicht einfach, das „beste“ Alignment zu konstruieren?

seqA TCAGACGATTG (11)  
seqB TCGGAGCTG (9)

I. TCAG-ACG-ATTG  
TC-GGA-GC-T-G

II. TCAGACGATTG  
TCGGAGCTG--

III. TCAG-ACGATTG  
TC-GGA--GCTG

Annahmen über den Ablauf  
der Sequenz-Evolution:

I. Keine mismatches

II. Keine internen Lücken

III. „Von beidem Etwas“

Aber was ist richtig?

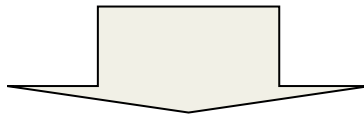
# Überlegungen...



Jede Sequenz lässt sich mit einer jeden anderen Sequenz alignen!

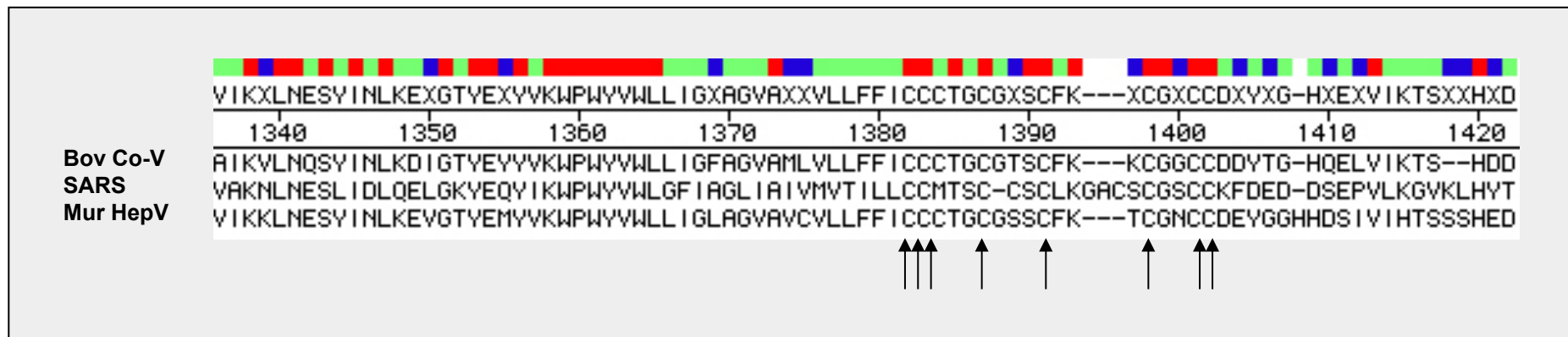
Aber macht das Alignment auch Sinn?

Also: Haben wir die richtigen Annahmen über den Verlauf der Evolution getroffen??



Wir brauchen „**evolutionäre Modelle**“, um ein möglichst richtiges Alignment zu erstellen!

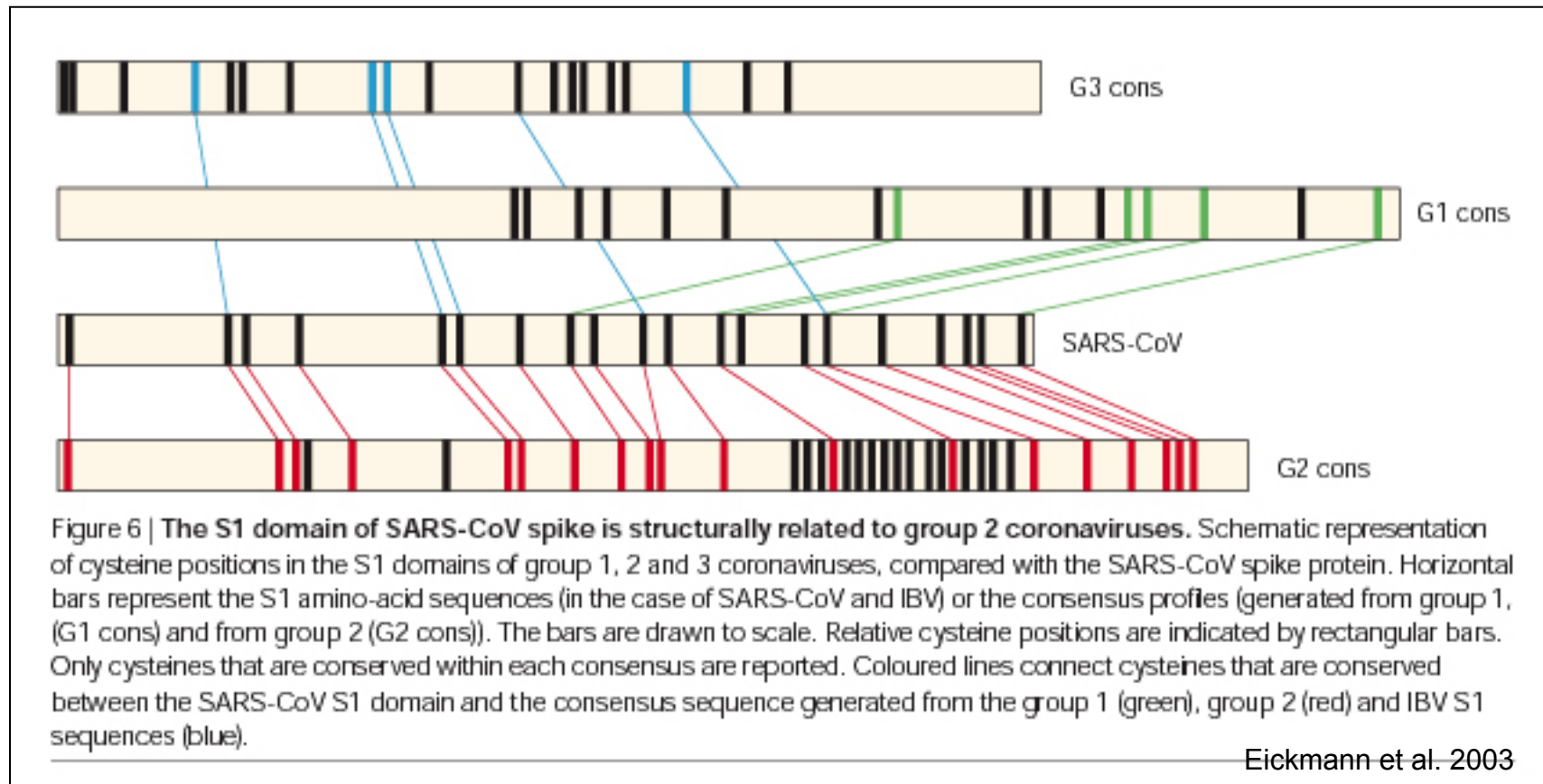
# Was bedeutet es, ein „Evolutionsmodell“ zu haben?



Ein ‚Evolutionsmodell‘ basiert auf empirischen Daten! Zum Beispiel:  
Ich weiß, die Aminosäure Cystein ist für die Proteinstruktur äußerst wichtig!

- Cysteine sind also **konserviert** während der Evolution von Proteinen!
- Cysteine können daher beim Alignment zweier Proteinsequenzen als **Ankerpunkte** dienen
- ein Alignment mit übereinstimmenden Cysteinen würde danach mit Pluspunkten **‚belohnt‘**

# SARS: konservierte Cysteine im Alignment des spike-Proteins



Resultat: Verwandtschaft von SARS zu Gruppe 2-Coronaviren?



# Ein einfacher Score-Wert zur Bewertung eines Alignments

$$S = Y - \sum W_k \times Z_k$$

$S$  = Similarity-Score („Belohnungspunkte“)

$Y$  = Anzahl an Matches

$Z_k$  = Anzahl der gaps mit Länge  $k$

$W_k$  = **gap penalty** für gaps der Länge  $k$

Mit Setzen der **gap penalty** trifft man Annahmen über die relative Häufigkeit von indel-Mutationen während der Evolution!

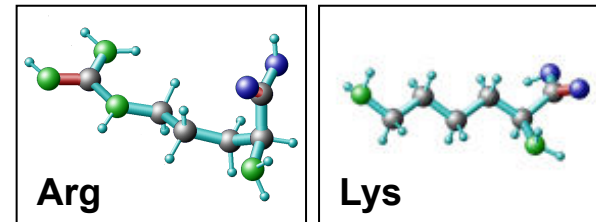
# Eine einfache Substitutionsmatrix für Nukleotidsequenzen

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

- alle Richtungen von Nt-Austauschen sind gleich wahrscheinlich
- bei jedem „match“ beider Sequenzen gibt es **1 Belohnungspunkt** für den Übereinstimmungs-Score

# Substitutions-Matrizen für Proteine

- **bei Proteinen gibt es 20 As!**

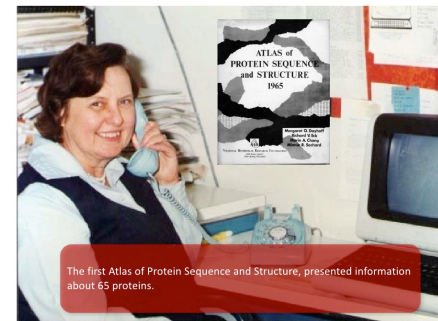


- chemisch-funktionelle Ähnlichkeit bestimmt Wahrscheinlichkeit eines Austauschs während der Evolution. Daher...
- ...sind die „Kosten“ für bestimmte Austausche (bzw. die Belohnung für gleiche As) unterschiedlich hoch!
- **Definition der „Kosten/Belohnungen“ erfolgt über Matrizen:**  
z. B. **PAM-Matrizen** (Dayhoff 1978)

# PAM-Matrizen...

	PAM-Matrix																					
Hydrophilic	C	12																				
	S	0	2																			
	T	-2	1	3																		
	P	-3	1	0	6																	
	A	-2	1	1	1	2																
	G	-3	1	0	-1	1	5															
Acid-amide	N	-4	1	0	-1	0	0	2														
	D	-5	0	0	-1	0	1	2	4													
	E	-5	0	0	-1	0	0	1	3	4												
	Q	-5	-1	-1	0	0	-1	1	2	2	4											
Basic	H	-3	-1	-1	0	-1	-2	2	1	1	3	6										
	R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6									
	K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5								
Hydro-phobic	M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6							
	I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	2	5							
	L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6					
	V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4				
Aromatic	F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9			
	Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10		
	W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17	
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W		

**Fig. 5.7** The PAM 250 matrix. For each pair of amino acids (see Table 3.1, p. 41, for key to the one-letter codes for amino acids) the matrix gives the ratio of the frequency at which the pair is observed in pairwise comparisons of proteins to that are expected due to chance alone, expressed as a ‘log odd’. Amino acids that regularly replace each other have a positive score, amino acids that rarely replace each other have negative scores. Note that replacements more often occur among chemically related amino acids (indicated on the left). From Dayhoff (1978: Fig. 84).



Margaret Dayhoff

...definieren ‚Belohnungswerte‘  
für zwei Aminosäuren, die sich  
in einem Alignment gegenüber-  
stehen:

- positiver Wert = Aminosäuren, die sich häufig in Alignments gegenüberstehen und somit ‚funktionell konserviert‘ sind

z.B. W-W	17
C -C	12

aber W-V - 6

**Wir haben also Kriterien (Substitutionsmatrizen, gap penalties), um Alignments zu bewerten.**

**Aber wie werden Alignments überhaupt erstellt?**



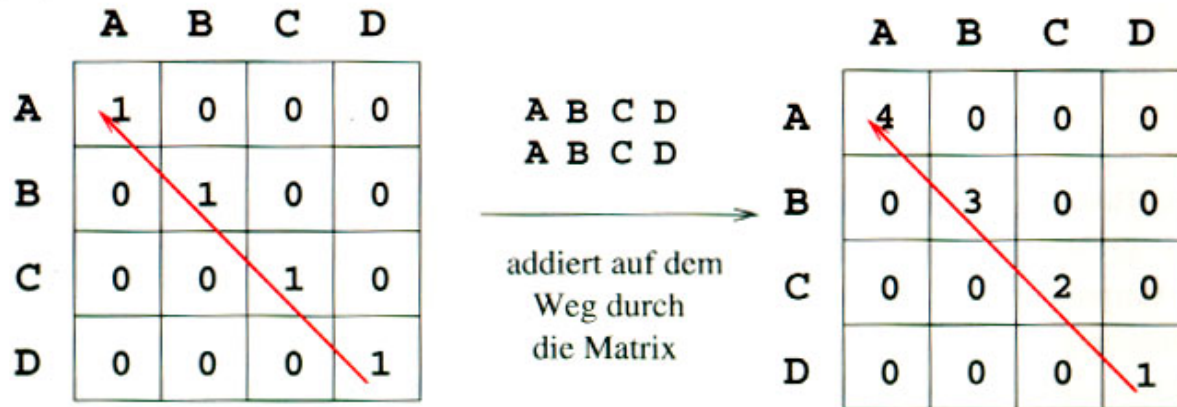
# Needleman-Wunsch (N-W) 1970

- Bei Erstellung des Alignments werden zunächst kleine Problem-Schritte gelöst. Dann wird aus den Teillösungen das Gesamtalignment rekonstruiert...
- Algorithmus: „dynamic programming“

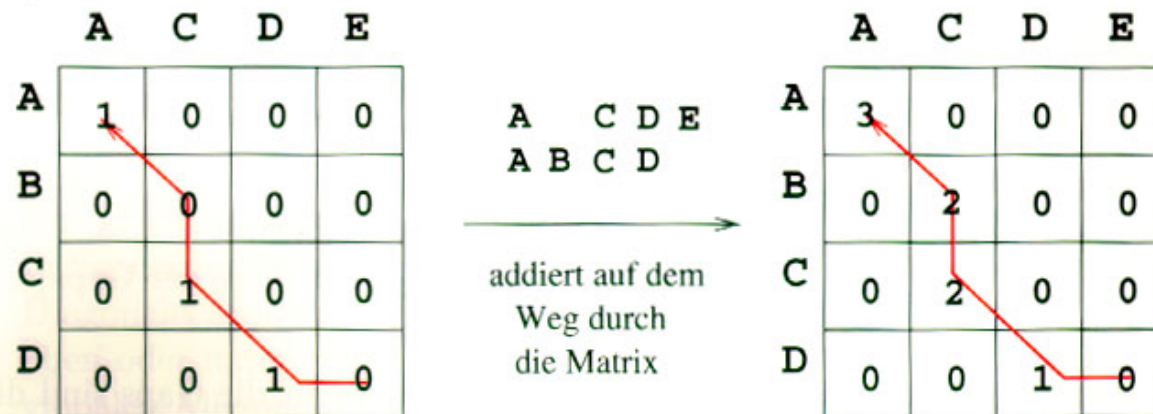


# Needleman-Wunsch

(1)



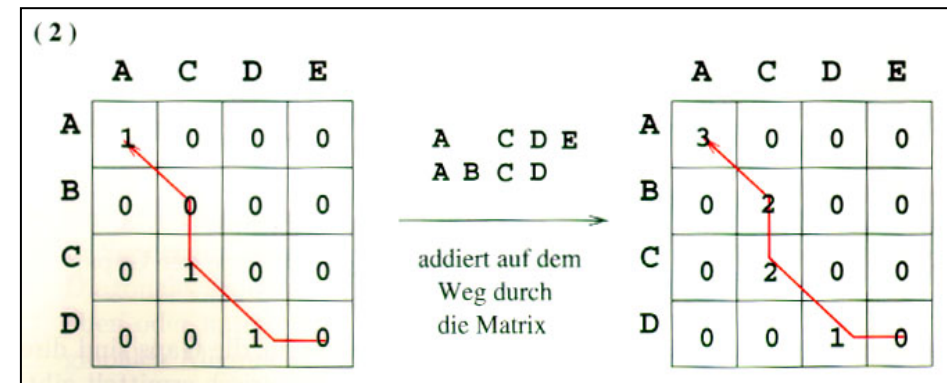
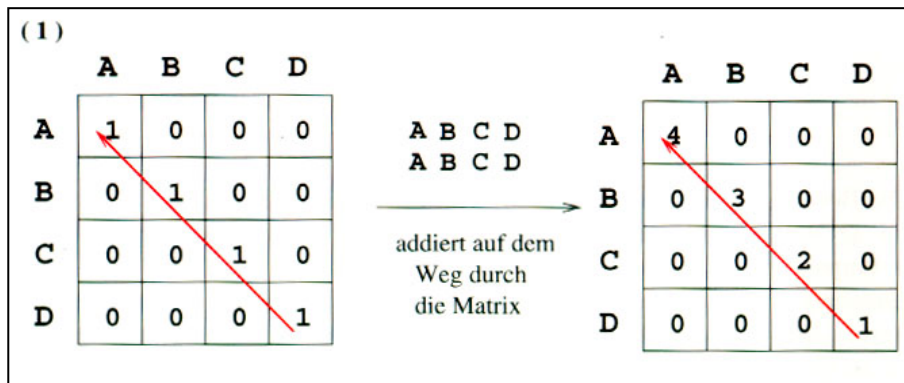
(2)



# Needleman-Wunsch



- es wird zunächst eine zweidimensionale Matrix mit den beiden zu vergleichenden Sequenzen erstellt
- in die Zellen der Matrix wird der Alignment-Score für die jeweils verglichenen Sequenzpositionen hineingeschrieben. Die Berechnung des Score-Werts erfolgt natürlich anhand einer Substitutionsmatrize.
- das Alignment ergibt sich als Pfad durch die Matrix. Der Pfad mit der höchsten Endsumme gewinnt...





N-W ist eine exakte Vorgehensweise und viel zu aufwändig für multiple Sequenzvergleiche!

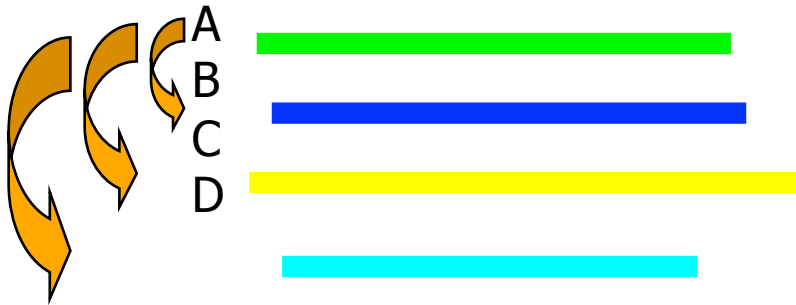
Wir brauchen....

**Heuristik** ([altgr.](#) εὕρισκω *heurísko* „ich finde“; von εὕρισκειν *heurískein* ‚auffinden‘, ‚entdecken‘) bezeichnet die Kunst, mit begrenztem Wissen ([unvollständigen Informationen](#)) und wenig Zeit dennoch zu wahrscheinlichen Aussagen oder praktikablen Lösungen zu kommen.<sup>[1]</sup> Es bezeichnet ein analytisches Vorgehen, bei dem mit begrenztem Wissen über ein System mit Hilfe [mutmaßender Schlussfolgerungen](#) Aussagen über das System getroffen werden. Die damit gefolgerten Aussagen können von der optimalen Lösung abweichen.

[www.wikipedia.de](http://www.wikipedia.de)

# Progressives Alignment

## 1) Sequenzvergleich



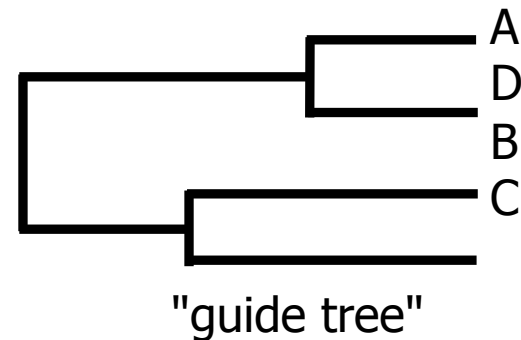
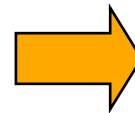
Alle Sequenzen werden miteinander verglichen  
(Option A: schnelles "quick and dirty" Alignment  
Option B: exaktes, langsames Needleman-Wunsch)  
=> Berechnen der Distanzen

# Progressives Alignment

2) Ähnliche Sequenzen werden gruppiert

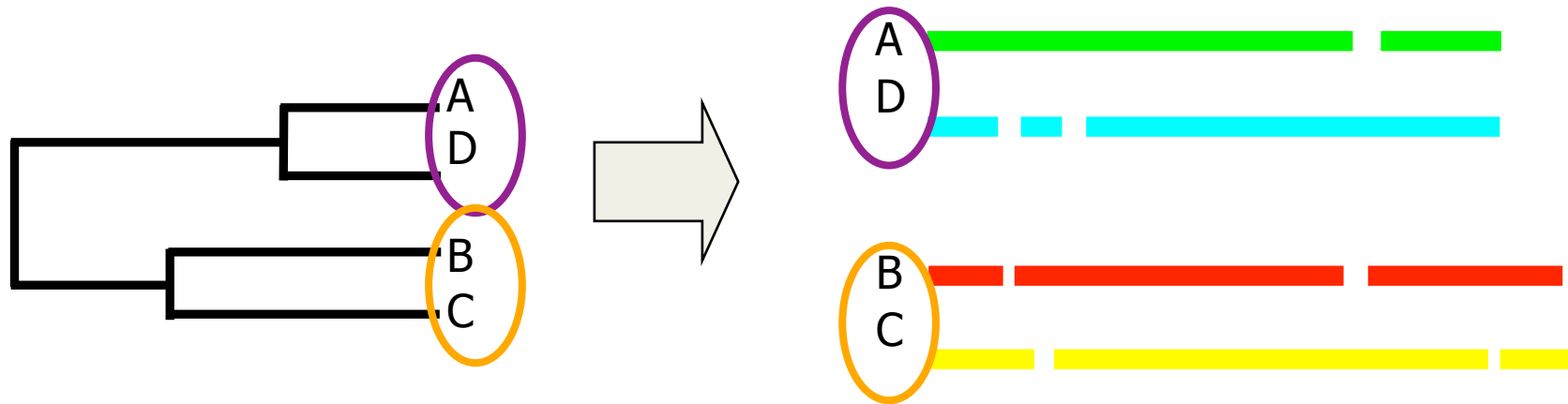
=> Cluster-Analyse = Erstellung eines hierarchischen Stammbaums ("guide tree").

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



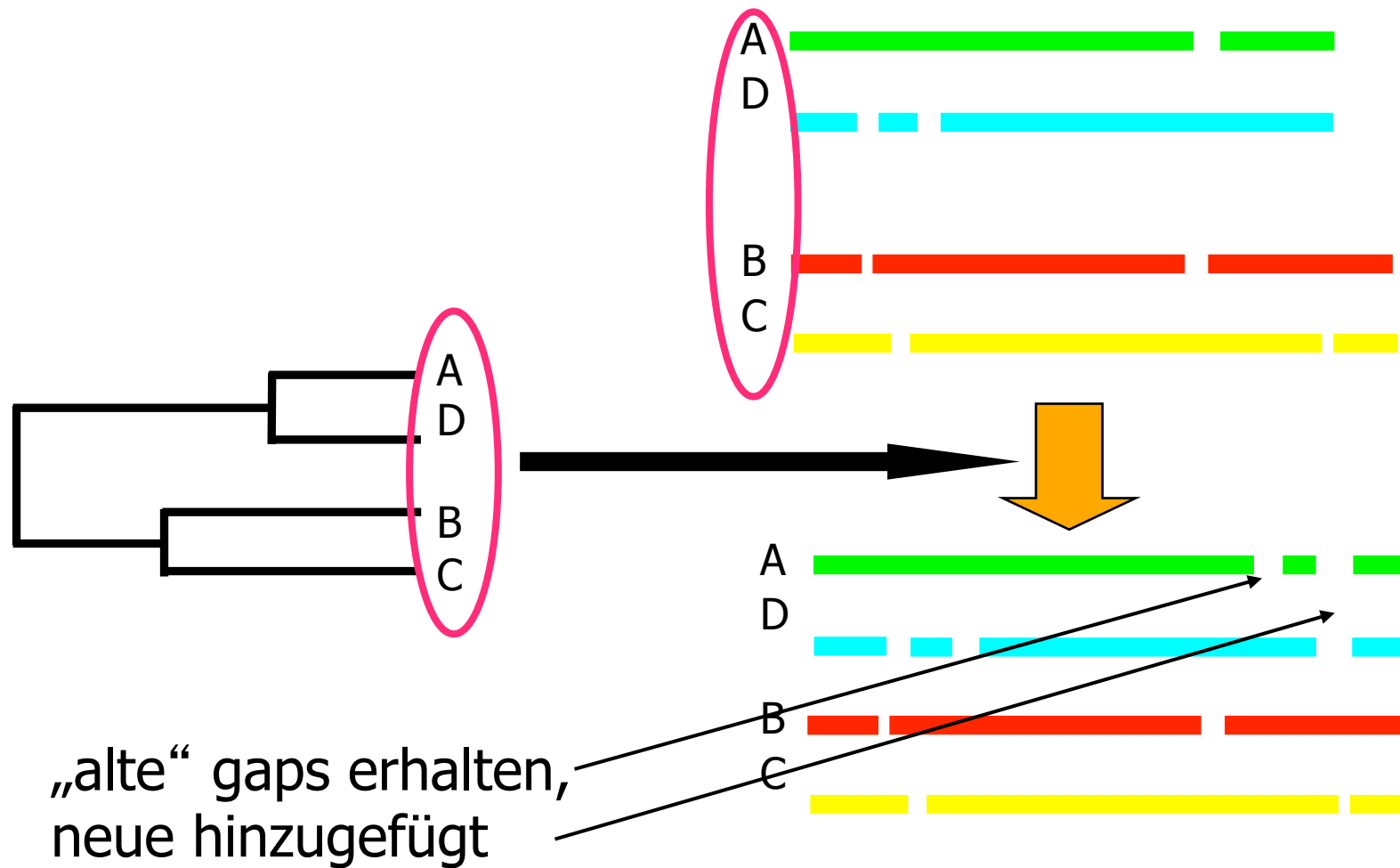
# Progressives Alignment

3) Alignment von nahe verwandten Sequenzen; die ähnlichsten zuerst.



# Progressives Alignment

## 4) Sukzessives globales Alignment



# Progressives Alignment



1. paarweiser Vergleich aller Sequenzen miteinander =>  
Berechnung der Distanzen zweier Sequenzen
2. gruppiert Sequenzen nach Ähnlichkeit (Cluster-Bildung)
3. Erstellung paarweiser Alignments
4. sukzessives Alignment nach Ähnlichkeit,  
dabei die ähnlichsten Sequenzpaare zuerst

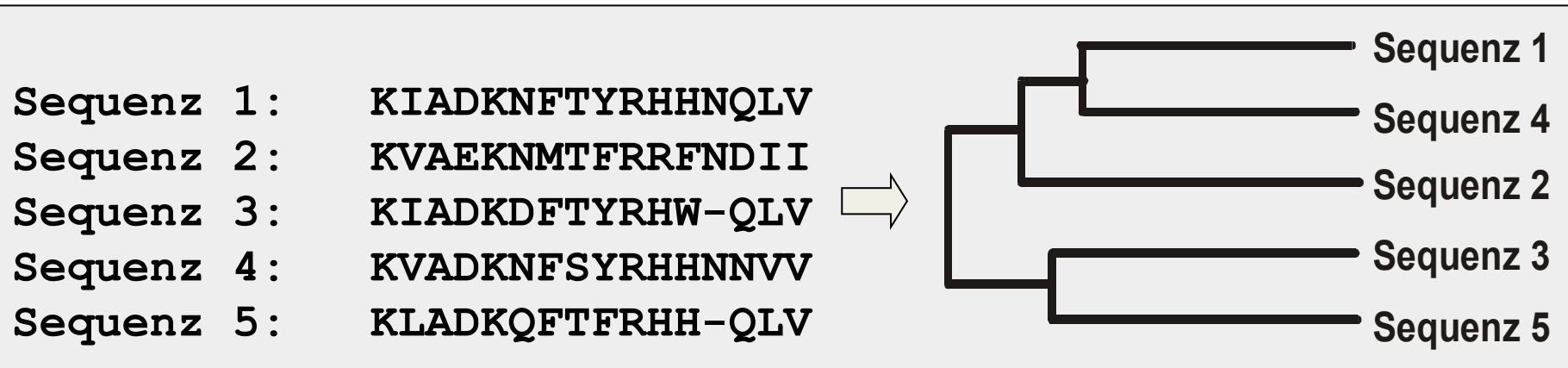
**Feng & Doolittle (1987): [PileUp](#) (GCG package)**

**Higgins and Sharp (1988), Thompson et al. (1994): [CLUSTAL](#)**

**Notredame et al. (2000): [T-COFFEE](#)**

Zur Erinnerung:

# Allgemeine Vorgehensweise...



- ➡ Multiples Sequenzalignment erstellen (**DNA oder Protein**)
- ➡ Sequenzen vergleichen > Ähnlichkeit bestimmen
- ➡ Aus Ähnlichkeitsmaß die Verwandtschaft rekonstruieren (Baum)

# Vom Alignment zu einem einfachen Baum-Rekonstruktionsverfahren...

Aus dem Alignment ergibt sich zunächst, wie ähnlich oder unähnlich die Sequenzen zueinander sind.

Meist wird eine **Distanzmatrix** erstellt:

		A	B	C	D
OTU <sup>*</sup>	A	0	6	10	18
	B		0	12	20
	C			0	19
	D				0

\* OTU = operational taxonomic unit: z. B. Spezies, Gen, Protein



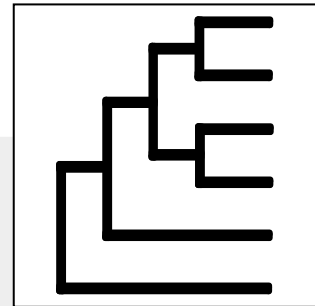
# Vom Alignment zu einem einfachen Baum-Rekonstruktionsverfahren...

Sokal and Michener 1967!

UPGMA

=

Unweighted Pair-Group Method using  
Arithmetic Means



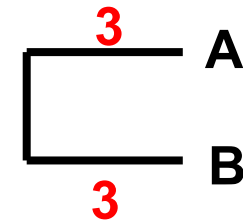
...eine Methode, die auf der Berechnung von Unähnlichkeiten  
(Distanzen) der alignierten Sequenzen beruht („Distanz-Methode“)

# UPGMA

## Ausgangs-Distanz-Matrix

1.

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

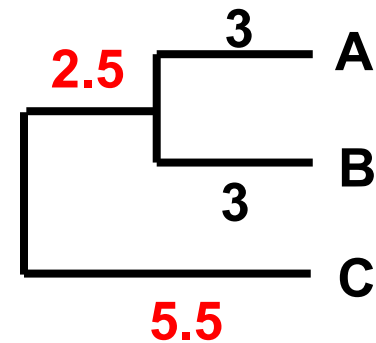


## Neu berechnete Distanz-Matrix

2.

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

$$(AC + BC) / 2$$

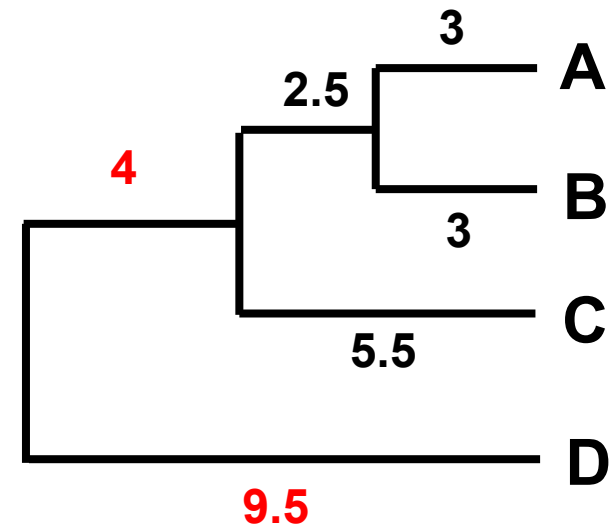


# UPGMA

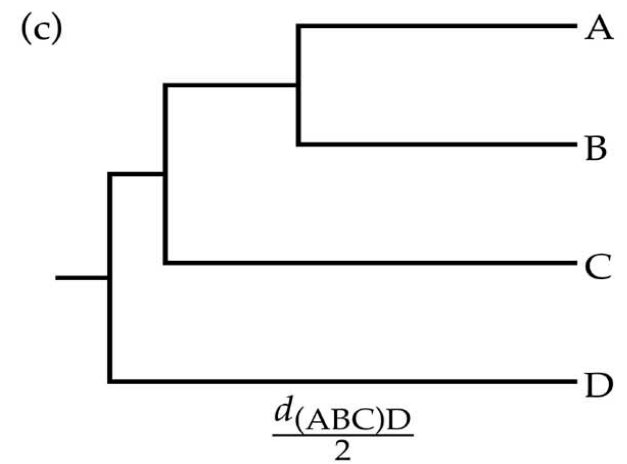
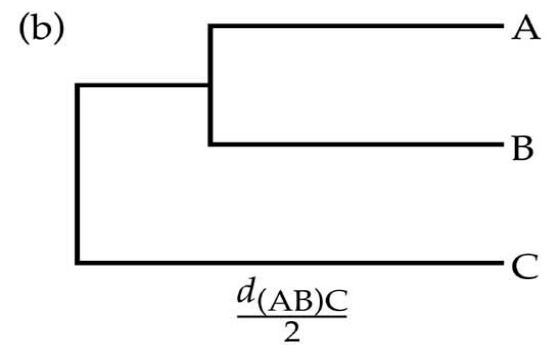
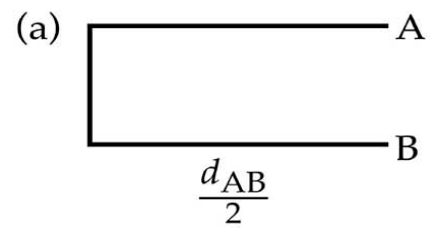
## Neu berechnete Distanz-Matrix

3.

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



# UPGMA



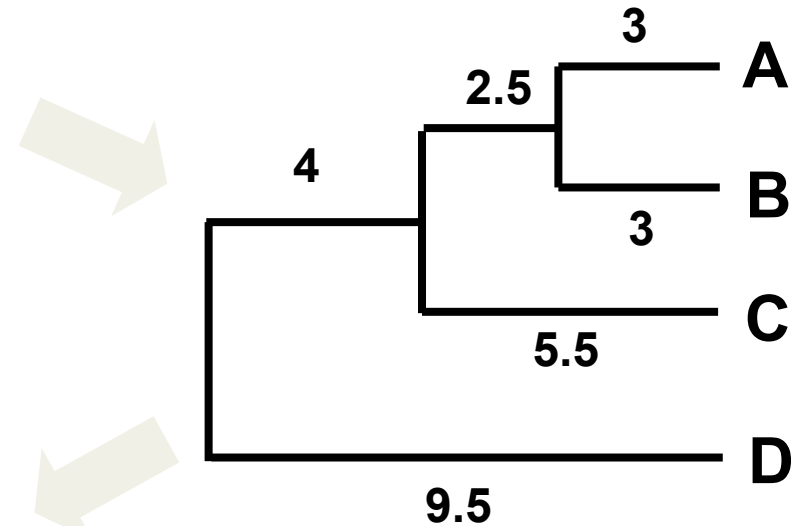
# UPGMA-Problem

Ausgangsmatrix

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

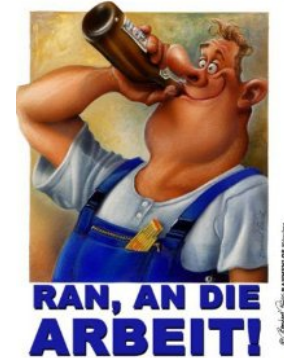
rekonstruierte Matrix

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



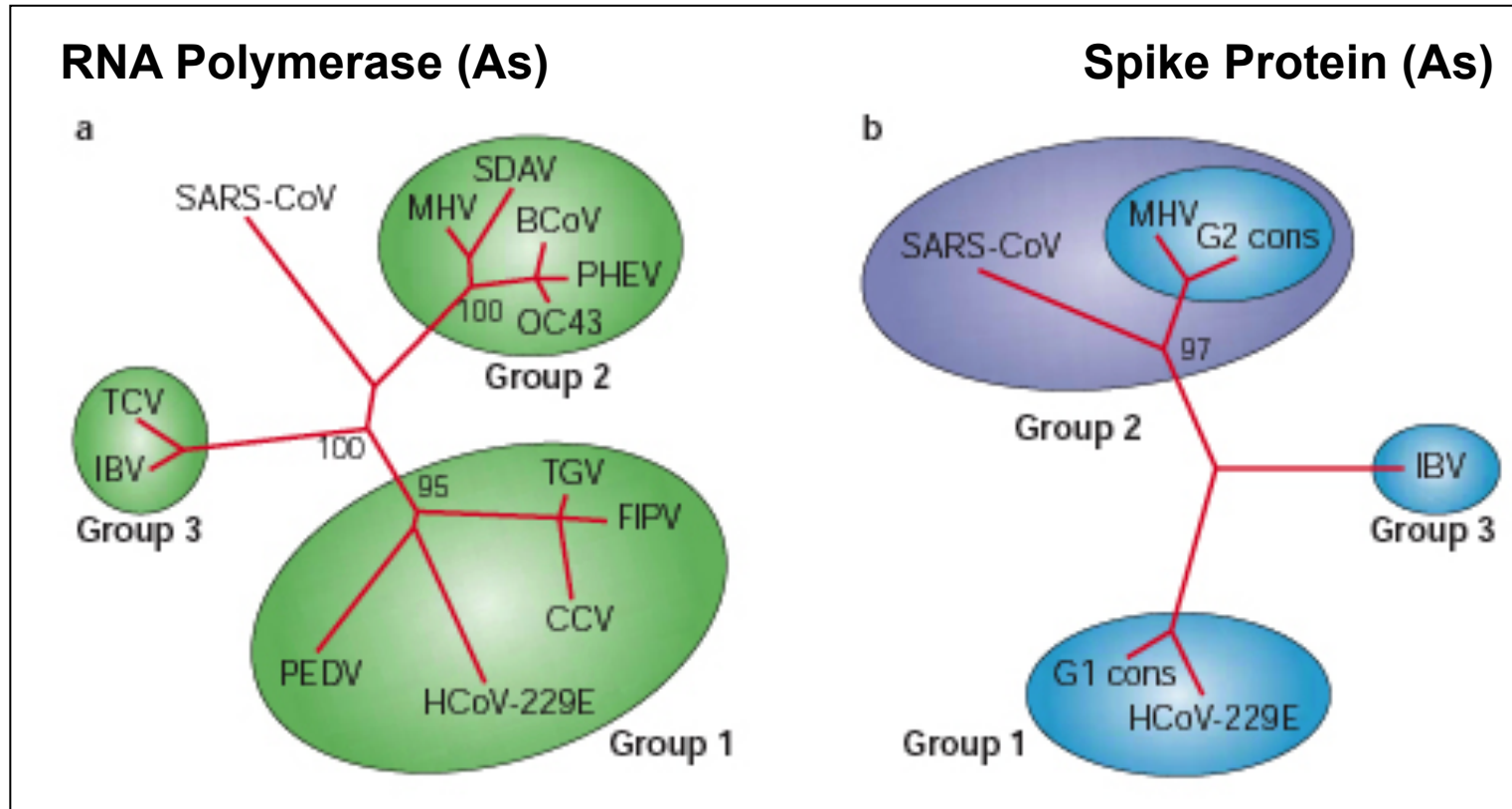
Astlängen des UPGMA-Baums passen nicht exakt zur Realität...

# Practical Exercises 4



- Build a phylogenetic tree manually using UPGMA
- Produce “multiple” nucleotide and amino acid sequence alignment of SARS genes and proteins
- Where did SARS originate and how did it spread in the population?
  - > Prepare a nt-based tree from different SARS genomes
- Which are the distant relatives of SARS coronavirus?
  - > Prepare an aa-based tree for the spike protein

# SARS-Verwandschaft



Unterschiedliche Datensätze und Rekonstruktionsmethoden können leicht unterschiedliche Baum-Topologien ergeben!!

**Aber: SARS Co-V ist alter Verwandter der Gruppe 2 Coronaviren**

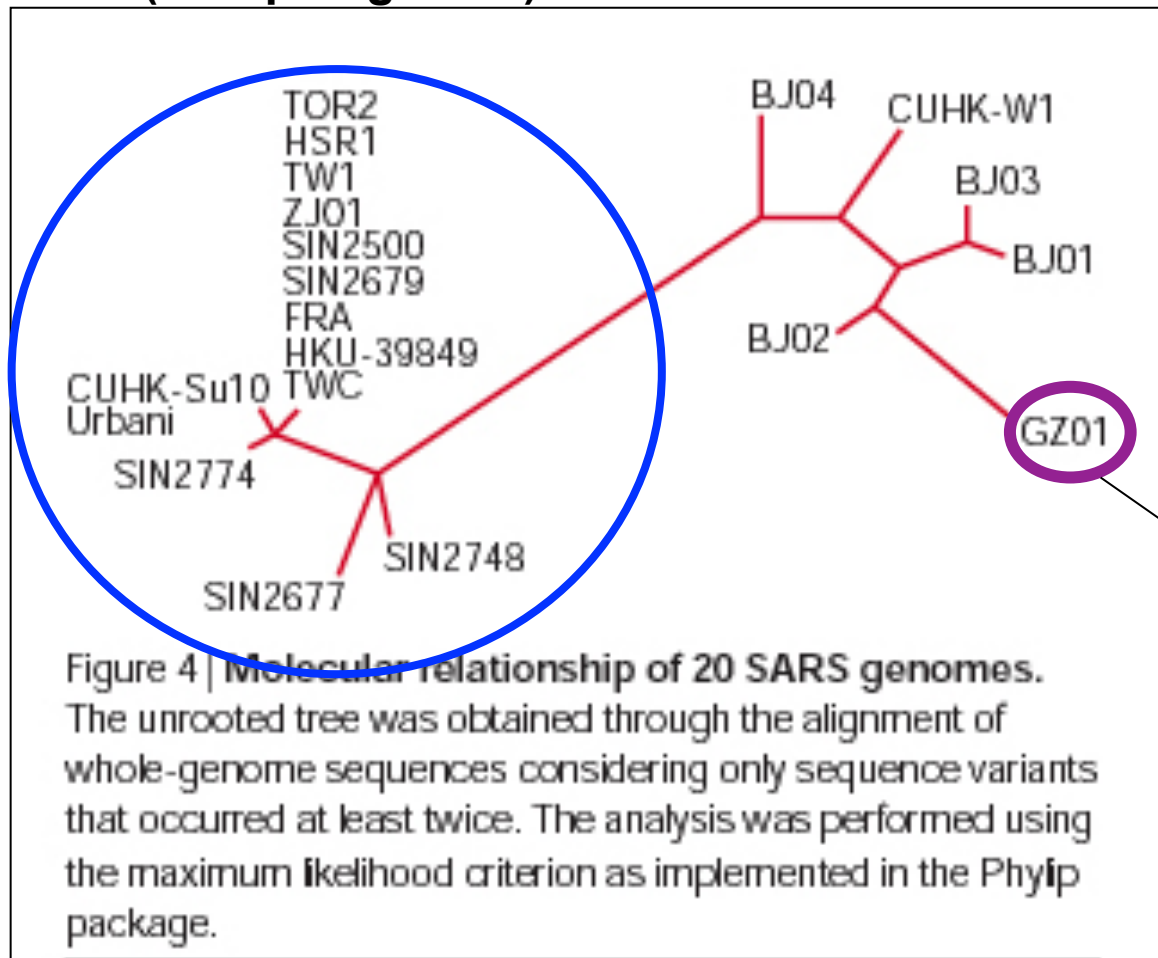
# SARS- Phylogenie: Ausbreitung





# SARS-Phylogenie: Herkunft?

## DNA (Komplettgenom)



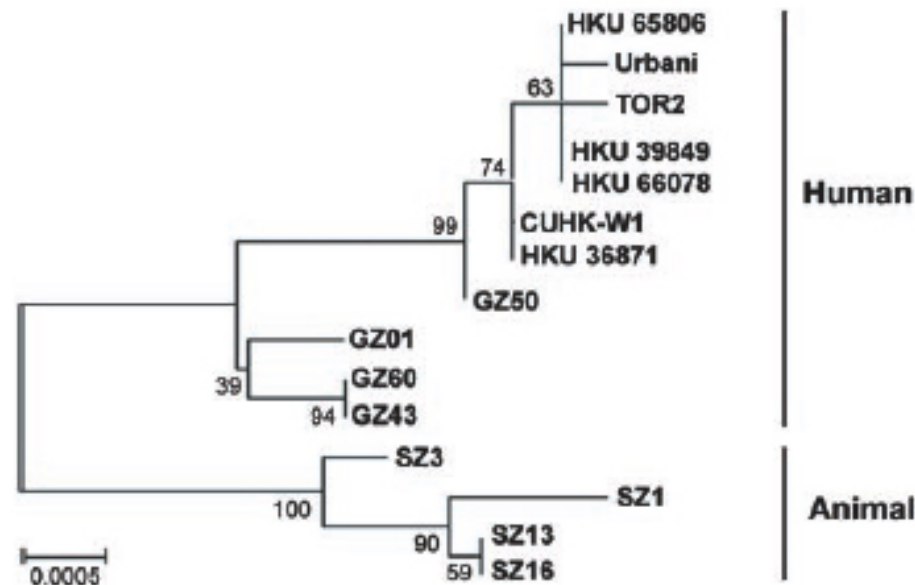
- Varianten sind >99% identisch. Dennoch ist eine geographische Zuordnung möglich.

**Sequenz zeigt Besonderheit:**

Sein Spike-Gen hat Insertion von 29 Bp, die sonst nur in tierischen SARS-Verwandten gefunden worden ist!

# SARS-Phylogenie: Herkunft?

**Fig. 2.** Phylogenetic analysis of the nucleotide acid sequence of the spike gene of SCoV-like viruses. Nucleotide sequences of representative SCoV S genes (S gene coding region 21477 to 25244, 3768 bp) were analyzed. The phylogenetic tree was constructed by the neighbor-joining method with bootstrap analysis (1000 replicates) using MEGA 2 (10). Number at the nodes indicates bootstrap values in percentage. The scale bar shows genetic distance estimated using Kimura's two-parameter substitution model (11). In addition to viruses sequenced in the present study, the other sequences used in the analysis could be found in GenBank with accession number: from AY304490 to AY304495, AY278741, AY278554, AY278491, AY274119, and AY278489.



**Table 2.** Prevalence of antibody to animal SCoV SZ16 in humans. Controls are serum specimens from patients hospitalized for nonrespiratory diseases in Guangdong made anonymous.

Occupation	Sample numbers	Antibody positive (%)
Wild-animal trader	20	8 (40)
Slaughterer of animals	15	3 (20)
Vegetable trader	20	1 (5)
Control	60	0 (0)

# SARS: zoonotischer Ursprung



Larvenroller - palm civet  
(*Paguma larvata*)



Marderhund - Raccoon dog  
(*Nyctereutes procyonoides*)





Ob diese unschuldig wirkenden Säugetiere für die Übertragung des tödlichen Coronavirus verantwortlich sind?

(Foto: AP)

Donnerstag, 22. August 2013

## **Genetisch 100-prozentige Übereinstimmung mit MERS Übertragen Fledermäuse das Coronavirus?**

**Ein Forscherteam entdeckt in Fledermauskot ein mysteriöses Virus, das dem Coronavirus Mers vollständig gleicht. Ob die Fledermäuse bei der Übertragung des Virus auf den Menschen eine Rolle spielen, ist noch unklar - die exakte Übereinstimmung jedoch eine Sensation.**

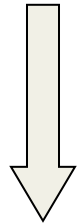
# Das Szenario ...ein neues tödliches Virus!

- Labor: Isolierung der Erbsubstanz und Sequenzierung



- Computer: Erkennen der Virusgene (*de novo* Genvorhersage)

Ähnlichkeit zu bekannten Genen? (Datenbanksuchen)



Verwandtschaft? Ausbreitung? Herkunft?  
(Phylogenetische Rekonstruktion)

Struktur der Proteine?  
(Struktur-Vorhersage, -Modellierung)  
Wirkstoff-Design

- Labor: Wirkstoff-Test

