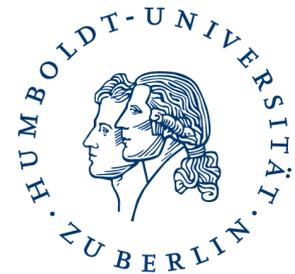


Algorithmische Bioinformatik

Einführung in die Phylogenie
(lat.: phylum = Stamm)



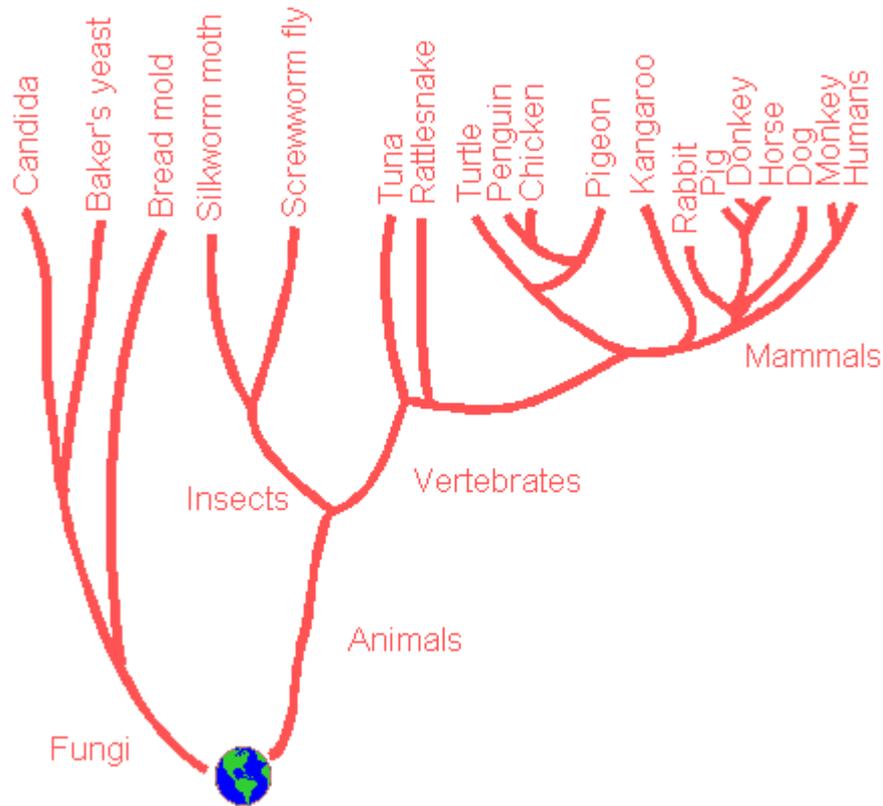
Ulf Leser
Wissensmanagement in der
Bioinformatik



Inhalt dieser Vorlesung

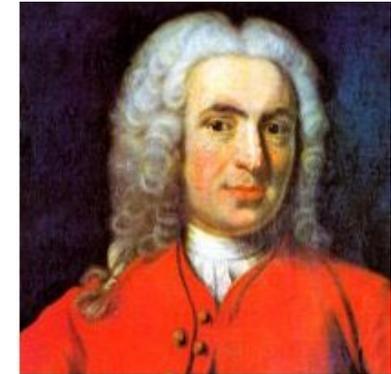
- Stammbäume
- Phylogenetische Bäume
- Evolutionsmodell

Tree of Life



Vererbung versus Klassifikation

- Zuerst war die **Klassifikation (Phänotyp)**
 - Carl Linnaeus, ca. 1740: *Systema Naturae*
 - Annahme: Arten verändern sich nicht
 - Hierarchische Einteilung aller Lebewesen
 - Kingdom, class, order, family, genera, species
 - Stamm, Klasse, Ordnung, Familie, Gattung, Art
 - Innere Knoten einer Klassifikation sind **abstrakte Klassen**
 - Definiert über gemeinsame, **manuell definierte Merkmale**
 - Haben keine eigenen Vertreter
- Auch: **Taxonomie**

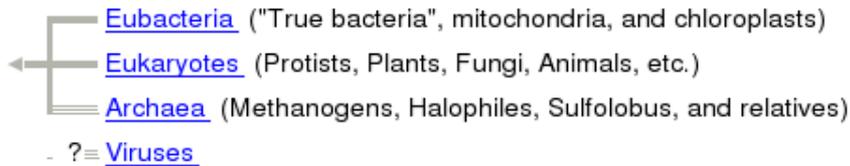


'Bridegroom portrait' of Linnaeus,
Clasif. LINNÆI, M. D.
METHODUS plantarum SEXUALIS
in SISTEMATE NATURÆ
detripa



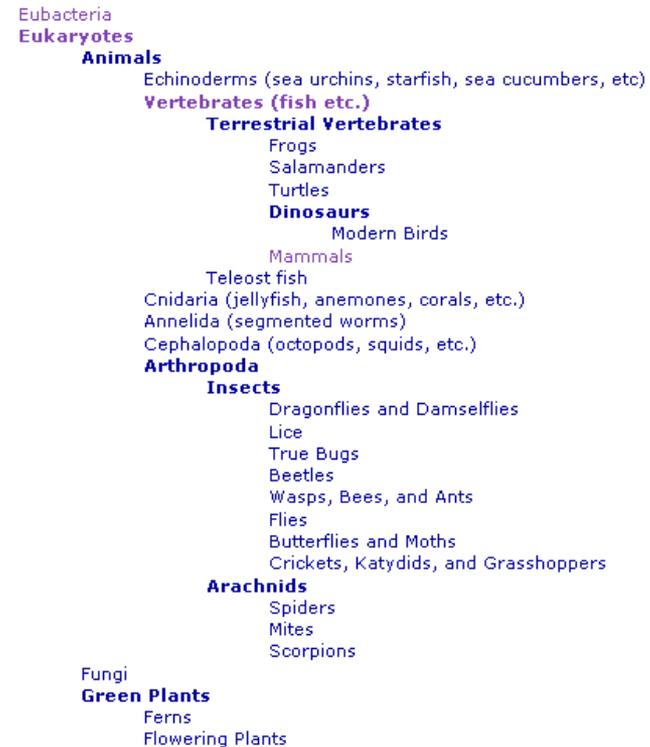
Linnaean sexual system
Illustration by Georg Dionysius Ehret of
Linnaeus' sexual system (1736)

Klassifikation



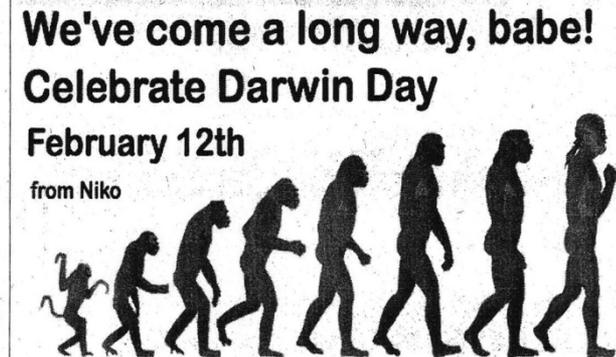
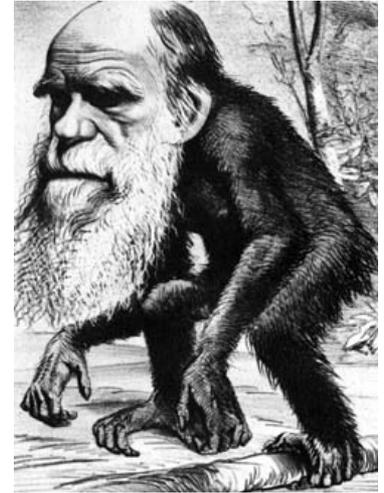
- Eukaryoten
- Tiere
- diverse Zwischenstufen
- Charniata (Schädelknochen)
- Vertebraten (Wirbeltier)
- Viele Zwischenstufen
- Mammals (Säugetiere)
- Eutheria (Placenta)
- Primaten (Affen)
- Catarrhini
- Hominidae (Mensch, Schimpanse, Orang-Utan, Gorilla)
- Homo (erectus, sapiens ...)
- Homo Sapiens

Popular Groups on the Tree of Life

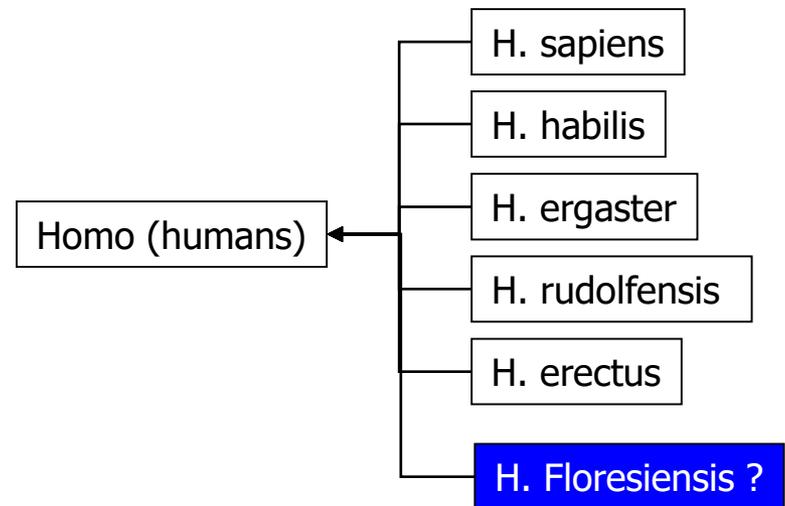


Vererbung – Stammbäume

- C. Darwin: „The origin of species“ 1859
 - Arten unterliegen **Wandel** mit der Zeit
 - „Survival of the fittest“
 - Lange war unklar, was sich da wie wandelt
- **Speziesstammbäume** (Genotyp)
 - Ergeben sich aus Annahme der Evolution
 - Wurden lange aus morphologischen Eigenschaften angenähert
 - Jeder innere Knoten (taxa) hat **als Art wirklich existiert**
- Was definiert eine Spezies?



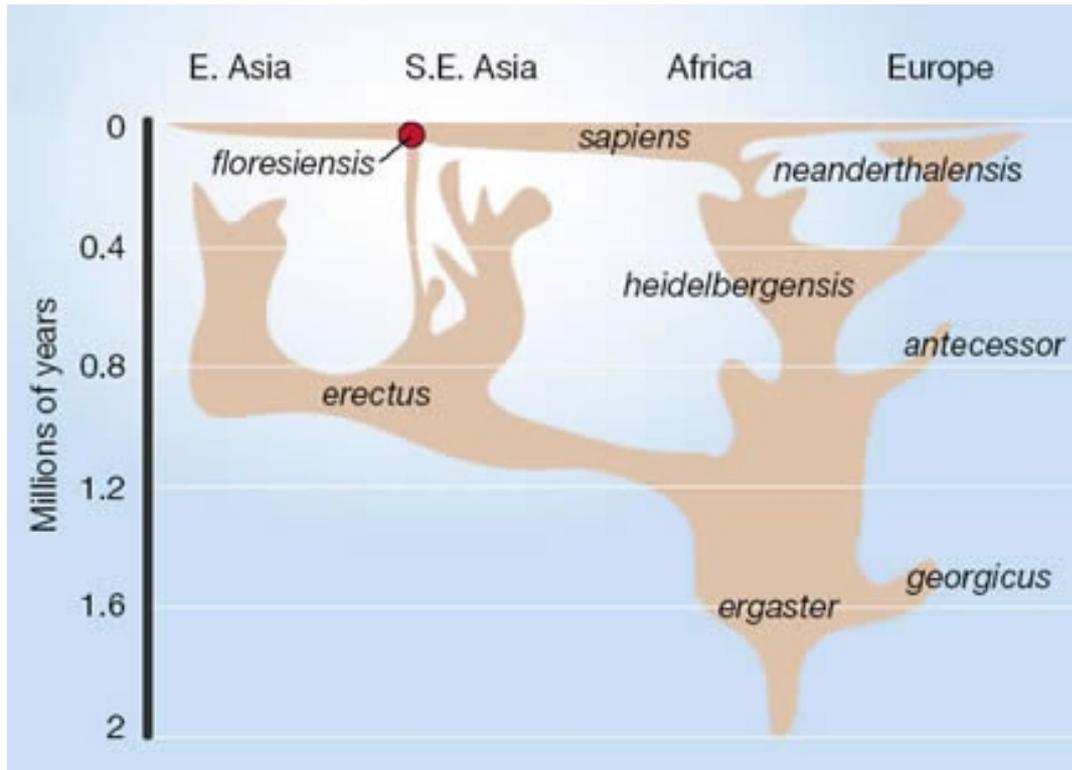
Homo floresiensis („Hobbit“)



- Entdeckung 2003
- Lebte ~95 000-12 000 v.C. auf Flores (Indonesien)
- Körperhöhe ca. 100 cm, Gehirnvolumen 380 cm³
 - Homo erectus: 600-1200, h. sapiens: 1400
- Rückentwicklung? Vorläufer?

Quelle:
Brown, P. *et al.* *Nature* 431, 1055-1061 (2004).

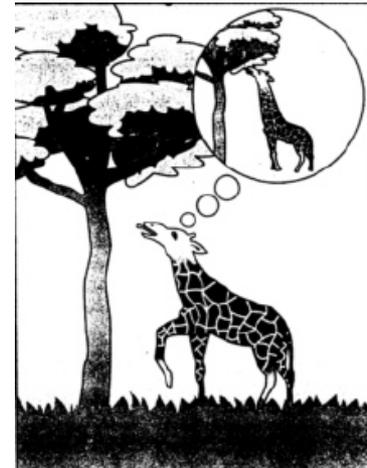
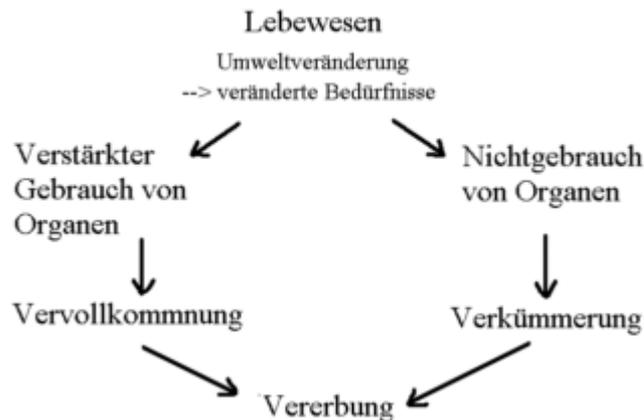
Evolutionsgeschichte neu geschrieben



- Stammbaum – welche „Spezies“ geht aus welcher hervor?

Theorie I: Lamarck

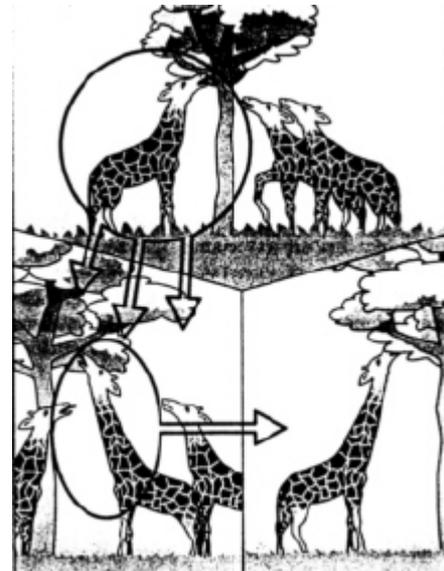
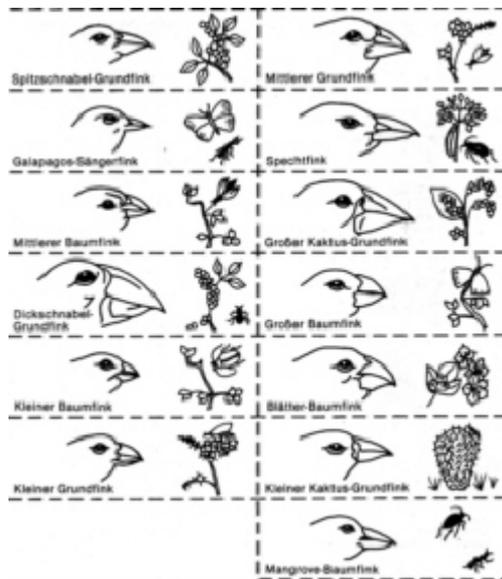
- Jean Baptiste de Lamarck (1744-1829)



- Häufiger Gebrauch von Organen führt zu ihrer Veränderung, die dann weitervererbt wird (nur wie?)
 - Z.B.: Hals der Giraffe
- Nicht gebrauchte Organe verkümmern
 - Z.B.: Augen des Maulwurf

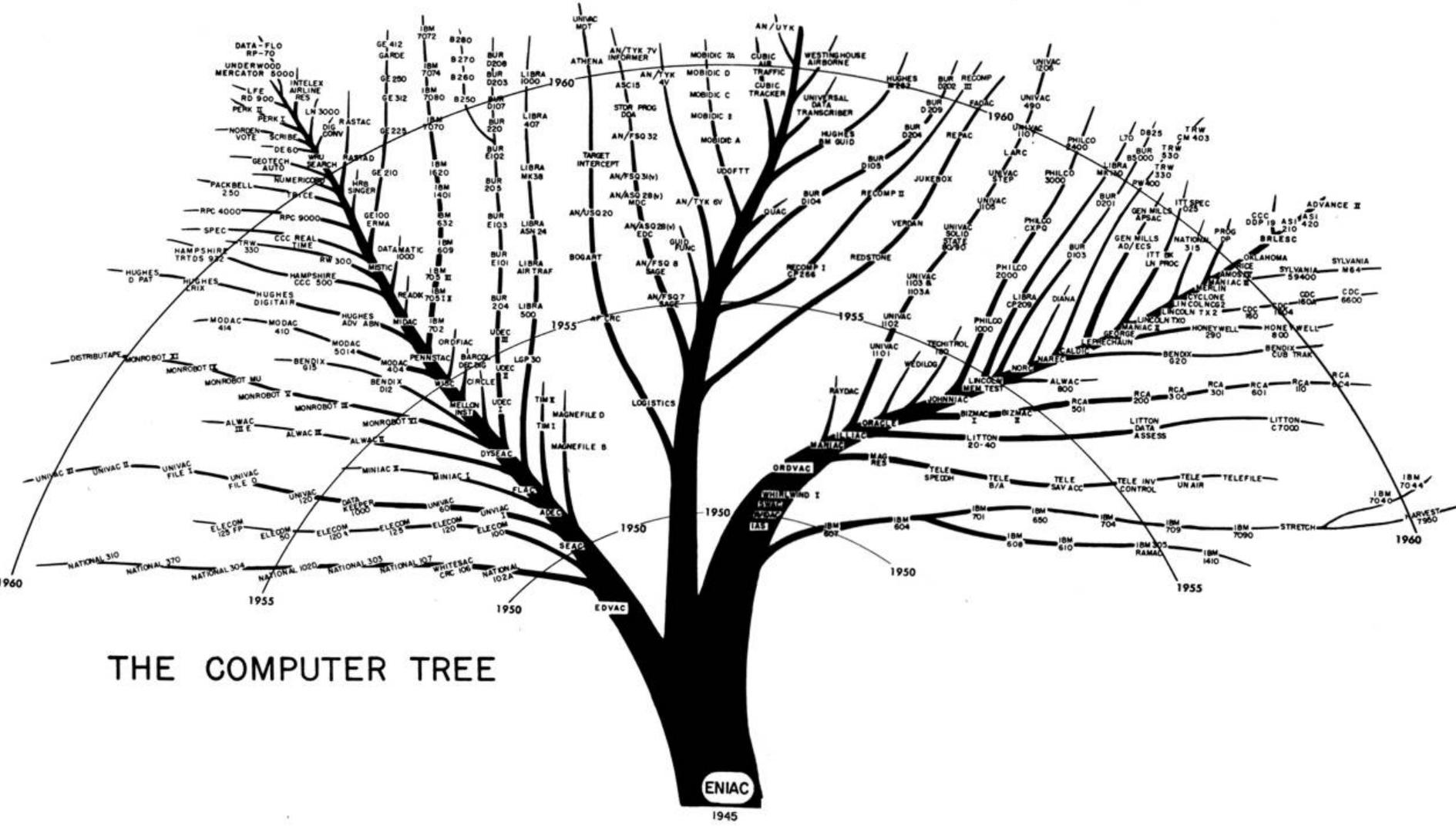
Theorie II: Darwin

- Charles Robert Darwin (1809-1882)
 - „Die merkwürdigste Tatsache ist die **vollkommene Abstufung** in der Größe des Schnabels ..., von einem Schnabel, der so groß ist wie der eines Kernbeißers bis zu dem eines Buchfinken“



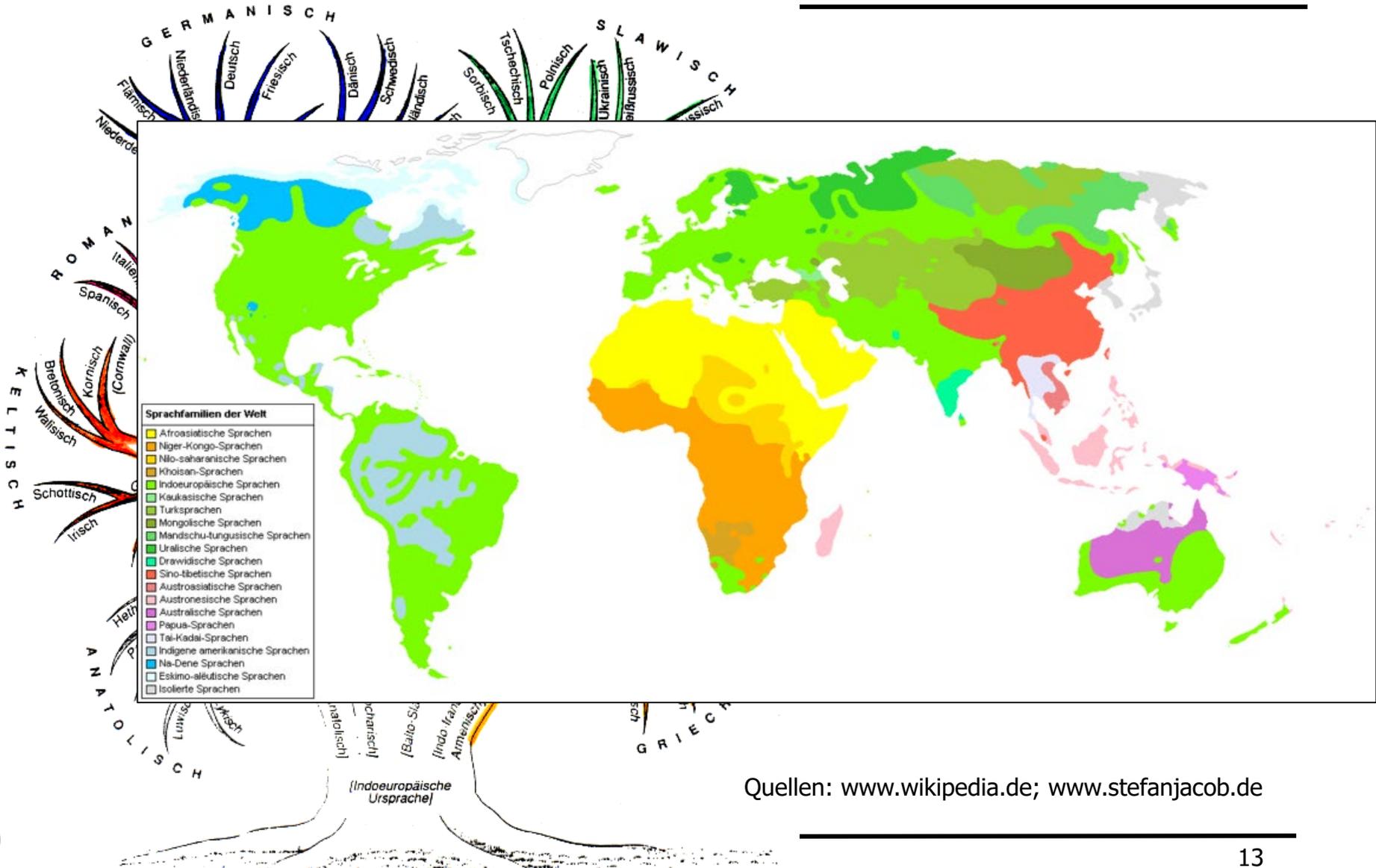
- Erklärung: **Zufällige Veränderungen** und **natürliche Auslese**

The Computer Tree



THE COMPUTER TREE

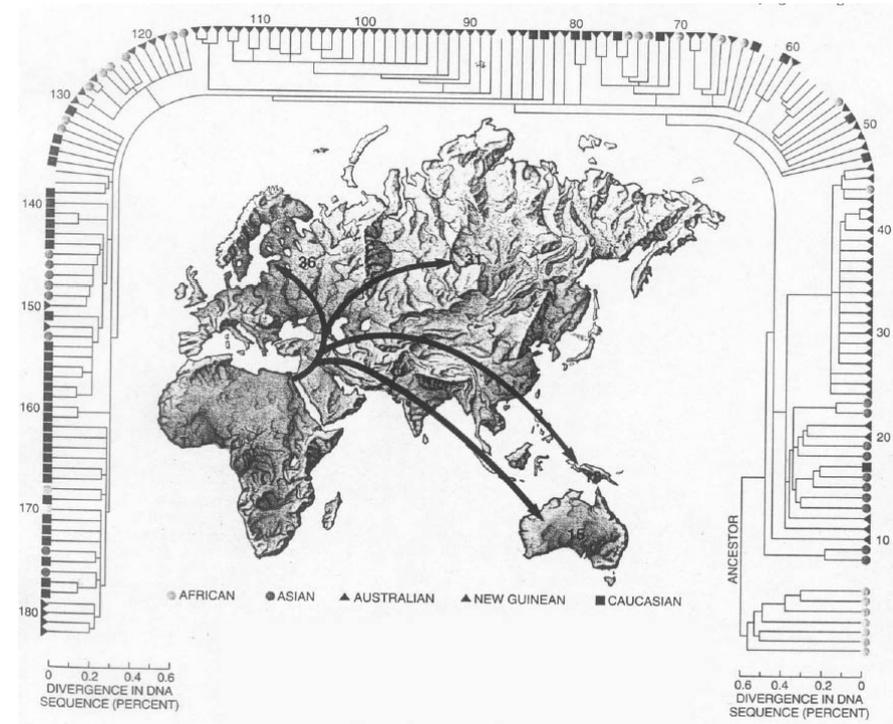
Sprachstammbäume



Quellen: www.wikipedia.de; www.stefanjacob.de

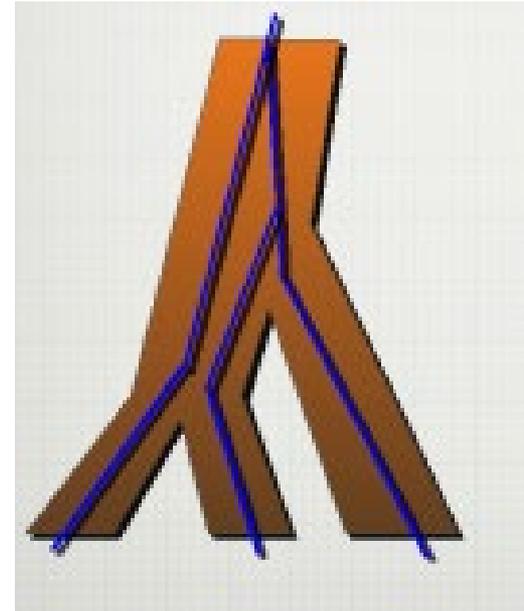
Moderne Stammbaumberechnung

- Mendel + Darwin: Das **Erbgut** unterliegt dem Wandel
- **Berechnung** von Stammbäumen aufgrund von DNA oder Proteinähnlichkeiten
 - *Molecular phylogeny*
 - Zuckerkandl und Pauling, 1965
- Annahme: **Evolution verläuft in kleinen Schritten**
- Wenn sich Sequenzen ähnlich sind, sind die Spezies evolutionär eng verwandt
 - Denn zufällige Ähnlichkeit ist unwahrscheinlich



Arten von Stammbäumen

- Abstammung einzelner Menschen
 - Stammbäume, Ahnentafeln
 - **Kein Baum**: Zwei Eltern
 - Rekombination
- Speziesstammbäume
 - Das ist ein Baum, wenn **Spezies nicht verschmelzen** können
- **Gene Trees**
 - Geschichte eines Sequenzabschnitts
 - Nicht leicht zu definieren
 - Baumförmig, wenn Gene nicht verschmelzen
 - Aber: 2 Allele jedes Gens vorhanden (Besser: **Haplotyp Tree**)



Quelle: <http://www.mailund.dk>

Wozu?

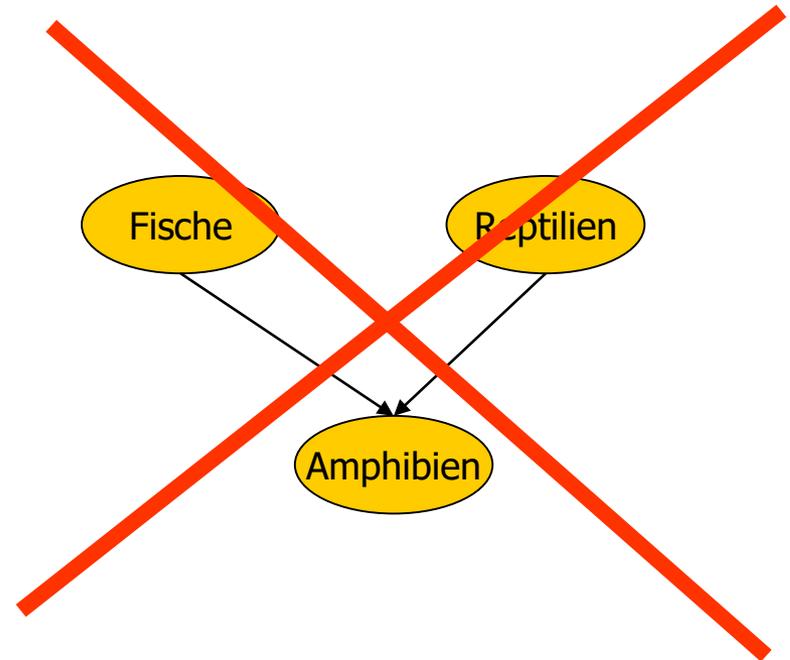
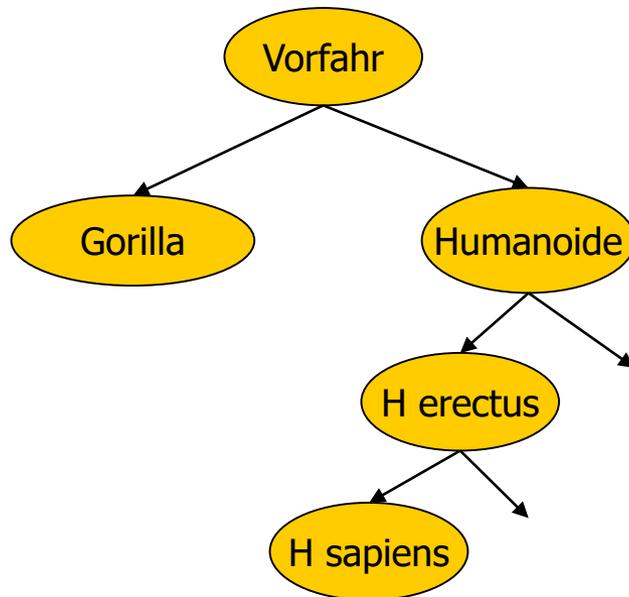
- Aufklärung der **evolutionären Verhältnisse**
- Verständnis Ausbreitung/Verkleinerung der **Artenvielfalt**
- Aufklärung von Infektionswegen
 - Bei **schnell mutierenden Viren** (z.B. HIV)
 - Varianten werden in verschiedenen Personen gefunden
 - Berechnung des Verbreitungsweges anhand der Abstammungsverhältnisse
- Phylogenetic Inference
 - Wenn alle nahe verwandten Spezies ein bestimmtes Gen haben, dann sollte ich das auch haben

Evolutionsmodell

- Lebewesen vermehren sich durch Kopieren mit Fehlern
 - Fehler: Mutationen
 - Führen (manchmal) zu veränderter Funktion
- **Selektion** – „Survival of the fittest“
 - Eingeschränkte Überlebensfähigkeit führen zur Ausrottung
 - Positive Mutationen: **mehr und lebensfähigere Nachkommen**
 - Positive Mutationen setzen sich in einer Population durch
- **Speziation**
 - Unterschiedliche Mutationen sind in unterschiedlichen Lebensräumen unterschiedlich vorteilhaft (ökologische Nischen)
 - Führt zur **umgebungsspezifischen Akkumulation von Änderungen**
 - Schließlich geht gemeinsame Fortpflanzungsfähigkeit verloren

Artenbildung

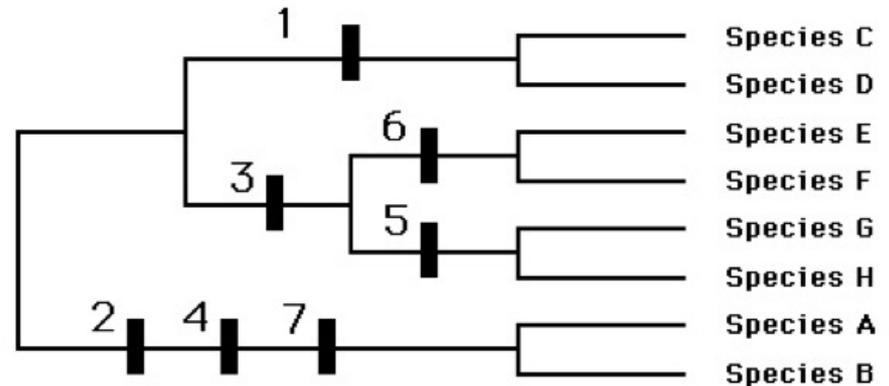
- Modell: Arten entstehen durch Mutationen aus anderen Arten
 - Anderes Modell: Phylogenetische Netzwerke



Gene Trees

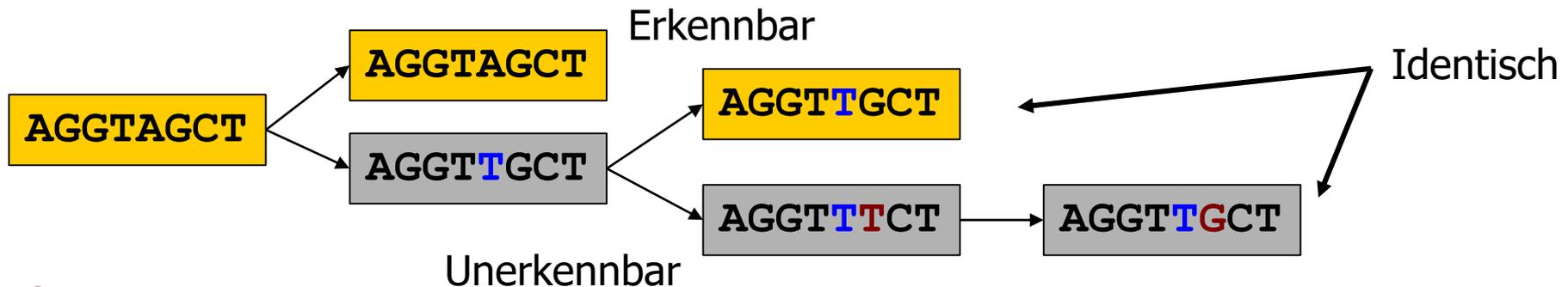
- Knoten = Arten
- Blätter = Lebende Arten
- Kanten
 - Länge idR (aber muss nicht) mit Zeit korrelieren
- Jeder Knoten hat exakt einen Vater
- Gewurzelte, binäre Bäume
- Ungeordnet: Reihenfolge der Kinder ist egal
- Viele Visualisierungsvarianten

	1	2	3	4	5	6	7
Species A	ACCAGCCTGTGCATCGATG	Δ	CGACTAAGTGATACCATAAA	Δ	GACT		
Species B	ACCAGCCTGTGCATCGATG	Δ	CGACTAAGTGATACCATAAA	Δ	GACT		
Species C	ACGAGCATGTGCATCGATGCCGACTAAGTGATACCATAA	T	GACT				
Species D	ACGAGCATGTGCATCGATGCCGACTAAGTGATACCATAA	T	GACT				
Species E	ACCAGCATGTGTATCGATGCCGACTAAGTGATACCA	Δ	AAATGACT				
Species F	ACCAGCATGTGTATCGATGCCGACTAAGTGATACCA	Δ	AAATGACT				
Species G	ACCAGCATGTGTATCGATGCCGACTAAGTGCTACCATAA	T	GACT				
Species H	ACCAGCATGTGTATCGATGCCGACTAAGTGCTACCATAA	T	GACT				

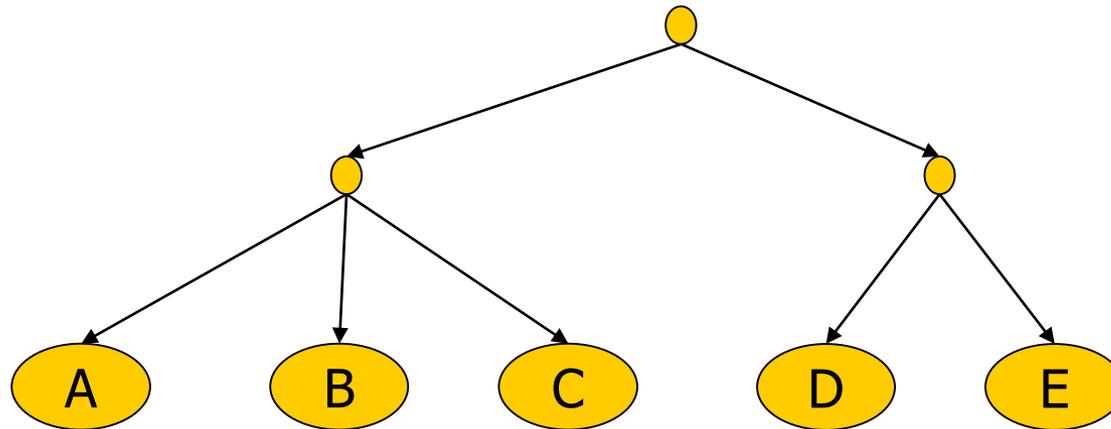


Daten

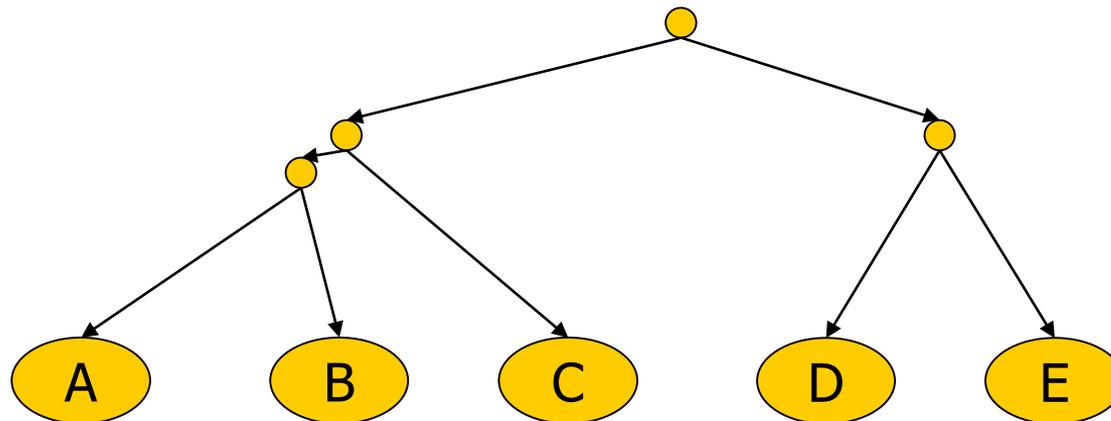
- Messen kann man nur die DNA **existierender Arten**
 - Und DNA von Fragmenten, die nicht zu alt sind
- Zwei mögliche Ziele
 - Rekonstruktion des **Stammbaums** der Arten
 - Rekonstruktion der **Ur-DNA** und aller Zwischenstufen
- Den **tatsächlichen Stammbaum** kann man nur annähern
 - Ausgestorbene Mutationen
 - Doppelmutationen



Binäre versus Multifurcation Trees



- Eines der Paare (A, B), (B,C), (A,C) wird **minimal ähnlicher** sein

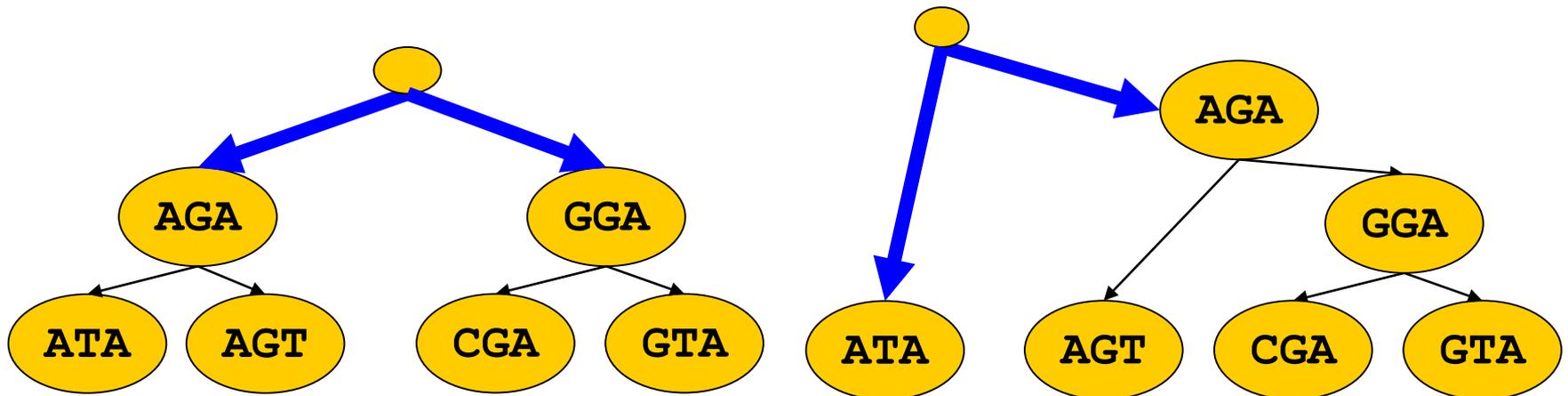
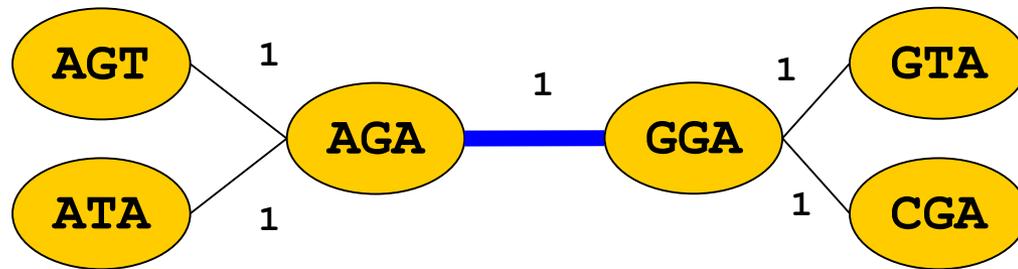


Weitere Probleme

- Gene Fusion
- Hybridisierung / Kreuzungen (bei Pflanzen)
- Rekombination
- Horizontal gene transfer
 - Insb. bei Viren / Bakterien bekannt
- Speziesbegriff unklar bei Lebewesen ohne sexuelle Reproduktion
- Homoplasy
 - Ähnliche Sequenzen, die nicht homolog sind
 - „[Convergent evolution](#)“ – Arten entwickeln Fähigkeiten oft unabhängig voneinander

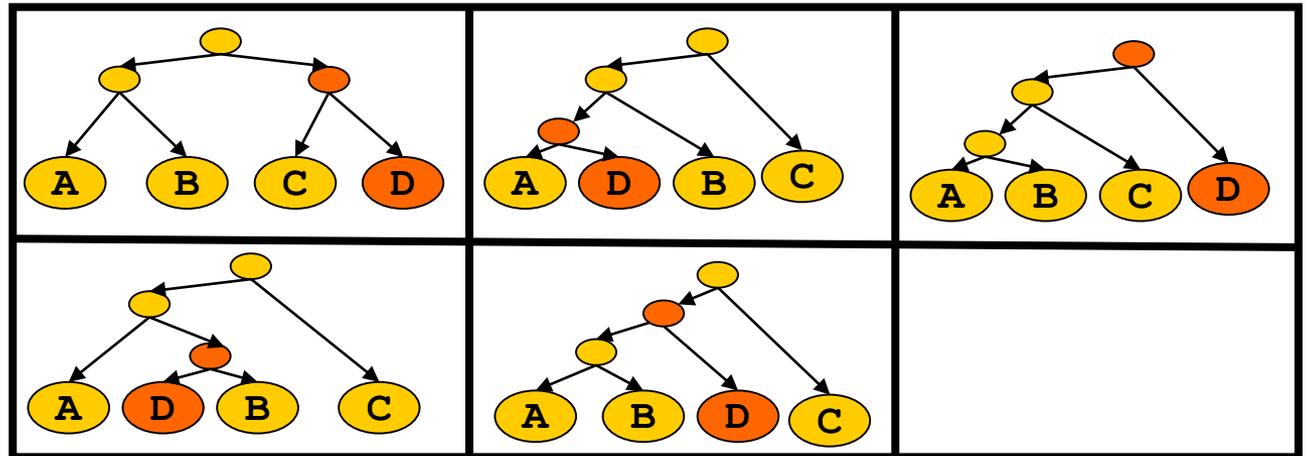
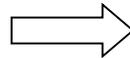
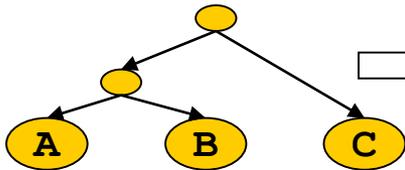
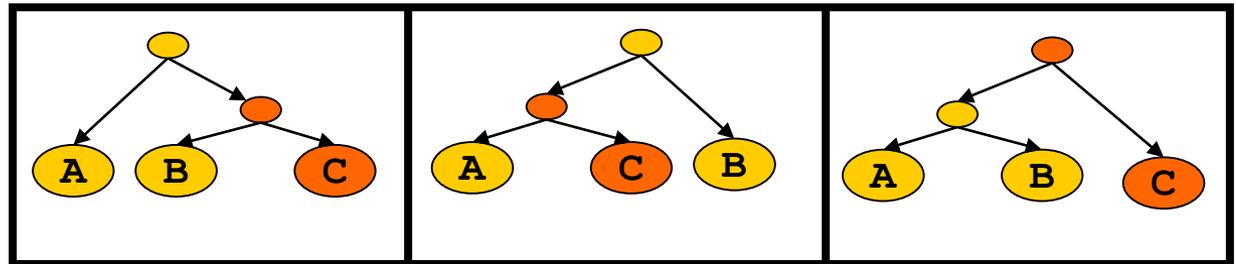
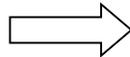
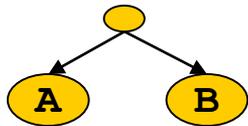
Bäume ohne Wurzeln

- Kanten symbolisieren Veränderungen
- Viele Methoden berechnen nur innere Knoten, aber können keine **Entwicklungsrichtung** ableiten



Wie schwierig wird das?

Wie viele binäre, ungeordnete Bäume für n Spezies gibt es?



Induktion über Zahl der Blätter

- Aus einem binären Baum mit n Blättern und m Kanten können **$m+1$ binäre Bäume mit $n+1$ Blättern** hervorgehen
 - Jede Kante kann gesplittet werden mit einem neuen Blattkind
- Wie viele Kanten hat ein binärer Baum?
 - Jeder Knoten außer der Wurzel hat genau eine eingehende Kante
 - Also sind es „Anz. Blätter“ + „Anz. innere Knoten“ -1 (Wurzel)
- Wie viele innere Knoten k hat ein binärer Baum mit n Blättern?

Von Blättern zu inneren Knoten

- Umgedreht: Wie viele Blätter (n) hat ein binärer Baum mit k inneren Knoten?
 - Induktionsanfang: Für $k=1$ ist $n=2$; $n(1)=2=k+1$
 - Sei $n(k)$ bekannt. Wo können wir neue innere Knoten hinzufügen?
An jeder Kante. Der neue Knoten teilt die Kante und muss als Kind den alten Teilbaum und ein neues Blatt haben. Also gilt:

$$n(k+1) = n(k) + 1 = n(k-1) + 1 + 1 = \dots = \sum_{i=1}^k 1 + 1 = k + 1$$

Von Knoten zu Bäumen

- Für die **Anzahl innerer Knoten** k (inkl. Wurzel) eines binären Baums mit n Blättern gilt: $n=k+1$; also $k=n-1$
- Ein binärer Baum mit n Blättern (und $n-1$ inneren Knoten) hat damit $n+(n-1)-1 = 2n-2$ **Kanten**
- Also
 - Aus einem Baum mit 2 Blättern können $(2*2-1)=3$ Bäume mit 3 Blättern hervorgehen
 - Aus einem Baum mit 3 Blättern können $(2*3-1)=5$ Bäume mit 4 Blättern hervorgehen
 - ...
 - Aus einem Baum mit n Blättern können $(2*n-1)$ Bäume mit $n+1$ Blättern hervorgehen

Ergebnis

- Sei $t(n)$ die Zahl **ungeordneter binärer Bäume** mit n Blättern

$$\begin{aligned}
 t(n) &= t(2) * t(3) * t(4) * \dots * t(n-1) = \\
 &= 1 * 3 * 5 * \dots * (2(n-1) - 1) = \\
 &= \frac{(2n-3)!}{2 * 4 * 6 * \dots * (2n-4)} = \\
 &= \frac{(2n-3)!}{2 \binom{2}{2} * 2 \binom{4}{2} * 2 \binom{6}{2} * \dots * 2 \binom{2n-4}{2}} = \\
 &= \frac{(2n-3)!}{2 * (1) * 2 * (2) * 2 * (3) * \dots * 2(n-2)} = \\
 &= \frac{(2n-3)!}{2^{n-2} * (n-2)!}
 \end{aligned}$$

1	1
2	1
3	3
4	15
5	105
6	945
7	10.395
8	135.135
9	2.027.025
10	34.459.425
11	654.729.075
12	13.749.310.575
13	316.234.143.225
14	7.905.853.580.625
15	213.458.046.676.875
16	6.190.283.353.629.370
17	191.898.783.962.511.000
18	6.332.659.870.762.850.000
19	221.643.095.476.700.000.000
20	8.200.794.532.637.890.000.000

Selbsttest

- Was ist Homoplasy? Wo gibt es horizontalen Gentransfer?
- Wann heißen zwei Gene paralog?
- Was macht Stammbaumberechnung beim Menschen kompliziert?
- Kann man Stammbäume auch aus nicht kodierenden Sequenzen berechnen?
- Was ist der Unterschied zwischen einem Gentree und einem Speziestree?
- Wie viele ungeordnete binäre Bäume mit n Blättern gibt es?