

# **Bioinformatik: Sequenzanalyse**

## **Anwendung statistischer Modelle in der Bioinformatik**

**Benedikt Brors**

**Abt. Intelligente Bioinformatiksysteme  
Deutsches Krebsforschungszentrum, Heidelberg**



# Häufigkeitsverteilungen und log-odds-Scores

- In der vorigen Vorlesung wurden Häufigkeitsverteilungen von Nukleotiden an bestimmten Positionen vorgestellt und log-odds-Scores erklärt.
- Wir hatten gesehen, daß es Schwierigkeiten gibt, wenn ein Nukleotid an einer Position nicht vorkommt, da dann die Wahrscheinlichkeit in einem Modell 0 wird und der log-odds-Score nicht mehr definiert ist.
- Um diese Schwierigkeit zu umgehen, kann man zu den beobachteten Anzahlen sogenannte *pseudocounts* addieren. Wenn man viele Sequenzen untersucht hat, ändert sich an den relativen Häufigkeiten der beobachteten Nukleotide wenig, aber für die nicht vorkommenden Nukleotide ergibt sich eine von 0 verschiedene Wahrscheinlichkeit.



# Andere Maße: Entropie

- Anstelle der log-odds-Scores kann man auch ein anderes Maß benutzen: Die Entropie.
- Die Entropie gibt an, wie viel Information aus der Verteilung an einer bestimmten Stelle erhalten werden kann.
- Ist viel Information vorhanden, d.h. man ist sich sicher über die Identität eines Nukleotids, ist die Entropie klein.



# Informationstheorie

- Wir stellen uns vor, wir hätten 64 umgedrehte Becher, und unter einem der Becher würde eine Kugel liegen. Was ist die minimale Anzahl an Fragen, die man stellen muß, um die Kugel sicher zu finden?■
- Die minimale Anzahl ist 6: Die erste Frage wäre, ob die Kugel in der ersten Hälfte zu finden ist, dann in der Hälfte der Hälfte, usw. Insgesamt braucht man  $\log_2(64)$  Fragen.
- Diesen Informationsgehalt bezeichnet man mit der Einheit *bit*. Der maximale Informationsgehalt bei 20 Aminosäuren ist  $\log_2(20) = 4.32$  bits, bei Nukleotiden  $\log_2(4) = 2$  bits.



# Entropie

- Die Entropie wird in bits gemessen und gibt das Maß der Unsicherheit für die Identität eines Nukleotids an.
- Die Entropie für eine Position ist definiert als

$$H_c = - \sum_i p_{i,c} \log_2(p_{i,c})$$

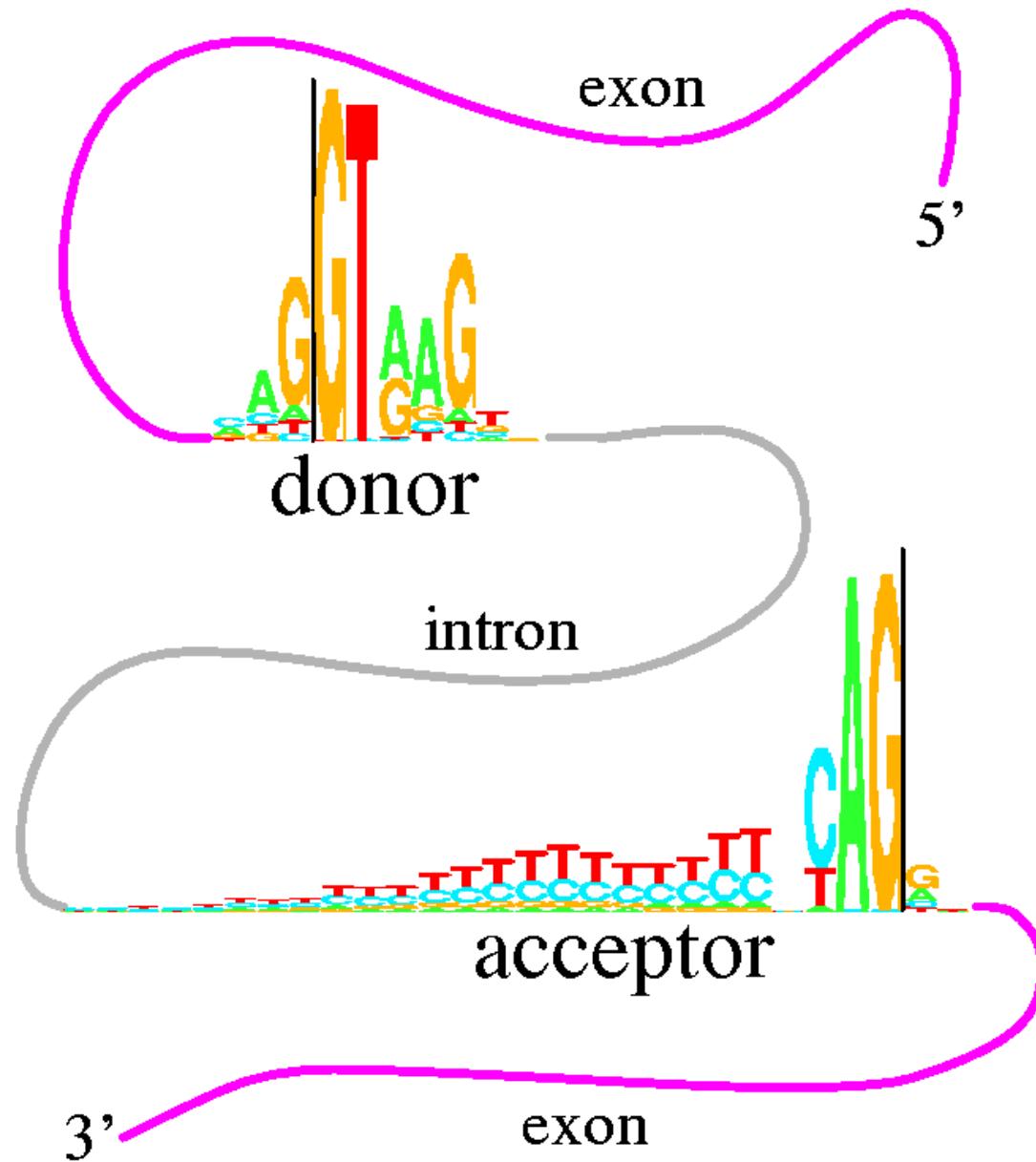
und für eine Sequenz bzw. ein Alignment als

$$H = \sum_c H_c$$

- Die Wahrscheinlichkeitsverteilung in Entropien (abgeleitet aus Häufigkeitsmatrizen, s.o.) kann man als Sequenzlogo darstellen.



# Sequenzlogo



# Konstruktion von Sequenzlogos

- Die Höhe der Buchstaben ist proportional zur *Reduktion* der Unsicherheit je Position; wenn die Entropie  $H_c$  ist, ist die Reduktion  $H_{\max} - H_i$ .
- Für Nukleinsäuren ist  $H_{\max} = \log_2(4) = 2$  bits, für Proteine  $H_{\max} = \log_2(20) = 4.32$  bits.
- Die Höhe der Buchstaben wird noch mit der relativen Häufigkeit des Nukleotids gewichtet, d.h.

$$\text{Höhe}_{i,c} = p_{i,c} (H_{\max} - p_{i,c} \log_2 p_{i,c})$$

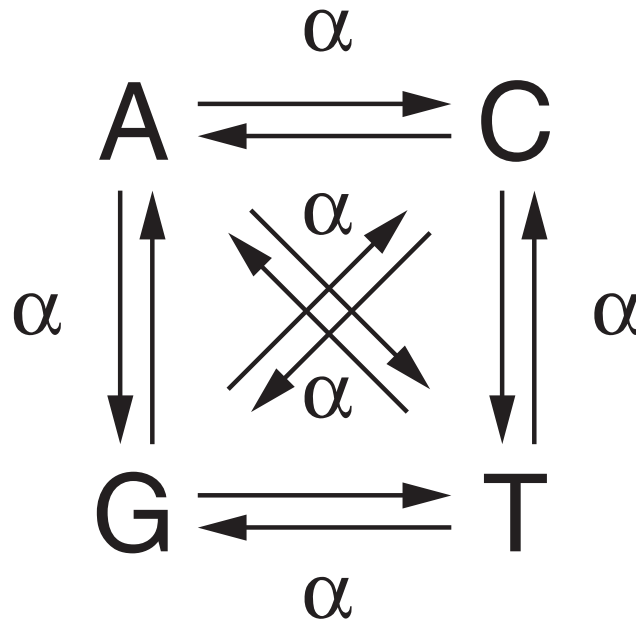


# Evolutionäre Distanzschätzung



# 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:

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

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

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

- 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

$$p_{t+1} = (1 - 3\alpha)p_t + \alpha(1 - p_t)$$

$$= (1 - 4\alpha)p_t + \alpha$$

$$p_{t+1} - p_t = -4\alpha p_t + \alpha$$

$$\Delta p_t = -4\alpha p_t + \alpha$$



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

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

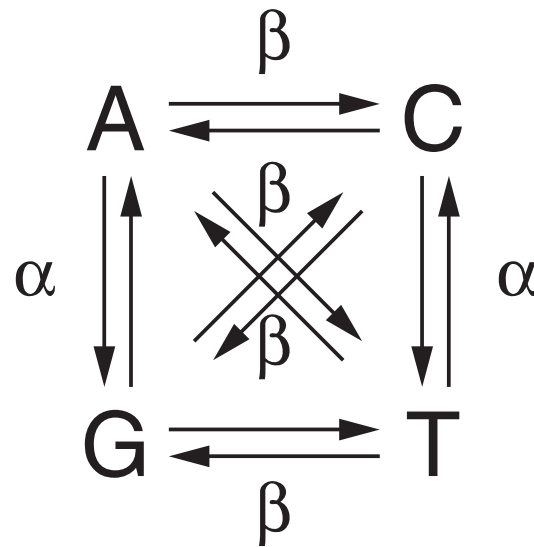
mit

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



# 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 im 2-Parameter-Modell

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

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



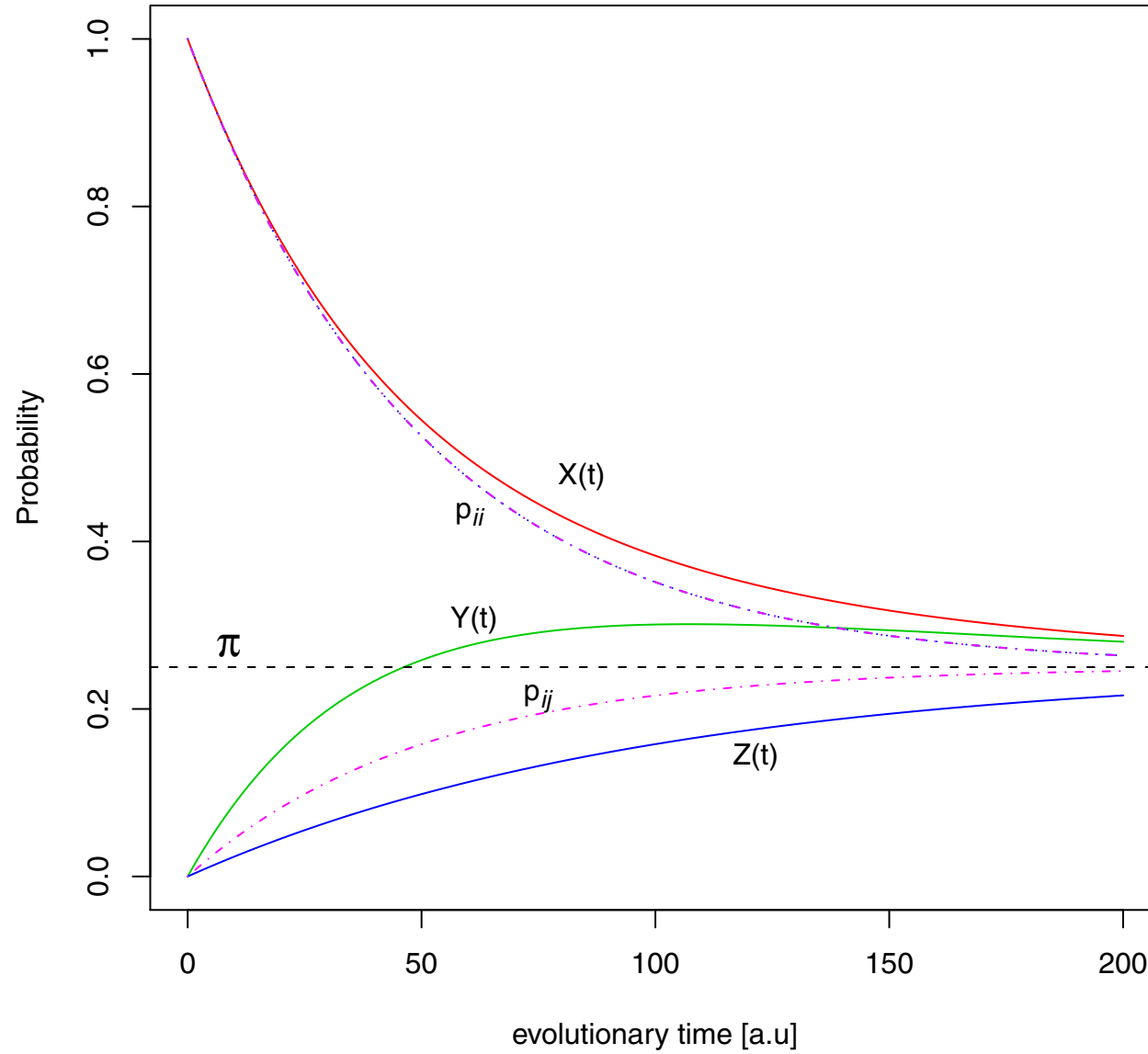
Die Lösung der Differentialgleichungen führt zu 3 unterschiedlichen Wahrscheinlichkeiten,  $X_t$  für Erhalt eines Nukleotids,  $Y_t$  für eine Transition, und  $Z_t$  für eine Transversion:

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

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



# Vergleich der Substitutionswahrscheinlichkeiten



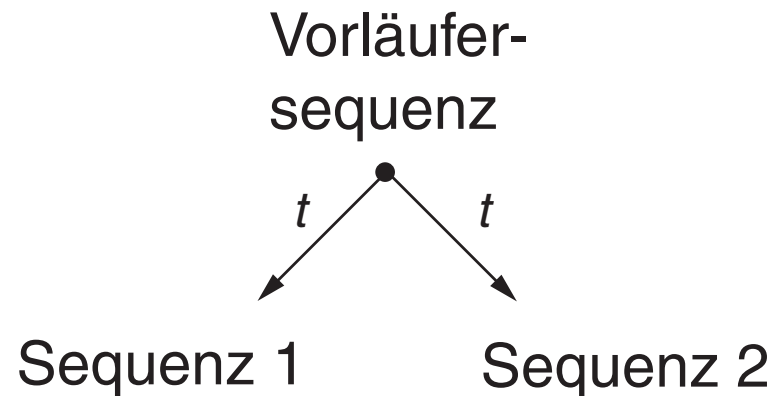
# andere Modelle

- 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.



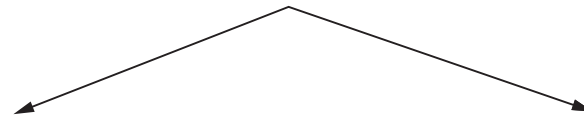
# Erwartungswert für Anzahl der Austausche zwischen Sequenzen

- 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



A  
C  
T  
G  
G  
T  
A  
C  
A  
C

Einzelsubstitution

multiple Subst.

koinzidierende Subst.

parallele Subst.

konvergente Subst.

Rücksubstitution

A  
C  
T  
G  
G  
T  
A  
C  
A  
C

C → A

T → A

A → C

A → C

C → G → 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

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

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

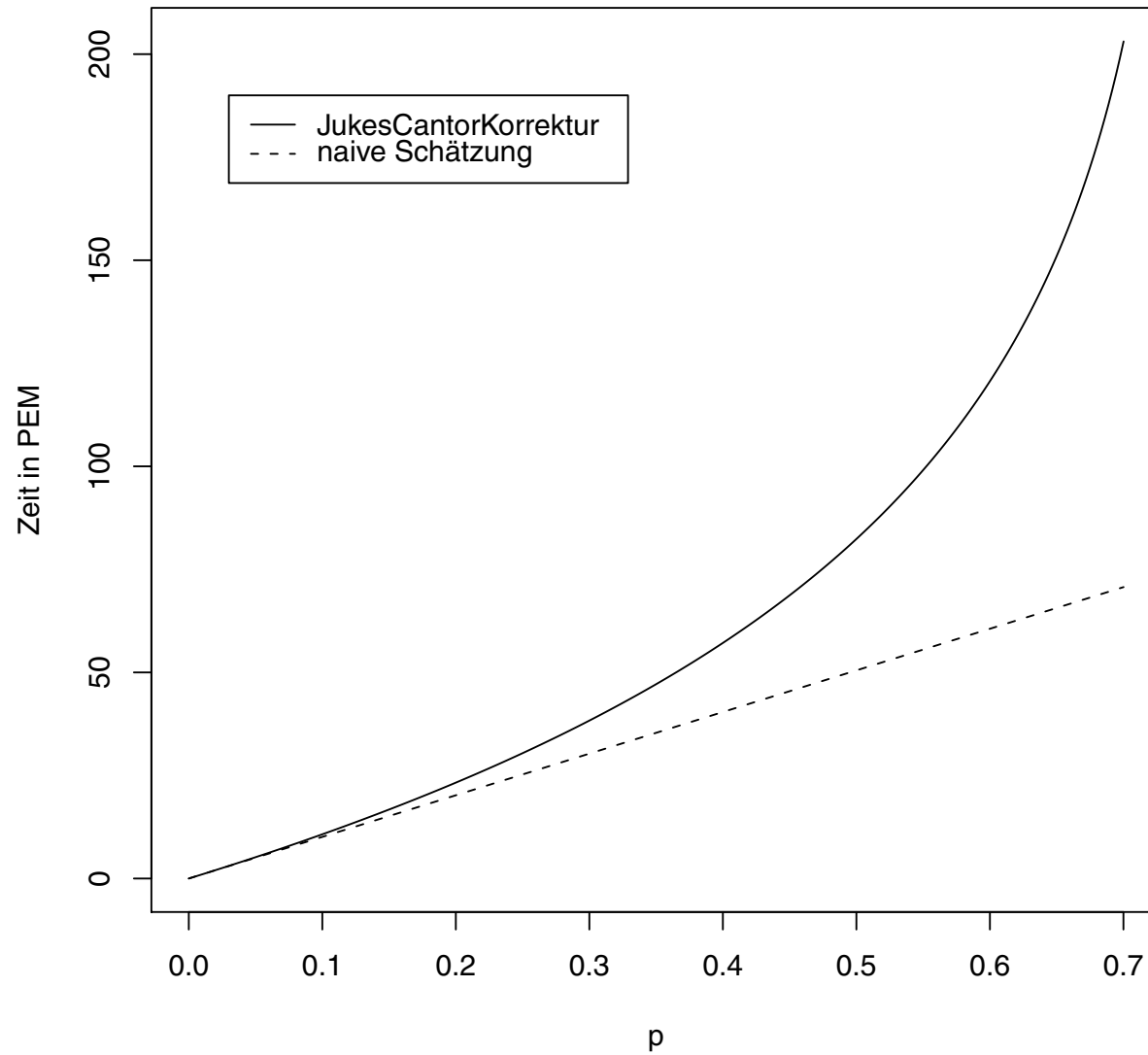


# Jukes-Cantor-Korrektur

- 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.

