

Bioinformatik

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



Ulf Leser
Wissensmanagement in der
Bioinformatik



Suche mit MSA

- Erinnerung: Erzeugung von Proteinfamilien
 - Starte mit Proteinen gleicher/ähnlicher Funktion
 - Finde das Gemeinsame durch MSA
 - Suche „damit“ nach weiteren Vertretern
 - Modifiziere Familie entsprechend
 - Iteriere, bis Zufriedenheit eintritt
- Wie sucht man mit einem MSA?
 - Wir müssen entscheiden, wie gut eine (neue) Sequenz S zu einem MSA M passt
 - Profiles

Profile

- Definition

Gegeben ein MSA M mit n Spalten, $\Sigma' = \Sigma \cup \{-\}$

- Das *Profil P* zu M ist eine Tabelle der Größe $n * |\Sigma'|$

- In der Zelle (i, j) steht die *relative Häufigkeit des Zeichens j in der Spalte i*

- Beispiel (Alphabet A, G, C)

| | | | | | |
|-------|------|------|------|------|------|
| S_1 | A | G | C | - | A |
| S_2 | A | G | A | G | A |
| S_3 | A | C | C | G | - |
| S_4 | C | G | - | G | C |
| A | 0.75 | 0 | 0.25 | 0 | 0.50 |
| G | 0 | 0.75 | 0 | 0.75 | 0 |
| C | 0.25 | 0.25 | 0.50 | 0 | 0.25 |
| - | 0 | 0 | 0.25 | 0.25 | 0.25 |

Optimale Profilalignments

- Wie finde ich das optimale Profilalignment?
 - Natürlich mit **dynamischer Programmierung**

- Theorem

Gegeben eine Substitutionsmatrix m , Profil P , Sequenz S

- *Sei $c(x,i)$ der Score für das Alignieren eines Zeichen x mit Spalte i in P , also*

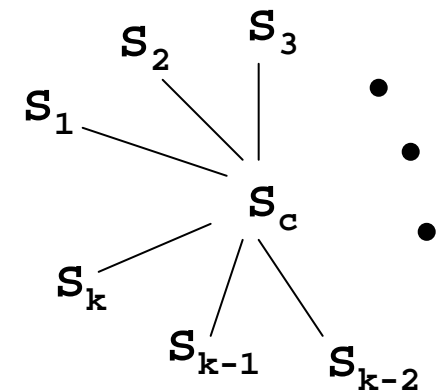
$$c(x,i) = \sum_{j=1}^{|\Sigma|} P[j,i] * m[c_j, x]$$

- *Sei $v(i,j)$ der Score für das optimale Alignment von $S[1..j]$ mit den ersten i Spalten von P . v berechnet sich als*

$$v(i, j) = \max \left(\begin{array}{l} v(i-1, j) + c(S[i], _) \\ v(i, j-1) + c(_, j) \\ v(i-1, j-1) + c(S[i], j) \end{array} \right)$$

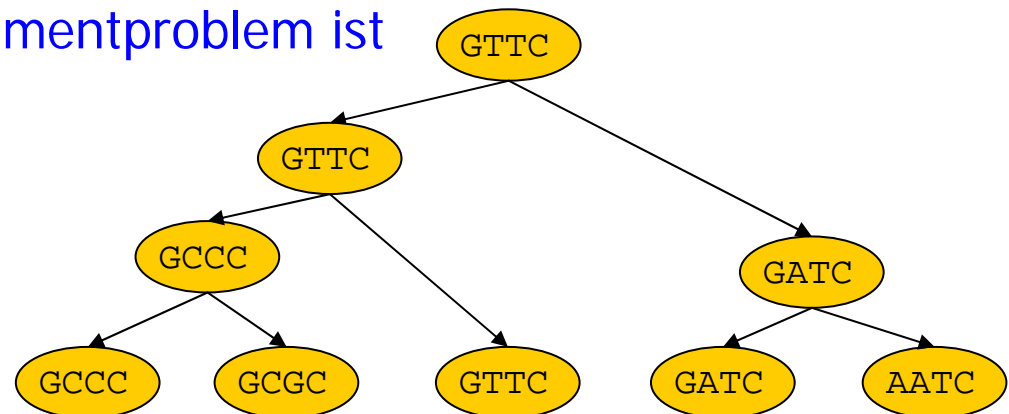
MSA mit Consensussequenz

- SP minimiert Summe aller paarweisen Alignments
- Alternativ: Minimierung der **Summe der Alignments aller Sequenzen S_1, \dots, S_k mit einer Consensussequenz S_c**
 - S_c kann eine der S_i sein, muss aber nicht
 - Konstruktion von S_c z.B. durch Untereinanderschreiben der S_i ohne Gaps und Wahl des häufigsten Buchstaben
 - MSA wird aus dem „Star“ abgeleitet
- Bei geeigneter Wahl von S_c und dem MSA – Ableitungsverfahren (gleich)
 - SP Score des MSA **höchstens doppelt so hoch wie der SP-optimale**



MSA mit phylogenetischen Bäumen

- Grundidee
 - Annahme: **Sequenzen durch Evolution** aus Ursequenz entstanden
 - Wir ordnen Sequenzen Knoten im Baum zu
 - Sequenzen haben sich entlang der Pfade entwickelt
- Zielfunktion
 - Suche den Baum T so, dass die **Summe aller Alignmentsscores von benachbarten Sequenzen in T minimiert** wird
 - Aus T kann man ein MSA ableiten (gleich)
- Leider
 - Das **phylogenetische Alignmentproblem ist NP vollständig**
 - Ähnlich zu Phylogenie mit Maximum Parsimony (später)



CLUSTAL W: Grundaufbau

- Gegeben k Sequenzen
- Drei Schritte
 - Berechne alle paarweisen Alignmentscores
 - Konstruiere „Guide Tree“ durch hierarchisches Clustering
 - Berechne und verschmelze Teil-MSA gemäß dem Guide Tree
- Idee dahinter
 - Aligniere erst sehr ähnliche Sequenzen – **Signale werden verstärkt**
 - Werden z.B. zwei sehr verschiedene Cluster von Sequenzen betrachtet, berechnet CLUSTAL automatisch erst zwei (homogene) MSA und verschmilzt diese am Ende
 - Hohe Chance, dass **konservierte Blöcke** erhalten bleiben
 - **Außenseiter** kommen erst spät dazu und können die Blockstruktur nicht mehr stören
 - Orientierung an der „tatsächlichen“ Entstehungsgeschichte, dem **phylogenetischen Baum**

Schritt 1 und 2

- Berechnen der Ähnlichkeitsmatrix M
 - Berechne die $O(k^2)$ paarweisen Alignmentsscores für die k Sequenzen
- Hierarchisches Clustering
 - Wähle Zelle (i,j) mit kleinstem Abstand aus Matrix M
 - Das ist das erste Paar
 - Erzeuge M': Lösche die Sequenzen i und j aus M und füge neue Spalte/Zeile (ij) ein
 - Für alle $k \neq ij$: $M'[ij,k] = (M[i,k] + M[j,k]) / 2$
 - Mittlerer Abstand zu i und j
 - Iteriere, bis Matrix nur noch 2x2 groß ist
- Tatsächlich benutzt CLUSTAL W heute Neighbour Joining
 - Details später: Ultrametrien und additive Bäume

Schritt 3: Progressive MSA Generierung

- Berechnung paarweiser Alignments in der Reihenfolge des Guide Trees
- Alignment eines MSA M_1 mit einem MSA M_2
 - Dynamische Programmierung mit linearem Gapscore
 - Wert eines Mismatches/Matches ist der **Durchschnittsscore aller Paare** mit einem Zeichen aus M_1 und dem anderen aus M_2
 - **Gaps** werden mit dem schlechtesten Score der verwendeten Substitutionsmatrix bestraft
 - Bei k Sequenzen sind das maximal $k/2 * k/2 = O(k^2)$ Scores

- Beispiel

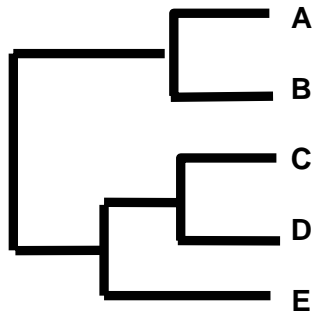
A ...P...
 B ...G...
 C ...P...

 D ...A...
 E ...A...
 F ...Y...

- Score

$$\begin{aligned}
 & - (2 * s(P, A) + s(P, Y) + \\
 & \quad 2 * s(G, A) + s(G, Y) + \\
 & \quad 2 * s(P, A) + s(P, Y)) / 9
 \end{aligned}$$

Beispiel



C PADKTNVKAAWGKVG**A**HAGEYGA

D AADKTNVKAAWSKVG**G**HAGEYGA

A P**E**EK**S**AVTALWGKVN**V**DEYGG

B G**E**E**K**AAVLALWDKV**N**EE**E**YGG

C PADKTNVKAAWG_ **K**V**G**A**H**A**G**EYGA

D AADKTNVKAAWS_ **K**V**G****G**HAGEYGA

E AA__TNV**K**TAWSS**K**V**G****G**HAPA__**A**

A PEEKSA**V**_TAL**W**G_**K**V**N**__VDEYGG

B GEEKA**A**V_**L**AL**W**D_**K**V**N**__EEYGG

C PADKTN**V**KAA_**W**G_**K**V**G**A**H**A**G**EYGA

D AADKTN**V**KAA_**W**S_**K**V**G****G**HAGEYGA

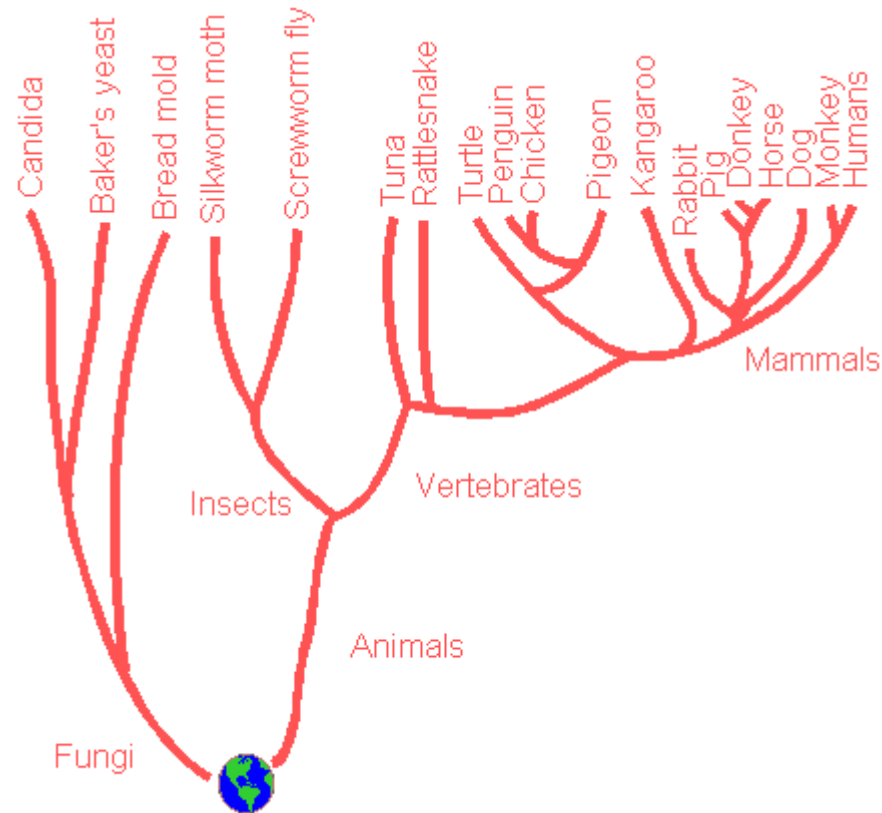
E AA__TN**V****K**T**A**_WSS**K**V**G****G**HAPA__**A**

Once a gap, always a gap

Inhalt dieser Vorlesung

- Stammbäume
- Phylogenetische Bäume
- Evolutionsmodell

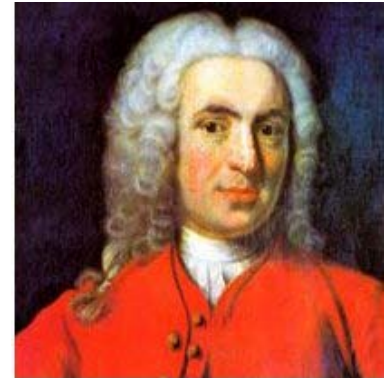
Phylogenetische Bäume



Stammbaum (*Phylogenetic Tree*)

Stammbäume versus Klassifikation

- Zuerst war die Klassifikation
 - Carl Linnaeus, ca. 1740: *Systema Naturae*
 - Annahme: Arten verändern sich nicht (Prä-Darwin)
 - Einteilung der Lebewesen in
 - Kingdoms - classes – orders – families – genera – species
 - SKOFGA (Stamm, Klasse, Ordnung, Familie, Gattung, Art)
 - **Innere Knoten** einer Klassifikation sind abstrakt

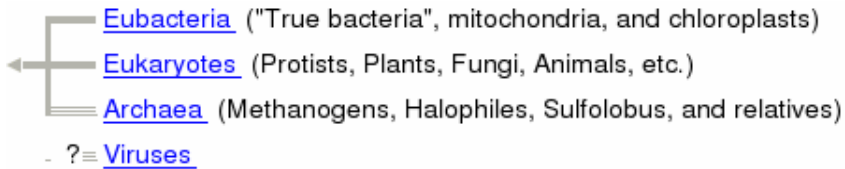


'Bridegroom portrait' of Linnaeus,
Clasif. LINNÆI, M. D.
METHODUS plantarum SEXUALIS
in SISTEMATE NATURÆ
de Ripa



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

Klassifikation



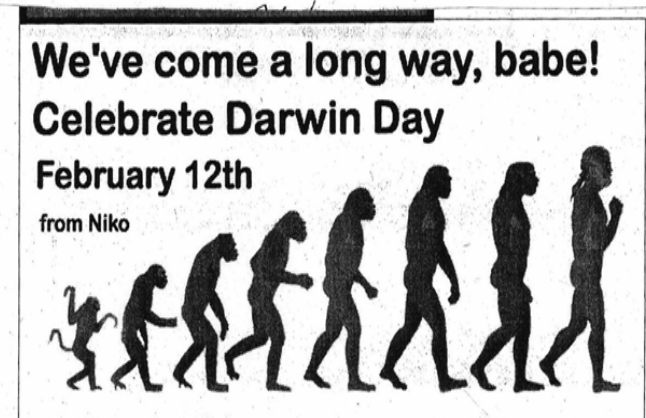
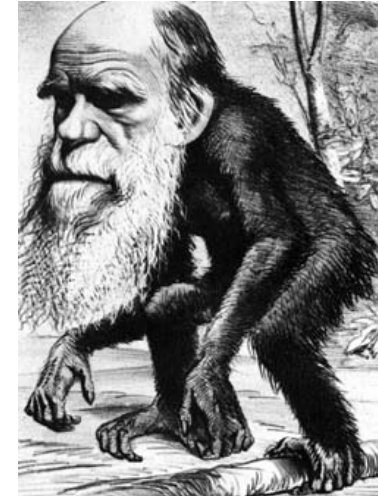
Popular Groups on the Tree of Life



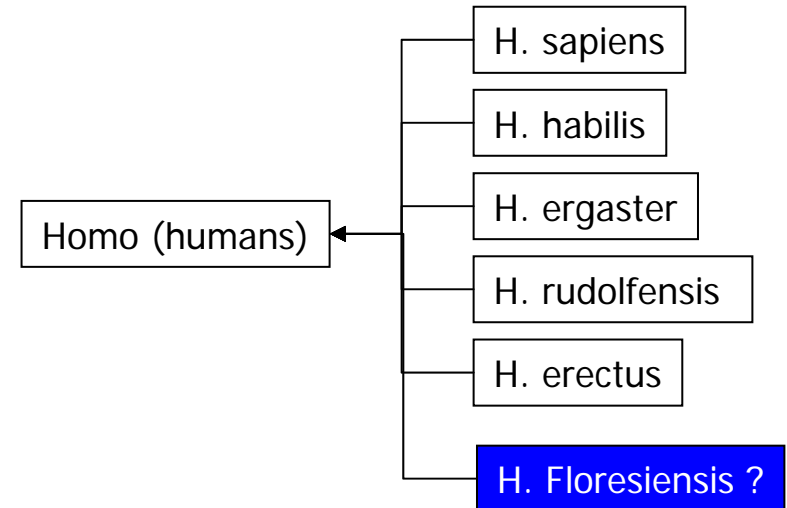
- Eukaryoten
- Tiere
- diverse Zwischenstufen
- Chraninata (Schädelknochen)
- Vertebraten (Wirbeltier)
- Viele Zwischenstufen
- Mammals (Säugetiere)
- Eutheria (Placenta)
- Primaten (Affen)
- Catarrhini
- Hominidae (Mensch, Schimpanse, Orang-Utan, Gorilla)
- Homo (erectus, sapiens ...)
- Homo Sapiens

Stammbäume

- Charles Darwin: „The origin of species“ (1859)
 - Arten sind nicht unveränderlich, sondern unterliegen im Laufe der Zeit einem Wandel
 - „Survival of the fittest“
 - Damals noch unklar, was sich eigentlich wandelt
- **Stammbäume** (Abstammungsbäume)
 - Ergeben sich aus der Annahme der Evolution
 - Auch Stammbäume wurden lange aus morphologischen Eigenschaften abgeleitet
 - Jeder Knoten in einem Stammbaum hat einmal als Art existiert
 - Knoten im Baum (also Spezies) heißen **taxa**
- Was definiert eine Spezies?



„Homo floresiensis“ (Hobbit?)

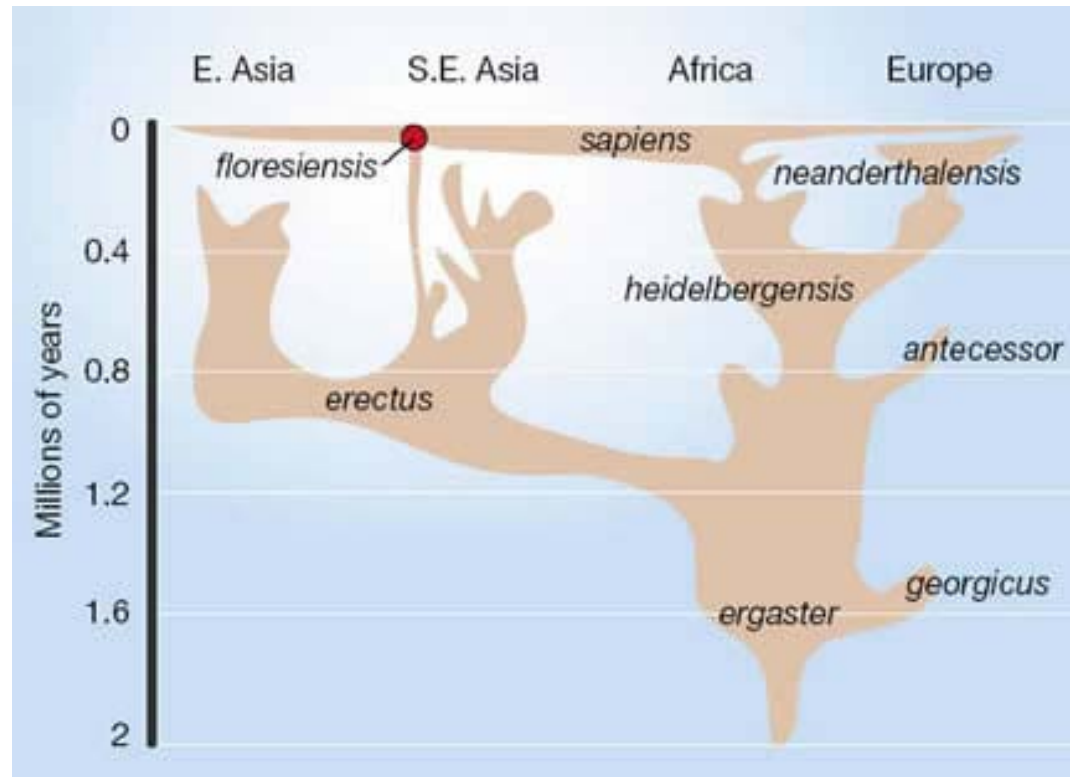


- Entdeckung 2003
- Lebte vermutlich vor 95 000 bis vor 12 000 Jahren auf der indonesischen Insel Flores
- Körperhöhe nur ca. 100 cm
- Gehirnvolumen nur 380 cm³
 - Homo erectus: 600-1200, h. sapiens: 1400
- Rückentwicklung? Vorläufer?

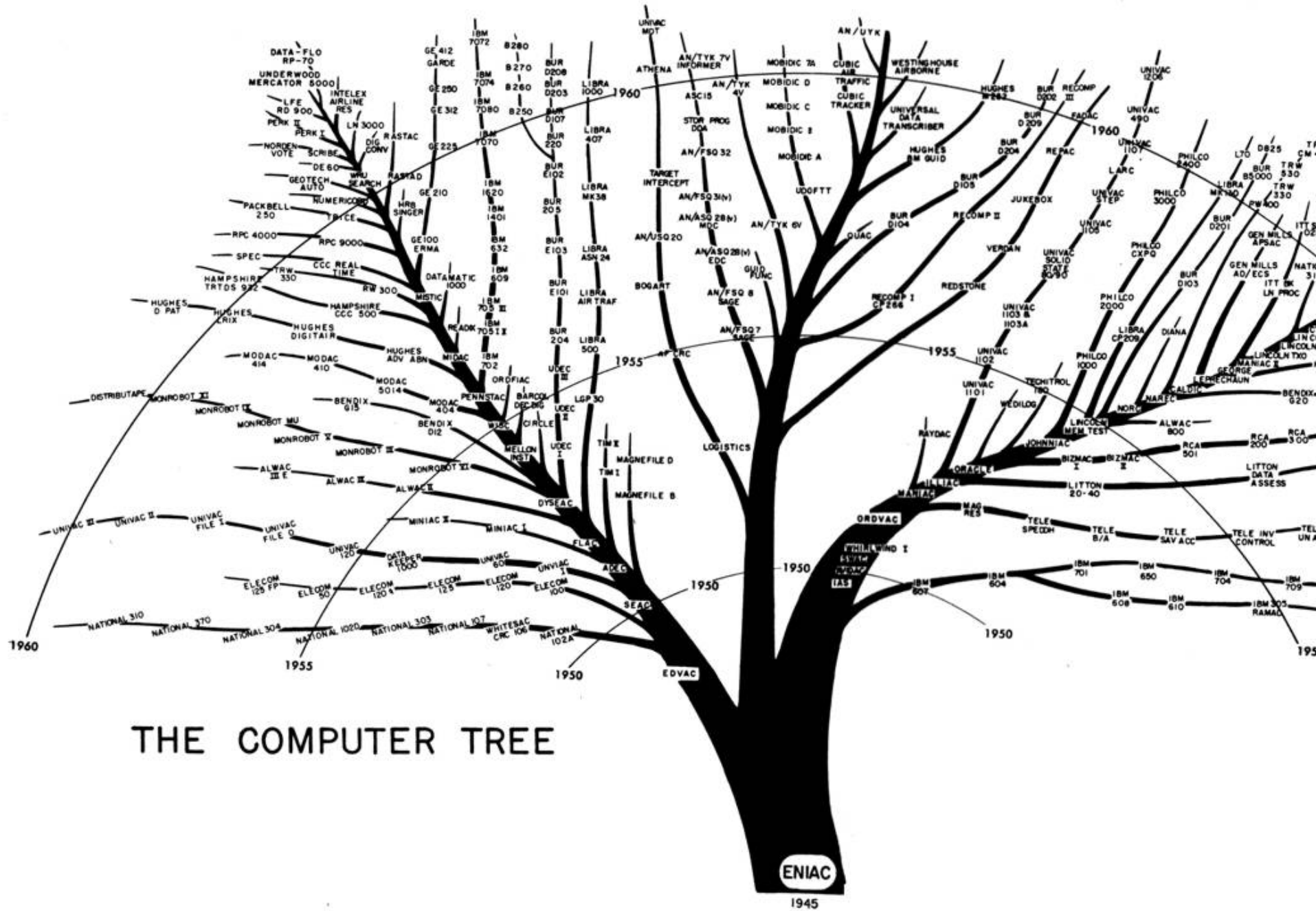
Quelle:

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

Evolutionsgeschichte neu geschrieben



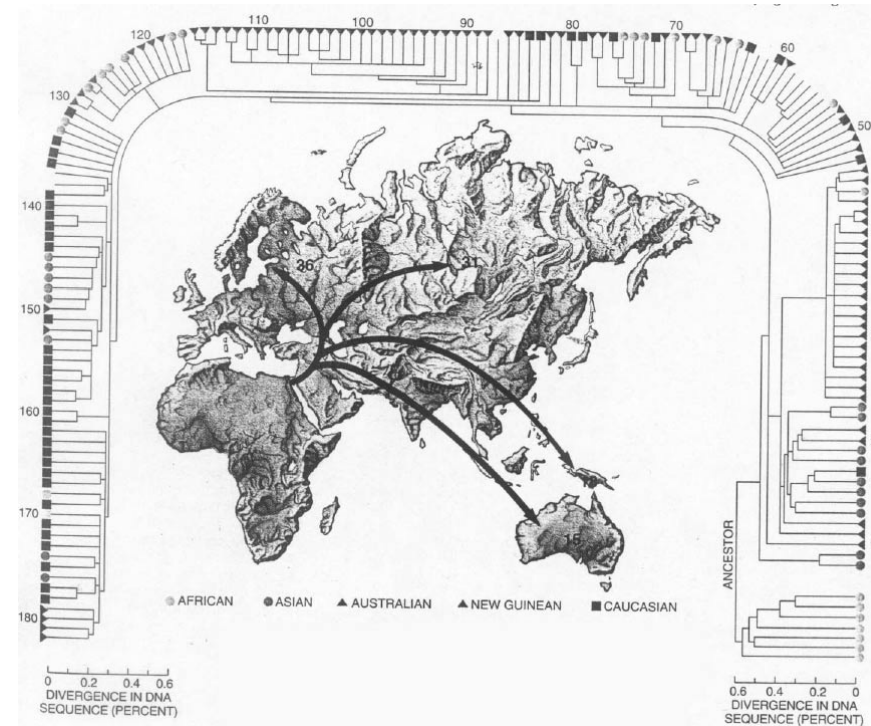
- Stammbaum – welche Art geht aus welcher hervor?



THE COMPUTER TREE

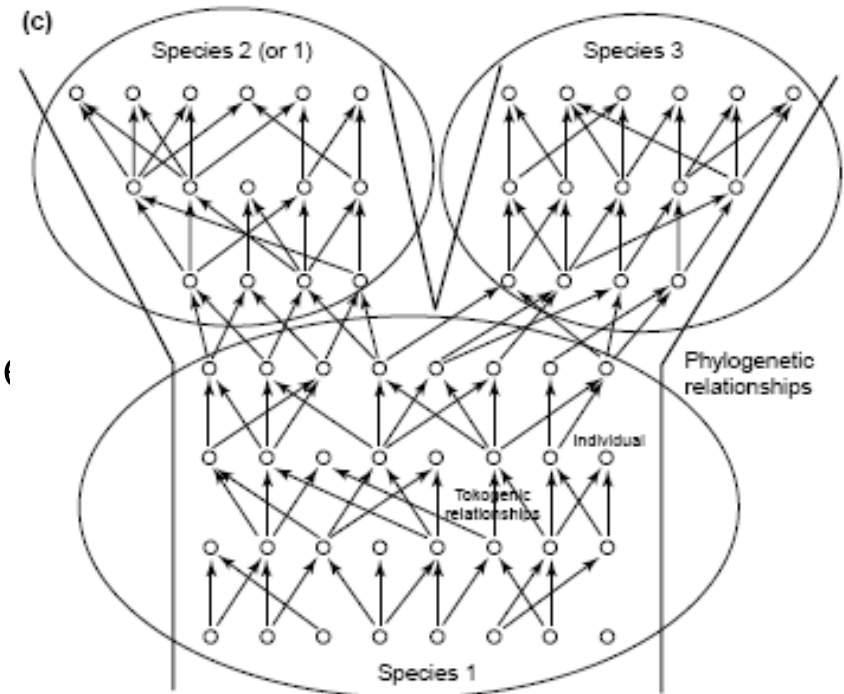
Moderne Stammbaumberechnung

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



Arten von Stammbäumen

- Individuelle Abstammung
 - Stammbäume, Ahnentafeln
 - Natürlich **kein Baum**: Zwei Eltern
 - Rekombination
- Speziesstammbäume
 - Ein Baum, wenn **Spezies nicht verschmelzen** können
 - Sprachen verschmelzen
- Gene Trees
 - Geschichte eines Sequenzabschnitts
 - Nicht leicht zu definieren
 - Baumförmig, wenn Gene nicht verschmelzen
 - Aber: 2 Allele jedes Gens vorhanden (Besser: **Haplotype Tree**)



TRENDS in Ecology & Evolution

Wozu?

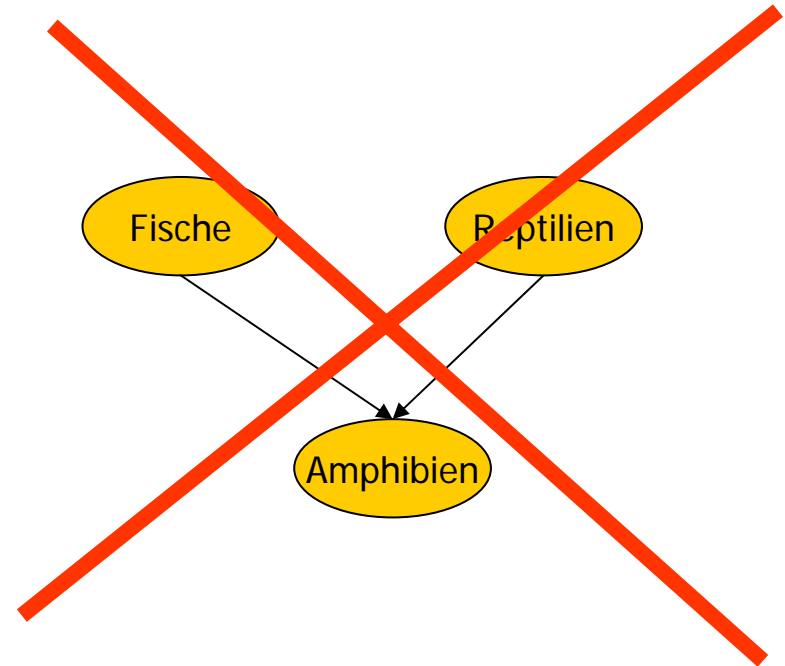
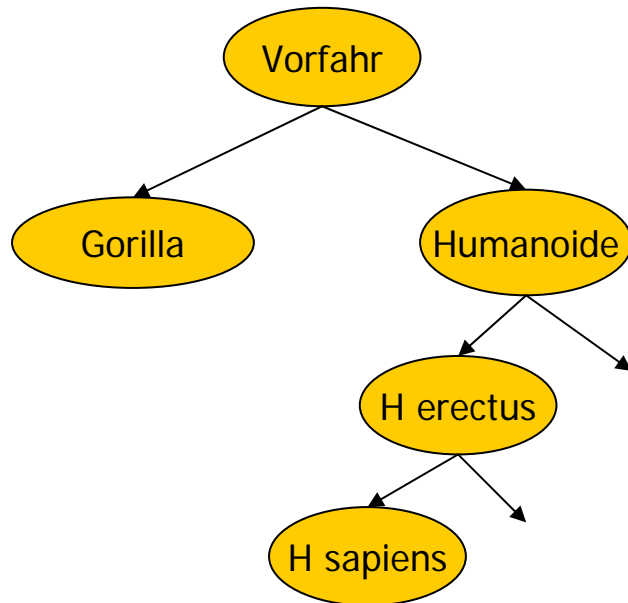
- Aufklärung der evolutionären Verhältnisse
- Aufklärung von familiären Verhältnissen
 - Wenn genügend Sequenzen vorhanden
- Aufklärung von Infektionswegen
 - Bei schnell mutierenden Viren (z.B. HIV)
 - Varianten werden in verschiedenen Personen gefunden
 - Herausrechnen des Verbreitungsweges anhand 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
- **Kopieren** mit kleinen Fehlern (Mutationen)
 - Führen (manchmal) zu veränderter „Funktion“
- **Selektion** – „Survival of the fittest“
 - Fatale Mutationen lassen Nachkommen gar nicht erst entstehen
 - Einschränkungen der Überlebensfähigkeit führen zur schnellen Ausrottung
- **Speziation**
 - Unterschiedliche Mutationen sind in unterschiedlichen Lebensräumen unterschiedlich vorteilhaft
 - **Anpassung** an ökologische Nischen
 - Führt zur Akkumulation von Änderungen
 - Schließlich geht die gemeinsame Fortpflanzungsfähigkeit verloren

Artenbildung

- Arten entstehen durch Veränderungen aus **einer** anderen Art



Mutationsfrequenzen

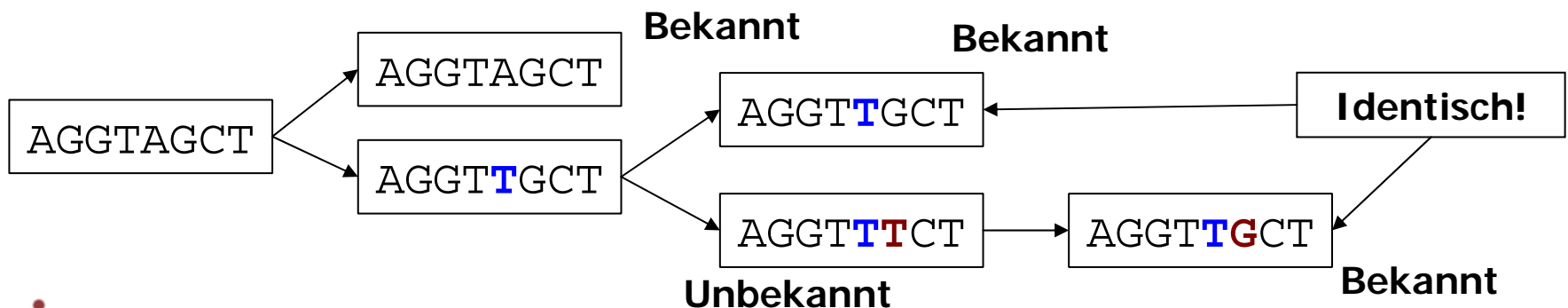
- Treten Mutationen in gleicher Frequenz auf?
 - Sind alle Mutationen gleich wahrscheinlich?
 - Substitutionsmatrizen
 - Gibt es Zeiten erhöhter **Mutationshäufigkeit**?
 - Sonneneruptionen, ...
 - Gibt es Zeiten erhöhten **Selektionsdrucks**?
 - Meteoriteneinschlag, Klimaverschiebungen, ...
 - Gibt es **Teilbäume**, die schneller mutieren?
 - Abhängig vom Anpassungsdruck oder der Güte der Reparaturmechanismen
 - Gibt es **Sequenzabschnitte**, die unterschiedlich schnell mutieren?
 - Coding versus non-coding Regions, House-Keeping genes, ...

Molecular Clock

- Häufige Annahme: **Molecular Clock**
 - Mutationen setzen sich bzgl. der Zeit immer mit gleicher Häufigkeit durch
 - Unabhängig von Teilbaum, Zeitpunkt, Ort und Art der Mutation
 - Anzahl der beobachteten Mutationen proportional zur vergangenen Zeit
- Die ist manchmal hilfreich, aber falsch
 - Unterschiedliche Zeiten erzeugen unterschiedlichen Selektionsdruck
 - Unterschiedliche Sequenzabschnitte unterliegen unterschiedlichem Selektionsdruck
 - Unterschiedliche Spezies unterliegen unterschiedlicher Frequenz erfolgreicher Mutationen
 - ...

Daten

- Messen kann man nur die DNA existierender Arten
- Zwei mögliche Ziele
 - Rekonstruktion des wahrscheinlichsten **Stammbaums** der Arten
 - Rekonstruktion der wahrscheinlichsten **Ur-DNA** und aller Zwischenstufen
- Den **tatsächlichen Stammbaum** kann man nicht berechnen
 - Man kennt die ausgestorbenen Arten nicht
 - Man kann ausgestorbene Mutationen nicht erkennen
 - Man kann Doppelmutationen nicht erkennen

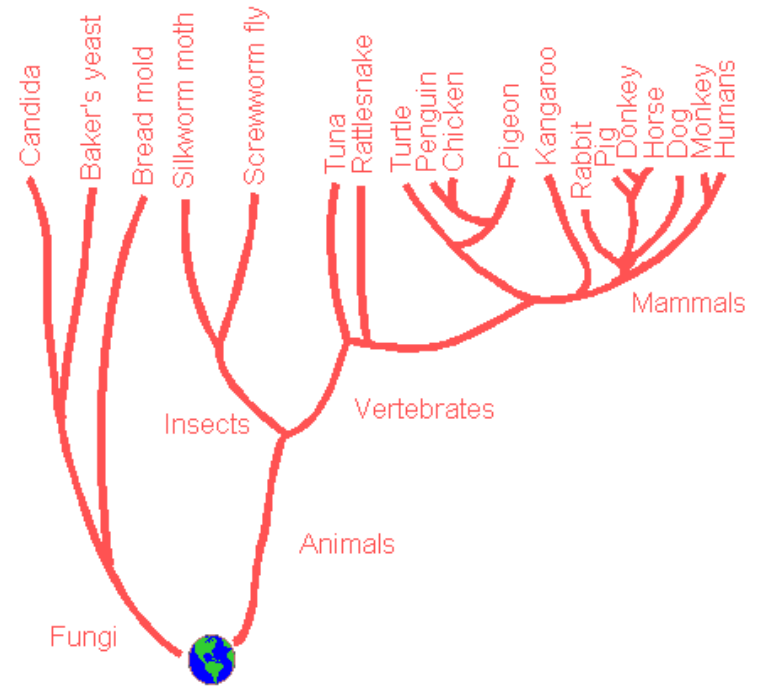


Weitere Probleme

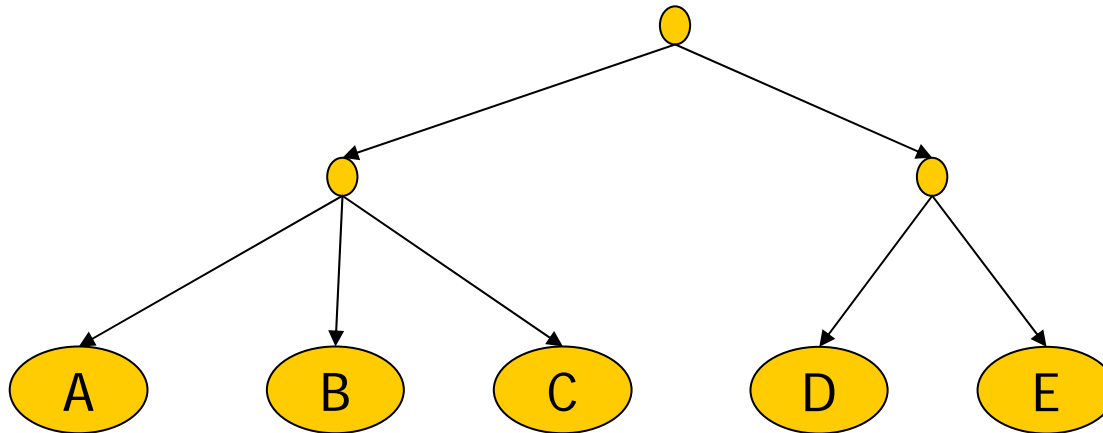
- Gene Fusion
 - Gene in einem Genom können verschmelzen
 - Also hat eine Sequenz zwei Vorfahren
- Hybridisierungskreuzungen
 - Pflanzengenome können sich kreuzen
 - Also hat eine Art zwei Vorfahren
- Horizontal gene transfer
 - Gerade im Viren / Bakterienbereich bekannt
 - Also stimmen Genetree und Speziestammbaum nicht überein
- Homoplasy
 - Ähnliche Sequenzen, die nicht miteinander verwandt sind
 - „**Convergent evolution**“ – Arten entwickeln Fähigkeiten oft unabhängig voneinander
 - Also kann man von Ähnlichkeit nicht auf Abstammung schliessen

Bäume

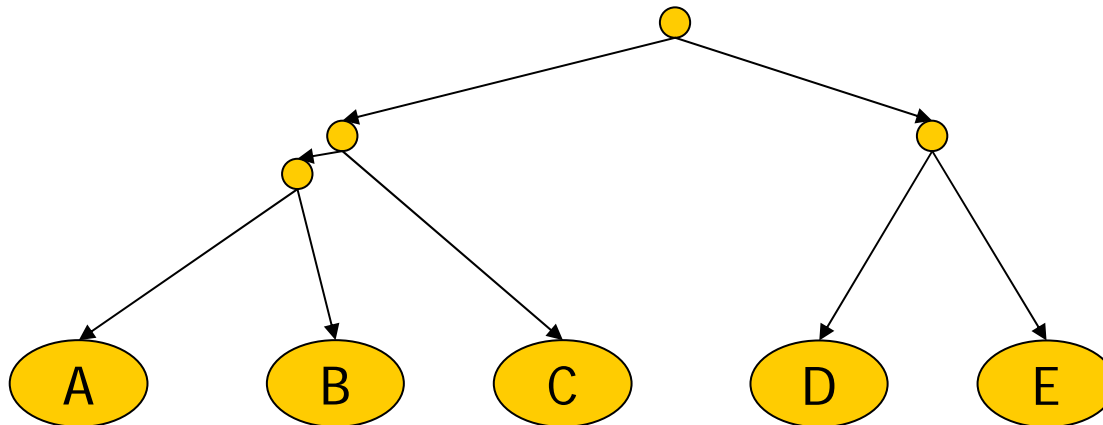
- Knoten = Arten
- Blätter = Lebende Arten
- Kanten
 - Länge kann (aber muss nicht) mit zeitlicher Entfernung korrelieren
 - Scaled trees
- Jeder Knoten hat exakt einen Vater
- Eine Wurzel
 - Problematisch (gleich)
- Binäre Bäume
 - Unproblematisch (gleich)
 - Reihenfolge der Kinder ist egal
- Viele Visualisierungsvarianten



Multifurcation Trees

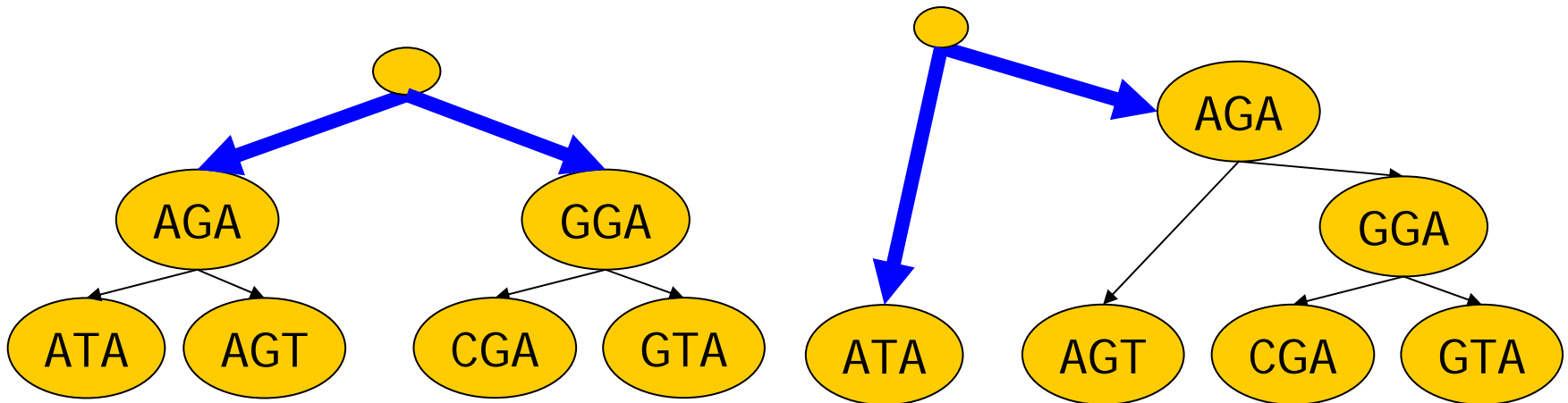
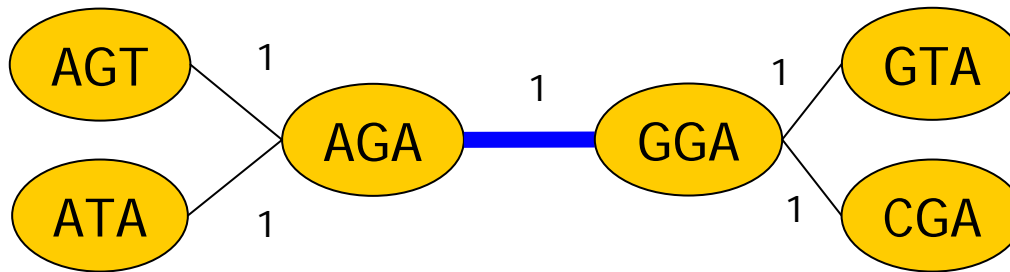


- Erscheinen als
 - Eines der Paare (A, B), (B,C), (A,C) wird **minimal ähnlicher** sein



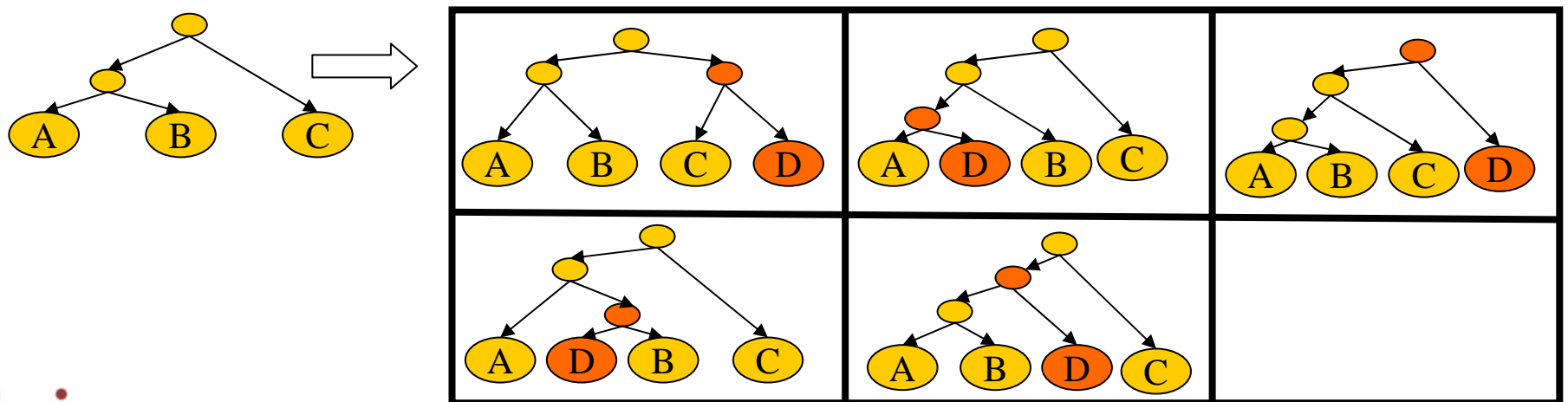
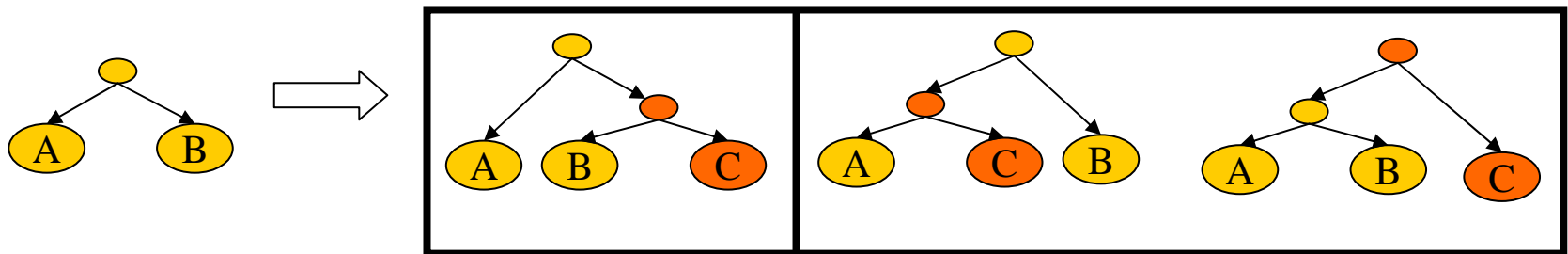
Bäume ohne Wurzeln

- Kanten symbolisieren Veränderungen
- Wir können nur Blätter vergleichen
- Damit ist idR keine Richtung der Entwicklung berechenbar



Wie schwierig wird das?

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



Von Blättern zu inneren Knoten

- Aus einem binären Baum mit n Blättern und k Kanten können $k+1$ binäre Bäume mit $n+1$ Blättern hervorgehen
- 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?
- 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 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 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 |