



# Evolutionäre Bioinformatik

**Benedikt Brors**

Abt. Theoret. Bioinformatik

Deutsches Krebsforschungszentrum

Heidelberg



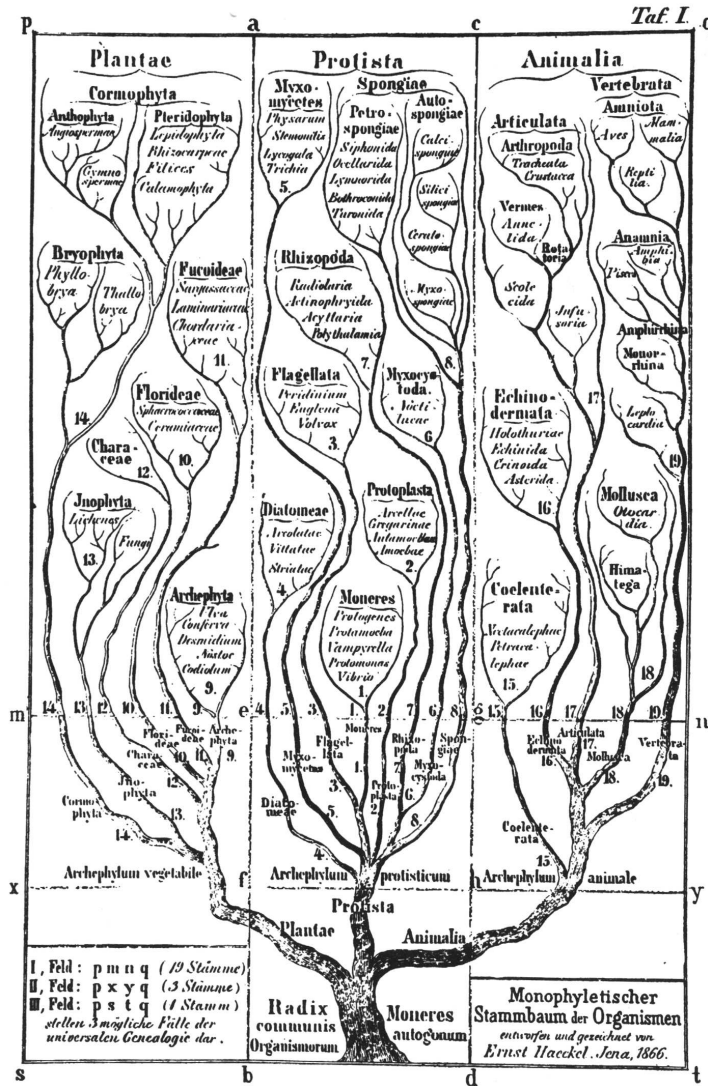
# Inhalt



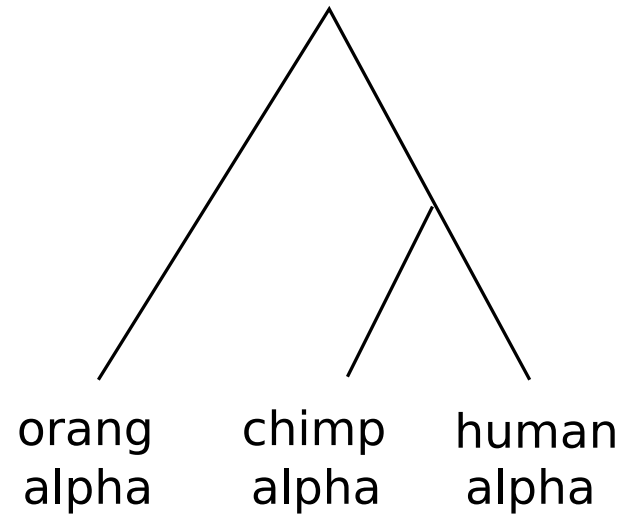
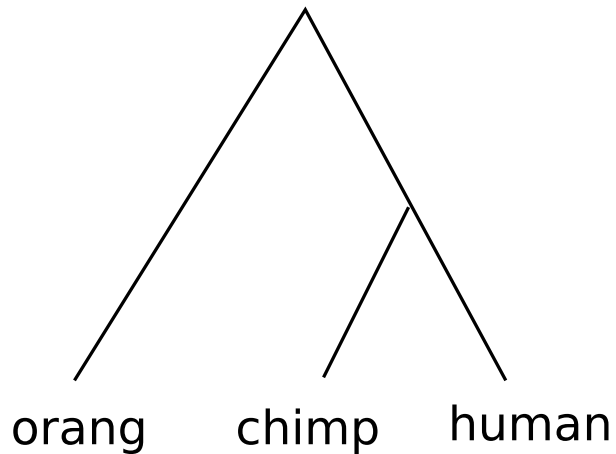
- Was ist und warum treibt man evolutionäre Bioinformatik?
- Einige grundlegende Dinge über Bäume als mathematische Objekte
- evolutionäre Distanzschätzung
- Methoden, um phylogenetische Bäume abzuleiten
  - Methoden, die auf Sequenzen basieren (Max. Parsimony, Max. Likelihood)
  - Methoden, die auf Distanzen basieren
- Seminarvorträge



# Der Baum des Lebens



# Gen- und Speziesbäume



- Ein Speziesbaum zeigt die Entwicklung der Arten
- Genbäume, die man üblicherweise aus Sequenzen (multiplen Alignments) ableitet, sind im allgemeinen mit Speziesbäumen **nicht** identisch



# Unterschiede



- Quellen der Sequenzdiversität, die nicht auf Artenbildung beruhen:
  - Genduplikation
  - partielle interne (Tandem-)Duplikationen
  - Rekombination (zwischen paralogen Genen)
  - partielle Duplikationen mit Translokation
  - horizontaler Gentransfer (z.B. Resistenzfaktoren gegen Antibiotika, mitochondriale Proteine)



# Anwendung evolutionärer Verfahren

- Schätzung von Evolutionsraten (z.B. von Viren)
- Datierung von Ereignissen (Trennung zw. Säugetieren und Vögeln, Auftreten der ersten Menschen etc.)
- Gewebespezifität z.B. von Viren
- geographische Korrelation (Epidemiologie, Geschichte des Menschen)
- Gen-Umwelt-Interaktionen (z.B. Klima: Eiszeiten)

# Evolutionäre Theorien



- Kimura (1968): Die allermeisten Aminosäureaustausche stehen nicht unter positiver Selektion, sondern sind neutral
- kann getestet werden; umgekehrt können konservierte Positionen in Proteinen gefunden werden
- Genomvergleich: Organismen sollten soweit entfernt sein, daß jeder neutrale Rest im Mittel einmal geändert wurde (ca. 180 Myrs)



# weitere Anwendungen

- neutrale Distanz f. Menschen: Beuteltiere



- Koevolution von Parasiten und Wirten
- Ursprung des genetischen Codes (tRNA Gene)
- Cenancestor aller Lebewesen



---

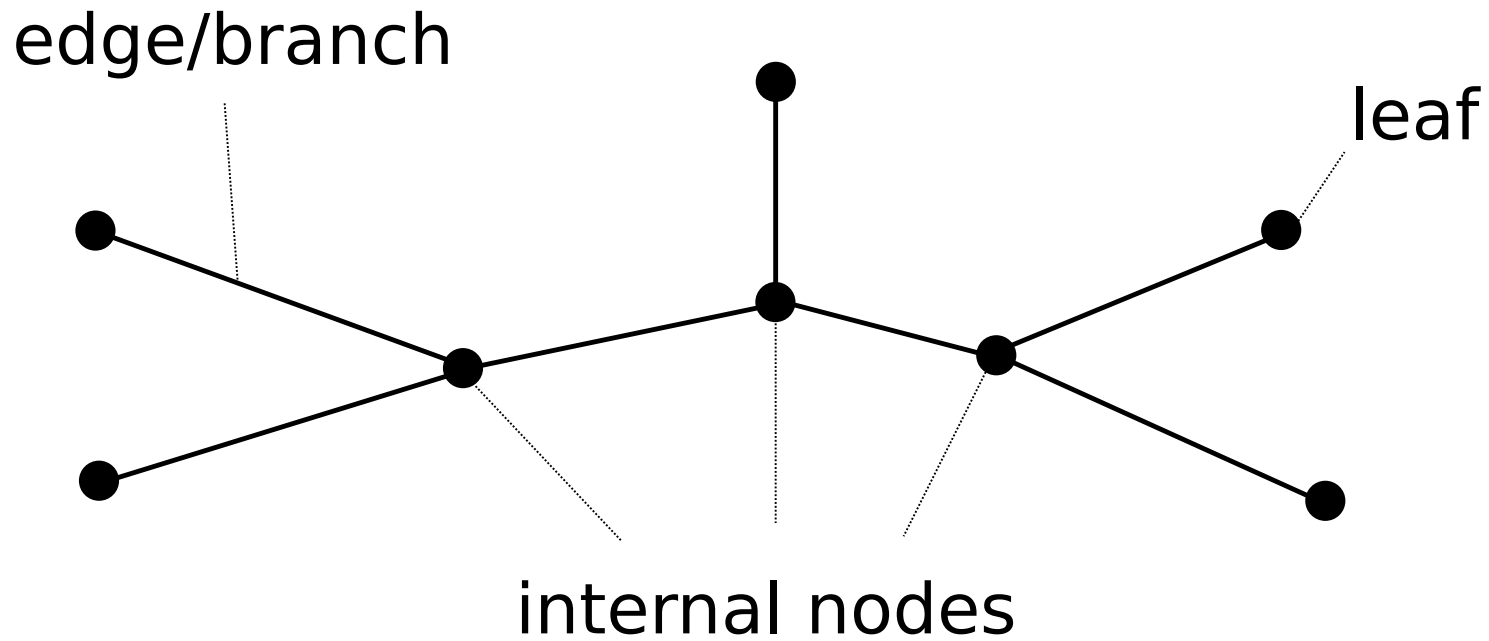
# Mathematische Grundlagen: Bäume



# Graphen und Bäume

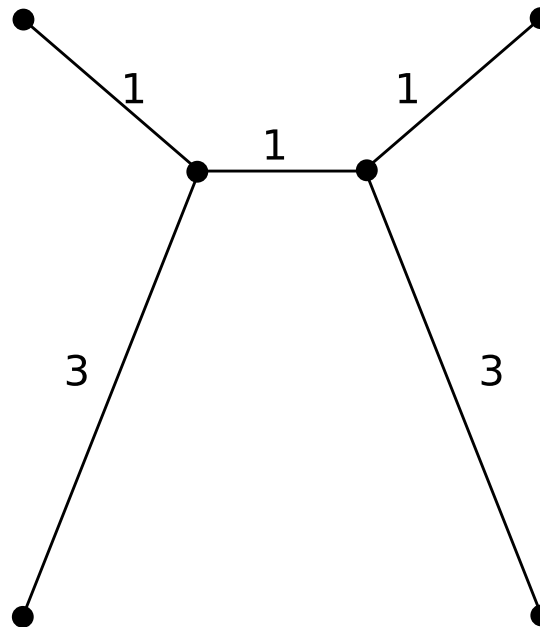
- Ein **Graph** ist ein Paar  $G = (V, E)$  von Knoten (*vertices*)  $V$  und Kanten (*edges*)  $E \subseteq V \times V$ , die die Knoten verbinden. Der **Grad** eines Knotens ist die Anzahl der Kanten, die von dem Knoten ausgehen.
- Ein **Pfad** ist eine Folge von Knoten  $v_1, v_2, \dots, v_n$ , bei der  $v_i$  und  $v_{i+1}$  durch eine Kante verbunden sind,  $\forall i$
- Ein **Zyklus** ist ein Pfad, bei dem Anfangs- und Endknoten identisch sind. Ein Graph ohne Zyklen heißt **acyclisch**.
- Ein **Baum** ist ein acyclischer Graph. Jedes Paar von zwei Knoten ist durch einen Pfad verbunden. Ein **binärer** Baum besitzt nur Knoten vom Grad 3 (interne Knoten) oder 1 (Blätter).

# Baum



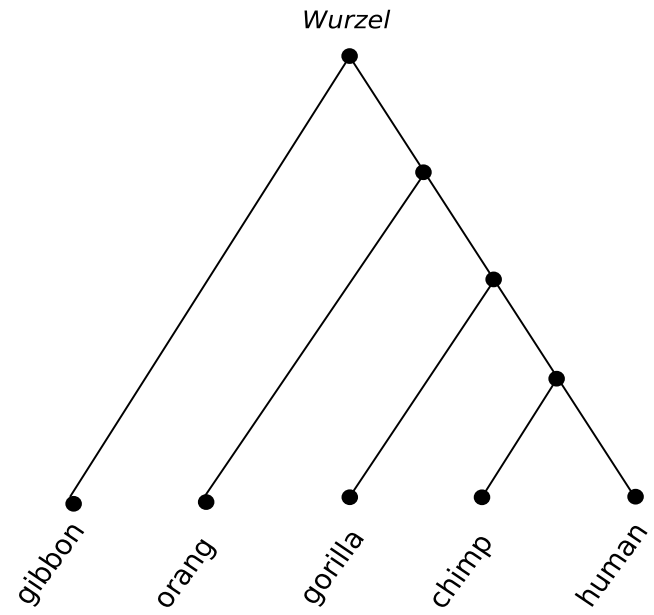
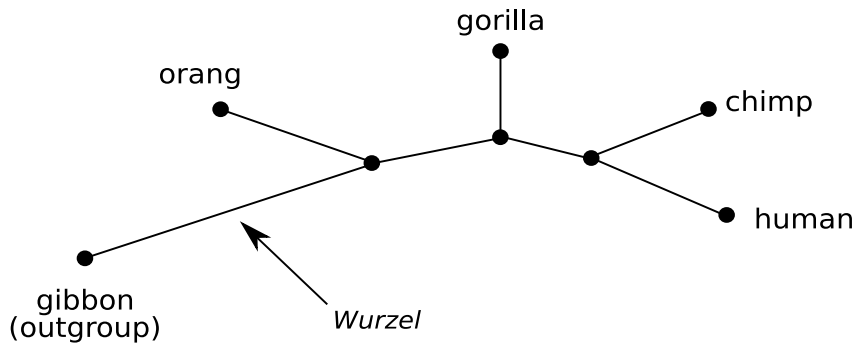
# gewichtete Bäume

- Rekonstruierte phylogenetische Bäume sind **gewichtete Bäume**, d.h. die Kanten haben Längen, die die Anzahl der Mutationen widerspiegeln, die man im Lauf der Evolution annimmt.



# gewurzelte Bäume

- Bei einem binären, ungewurzelten Baum ist nicht klar, welcher interne Knoten der Vorläufer oder Nachfolger eines benachbarten Knotens ist
- Manchmal gibt es externe Informationen darüber, daß ein Taxon weit weniger mit den anderen Taxa verwandt ist als diese untereinander. Solch ein Taxon heißt **outgroup**. Die Wurzel fügt man an der Kante an, die zur Outgroup führt; so kann man die zeitliche Abfolge der Splits interpretieren.



# Anzahl der Topologien

Die Anzahl möglicher Topologien kann wie folgt abgeleitet werden: Man fängt mit einem Baum zweier Spezies an und fügt sukzessive andere Spezies hinzu. Hierzu gibt es  $2n - 3$  Möglichkeiten (Anz. d. Kanten). Damit gibt es

$$U_n = \prod_{i=3}^n (2i - 5)$$

ungewurzelte und

$$R_n = \prod_{i=3}^{n+1} (2i - 5)$$

gewurzelte Bäume.





$n$	$U_n$	$R_n$
2	1	1
3	1	3
4	3	15
5	15	105
6	105	954
7	945	10,395
8	10,395	135,135
9	135,135	2,027,025
10	2,027,025	34,459,425







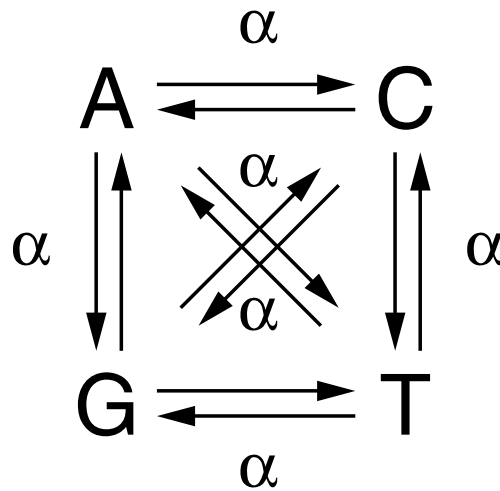
---

# Evolutionäre Distanzschätzung



# Modelle für Mutationsraten

- Das einfachste mögliche Modell ist, daß die Mutationsrate für alle Nukleotide gleich ist, nämlich  $\alpha$ . Man bezeichnet es als das Jukes-Cantor-Modell:



# Mutationswahrscheinlichkeit



- Nehmen wir an, zur Zeit  $t = 0$  hat eine Sequenz an einer bestimmten Position ein A. Dann ist  $P(A)_0 = 1$ .  
Dann ist:

$$\text{bei } t = 1 \quad : \quad P(A)_1 = 1 - 3\alpha$$

$$\text{bei } t = 2 \quad : \quad P(A)_2 = (1 - 3\alpha)P(A)_1 + \alpha [1 - P(A)_1]$$

$$\text{bei } t + 1 \quad : \quad P(A)_{t+1} = (1 - 3\alpha)P(A)_t + \alpha [1 - P(A)_t]$$

- Dieses Modell bezeichnet man als Markovkette



# Markovketten

- Markovketten sind Folgen von Ereignissen, bei denen jedes Ereignis vom vorhergehenden, aber nur vom vorhergehenden abhängt. Sie haben bestimmte Eigenschaften, insbesondere in diesem Modell:
  - Sie sind vollständig definiert durch die Startverteilung und eine Matrix  $P$  an Übergangswahrscheinlichkeiten, die Übergangsmatrix;
  - es gibt genau eine stationäre Verteilung  $\pi$  mit

$$\pi P = \pi$$

$\pi$  ist auch die Startverteilung;

- es existiert eine Ratenmatrix  $Q$  mit  $P = \exp(Q)$ ;
- Der Prozeß ist zeitreversibel.

# Jukes-Cantor-Modell

Für das Jukes-Cantor-Modell lautet die Ratenmatrix:

$$Q = \begin{pmatrix} 1 - 3\alpha & \alpha & \alpha & \alpha \\ \alpha & 1 - 3\alpha & \alpha & \alpha \\ \alpha & \alpha & 1 - 3\alpha & \alpha \\ \alpha & \alpha & \alpha & 1 - 3\alpha \end{pmatrix}$$

Wir schreiben jetzt kürzer für  $P(A)_t = p_t$ . Es war

$$\begin{aligned} p_{t+1} &= (1 - 3\alpha)p_t + \alpha(1 - p_t) \\ &= (1 - 4\alpha)p_t + \alpha \end{aligned}$$

$$\begin{aligned} p_{t+1} - p_t &= -4\alpha p_t + \alpha \\ \Delta p_t &= -4\alpha p_t + \alpha \end{aligned}$$



Wenn wir zu sehr kleinen Zeitabständen gehen, gehen wir von einem Modell mit diskreten Zeiten ( $t = 0, 1, 2, \dots$ ) zu einer stetigen Zeitskala über. Damit wird:

$$\frac{dp}{dt} = -4\alpha p + \alpha$$

Das ist eine lineare inhomogene Differentialgleichung erster Ordnung. Die Lösung ist:

$$p_t = \frac{1}{4} + \left( p_0 - \frac{1}{4} \right) e^{-4\alpha t}$$

Da wir mit A angefangen haben, war  $p_0 = 1$ , und wir erhalten

$$p_t = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t}$$





Hätten wir nicht mit A angefangen, wäre  $p_0 = 0$ , und wir erhalten:

$$p_t = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t}$$

Die Übergangsmatrix  $P$  ist damit

$$P = \begin{pmatrix} 1/4 + 3a_t & 1/4 - a_t & 1/4 - a_t & 1/4 - a_t \\ 1/4 - a_t & 1/4 + 3a_t & 1/4 - a_t & 1/4 - a_t \\ 1/4 - a_t & 1/4 - a_t & 1/4 + 3a_t & 1/4 - a_t \\ 1/4 - a_t & 1/4 - a_t & 1/4 - a_t & 1/4 + 3a_t \end{pmatrix}$$

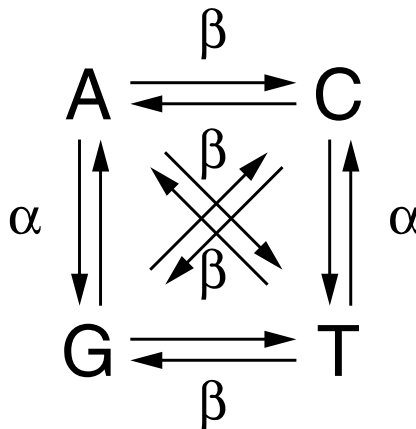
mit

$$a_t = \frac{e^{-4\alpha t}}{4}$$



# Kimura-Modell

- Von Kimura stammt ein Modell, das zwei unterschiedliche Mutationsraten kennt. Es ist nämlich chemisch leichter, ein Purin (A,G) in ein anderes Purin umzuwandeln, und ein Pyrimidin (C,T) in ein anderes Pyrimidin, als umgekehrt.
- Übergänge Pur→Pur und Pyr→Pyr bezeichnen wir als Transitionen, Pur→Pyr und Pyr→Pur als Transversionen.



# Übergangswahrscheinlichkeiten



Für die Übergangswahrscheinlichkeiten gilt:

$$p_1 = 1 - \alpha - 2\beta$$

$$p_2 = (1 - \alpha - 2\beta)p_A(1) + \beta p_T + \beta p_C + \alpha p_G$$

⋮

$$p_{t+1} = (1 - \alpha - 2\beta)p_A(t) + \beta p_T(t) + \beta p_C(t) + \alpha p_G(t)$$

Sowie 3 weitere Gleichungen für  $p_G(t)$ ,  $p_C(t)$  und  $p_T(t)$ .





Dieses Gleichungssystem löst man durch Berücksichtigen der Anfangsbedingungen, z.B.

$$\pi_A^0 = (1, 0, 0, 0)$$

Die stationäre Verteilung ist hier, wie im Jukes-Cantor-Modell, die Gleichverteilung

$$\pi = \left( \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4} \right)$$





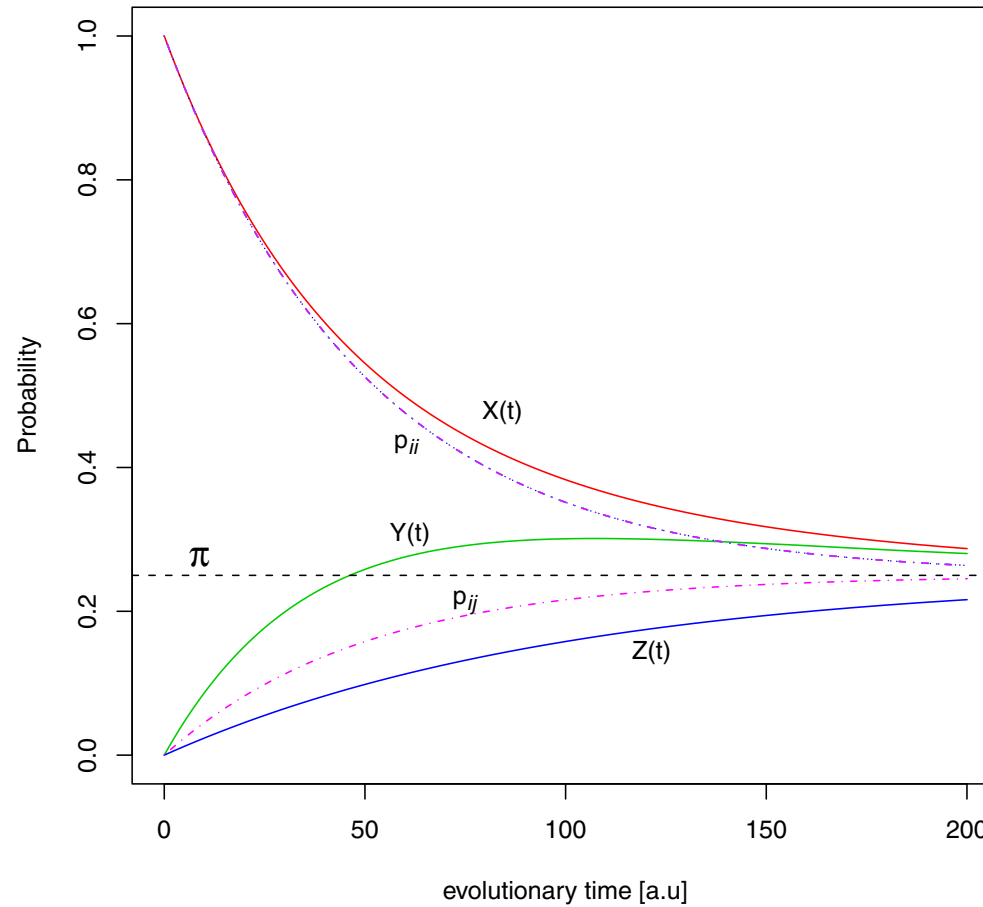
Die Lösung der Differentialgleichungen führt zu 3 unterschiedlichen Wahrscheinlichkeiten,  $X_t$  für Erhalt eines Nukleotids,  $Y_t$  für eine Transition, und  $Z_t$  für eine Transversion:

$$\begin{aligned} X(t) &= p_{AA}(t) + p_{GG}(t) + p_{CC}(t) + p_{TT}(t) \\ &= \frac{1}{4} + \frac{1}{4} e^{-4\beta t} + \frac{1}{2} e^{-2(\alpha+\beta)t} \\ Y(t) &= \frac{1}{4} + \frac{1}{4} e^{-4\beta t} - \frac{1}{2} e^{-2(\alpha+\beta)t} \\ Z(t) &= \frac{1}{4} - \frac{1}{4} e^{-4\beta t} \end{aligned}$$

Es gilt  $X_t + Y_t + 2Z_t = 1$ .



# Vergleich der Modelle

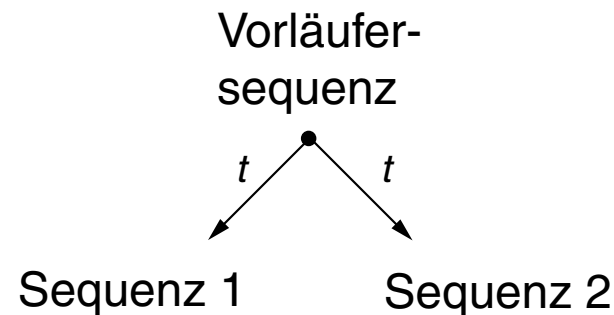


# andere Modelle

- Andere Modelle umfassen mehr Parameter, z.B.
  - 4-Parameter-Modell (Blaisdell, 1985)
  - 6-Parameter-Modell (Kimura, 1981)
  - 9-Parameter-Modell
  - Generelles Modell (12 Parameter)
- Modelle mit mehr als 6 Parametern sind nicht mehr zeitreversibel!

# Maße für Sequenzähnlichkeit

- Wir wollen für die beiden Modelle formulieren, wie die Wahrscheinlichkeit für den Erhalt eines Nukleotids in einer Position und für eine Mutation ist. Wir nehmen wieder an, daß wir bei  $t = 0$  A haben.



- Die Wahrscheinlichkeit, daß A erhalten wurde, ist  $p_{AA}(t)^2$ .



- Für ein beliebiges Nukleotid N ist dann

$$I(t) = p_{AA}^2(t) + p_{GG}^2(t) + p_{CC}^2(t) + p_{TT}^2(t)$$

- Setzen wir die Gleichungen aus dem Jukes-Cantor-Modell ein, erhalten wir

$$I(t) = \frac{1}{4} + \frac{3}{4} e^{-8\alpha t}$$

- Für das Kimura-Modell erhalten wir:

$$I(t) = \frac{1}{4} + \frac{1}{4} e^{-8\beta t} + \frac{1}{2} e^{-4(\alpha+\beta)t}$$



# Wahrscheinlichkeit für eine Mutation

- Beobachten wir eine Differenz zwischen 2 Sequenzen, z.B. in der einen Sequenz ein G und in der anderen ein A, können wir das wegen der Zeitreversibilität als Prozess über  $2t$  betrachten. Damit wird im Jukes-Cantor-Modell

$$p_{AG}(2t) = \frac{1}{4} - \frac{1}{4} e^{-4\alpha(2t)} = \frac{1}{4} - \frac{1}{4} e^{-8\alpha t}$$

- Für das Kimura-Modell berechnet man das analog.



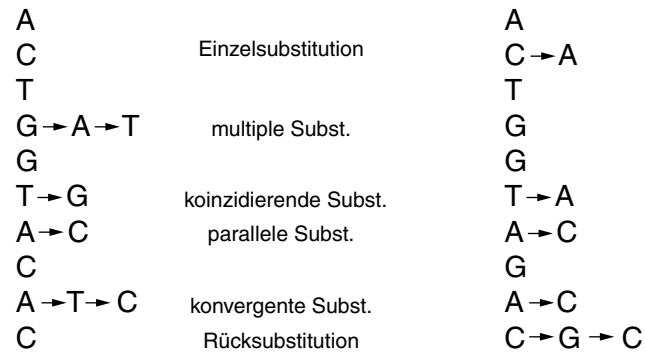
# Anzahl der Austausche

- Für die evolutionäre Zeitrechnung definiert man sich Einheiten:
  - 1 PEM (1 percent expected mutations): Zeit, in denen 1 Austausch pro 100 Positionen *erwartet* wird
  - 1 PAM (1 percent accepted mutations): Zeit, in denen 1 Austausch pro 100 Positionen *beobachtet* wird
- 1 PEM ist eine etwas kleinere Einheit als 1 PAM, da Rücksubstitutionen stattfinden können; man beobachtet also keinen Austausch (0 PAM), während in Wirklichkeit 2 Austausche stattgefunden haben (2 PEM).
- PEM und PAM haben nichts mit der Realzeit zu tun.



Vorläufersequenz

A  
C  
T  
G  
G  
T  
A  
C  
A  
C



# Jukes-Cantor-Modell



- Wahrscheinlichkeit für Identität:

$$I(t) = \frac{1}{4} + \frac{3}{4} e^{-8\alpha t}$$

- Wahrscheinlichkeit für Differenz:

$$p(t) = 1 - I(t) = \frac{3}{4} (1 - e^{-8\alpha t})$$

$$8\alpha t = -\ln \left( 1 - \frac{4p}{3} \right)$$





- Da die evolutionäre Zeit zwischen zwei Sequenzen in der Regel unbekannt ist, kann man  $\alpha$  nicht schätzen. Statt dessen betrachten wir  $K$ , die Anzahl der Austausche je Position. Es ist

$$K = 3\alpha(2t)$$

$$K = -\frac{3}{4} \ln \left( 1 - \frac{4p}{3} \right)$$

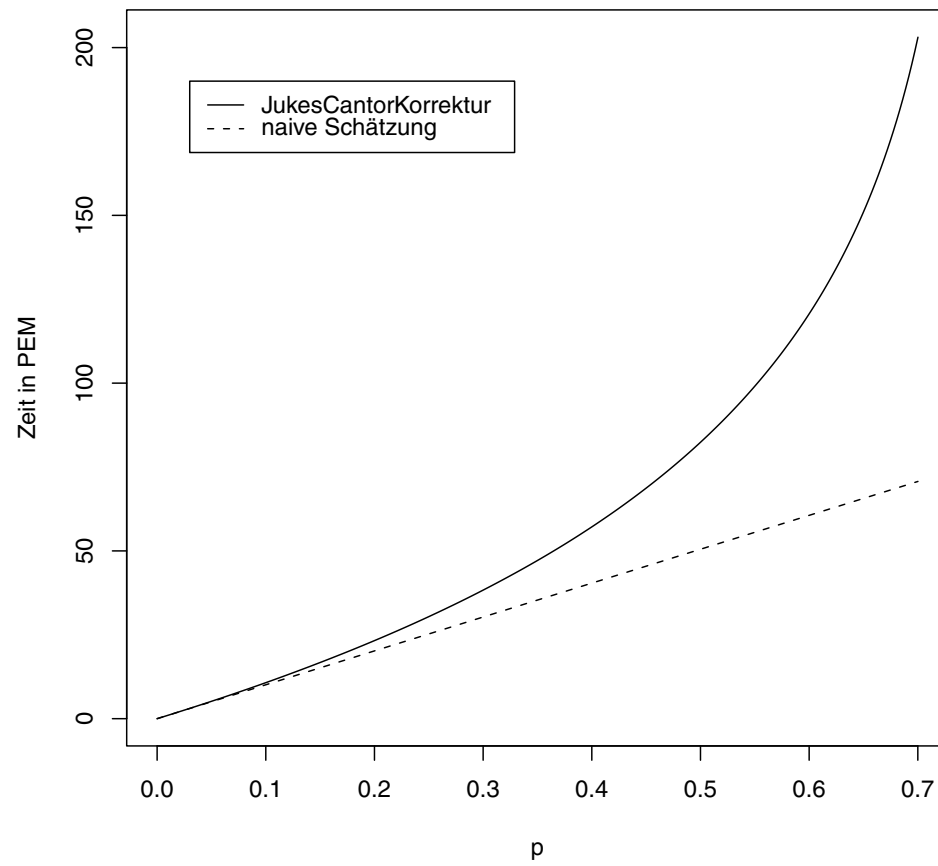


# Jukes-Cantor-Korrektur

- Diese Formel heißt Jukes-Cantor-Korrektur, da sie die naive Schätzung  $K \approx p$  für Rücksubstitutionen korrigiert. Tatsächlich ist die Korrektur klein, wenn Sequenzen nahe verwandt sind, also nur wenig Zeit vergangen ist, seit sie sich vom gemeinsamen Vorläufer getrennt haben. Für sehr unterschiedliche Sequenzen ist der Unterschied aber beachtlich.



### JukesCantorKorrektur



# Kimura-Modell



- Für das Kimura-Modell lauten die Formeln:

$$K = \frac{1}{2} \ln a + \frac{1}{4} \ln b$$

mit

$$a = \frac{1}{1 - 2P - Q} \quad b = \frac{1}{1 - 2Q}$$

wenn  $P$  der Anteil der Transitionen und  $Q$  der Anteil der Transversionen ist.



# Literatur

- Durbin, Eddy, Krogh & Mitchison: Biological Sequence Analysis. CUP 1998.

- Harvey, Brown, Smith & Nee (eds.): New uses for new phylogenies. OUP 1996.

- Hillis, Moritz & Mable (eds.): Molecular Systematics. 2nd ed. Sinauer 1996.

- H. Luz. Skript zur Vorlesung “Algorithmische Bioinformatik”, FU Berlin WS 03/04.

<http://www.inf.fu-berlin.de/inst/ag-bio/FILES/ROOT/Teaching/Lectures/WS03/04/>

- S. Rahmann. Spezielle Methoden und Anwendungen der Statistik in der Bioinformatik. Skript, MPI f. Mol. Genetik und FU Berlin, 2003.

<http://www.molgen.mpg.de/~rahmann/afw-rahmann.pdf>