Phylogenien und Methoden zu ihrer Rekonstruktion

Seminar Bioinformatik: Algorithmische und statistische Verfahren der strukturellen Genomanalyse

Niklas von Öhsen, 30.10.00

Übersicht

- Geschichte und Beispiele
  - Darwin und Kreationismus
  - Geschichte des Organismenstammbaums
  - Geschichte der Phylogenierekonstruktionsmethoden
- Charakterbasierte Methoden
  - Perfekte Phylogenien
  - Binäre perfekte Phylogenien
  - Parsimoniekriterien
- Metrikbasierte Methoden
- Maximum Likelihood
  - Evolutionsmodelle:
    - General Time Reversable
    - Jukes-Cantor

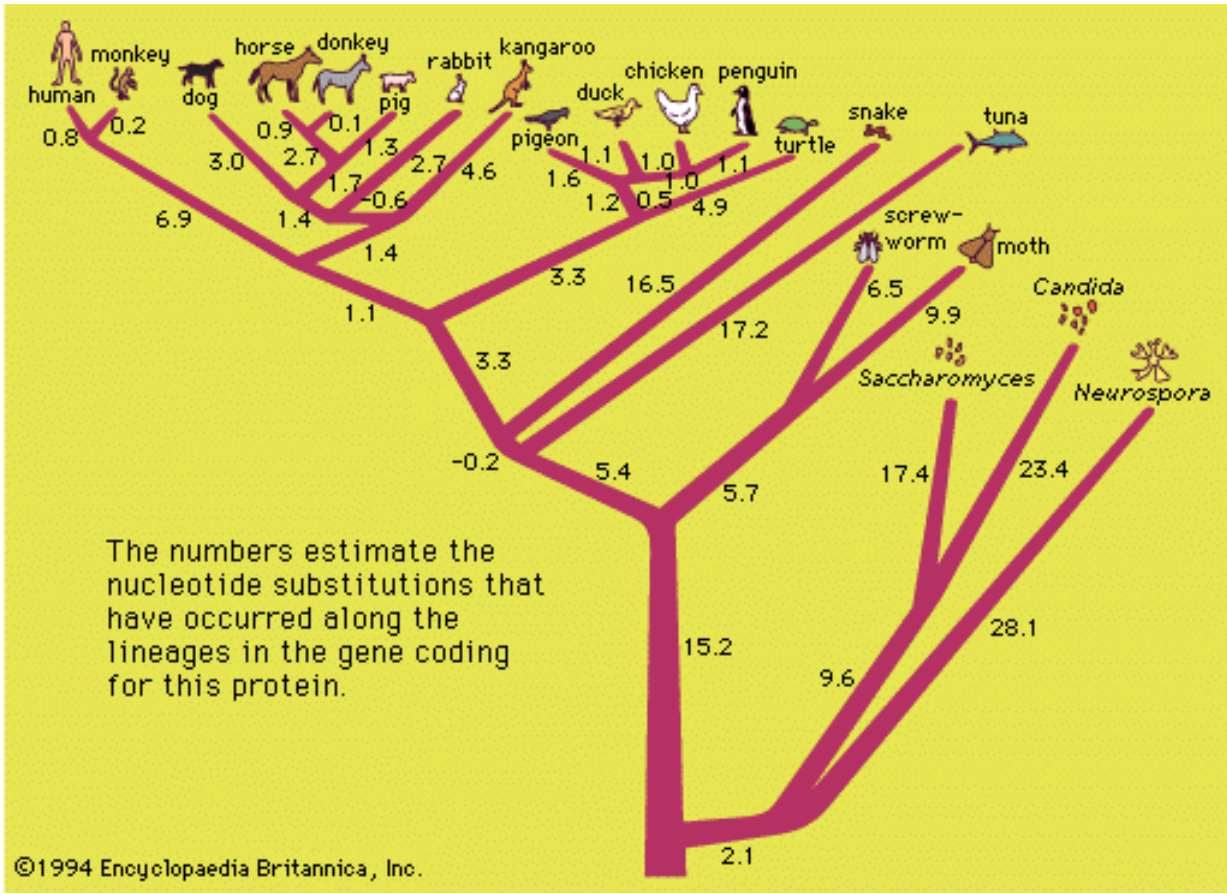
Phylogenie: (lt. Encyclopedia Britannica)

- Die Evolutionsgeschichte einer Spezies oder Gruppe, insbesondere mit Blick auf die Abstammung und Beziehungen zwischen größeren Organismengruppen

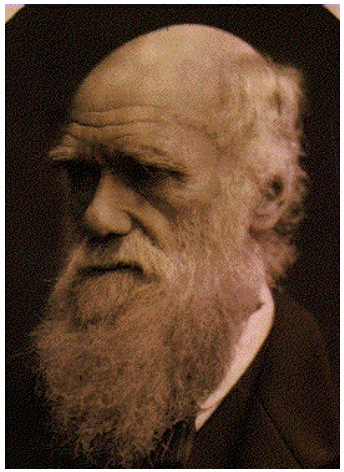
Fitch/Margoliash

Cytochrome c – Baum (1967):

(Protein mit wichtiger Funktion im Zellstoffwech-sel)



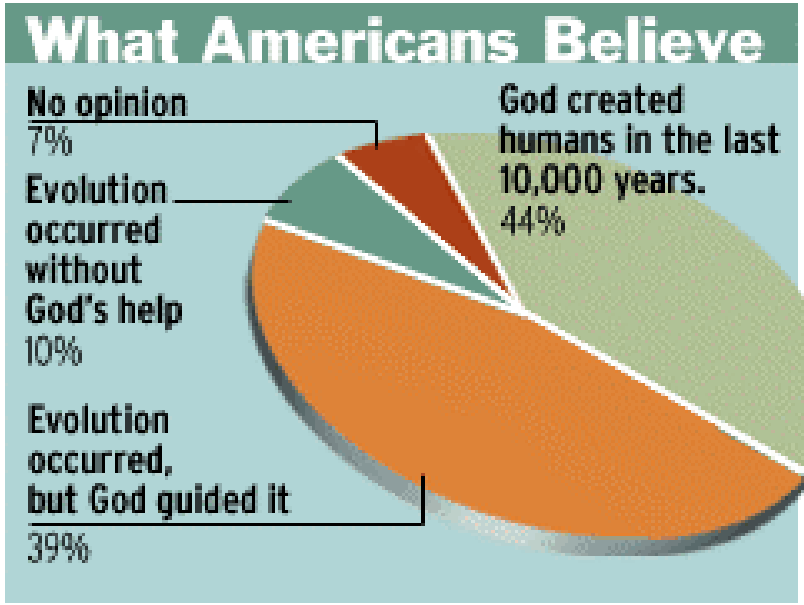
- Carl von Linné 1758: Systema Naturae
  - Formale Klassifikation der Organismen (zweiteilige lateinische Standardnamen)

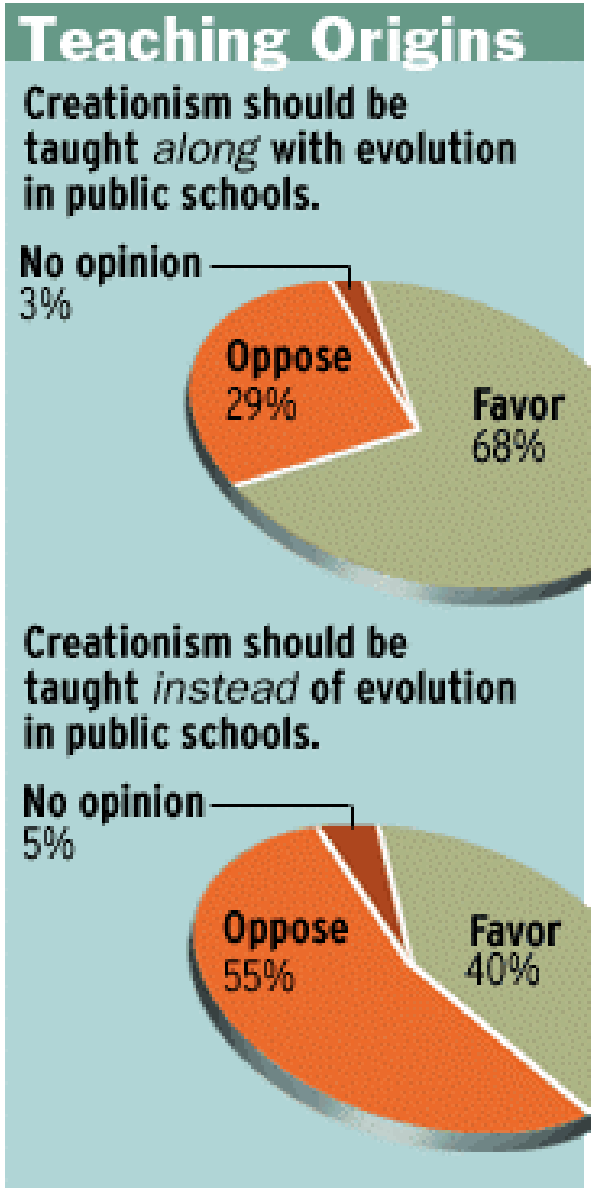


- Charles Darwin 1859: The Origin of Species
  - Theorie der Entstehung der Arten aus gemeinsamen Vorfahren

- USA: Seit über hundert Jahren Debatte über die Unvereinbarkeit von Evolutionstheorie mit der Schöpfungsgeschichte (starke Fraktion christlicher Fundamentalisten, die die Bibel wörtlich interpretieren)
- Kreationismus: Die Erde wurde ca. 6000 vor Christus von Gott erschaffen, inklusive aller Arten (keine Theorie im wissenschaftlichen Sinne, da prinzipiell nicht widerlegbar)
- Scopes trial 1925: Ein Lehrer (J. Scopes) wurde in einem Showprozess verklagt, gegen ein Gesetz verstoßen zu haben, das das Unterrichten von Darwins Evolutionstheorie untersagt.
- Dies Gesetz von Tennessee hatte 42 Jahre bestand.

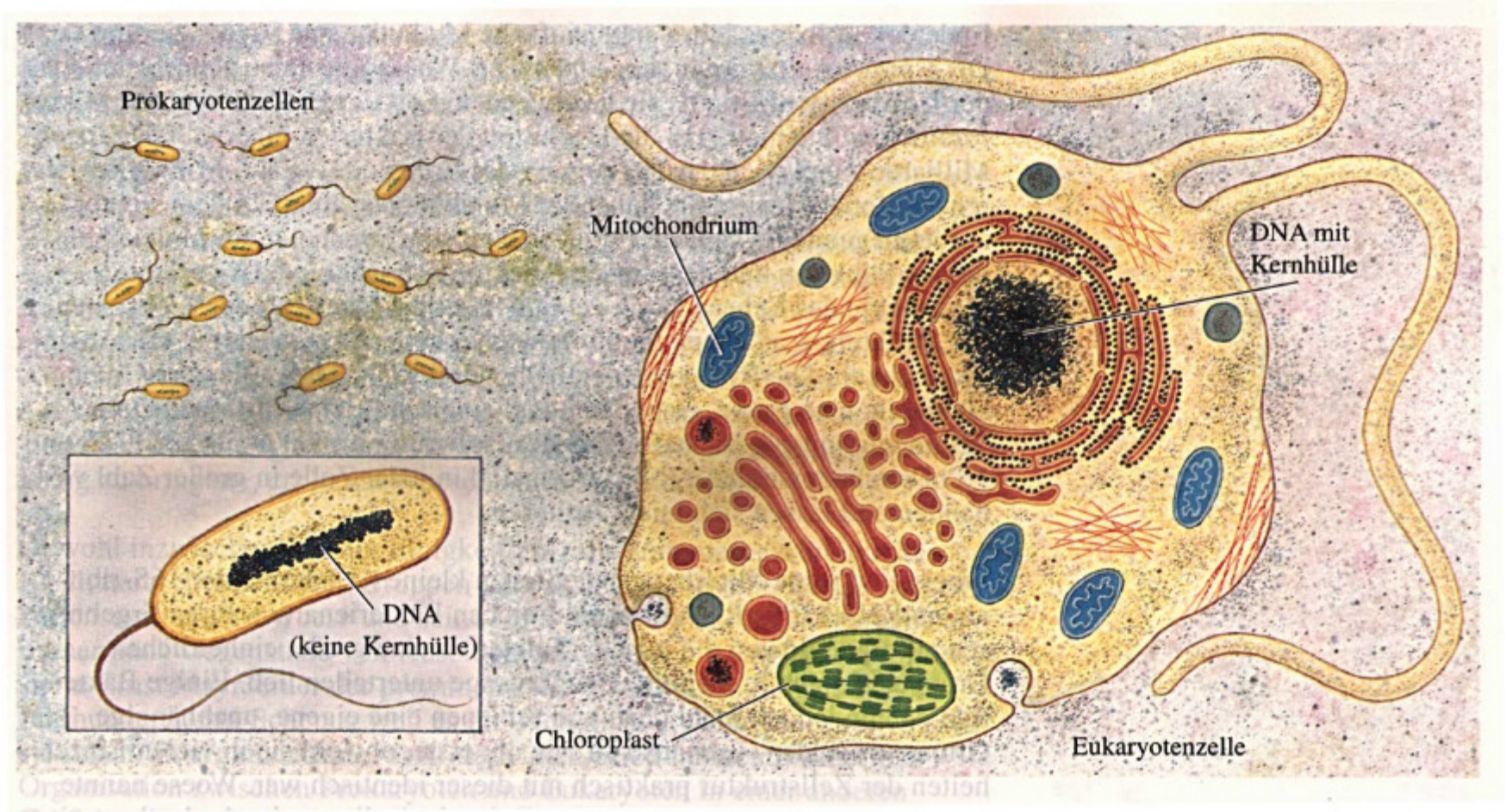
- 1987: Supreme Court urteilt: Kreationismus darf an US-Schulen nicht unterrichtet werden.
- 1999: Kansas State Board of Education streicht „Makroevolution“ aus dem Standard für high school-Absolventen
- 1997 Gallup-Studie:



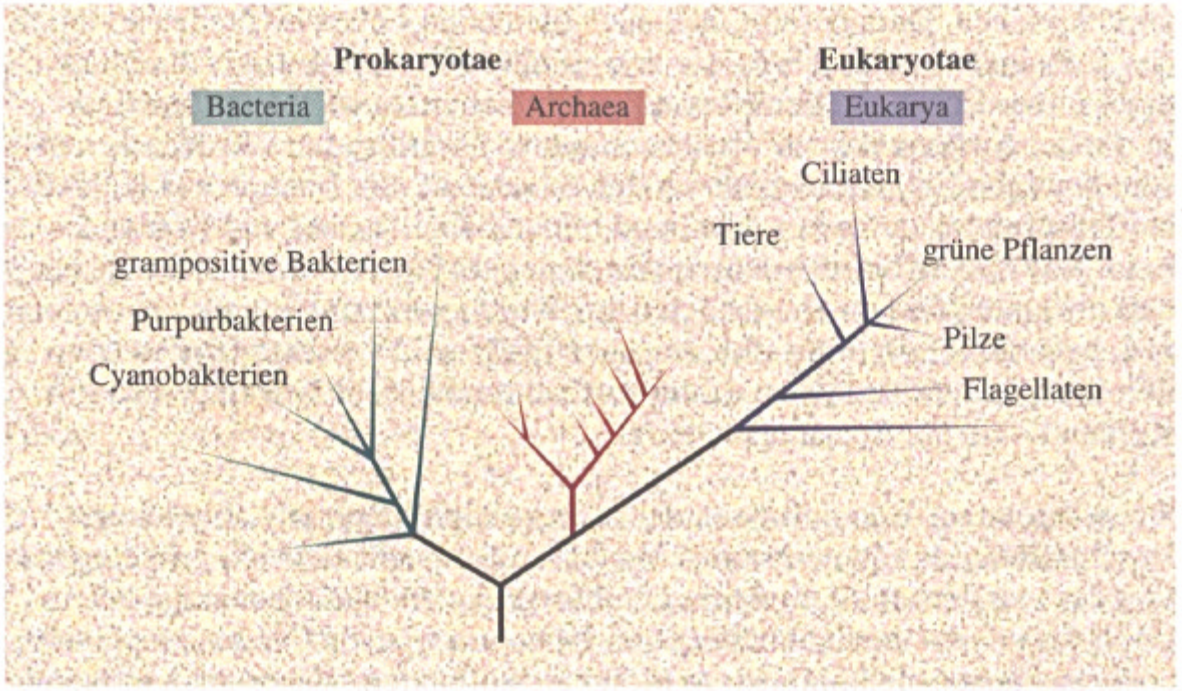


- Frühe Naturphilosophen:
  - Zwei Klassen: Animalia, Plantae
- Vor ca. drei Jahrhunderten:
  - Entdeckung der Mikroorganismen und Einpassen ins Zwei-Klassen-Schema
- Mitte 19. Jh (Ernst Haeckel):
  - Protisten (Einzeller bis auf Bakterien) können so nicht klassifiziert werden (Photosynthese und Bewegung = ?) also:
  - Drei Klassen: Animalia, Plantae, Protista
- Anfang 20. Jh:
  - Bakterien (Monera) werden gesondert behandelt
  - Vier Klassen: Animalia, Plantae, Protista, Monera
- 1959 (Whittaker)?:
  - Pilze (Fungi) werden gesondert klassifiziert
  - Fünf Klassen: Animalie, Plantae, Fungi, Protista, Monera

- Parallel dazu:
  - Zellbiologen entwickeln Zwei-Reiche-Schema:
  - Prokaryoten, Eukaryoten

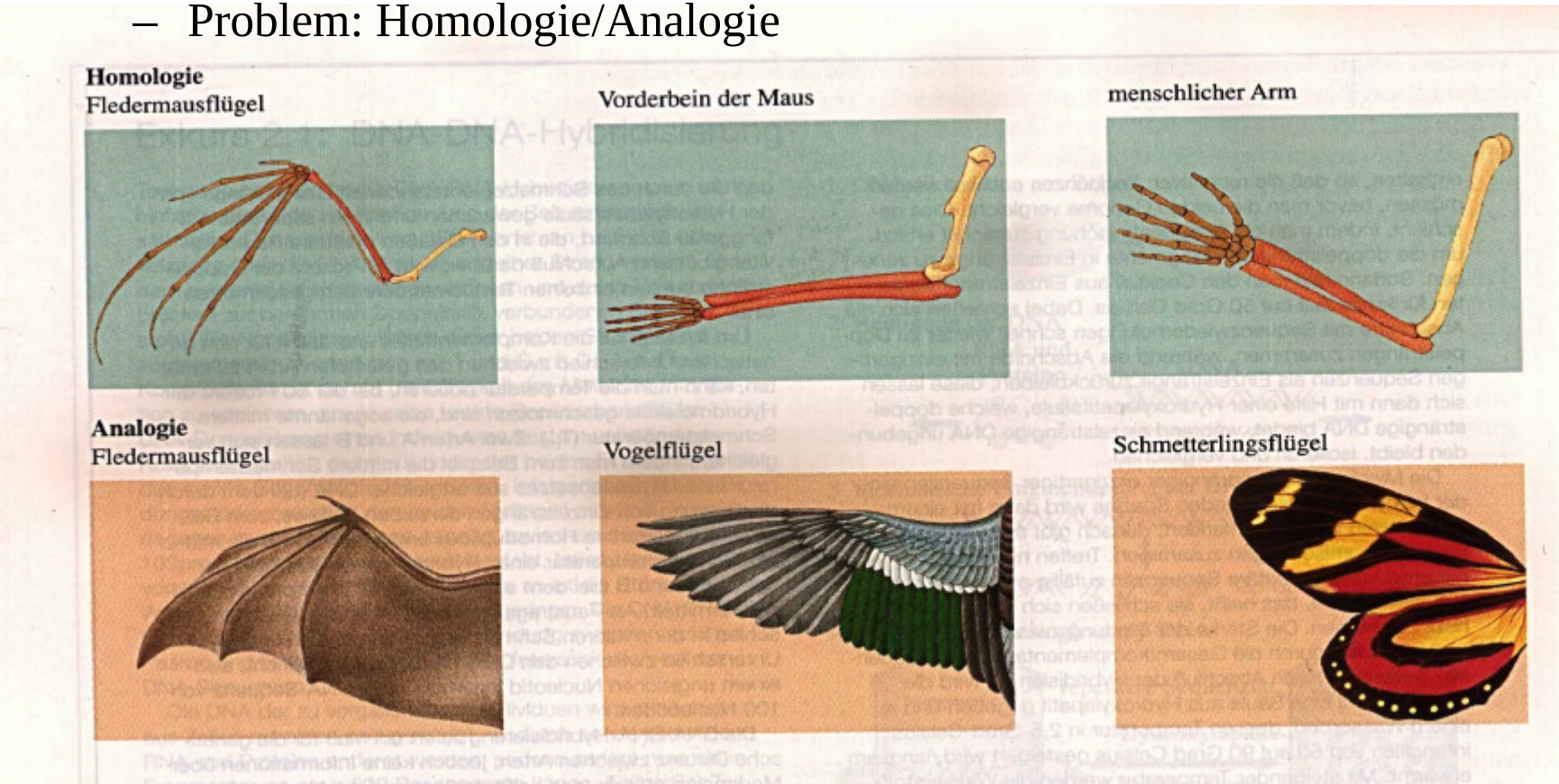


- Ende 60er Jahre (Woese):
  - Prokaryoten müssen in Archae und Bacteria unterteilt werden (Untersuchung von 16S-rRNA)
  - Archae, Bacteria, Eukaryotae



- 1999 (Doolittle):
  - Rekonstruktion der ursächlichen Abstammungsbeziehungen evtl. unmöglich, da durch vielfältigen lateralen Gentransfer nicht baumförmig

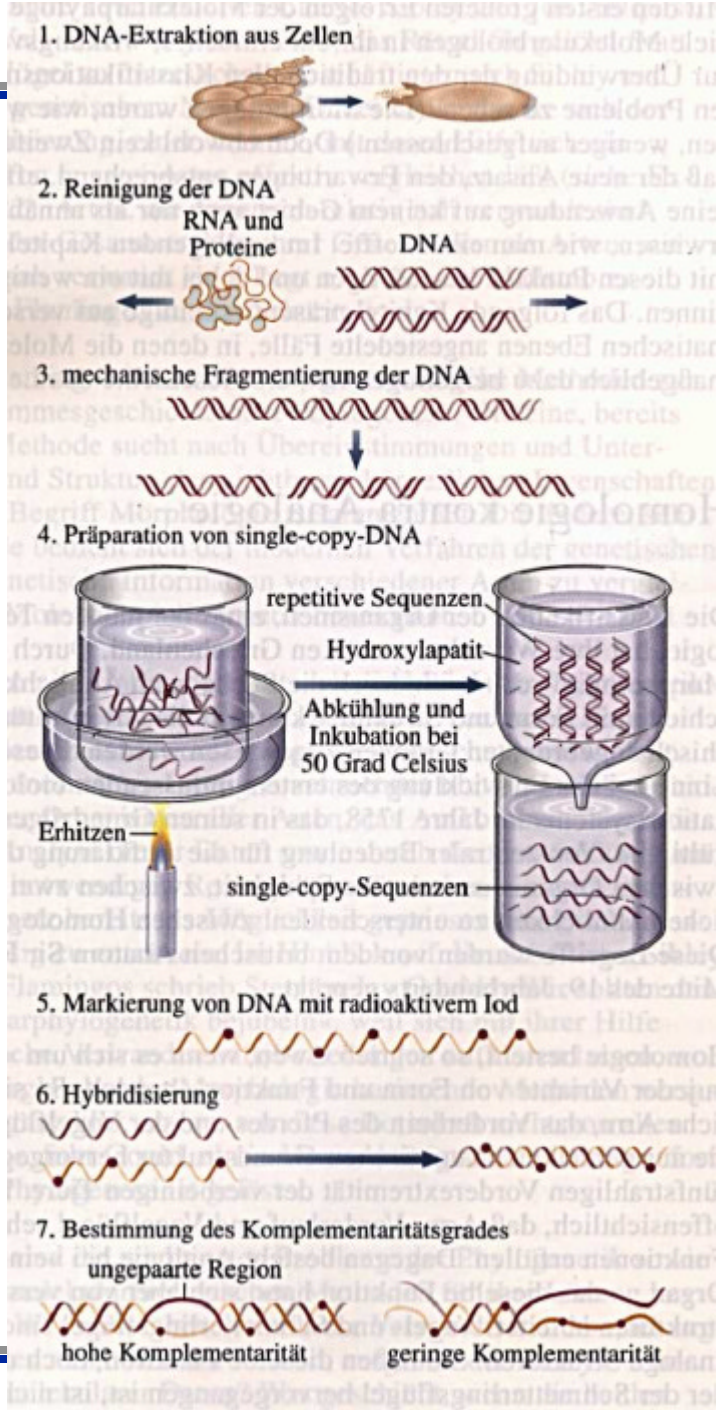
- Vor 1950: „Vergleichende Morphologie“
  - keine standardisierten Methoden, größtenteils „intuitive“ Rekonstruktion
  - Problem: Homologie/Analogie



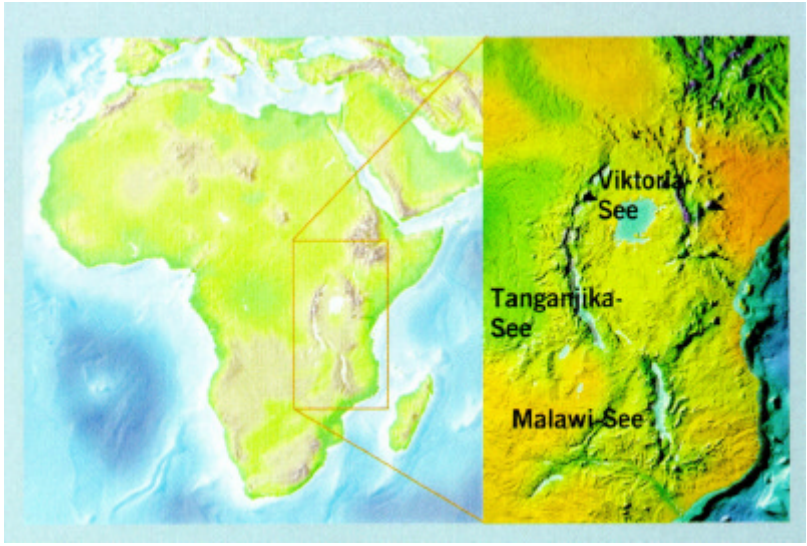
- 1950:
  - Willi Hennig (deutscher Entomologe) beschreibt in seinem Buch „Grundzüge einer Theorie der **Phylogenetischen Systematik**“ den **kladistischen** Ansatz.
    - Ziel: Rekonstruktion der „wahren“ Abstammungsbeziehungen
    - Für jede Klasse von Organismen werden Merkmale („X ist (nicht) vorhanden“) in primitive und abgeleitete unterteilt. Anhand dieser wird die Phylogenie bestimmt.
    - Subjektivität wird beschränkt auf die Auswahl der Merkmale
- 60er Jahre:
  - Hennigs Buch wird ins Englische übersetzt und findet Verbreitung
  - Parallel dazu wird der **phänetische** Ansatz entwickelt
    - Phänetik oder **numerische Taxonomie** versucht eine Gruppierung aufgrund von morphologischer Ähnlichkeit zu finden (nicht unbedingt den Stammbaum)
    - morphologische Merkmale sind reelle Größen (z. B. bestimmte Knochenlängen)
- => „Phänetik-Kladistik-Krieg“

Phylogenierekonstruktion III

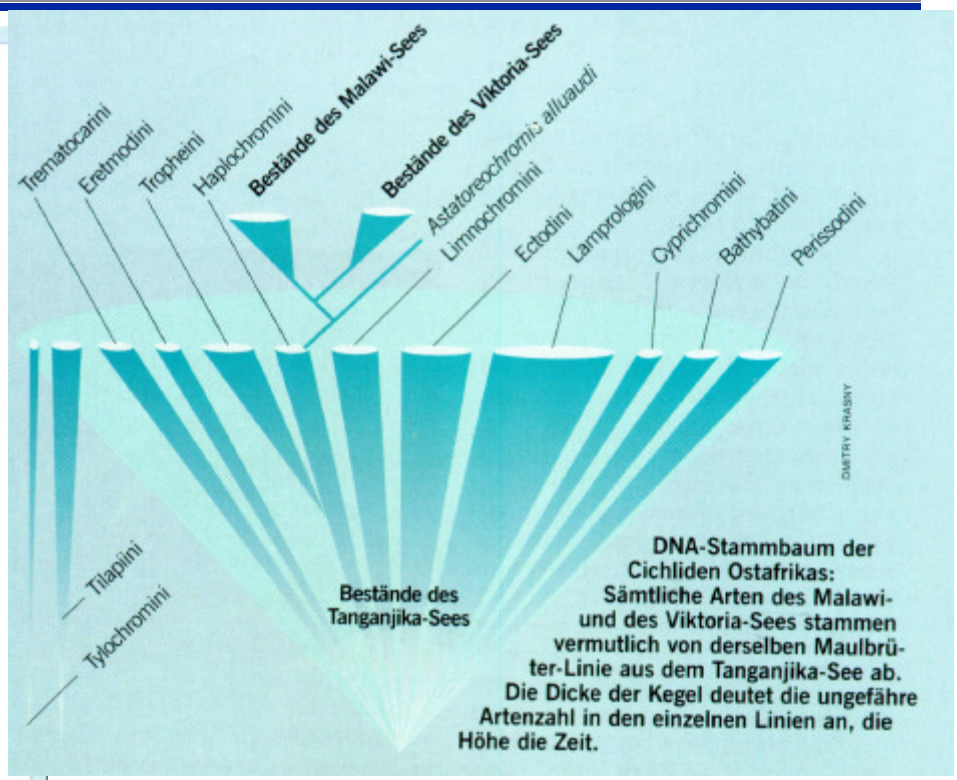
- Ab 60er Jahre:
  - Molekularbiologische Daten erhalten Einzug in die Phylogenierekonstruktion.
  - Einige Verfahren, die die aufwändige Sequenzierung umgehen, liefern numerische Ähnlichkeitsmaße für die Verwandtschaft zweier Spezies
  - z. B. DNA-DNA-Hybridisierung
- Akzeptanz molekularer Methoden wird teilweise behindert durch zwangsweise phänetischen Ansatz zur Phylogenierekonstruktion



- 1985 Fitch/Atchley:
  - Vergleich der morphologischen und molekularbiologischen Methoden zur Phylogenierekonstruktion anhand des über 70 Jahre dokumentierten Stammbaums von Labormäusen
  - Molekularbiologische Daten waren den morphologischen Daten (Unterkiefermaße 10 Wochen alter Mäuse) überlegen
- 1990 Wilson, Meyer.:
  - Molekularbiologische Analyse von afrikanischen Buntbarschen liefert verblüffendes Ergebnis



Buntbarsche in Afrika



=>Rein morphologische Betrachtungen können erheblich fehlleiten!

- Bei den Daten wird unterschieden zwischen
- Diskreten **Charakteren** (mit endlich vielen Zuständen) z. B.
  - Schnabelform
  - Anzahl der Finger
  - Nukleotid an einer gewissen DNA-Position
  - Hier kann noch zwischen **geordneten** und **ungeordneten** Charakteren unterschieden werden, je nachdem ob eine natürliche Ordnung der Zustände existiert (Zustandsübergänge müssen entlang dieser Ordnung stattfinden):
    - Anzahl der Finger: geordnet (natürliche Zahlen)
    - Nukleotide: ungeordnet
- Ähnlichkeits- oder Distanzmaßen z. B.
  - Länge eines Fingers
  - DNA-DNA-Hybridisierungstemperatur
  - ...

Baum/Phylogenie: zyklenfreier, zusammenhängender, ungerichteter Graph

Kompatibler Charakter: Ein Charakter heißt **kompatibel** mit einem Baum T, wenn die Knoten des Baums so auf die Zustände des Charakters abgebildet werden können, daß jeder Charakterzustand einen Subbaum induziert (insb. zusammenhängend).

Perfekte Phylogenie für eine Menge von Charakteren C: Ein Baum T heißt **perfekte Phylogenie** für C, wenn jeder Charakter aus C mit T kompatibel ist.

Problem der perfekten Phylogenie (PPP): Gegeben eine Menge O von n Objekten sowie eine Menge von m Charakteren C mit jeweils maximal r Zuständen, existiert eine perfekten Phylogenie für C?

- Konstruktion?

- Für ungeordnete Charaktere ist das PPP NP-vollständig
  - Bodlaender, Fellows & Warnow 1992
  - Steel, 1992

- r=4: $O(n^2 m)$ Kannan, Warnow 1994
  - Wichtig für Nukleotiddaten
- r=3: $O(n m^2)$ Dress, Steel 1992

- r=2: $O(n m)$ Gusfield 1991

Satz: Sei C eine Menge binärer Charaktere,

$$j \in C : O_j := \{ o \in O \mid o(j) = 1 \}$$

Dann gilt: C hat eine PP <=> für alle i, j gilt mindestens eine der Aussagen:

$$O_j \subseteq O_i \qquad O_i \subseteq O_j \qquad O_j \cap O_i = \emptyset$$

=>Einfacher Test für PP kann in O(n m²) sofort durchgeführt werden

- Beispiel:

Obj\Chr	1	2	3	4	5
A	1	1	0	0	0
B	0	0	1	0	0
C	1	1	0	0	1
D	0	0	1	1	0
E	0	1	0	0	0

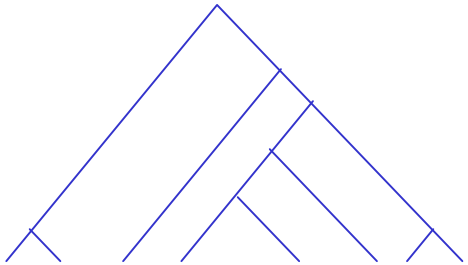
- engl. parsimony: Sparsamkeit, Geiz
- Wenn keine perfekte Phylogenie existiert, finde eine möglichst wenig „unperfekte“
  - Alternativ: Kompatibilität, lasse (möglichst wenig) Charaktere fallen, bis wieder eine perfekte Phylogenie existiert
- Optimalitätskriterien für die Bewertung von Phylogenien. Gesucht wird im allgemeinen der „most parsimonious tree“
- Alle Kriterien bewerten die Anzahl von Zustandsübergängen, die ein Baum benötigt, um die vorhandenen Daten zu erklären.
- Unterschiede ergeben sich durch die verschiedene Wertung / Zulassung von Zustandsübergängen

Fitch und Wagner Parsimonie

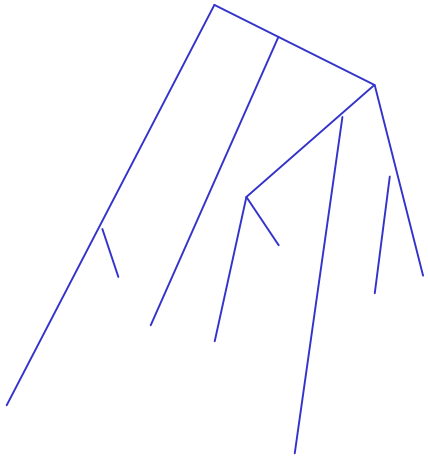
- Wagner:
  - Binäre oder geordnete Charaktere
  - Zustandsübergänge werden symmetrisch behandelt
  - Zustandsübergänge können nur entlang der Ordnung erfolgen
- Fitch:
  - Auch ungeordnete Charaktere
  - Zustandsübergänge ebenfalls symmetrisch
  - Zustandsübergänge zwischen allen Zuständen direkt möglich
- Diverse ... (Dollo, Camin-Sokal, Transversion)

Fitch-Kriterium: Beispiel-Algorithmus

- Einfacher Algorithmus zum Ausrechnen der Anzahl nötiger Übergänge bei gegebenem Baum (Fitch):
  - Baum an einem Blatt beliebig rooten
  - Jedem inneren Knoten werden Teilmengen der Zustandsmenge des Charakters zugordnet:
  - Einem inneren Knoten, dessen beide Tochterknoten schon bearbeitet wurden (post-order-traversal) ordne
    - Die Vereinigung der beiden Tochter-Zustandsmengen zu, falls diese disjunkt sind. In diesem Fall Zähler um eins erhöhen
    - Den Durchschnitt beider zu, falls nicht. In diesem Fall erhöhe den Zähler nicht.
  - Falls der root-Zustand am Ende nicht in der letzten Menge drin ist, Zähler um eins erhöhen.



Ultrametrik



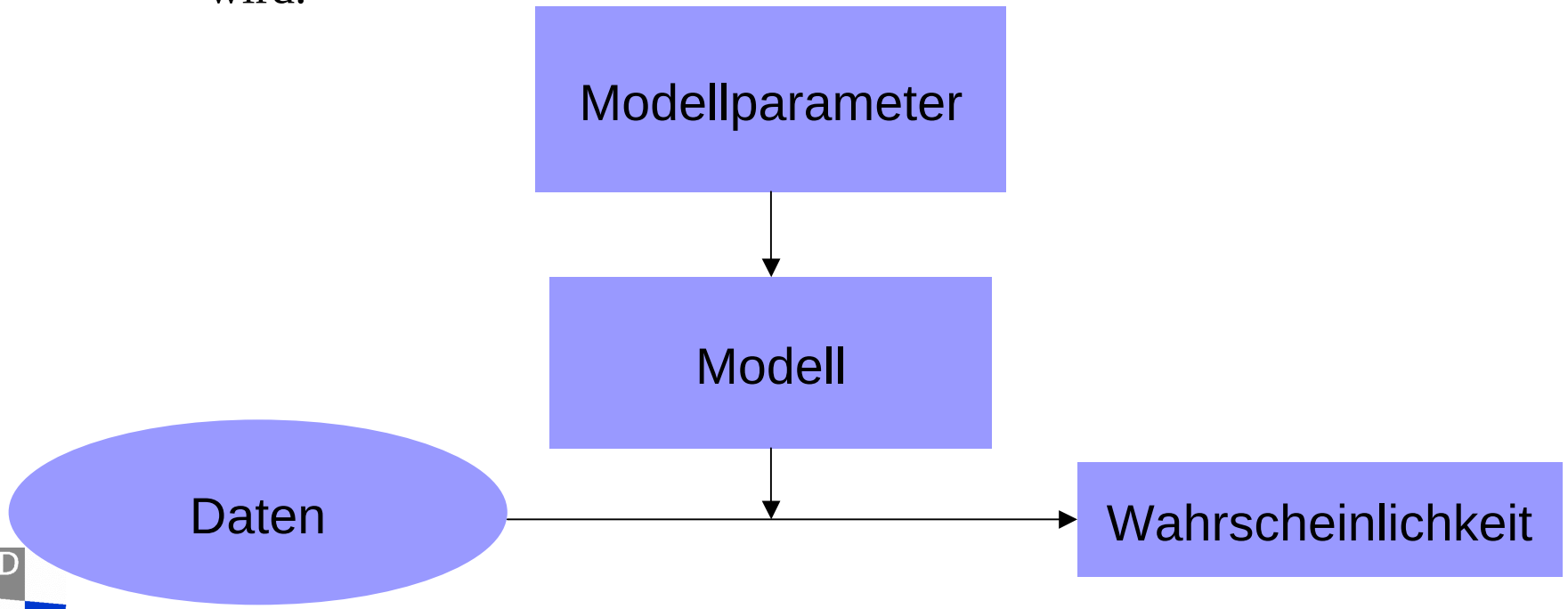
Additive Metrik

- Alle Verfahren basieren darauf, daß sie Ultrametriken oder additive Metriken aus gegebenen paarweisen Distanzen rekonstruieren können
- Clustermethoden:
  - Single linkage, complete linkage
  - UPGMA
  - Nearest Neighbor
- Vier-Punkt-Verfahren
  - Buneman-Tree
  - Split decomposition
- Minimierung des Abstands der rekonstruierten Metrik von der ursprünglichen:
  - Fitch-Margoliash
  - L^p-Normen

- Problem: Um eine Metrik zu bekommen, muss aus den Daten das Alter des LCA (least common ancestor – letzter gemeinsamer Vorfahre) rekonstruiert werden – oder zumindest eine dazu proportionale Größe

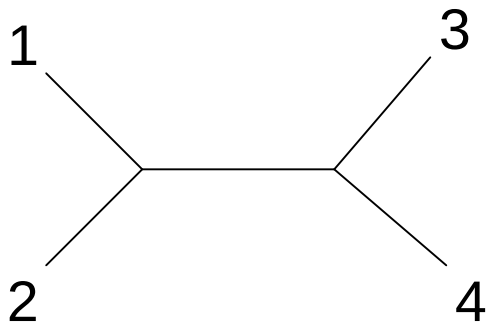
- Vorteil: Geschwindigkeit (insbesondere Clustermethoden)
  - Für viele Datensätze die einzige Möglichkeit in „endlicher“ Zeit zum Ergebnis zu kommen

- (ungefähr: größte Plausibilität) Allgemeines stochastisches Prinzip, um Parameter eines stochastischen Modells zu schätzen, aus dem ein gegebener Datensatz entstanden sein soll.
  - Schätze den/die Parameter so, daß die Wahrscheinlichkeit, daß die Daten aus dem so parametrisierten Modell entstehen, maximiert wird.



- Erste Benutzung von max. Likelihood in diesem Zusammenhang 1967 (Cavalli-Sforza & Edwards)
- Anwendung auf Nukleotiddaten: Felsenstein 1981 (auch Autor des PHYLIP- Pakets)
- Beispiel: Berechnung des Likelihood-wertes:

```
1:...CA C GT...
2:...CA C CT...
3:...TA A GT...
4:...TA G GC...
```



1. Beliebig an innerem Knoten rooten
2. Alle möglichen Zustände der inneren Knoten erstellen und die Wahrscheinlichkeiten der Konstellationen aufaddieren
3. Log-likelihoods der verschiedenen Sequenzpositionen aufaddieren

- Beschreibung der Übergänge zwischen Nukleotiden durch eine 4 x 4 Rate-Matrix Q, in der die Übergangsgeschwindigkeiten stehen (Substitutionen pro Zeiteinheiten)
- Allgemeines Modell:

$$Q = \begin{pmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu g\pi_A & -\mu(g\pi_A + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu h\pi_A & \mu j\pi_C & -\mu(h\pi_A + j\pi_C + f\pi_T) & \mu f\pi_T \\ \mu i\pi_A & \mu k\pi_C & \mu l\pi_G & -\mu(i\pi_A + k\pi_C + l\pi_G) \end{pmatrix}$$

- GTR: General Time Reversible Model

$$Q = \begin{pmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu a\pi_A & -\mu(a\pi_A + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu b\pi_A & \mu d\pi_C & -\mu(b\pi_A + d\pi_C + f\pi_T) & \mu f\pi_T \\ \mu c\pi_A & \mu e\pi_C & \mu f\pi_G & -\mu(c\pi_A + e\pi_C + f\pi_G) \end{pmatrix}$$

Der Evolutionsmodelle-Zoo

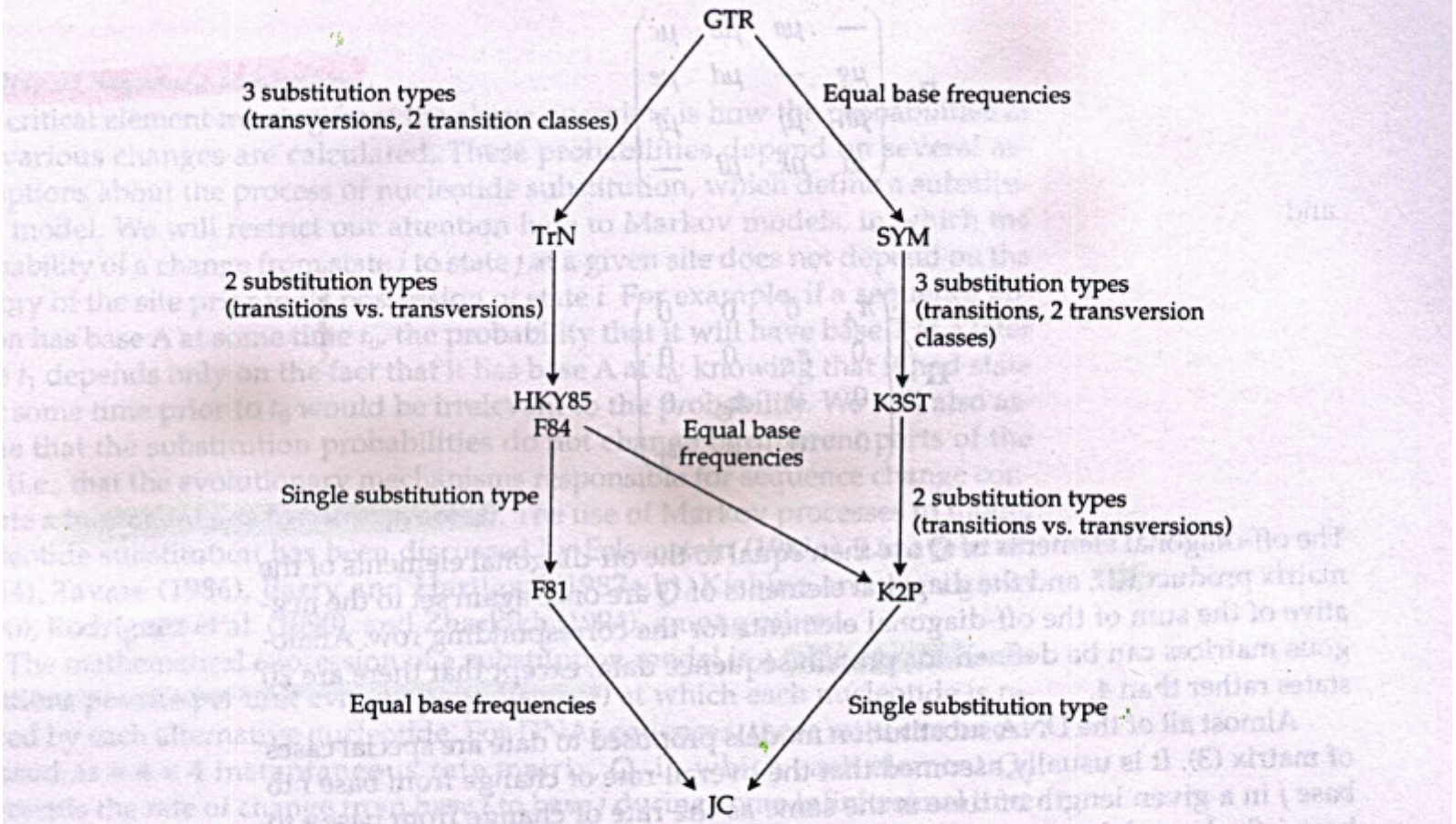


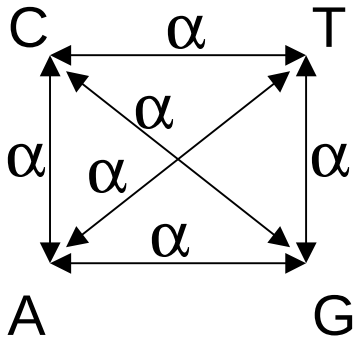
Figure 11 Relationship between special cases of the general time-reversible family of substitution models. Arrow labels indicate restrictions that convert a more general model to a more specific one. Model abbreviations: F81, model of Felsenstein, 1981a (equivalent to the "equal input" model of Tajima and Nei, 1982); F84, model used in versions 2.6 and later of PHYLIP (Felsenstein, 1993; Kishino and Hasegawa, 1989); GTR, General time-reversible (Lanave et al., 1984; Tavaré, 1986; Rodríguez et al., 1990); HKY85, Hasegawa-Kishino-Yano model (Hasegawa et al., 1985b); JC, Jukes and Cantor (1969) model; K2P, Kimura (1980) two-parameter model; K3ST, Kimura (1981) three-substitution-type model; SYM, model described by Zharkikh (1994); TrN, Tamura and Nei (1993) model.

$$Q = \begin{pmatrix} -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu \end{pmatrix}$$

The base frequency and substitution rate are typically combined into a single parameter $\alpha = \mu/4$, leading to the simpler form:

$$Q = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}$$

Identische Substitutionsraten:



- Aus den Rate-Matrizen errechnen sich die Übergangswahrscheinlichkeiten (s. Markov-Ketten):

$$P(t) = e^{Qt}$$

(Matrixexponentialfkt.)

Jukes-Cantor-Modell II

- Für das JC-Modell erhält man Übergangswahrscheinlichkeiten:

$$\text{JC:}\quad P_{ij}(t) := \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-\mu t} & (i = j) \\ \frac{1}{4} - \frac{1}{4}e^{-\mu t} & (i \neq j) \end{cases}$$

- Die Zeit-Parameter t (Zweiglängen) müssen zusätzlich durch den max. likelihood Ansatz geschätzt werden!
- Insgesamt:
  - Max. Likelihood liefert sehr präzise Resultate
  - Aufwändigste Methode (Optimierung der Zielgröße über alle möglichen Baumtopologien und alle Zweiglängen)