

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

Variante 1: Reguläre Ausdrücke

A C A - - - A T G

T C A A C T A T C

A C A C - - A G C

A G A - - - A T C

A C C G - - A T C

Beispiel aus [Kro98]

Krogh, A. (1998). An Introduction to HMM for Biological Sequences.

$[AT] [CG] [AC] [ACGT]^* A [TG] [GC]$

- Vorteil: Schnell berechnet, schnelle Matchingalgorithmen
- Nachteil: Keine Unterscheidung der **Güte eines Matches**

Variante 2: 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

Bewertung eines Alignments

- Definition

Gegeben ein Profil P mit n Spalten, eine Sequenz S und eine Substitutionsmatrix m .

- Ein *Alignment* A von P und S ist ein Untereinanderschreiben von P und S , wobei immer eine Spalte von P (oder ein Leerzeichen) über einem Zeichen von S (oder einem Leerzeichen) steht.
 - Aber niemals zwei Leerzeichen untereinander stehen
- Wir erzeugen aus P ein P' , in dem wir an den betreffenden Stellen leere Spalten einfügen; dito ein S' aus S
- Der *Score* $s(A)$ von A berechnet sich als

$$s(A) = \sum_{i=1}^{|A|} \sum_{c_k \in \Sigma'} \begin{cases} (P'[c_k, i] * m[c_k, S'[i]]), & \text{wenn } i \text{ keine Leerspalte} \\ m[_, S'[i]], & \text{sonst} \end{cases}$$

Optimale Profilalignments

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

- Theorem

Gegeben eine Substitutionsmatrix m , Profil P , Sequenz S

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

$$c(x, j) = \sum_{c_k \in \Sigma'} P[c_k, j] * m[c_k, x]$$

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

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

MSA – HMM

```

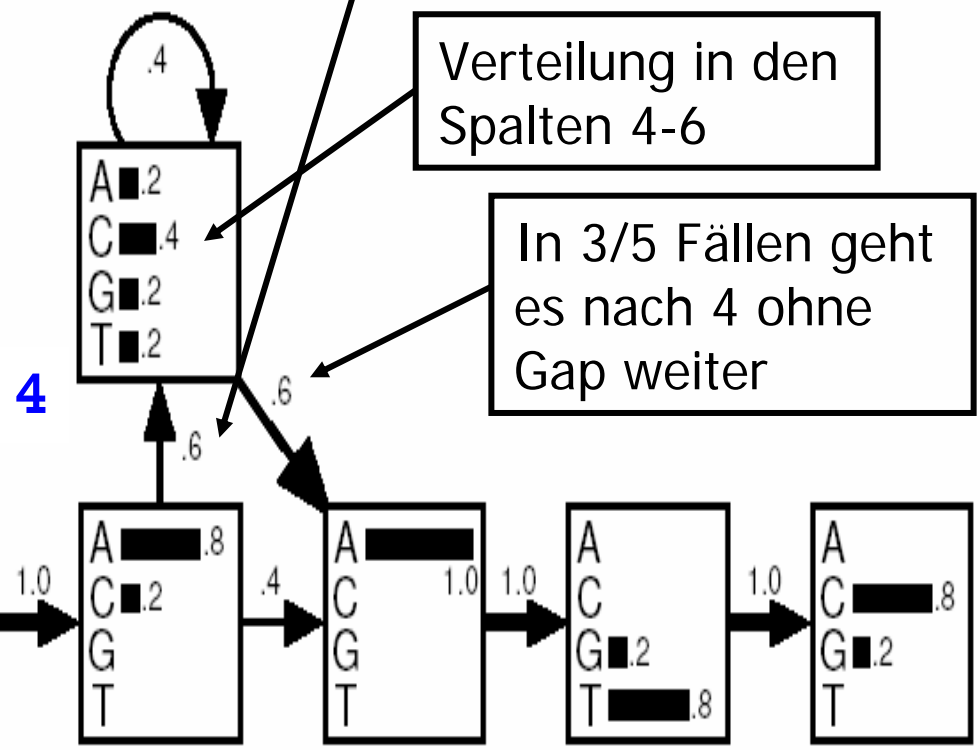
A C A - - - A T G
T C A A C T A T C
A C A C - - A G C
A G A - - - A T C
A C C G - - A T C
1 2 3 4 5 6 7 8 9
    
```

Nach Spalte 3 betritt man in 3/5 Fällen den INS State (und macht nicht gleich mit 7 weiter)

Insertion

Verteilung in den Spalten 4-6

In 3/5 Fällen geht es nach 4 ohne Gap weiter

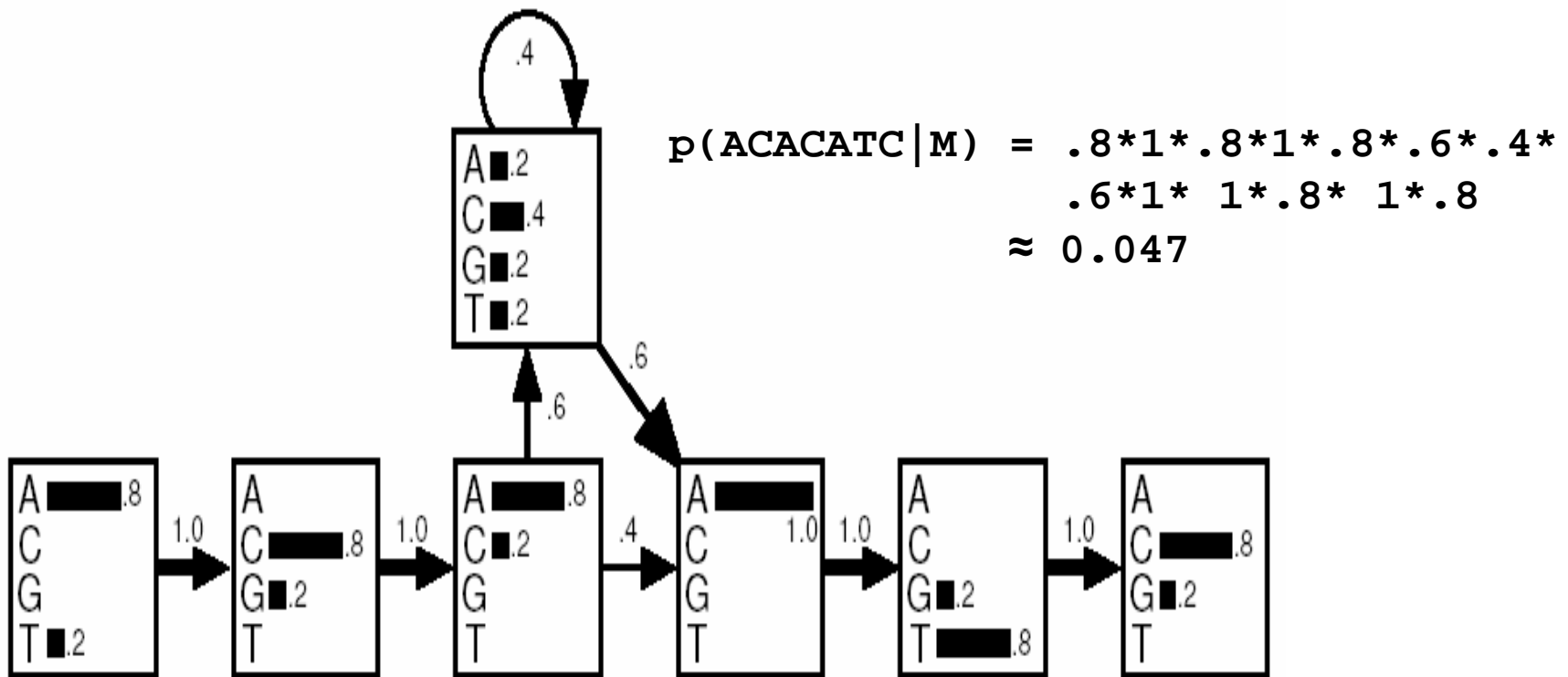


Matches

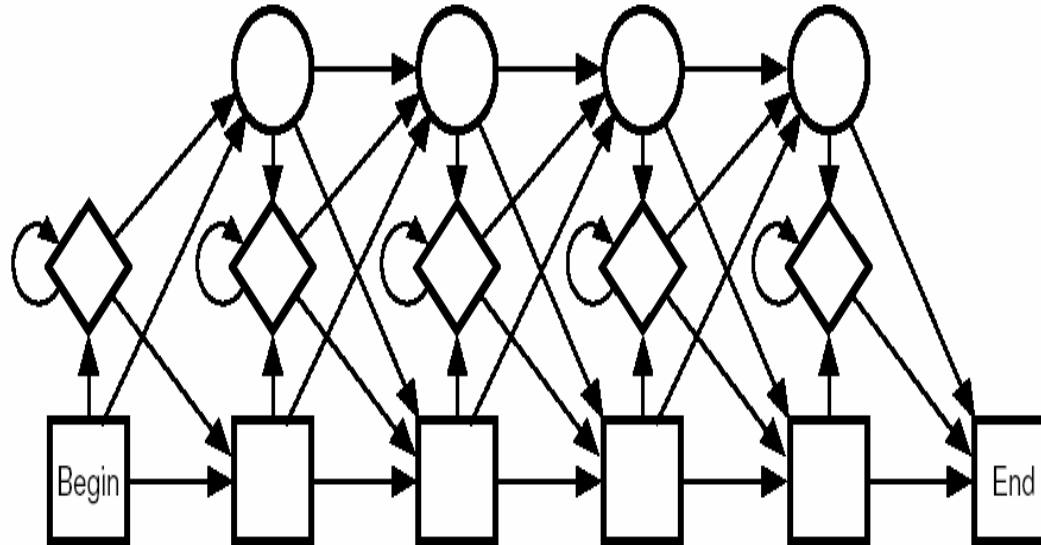
[Kro98]

Scoring

- Wir können nun die maximale Wsk einer neuen Sequenz, gegeben das Modell, berechnen
 - Sprich: Zu der Sequenz-/ Proteinfamilie zu gehören



Struktur

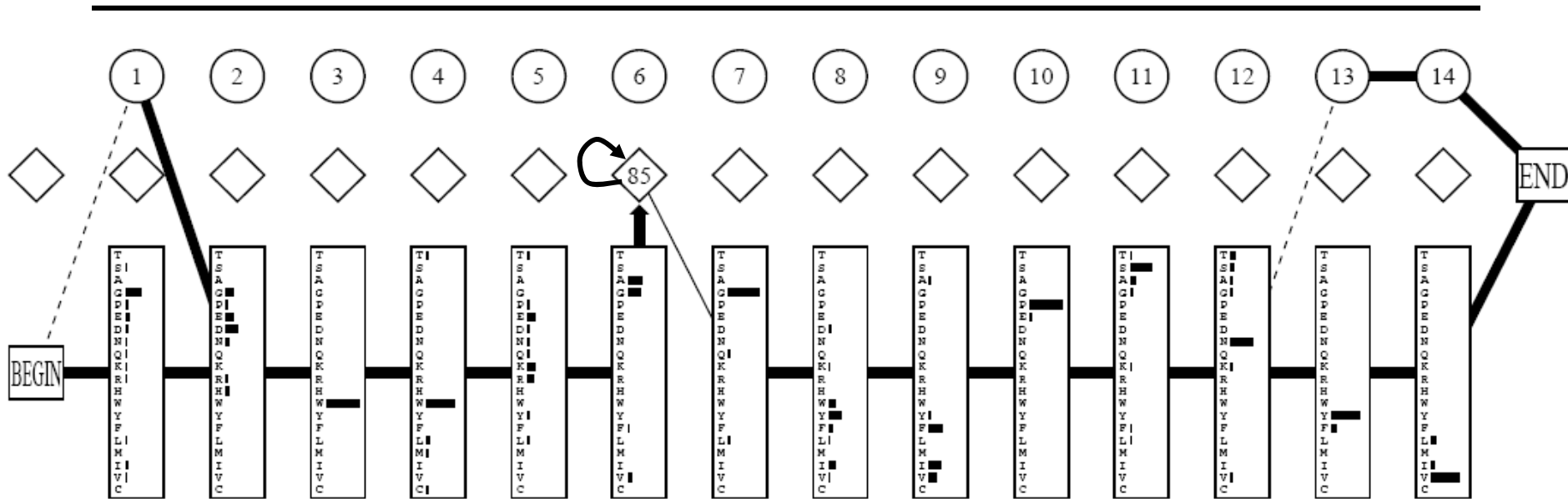


- Rechtecke: Match-Zustände
 - Symbolisieren relativ volle Spalten (Typisch: >50%)
- Rauten: Insertion-Zustände
 - Symbolisieren **Spalten/Bereiche mit vielen Gaps**
 - Können beliebig oft wiederholt werden
- Kreise: Deletion-Zustände
 - Überspringen jeweils genau einen Match-Zustand, aber können verkettet sein
 - „**Silent States**“: Emittieren keine Zeichen

Beispiel

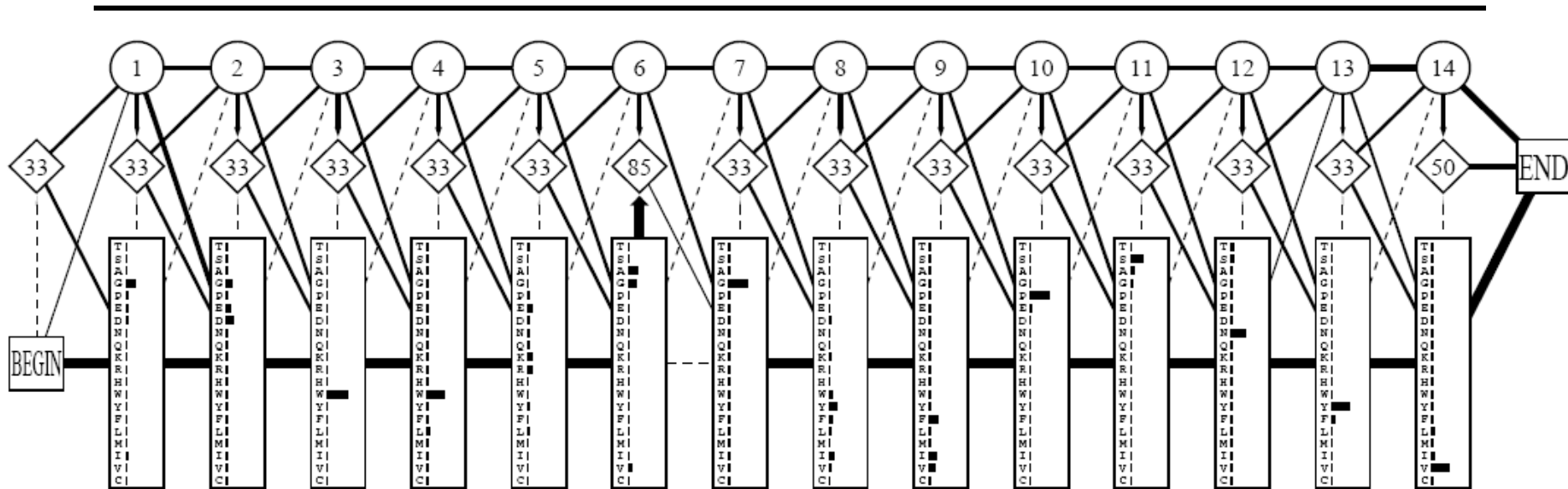
```
GGWWRRGdy.ggkkkqLWFPSSNYYV
IGWLNNGdyne.ttgkerLGDFPSTYYV
PNWWEQGql..nnrrrGIFPSTYYV
DEWWEQAarr..deqqiGIVPSK--
GEWWEKAqrs..tggqqeGFI PFNFV
GDWWLArss..sggqqtGYIPSNYYV
GDWWDAel..kqgrrrGKVPSNYYL
-DWWEArsslssghrGYVPSNYYV
GDWWYArsslitnseGYIPSTYYV
GEWWEKArsllatrkGYIPSNYYV
GDWWLArsslvtgreGYVPSNFV
GEWWEKAksllsskreGFI PSNYYV
GEWCEAqt.knngq.GWVPSNYYI
SDWWRVvnltttrqqeGLIPLNFV
LPWWRARrd.knngqqeGYIPSNYYI
RDWWEFRsktvtytpGYYESGYV
EHWWEKVkd.algnvGYIPSNYYV
IHWWRVqrd.rnqhGYVPSNYYL
KDWWEKVe.v..ndrqqGFVPAAYV
VGWMPGlnerttrqrrGDFPSTYYV
PDWWEGgel..nggqrrGVFPASYYV
ENWWEWNGeie..gnrkGIFPATYYV
EEWLEGEcec..kqkvGYIFPKVFFV
GGWWEKGDy.gttriqqQYFPSNYYV
DGWWEWRGsy..ngqrvGWFPSTYYV
QGWWEWRGeie..ygrvGWFPANYYV
GRWWEKArrr..angetGIIPSNYYV
GGWWEWQGel.ksggqkGWAPTNYL
GDWWEArsn.tggenGYIPSNYYV
NDWWEWTrt..nqkeGIFPANYYV
```

Das Profil-HMM dazu



- Schlecht konservierter Block wird ein einziger INS Zustand
 - „85“: Wsk des Übergangs zum selben Zustand (Kreis)
 - Schätzen der Parameter: Später
- Offensichtliches Problem: **Overfitting**
 - Deletions sind praktisch überall „verboten“

Profil-HMM mit Pseudo-Counts



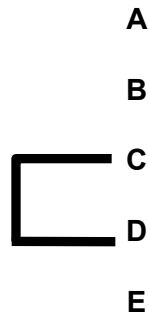
- Profil-HMM, gelernt mit einem Pseudo-Count von 1
- **Smoothing** ist sehr wichtig, da Sequenzfamilien selten 1000nde Sequenzen umfassen
 - Und als Profil-HMM sehr viele Zustände hat

CLUSTAL W: Grundaufbau

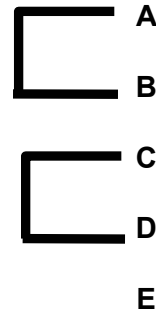
- Gegeben k Sequenzen
- Drei Schritte
 - **Ähnlichkeitsmatrix**: Berechne alle paarweisen Alignmentsscores
 - Konstruiere einen „**Guide Tree**“ durch hierarchisches Clustering
 - Berechne und verschmelze Teil-MSA gemäß dem Guide Tree
- Idee dahinter
 - Aligniere erst sehr ähnliche Sequenzen – **Signale 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**

Konstruktion des Guide Trees

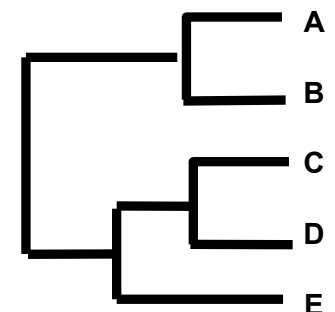
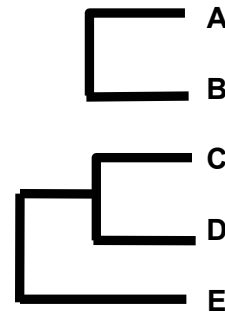
	A	B	C	D	E
A		17	59	59	77
B			37	61	53
C				13	41
D					21



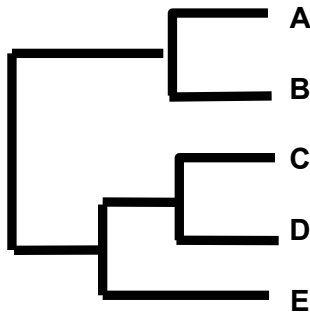
	A	B	E	CD
A		17	77	59
B			53	49
E				31



	E	CD	AB
E		31	65
CD			54



Beispiel



C PADKTNVKAAWGKVG**A**HAGEYGA
D AADKTNVKAAWSKVGGHAGEYGA

A PEEKSAVTALWGKVN**V**DEYGG
B GEEKA**A**VLALWDKVN**E**EEYGG

C PADKTNVKAAWG_**K**VGAHAGEYGA
D AADKTNVKAAWS_**K**VGGHAGEYGA
E AA__TNVKTAWSSKVGGHAPA__**A**

A PEEKSA**V**_TALWG_**K**VN__VDEYGG
B GEEKA**A**_LALWD_**K**VN__EEYGG
C PADKTNV**K**AA_**W**G_**K**VGAHAGEYGA
D AADKTNV**K**AA_**W**S_**K**VGGHAGEYGA
E AA__TNV**K**TA_**W**SS**K**VGGHAPA__**A**

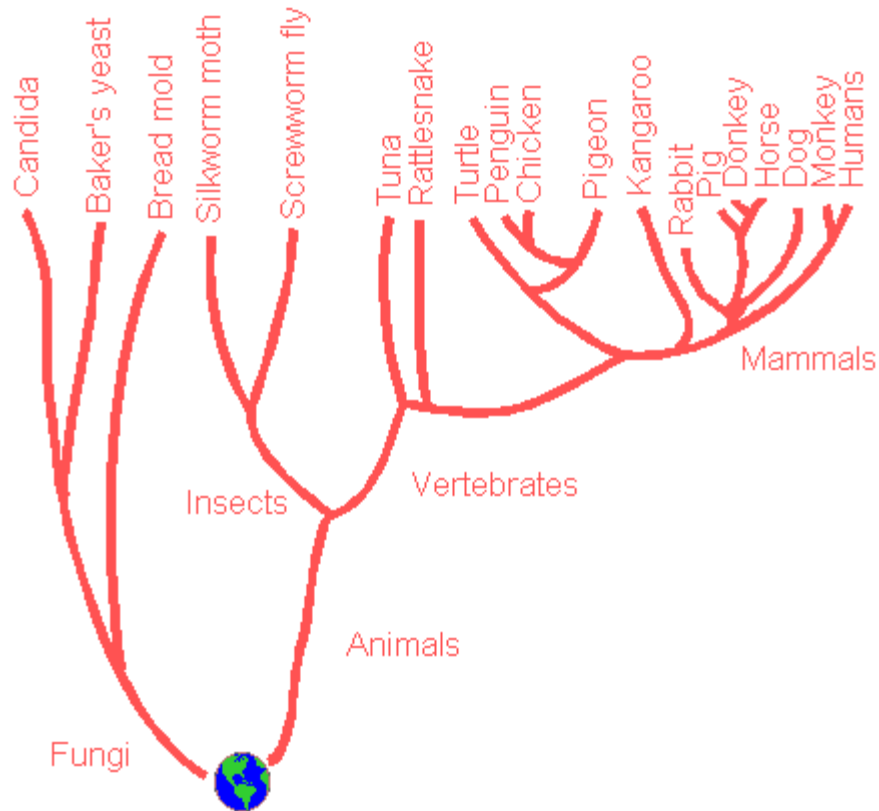
Once a gap, always a gap

Inhalt dieser Vorlesung

- Stammbäume
- Phylogenetische Bäume
- Evolutionsmodell

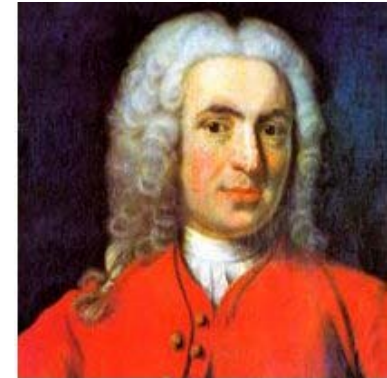


Tree of Life



Stammbäume versus Klassifikation

- Zuerst war die **Klassifikation**
 - Carl Linnaeus, ca. 1740: *Systema Naturae*
 - Annahme: Arten verändern sich nicht
 - Wir sind Prä-Darwin
 - Einteilung der Lebewesen in
 - Kingdoms - classes – orders – families – genera-species
 - SKOFGA (Stamm, Klasse, Ordnung, Familie, Gattung, Art)
 - **Innere Knoten** einer Klassifikation sind abstrakte Klassen, die i.d.R. keine eigenen Vertreter haben können

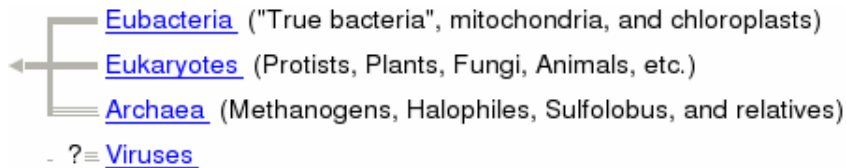


'Bridegroom portrait' of Linnaeus,
Clavis: LINNÆI, M. D.
METHODUS plantarum SEXUALIS
in SISTEMATE NATURÆ
del. ripa



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

Klassifikation



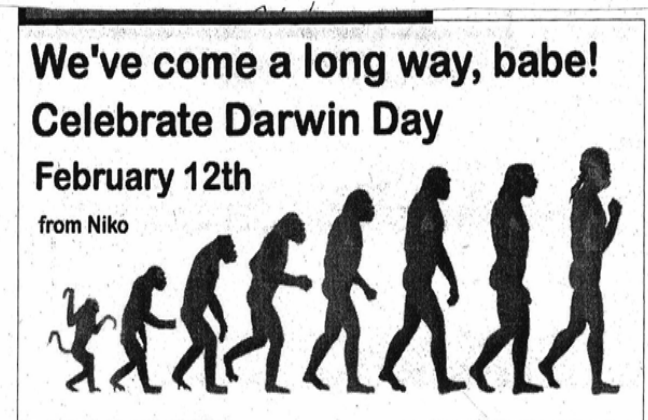
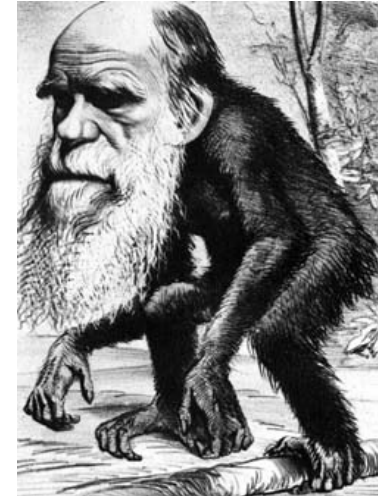
- Eukaryoten
- Tiere
- diverse Zwischenstufen
- Charniata (Schädelknochen)
- Vertebraten (Wirbeltier)
- Viele Zwischenstufen
- Mammals (Säugetiere)
- Eutheria (Placenta)
- Primaten (Affen)
- Catarrhini
- Hominidae (Mensch, Schimpanse, Orang-Utan, Gorilla)
- Homo (erectus, sapiens ...)
- Homo Sapiens

Popular Groups on the Tree of Life

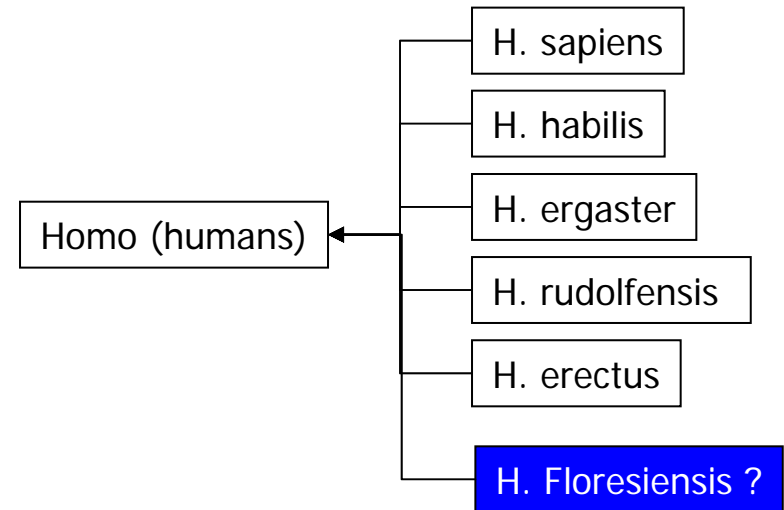


Stammbäume

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



„Homo floresiensis“ (Hobbit?)

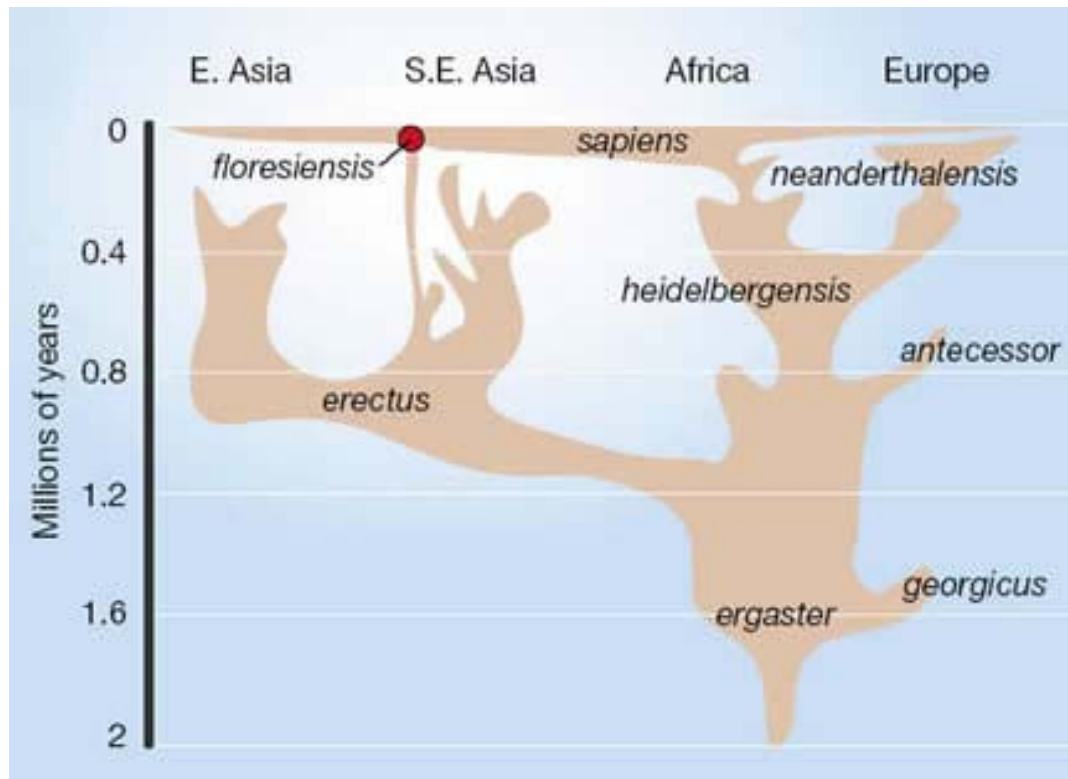


- Entdeckung 2003
- Lebte vermutlich vor 95 000 bis vor 12 000 Jahren auf der indonesischen Insel Flores
- Körperhöhe nur ca. 100 cm
- Gehirnvolumen nur 380 cm³
 - Homo erectus: 600-1200, h. sapiens: 1400
- Rückentwicklung? Vorläufer?

Quelle:

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

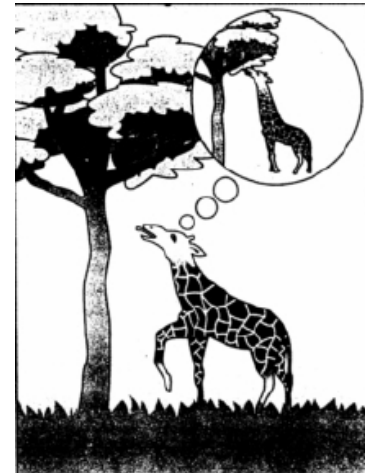
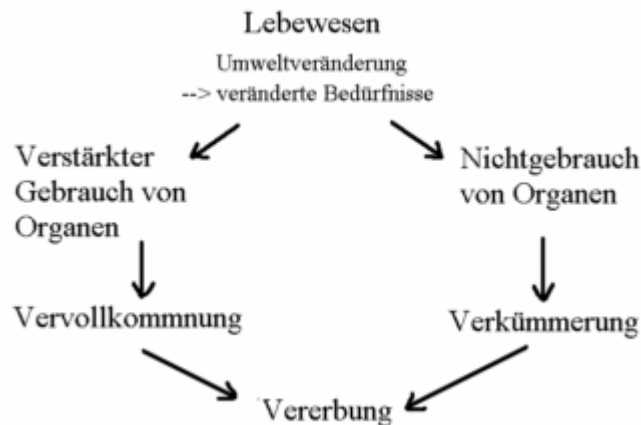
Evolutionsgeschichte neu geschrieben



- Stammbaum – welche Art geht aus welcher hervor?

Theorie I: Lamarck

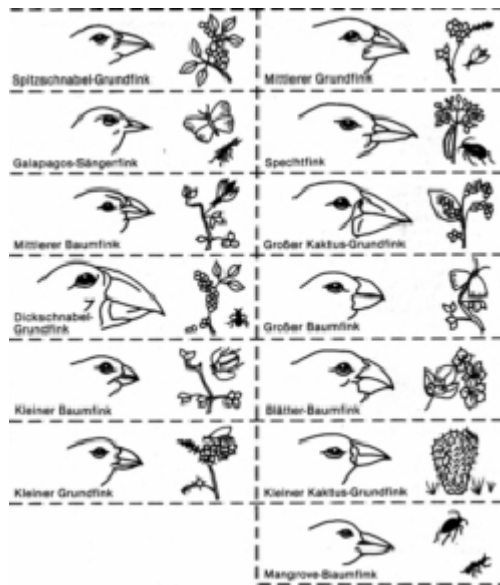
- Jean Baptiste de Lamarck (1744-1829)



- 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

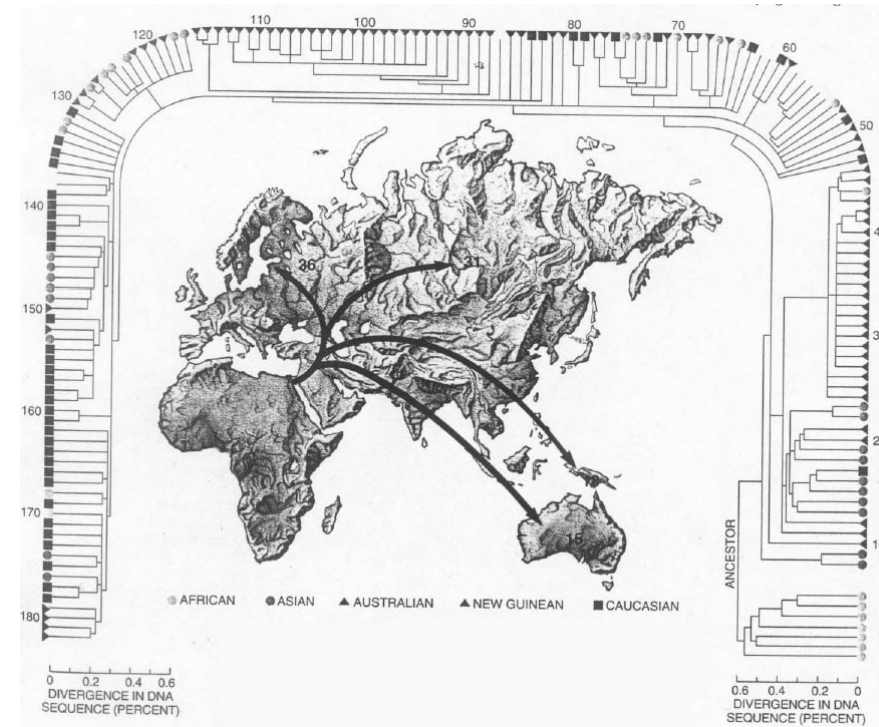
- Charles Robert Darwin (1809-1882)
 - „Die merkwürdigste Tatsache ist die **vollkommene Abstufung** in der Größe des Schnabels ..., von einem Schnabel, der so groß ist wie der eines Kernbeißers bis zu dem eines Buchfinken“



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

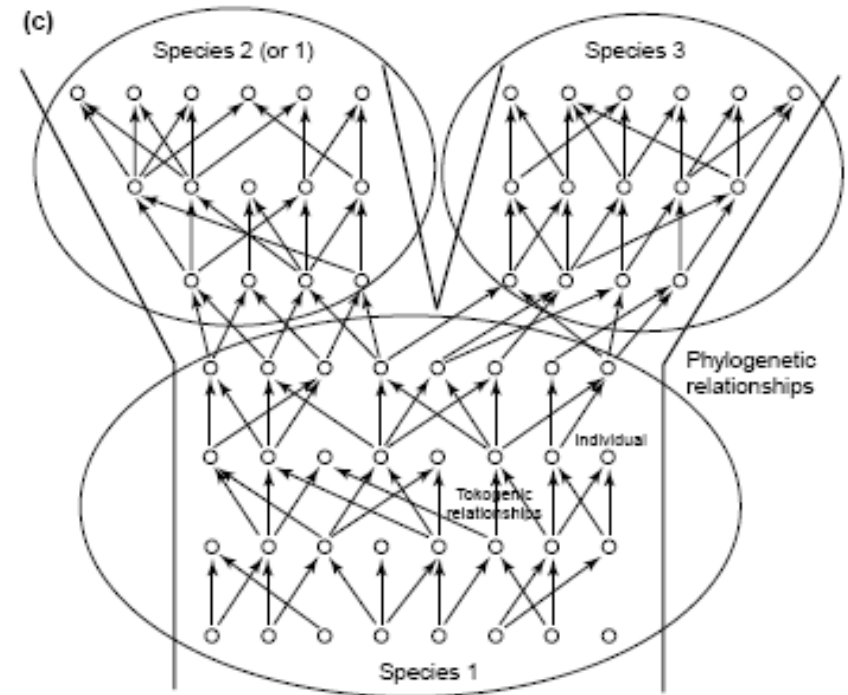
Moderne Stammbaumberechnung

- *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
 - **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**)



TRENDS in Ecology & Evolution

Wozu?

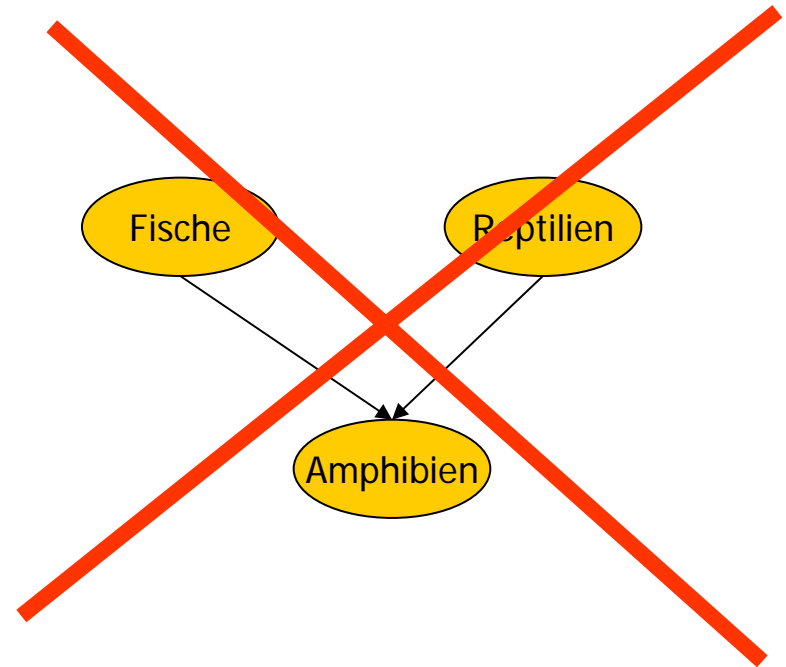
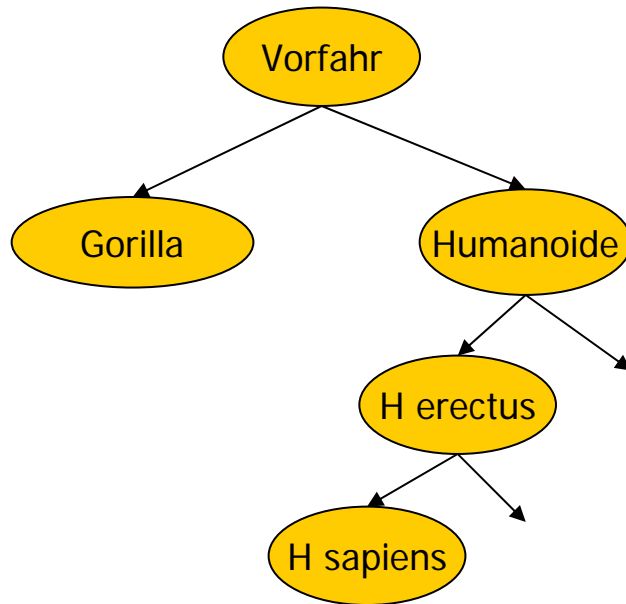
- Aufklärung der evolutionären Verhältnisse
- Aufklärung von familiären Verhältnissen
 - Wenn genügend Sequenzen vorhanden
- Aufklärung von Infektionswegen
 - Bei schnell mutierenden Viren (z.B. HIV)
 - Varianten werden in verschiedenen Personen gefunden
 - Herausrechnen des Verbreitungsweges anhand 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
 - Positive Mutationen führen zu **mehr und lebensfähigeren Nachkommen**
 - Positive Mutationen setzen sich nach und nach in einer Population durch
- **Speziation**
 - Unterschiedliche Mutationen sind in unterschiedlichen Lebensräumen unterschiedlich vorteilhaft
 - Anpassung an ökologische Nischen
 - Führt zur **umgebungsspezifischen Akkumulation von Änderungen**
 - Schließlich geht die gemeinsame Fortpflanzungsfähigkeit verloren

Artenbildung

- Unser Modell: Arten entstehen durch Veränderungen aus **einer** anderen Art
 - Es gibt massenhaft Ausnahmen: Phylogenetische Netzwerke



Mutationsfrequenzen

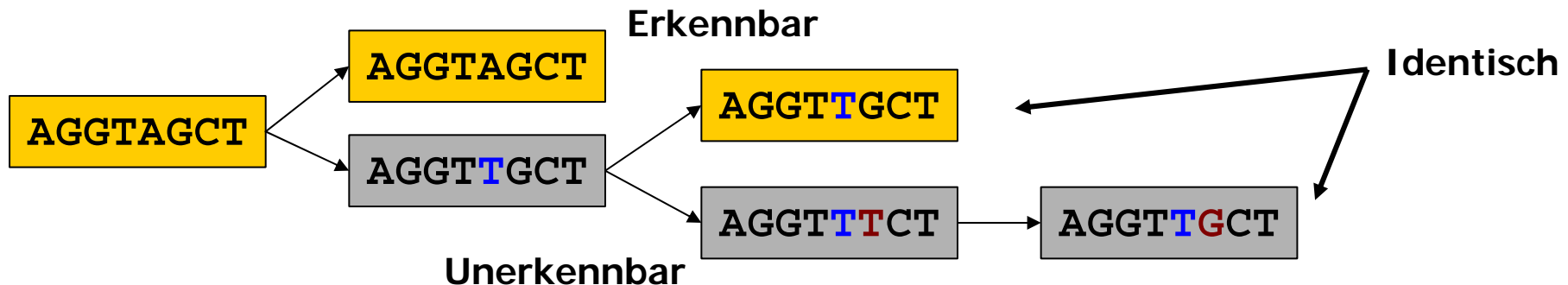
- Treten Mutationen in gleicher Frequenz auf?
 - Sind alle Mutationen gleich wahrscheinlich?
 - Substitutionsmatrizen
 - Gibt es Zeiten erhöhter **Mutationshäufigkeit**?
 - Sonneneruptionen, ...
 - Gibt es Zeiten erhöhten **Selektionsdrucks**?
 - Meteoriteneinschlag, Klimaverschiebungen, ...
 - Gibt 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 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 i.d.R. nicht berechnen
 - Man kennt die ausgestorbenen Arten nicht
 - Man kann ausgestorbene Mutationen nicht erkennen
 - Man kann Doppelmutationen nicht erkennen

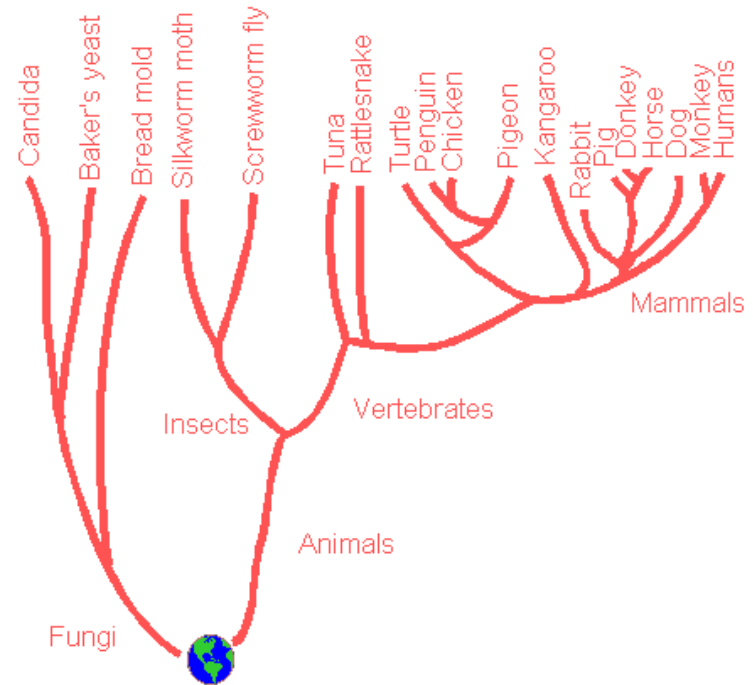


Weitere Probleme

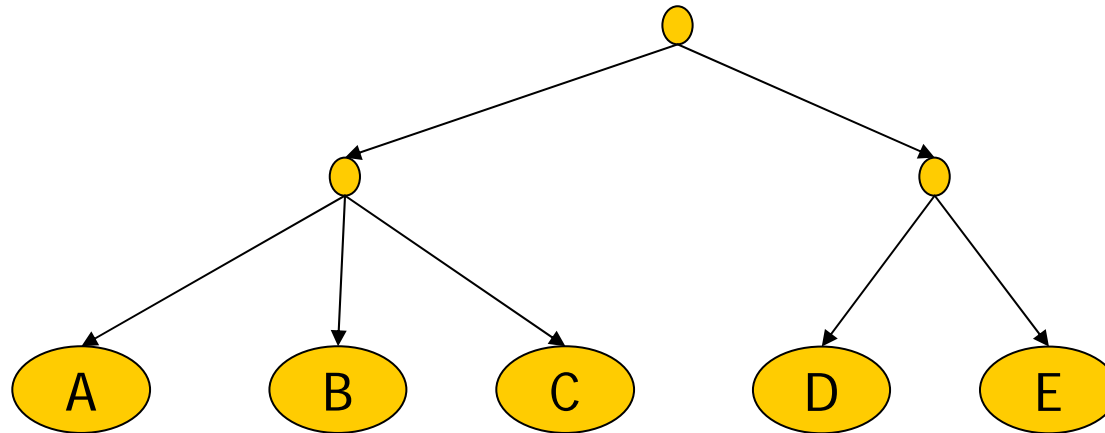
- **Gene Fusion**
 - Gene in einem Genom können verschmelzen
 - Also hat eine Sequenz zwei Vorfahren
- **Hybridisierung / Kreuzungen**
 - Pflanzengenome können sich kreuzen
 - Erzeugt Spezies mit zwei Vorfahren
- **Horizontal gene transfer**
 - Gerade im Viren / Bakterienbereich bekannt
 - Genetree und Speziesstammbaum stimmen nicht mehr überein
 - Speziesbegriff ist sehr problematisch bei Lebewesen ohne sexuelle Reproduktion
- **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

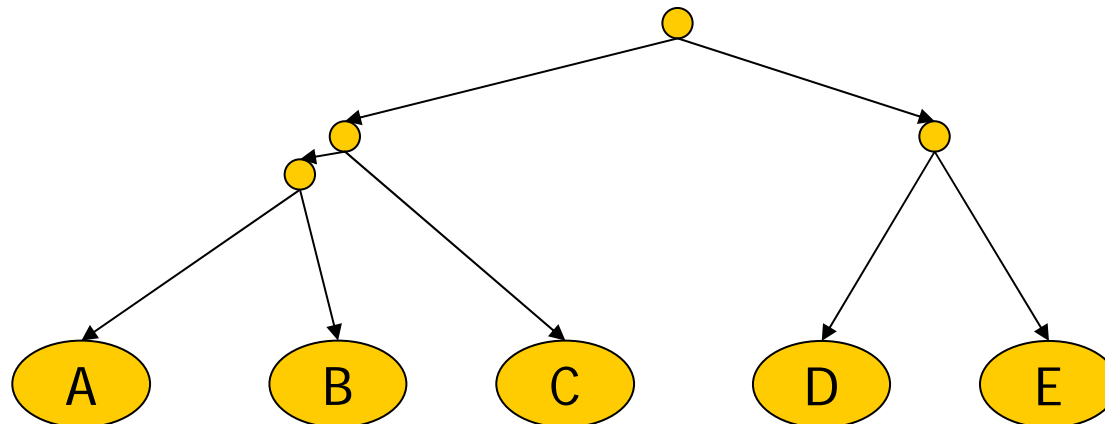
- Knoten = Arten
- Blätter = Lebende Arten
- Kanten
 - Länge kann (aber muss nicht) mit zeitlicher Entfernung korrelieren
 - Scaled trees
- Jeder Knoten hat exakt einen Vater
- Wurzel oder nicht
- Binäre Bäume
 - Unproblematisch (gleich)
- Reihenfolge der Kinder ist egal
- Viele Visualisierungsvarianten



Binäre versus Multifurcation Trees

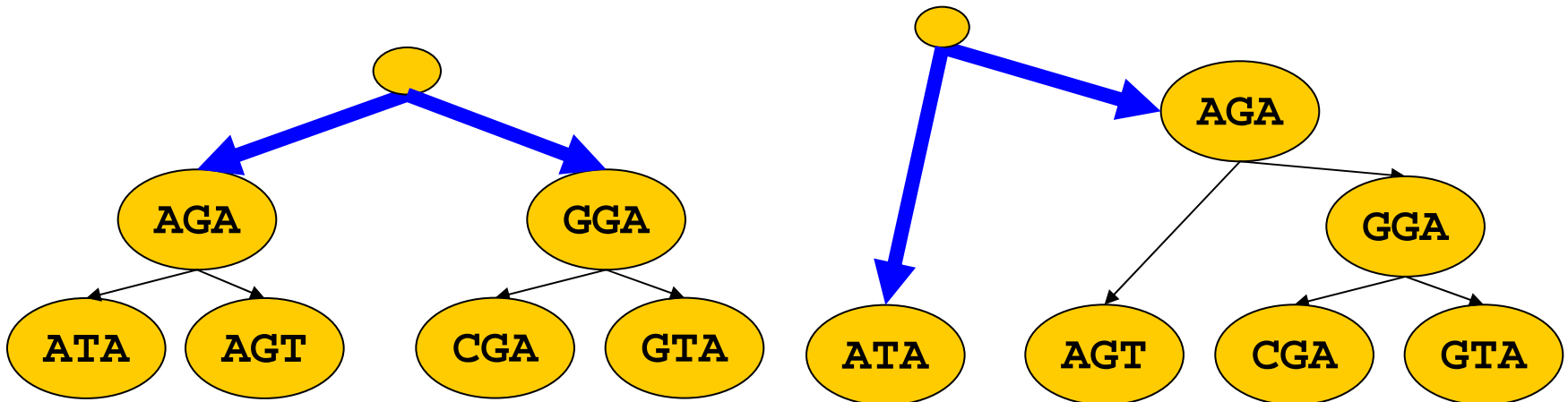
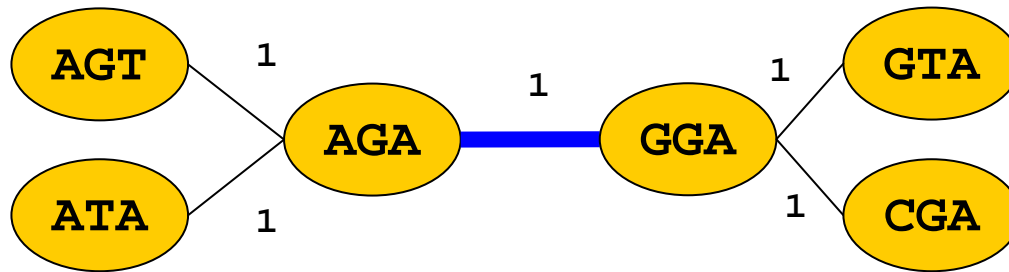


- Erscheinen als
 - Eines der Paare (A, B), (B,C), (A,C) wird **minimal ähnlicher** sein



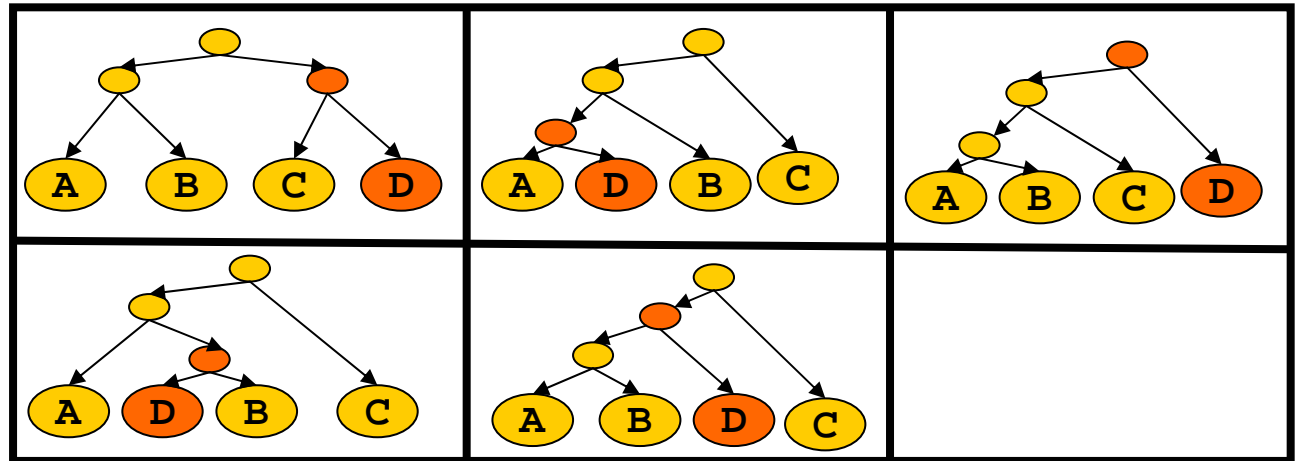
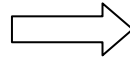
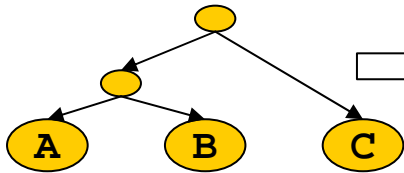
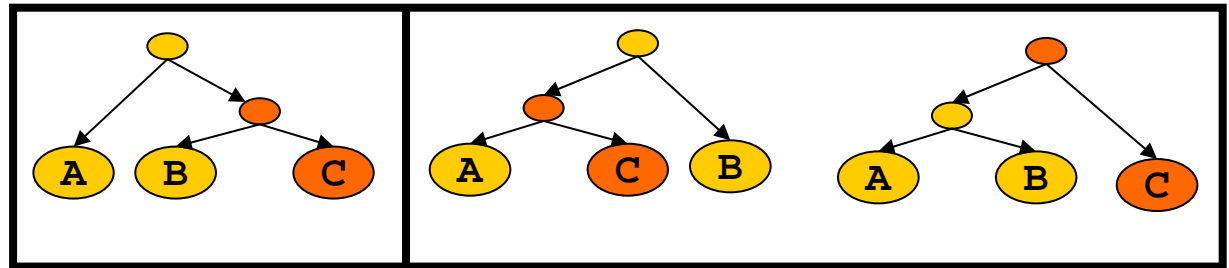
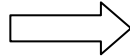
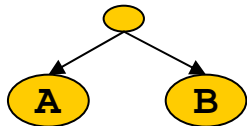
Bäume ohne Wurzeln

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



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 (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 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