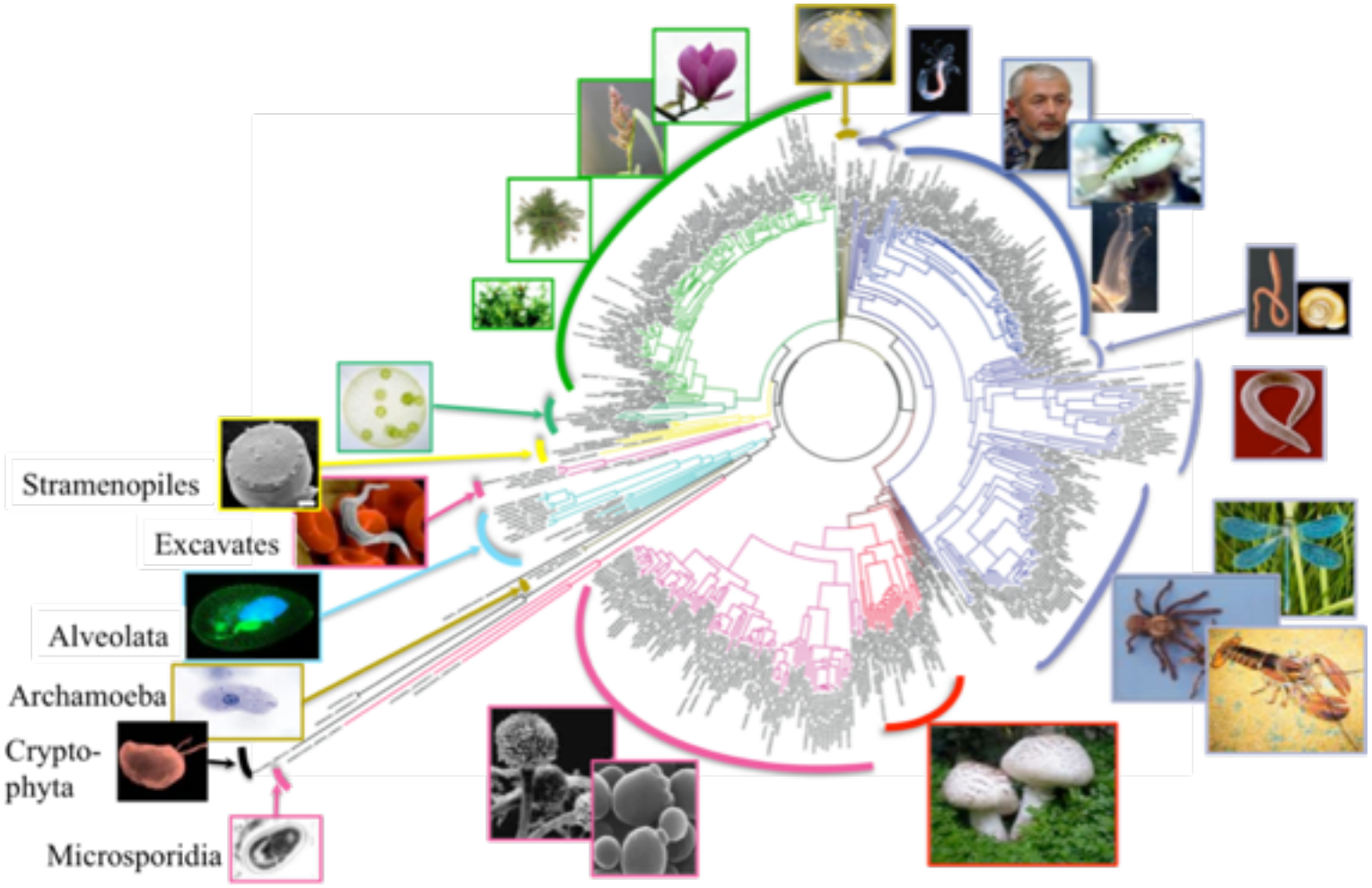
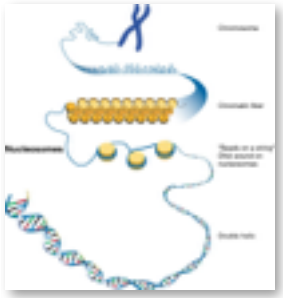
Ein Baum beschreibt evolutionäre Geschichte

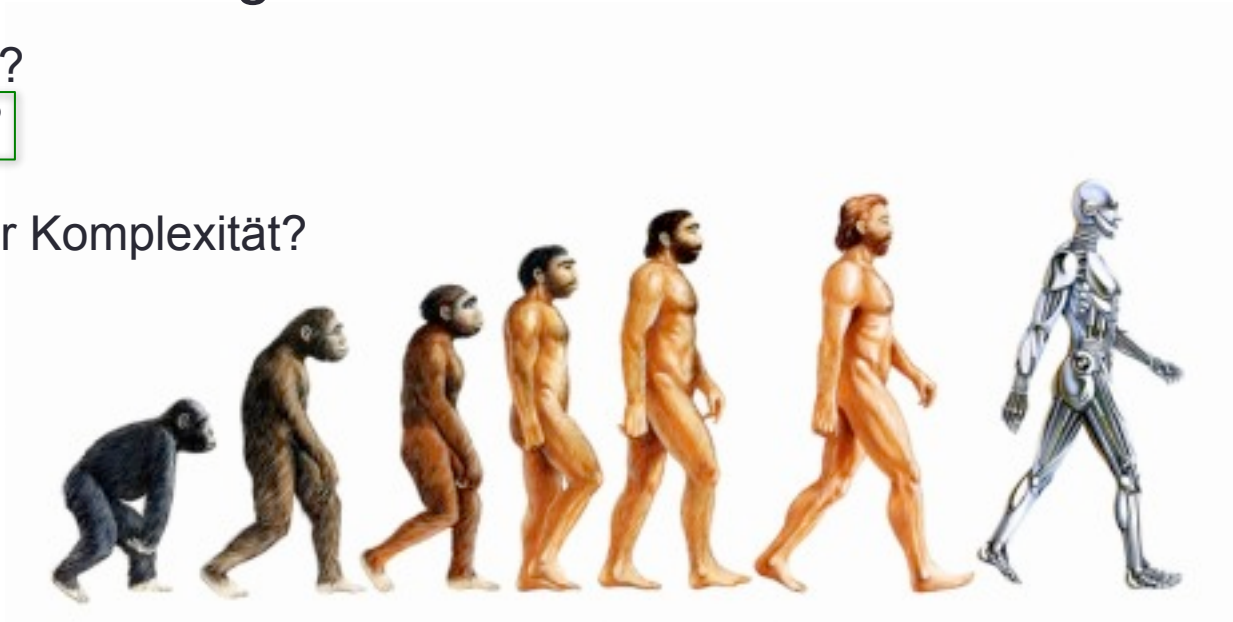


Zwei Fragen vorab:



Was bedeutet eigentlich Evolution?

Verbesserung?
Veränderung?
Anpassung?
Steigerung der Komplexität?



Zeit

SO

ANDERS

Zwei Fragen vorab:

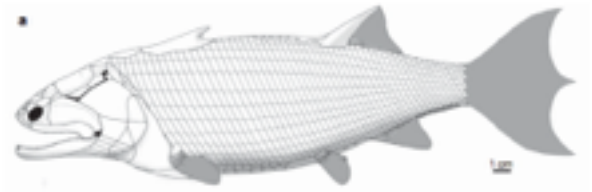
Was bedeutet eigentlich ‘primitiv’?

Primitivität

Primitivität (latein. *primitivus* „der Erste in seiner Art“) ist eine Bezeichnung für besondere Einfachheit.[1] Im sozialen Zusammenhang steht *primitiv* für einen empfundenen Mangel an Zivilisiertheit, oder auf eine Person bezogen für geringe Intelligenz.

In der Biologie, speziell in der Entwicklungsbiologie, der Paläontologie und der Paläoanthropologie, wird die Bezeichnung *primitiv* für anatomische Merkmale jedoch wertneutral im Sinne von *ursprünglich*, *urtümlich* und *alt* verwendet (siehe Plesiomorphie), das heißt als Gegensatz zu – gleichfalls wertneutral beschriebenen – *neuartigen*, *fortschrittlichen*, vom ursprünglichen Zustand *abgeleiteten* Merkmalen (siehe Apomorphie). Häufig kann das Gegensatzpaar „primitiv“ / „abgeleitet“ auch im Sinne von „einfach“ / „komplex“ (vielschichtig) verstanden werden.

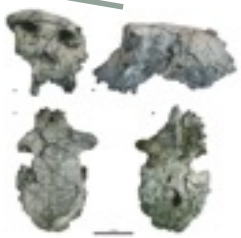
Mit manchen Methoden können wir direkt zurück in die Zeit blicken!



Guyu oneiros
(~412 MYBP)

~5,000

~30,000



Sahelanthropus
(~6.5 MYPB)

Tyrannosaurs rex
~63 MYBP)

Part 2: Modellierung von Sequenzvolution, i.e. die Zeit-abhängige Veränderung biologischer Sequenzen

Sequenzevolution

heute ißt ingo, gegebenenfalls, ein schwein

Sequenzevolution

heute ißt ingo, gegebenenfalls, ein schwein

1) Copy the line
2) Replace original with copy
3) Goto 1

Sequenzevolution

heute ißt ingo, gegebenenfalls, ein schwein

1) Copy the line
2) Replace original with copy
3) Goto 1

Iteration 1: Heute ist ingo, gegebenenfalls, ein schwein

- Substitution

Iteration 2: Heute ist ingo, gegebenenfalls, ein scheein

- Substitution

Starting positions of words remain unchanged!!

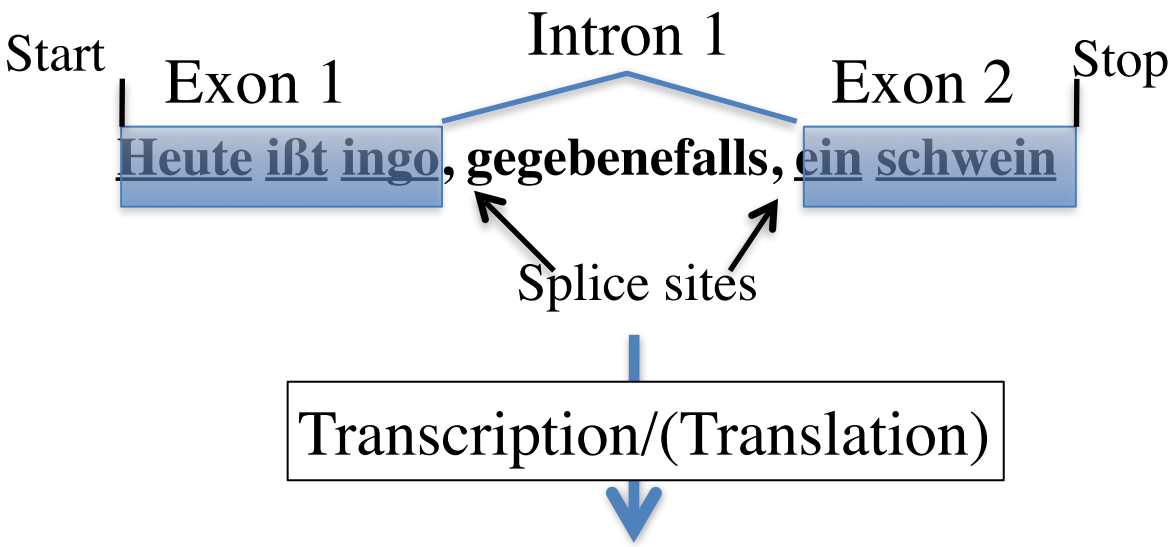
Iteration 3: Heuti sti ngo, gegebenenfalls, e ins cheein

- Deletion and shift of reading frame

Iteration 4: Heuti sti Nngo, gegebenenfalls, ein scheein

- Insertion and shift of reading frame

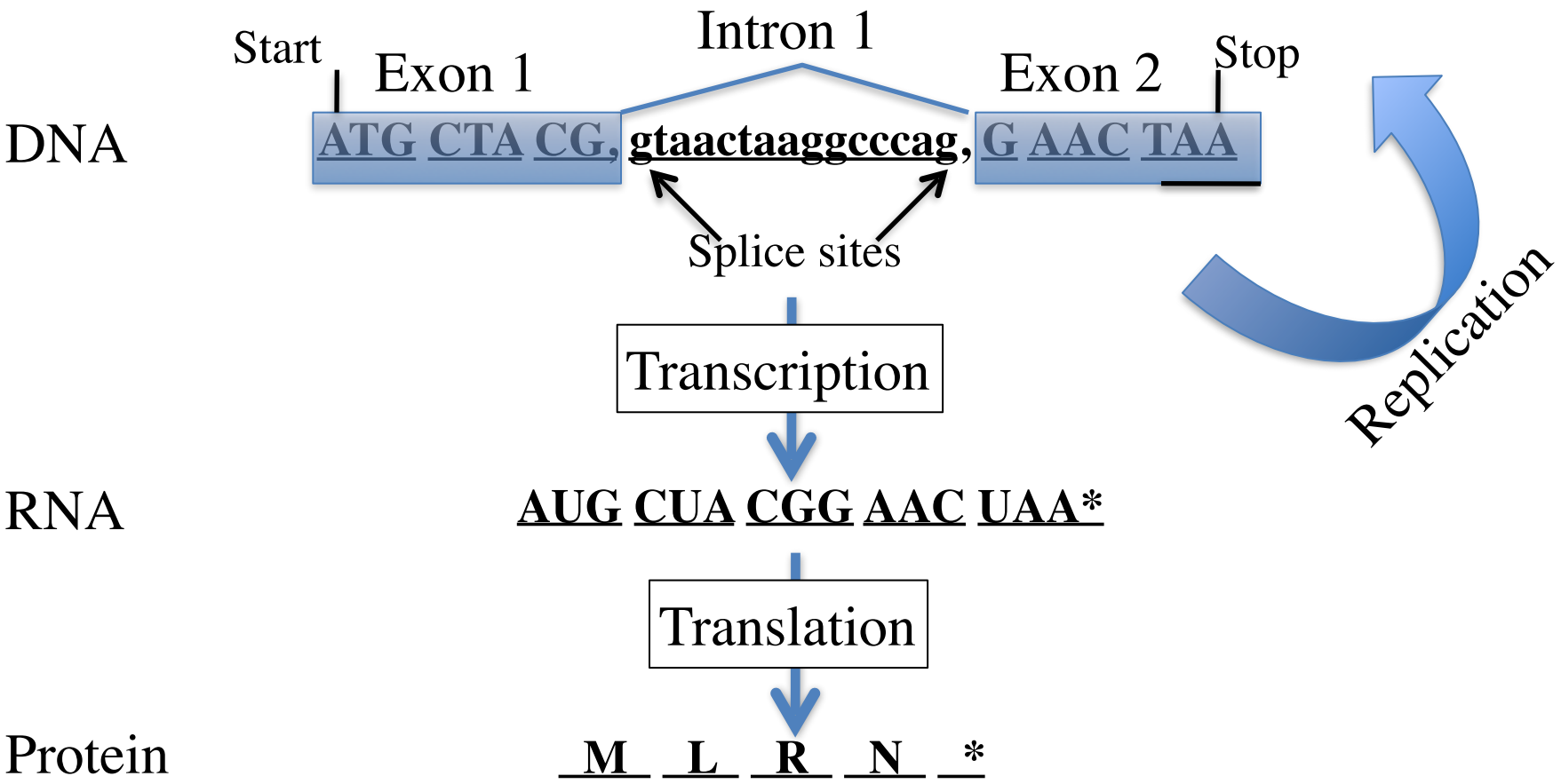
Projezieren wir das ganze auf ein Genmodell



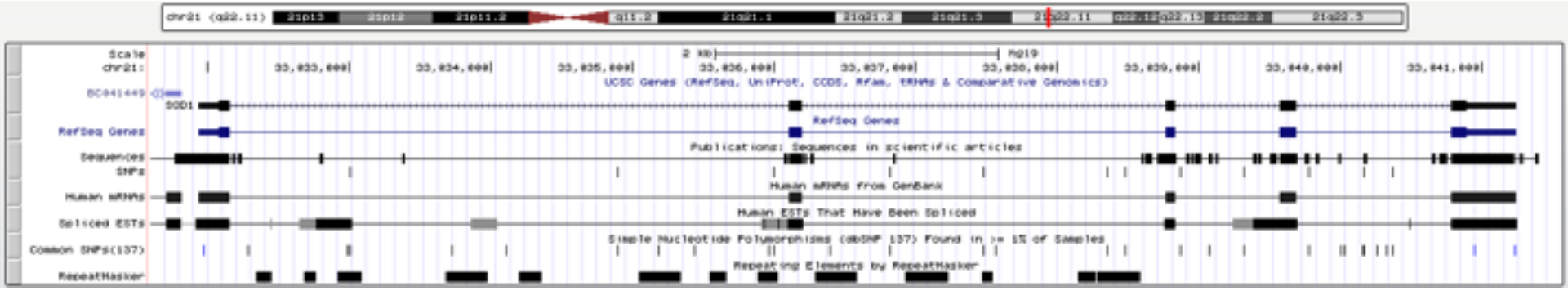
Heute ißt ingo ein schwein

In Kürze: Die Auswirkung einer Mutation hängt ab von **(a)** der Art der Mutation, i.e. Substitution, Insertion oder Deletion und **(b)** wo im Gen die Mutation geschieht. Die Wahrscheinlichkeit, dass eine Mutation "überlebt" hängt davon ab, welchen Effekt sie für die Information hat.

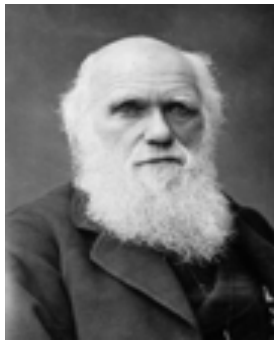
Projizieren wir das ganze auf ein Genmodell



Welchen Einfluss hat “Funktion” nun wirklich auf die Evolution von DNA?



Die überwiegende Anzahl aller Mutationen geschieht in Introns oder intergenischen Regionen und haben vermutlich keinerlei Einfluss auf eine molekulare Funktion.



?

Charles Darwin:
Evolution is driven by positive selection

Motoo Kimura:
Neutral model of molecular evolution

Die neutrale Theorie der Molekularen Evolution

Motoo Kimura

This book represents my attempt to convince the scientific world that the main cause of evolutionary change at the molecular level – changes in the genetic material itself – is random fixation of selectively neutral or nearly neutral mutants rather than positive Darwinian selection. This thesis, which I here call the neutral theory of molecular evolution, has caused a great deal of controversy since I proposed it in 1968 to explain some then new findings in evolution and variation at the molecular level. The controversy is not surprising, since evolutionary biology has been dominated for more than half a century by the Darwinian view that organisms become progressively adapted to their environments by accumulating beneficial mutants, and evolutionists naturally expected this principle to extend to the molecular level. The neutral theory is not antagonistic to the cherished view that evolution of form and function is guided by Darwinian selection, but it brings out another facet of the evolutionary process by emphasizing the much greater role of mutation pressure and random drift at the molecular level.

M. Kimura "**The Neutral Theory of molecular evolution**" 1983, Cambridge University Press

Die neutrale Theorie der Molekularen Evolution

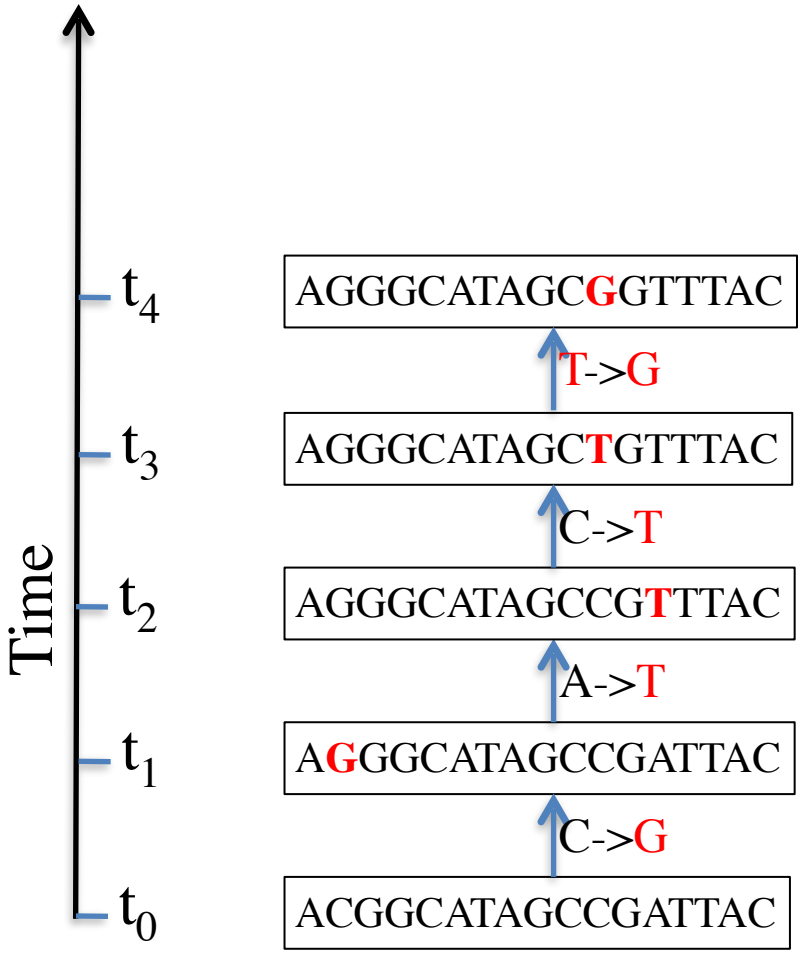
Motoo Kimura

In the decades since its introduction, the neutral theory of evolution has become central to the study of evolution at the molecular level, in part because it provides a **way to make strong predictions that can be tested against actual data**.

Bitte erinnern Sie sich an diese Aussage wann immer sie einen Selektionstest durchführen sollen!

L. Duret "**Neutral Theory: The Null Hypothesis of Molecular Evolution**" Nature Education 1(1)

Die Simulation von Zeit-abhängiger Sequenzveränderung im Zusammenhang mit drei Grundannahmen

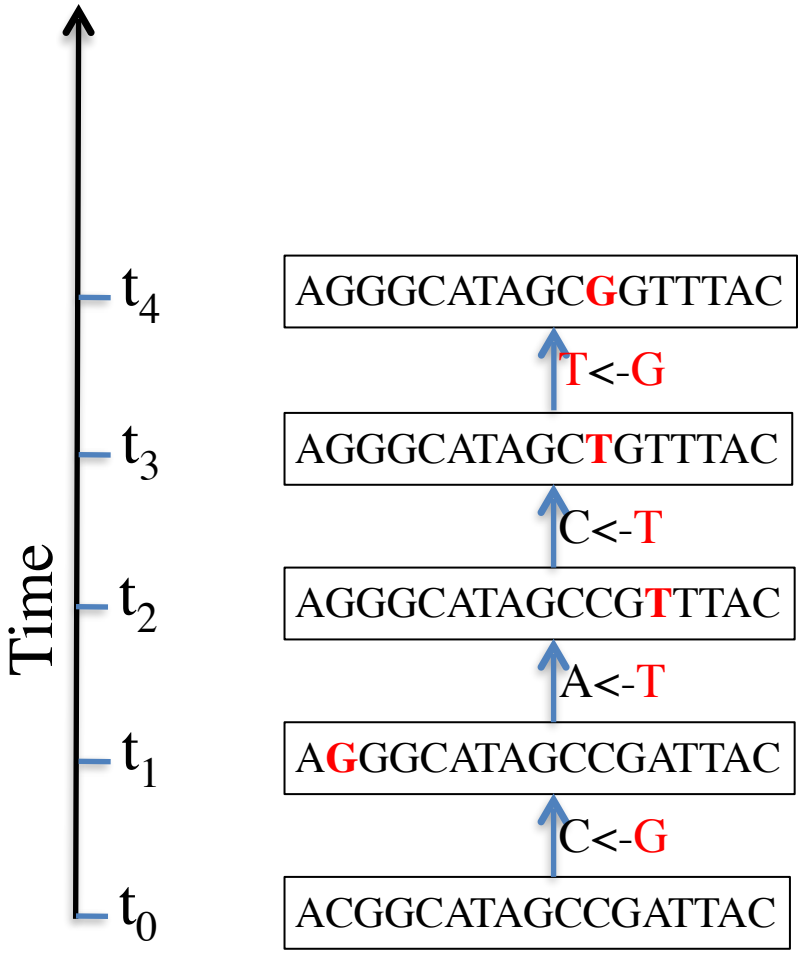


- Achtung, eine Position kann sich im Laufe der Zeit mehrfach verändern (multiple Substitution)

Dieses sind die Grundbedingungen, um Sequenzevolution mittels einer Zeit-kontinuierlichen Markov-Kette zu modellieren!

- **Stationär**, d.h. die relative Häufigkeit der Buchstaben ändert sich nicht
- **Gedächtnislos**, d.h. die Wahrscheinlichkeit einer Veränderung hängt nur vom unmittelbar vorhergehenden Zustand ab
- **Zeit-kontinuierlich**, d.h. Veränderungen können zu jeder Zeit *t* stattfinden.

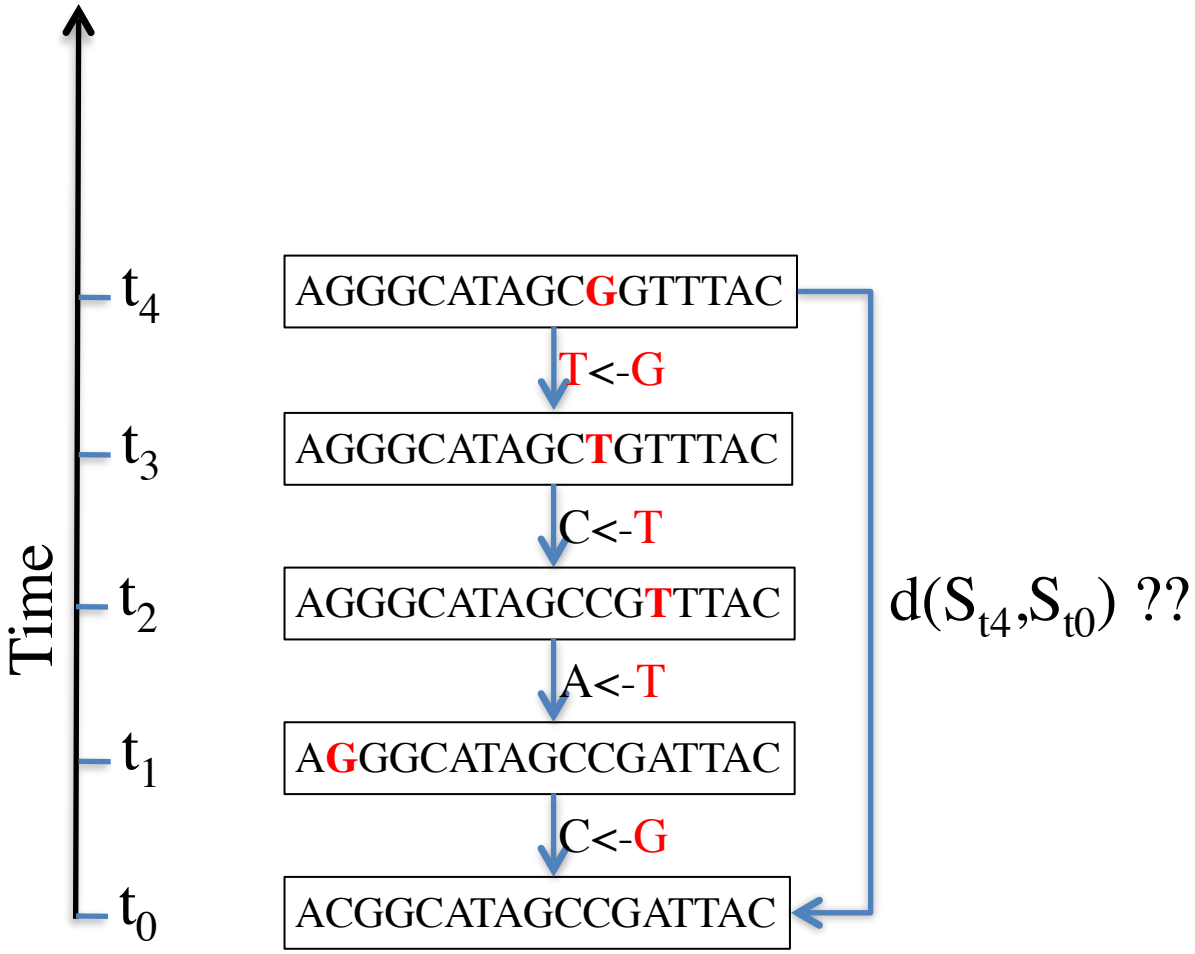
Die Simulation von Zeit-abhängiger Sequenzveränderung im Zusammenhang mit drei + einer Grundannahme



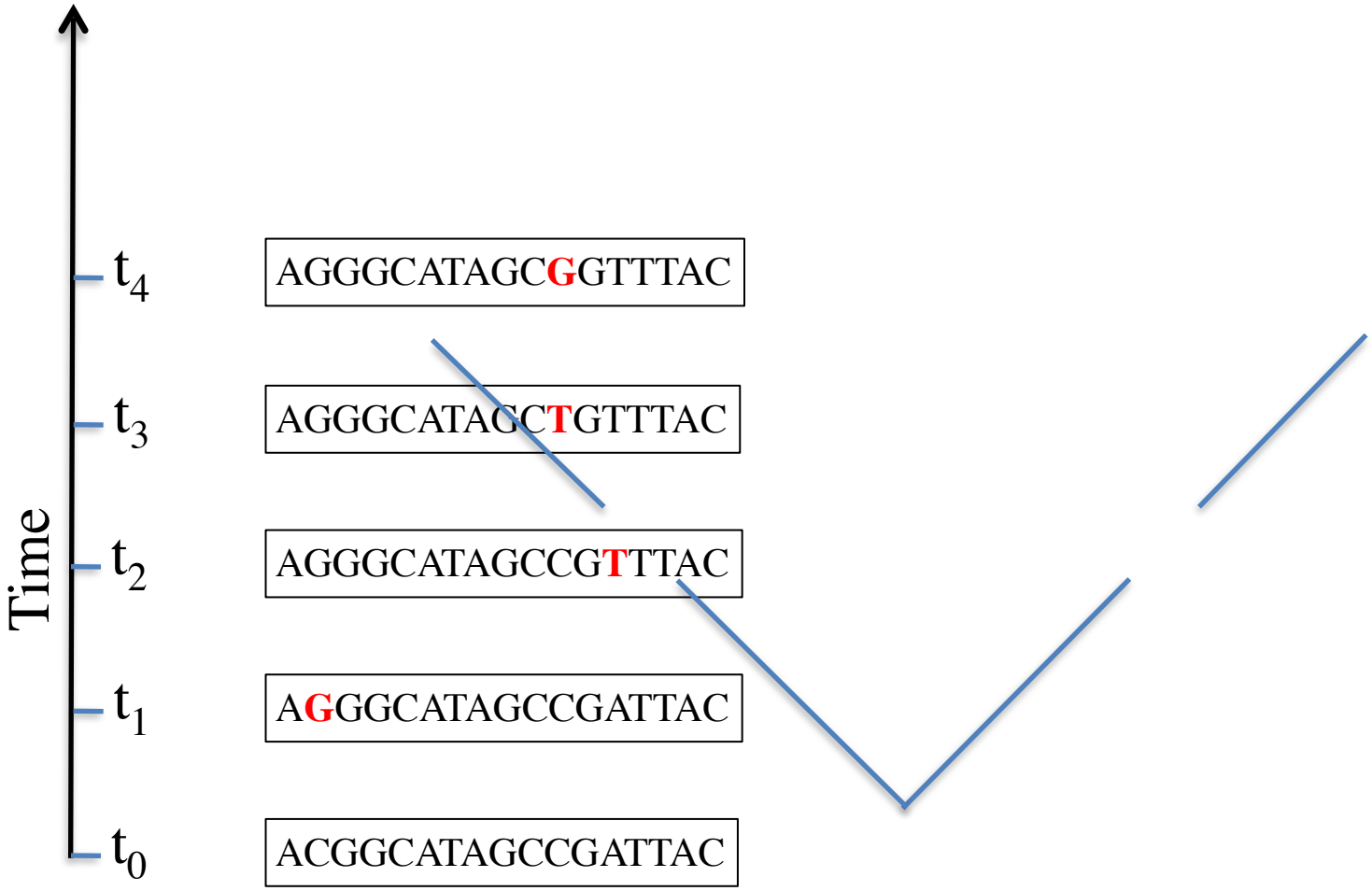
Typischerweise machen wir noch eine vierte Annahme:

- **Zeit-reversibel**, d.h. es macht keinen Unterschied, ob der Prozess von t_0 nach t_4 läuft oder von t_4 nach t_0! (das ist wichtig!!!)
- **Stationär**, d.h. die relative Häufigkeit der Buchstaben ändert sich nicht
- **Gedächtnislos**, d.h. die Wahrscheinlichkeit einer Veränderung hängt nur vom unmittelbar vorhergehenden Zustand ab
- **Zeit-kontinuierlich**, d.h. Veränderungen können zu jeder Zeit *t* stattfinden.

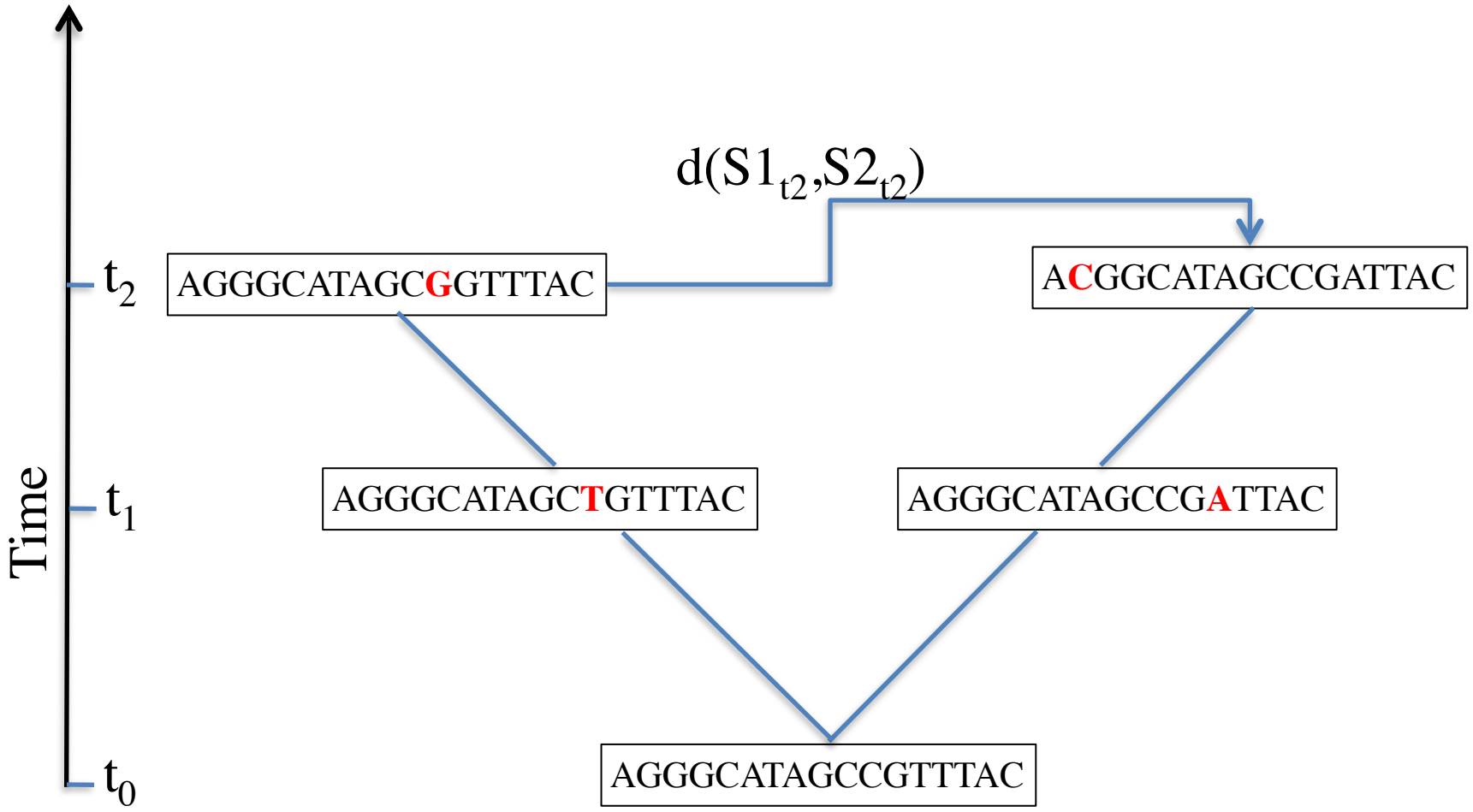
Warum sollte uns das nun interessieren? Nun, manchmal interessiert uns die Distanz zwischen zwei Sequenzen



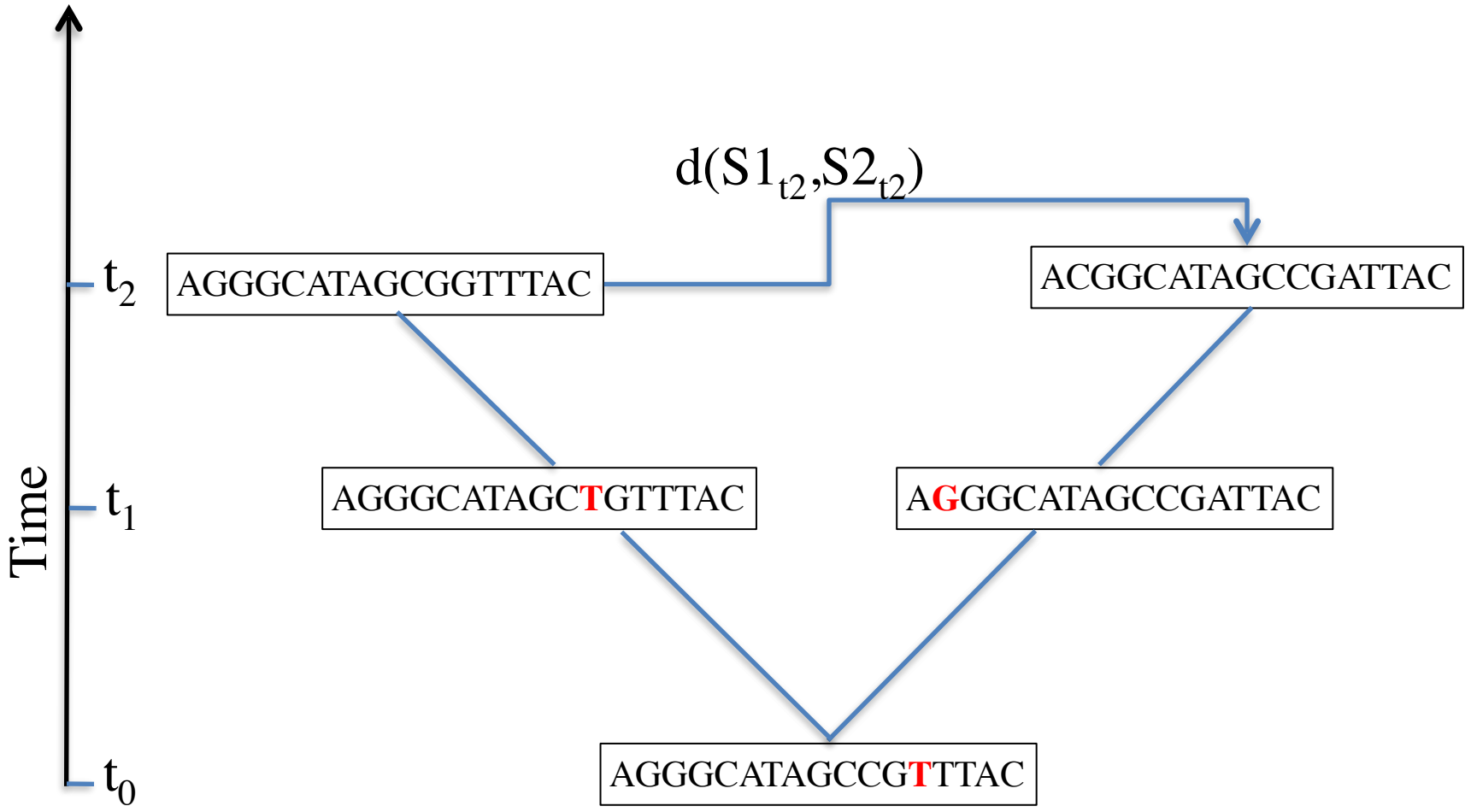
Bisher gestaltet sich das schwierig, da wir ja nicht zurück in die Zeit blicken können, also wenden wir einen kleinen Trick an.
(thanks to time-reversibility).



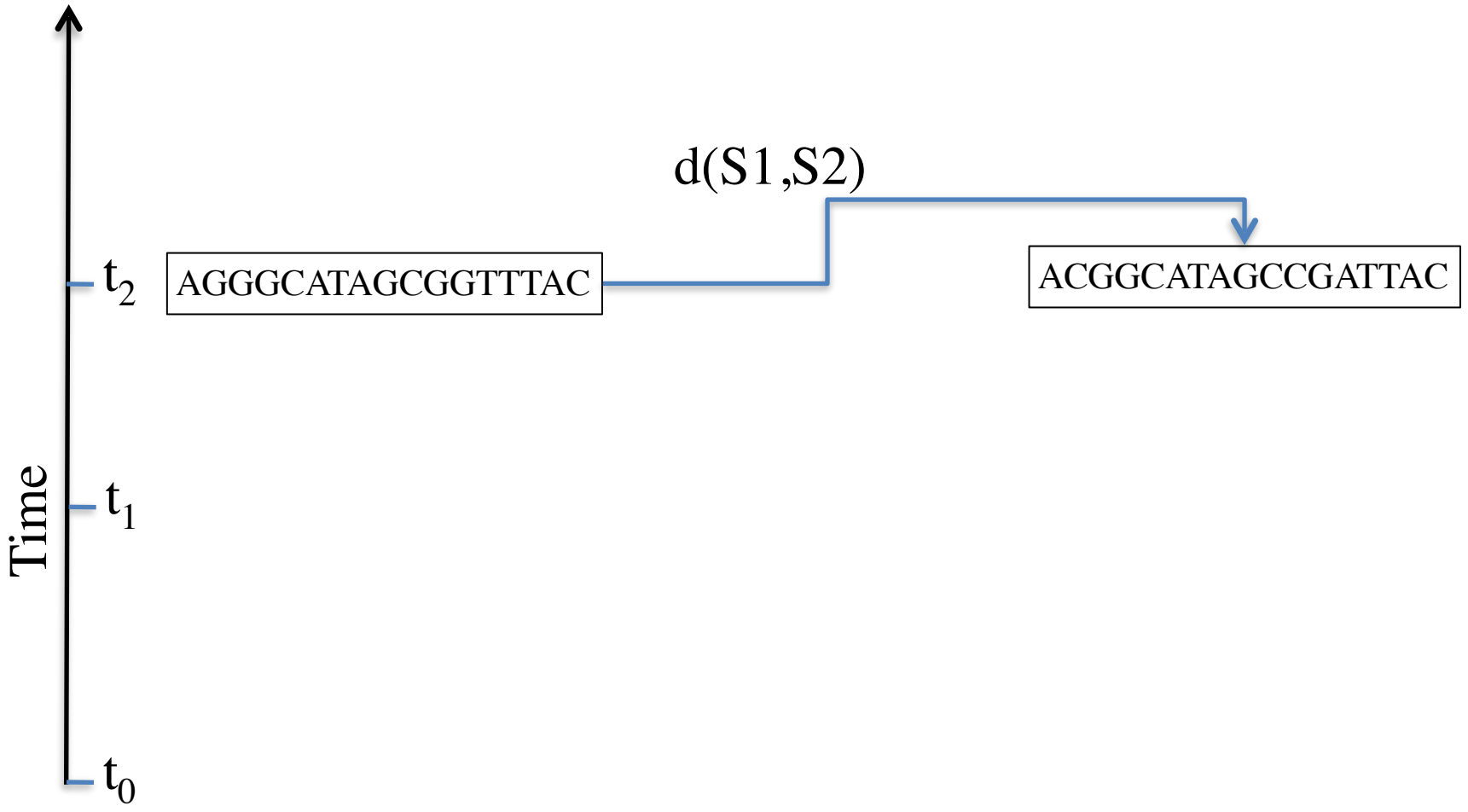
Wir können jetzt zwei heutige Sequenzen miteinander vergleichen, die sich einen gemeinsamen Vorfahren teilen



In der Zeit können wir immer noch nicht zurückblicken, wissen also nichts über die evolutionäre Geschichte der Sequenzen!



Letztlich können wir also nur Unterschiede zwischen heutigen Sequenzen zählen.

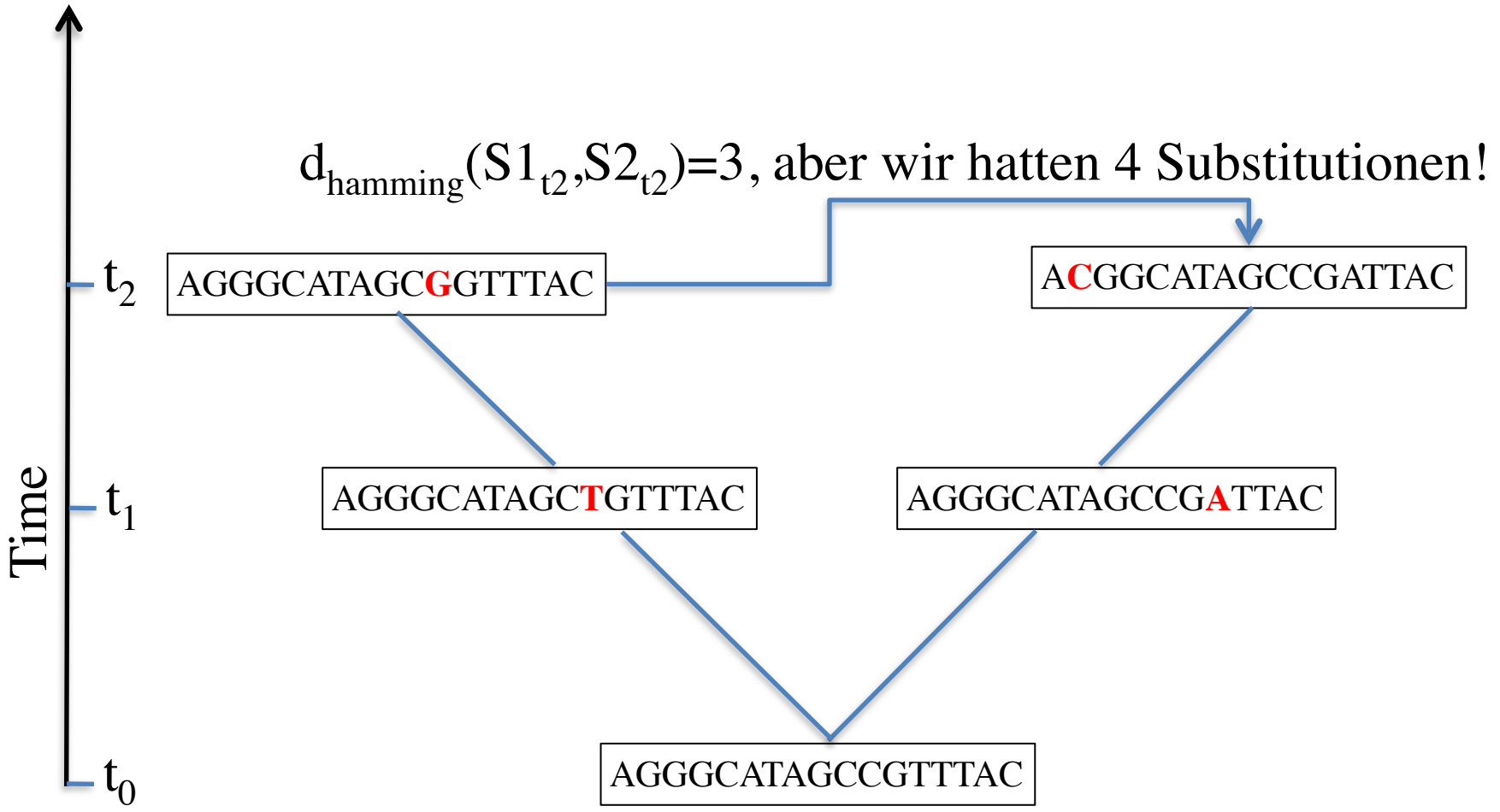


Die Hamming Distanz ist die Anzahl der Editier-Schritte, die ich benötige um die eine Sequenz in die andere zu überführen

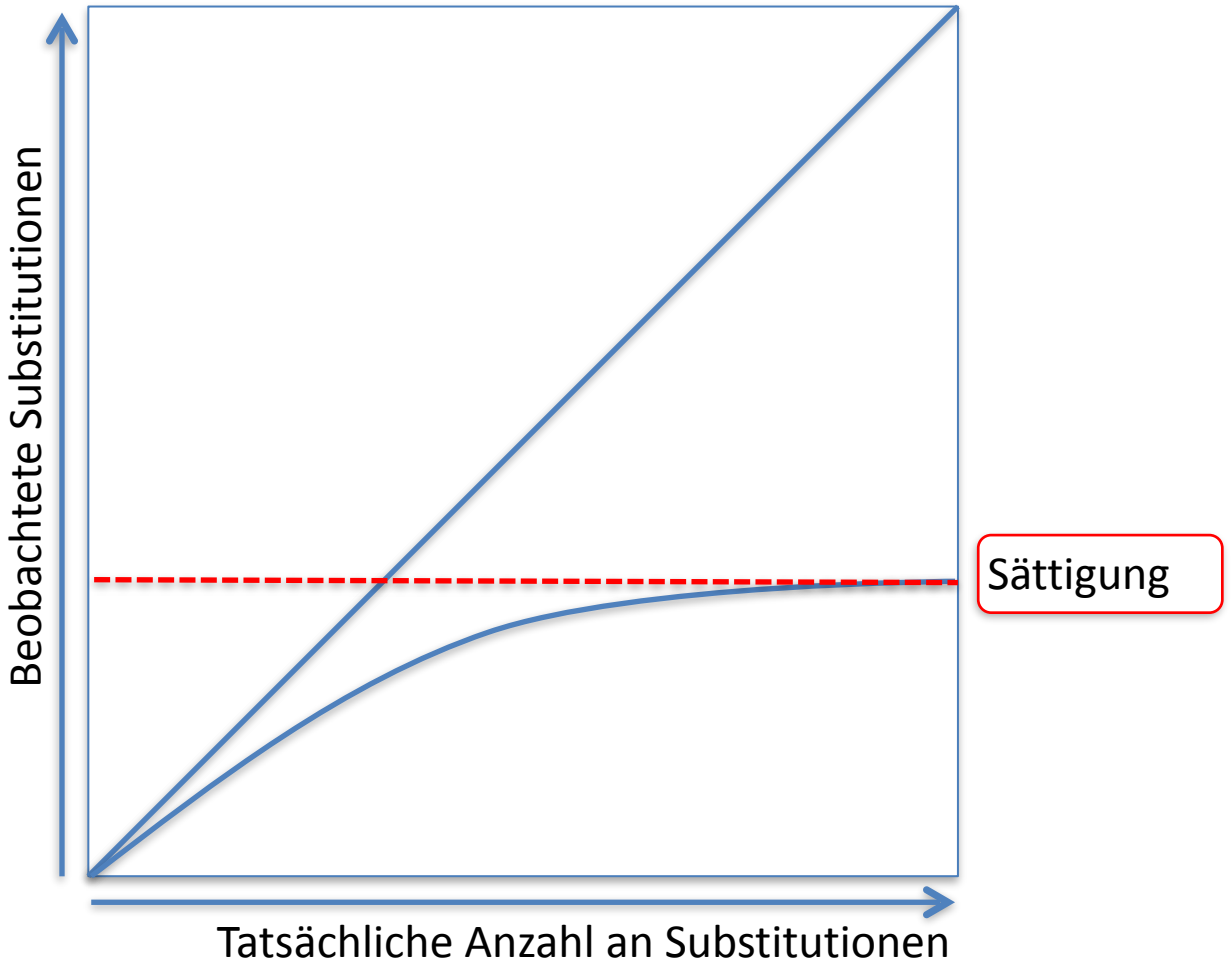
$$d_{hamming}(S1,S2)=3$$

Goal accomplished!?

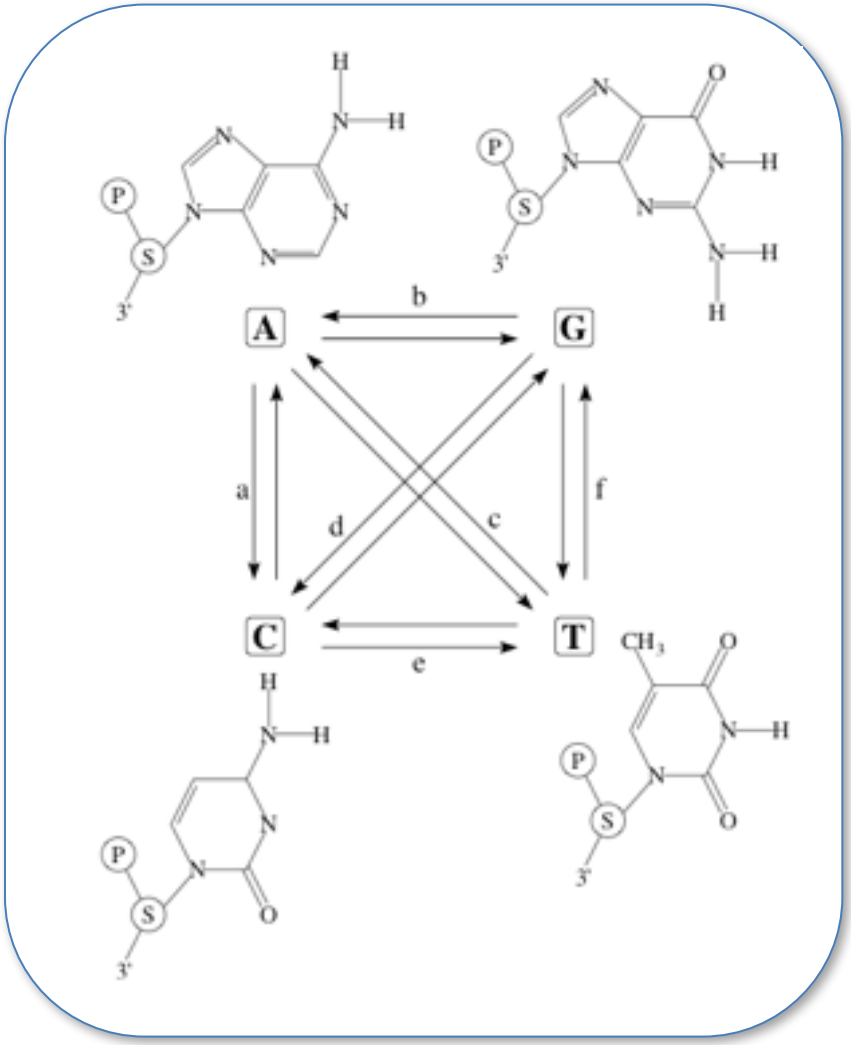
Die Hamming Distanz ist in der Regel kleiner als die wahre Distanz (tatsächliche Anzahl der Veränderungen)



Der Unterschied zwischen der Hamming Distanz und der wahren Distanz steigt mit der Anzahl der Substitutionen



Lösung: Wir brauchen eine **Erwartung** wie viele Substitutionen tatsächlich stattgefunden haben*. Wir müssen uns von der Beobachtung von Veränderung weg- und zur Modellierung von Sequenzveränderung hinbewegen.

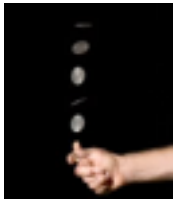


Transition (Ti): Substitution Pyrimidine -> Pyrimidine or Purin -> Purin

Transversion (Tv): Substitution Pyrimidine <-> Purin

*Das geschieht i.d.R. nur für Substitutions und nicht für Insertionen und Deletionen. Warum ist das so??

Evolution wirft eine Münze…



Um in unserem Bild zu bleiben: Der Mönch wirft für jeden zu kopierenden Buchstaben eine Münze ob er ihn richtig oder falsch kopiert. Es handelt sich also um ein **Bernoulli** experiment.

Nehmen wir an, die Wahrscheinlichkeit für einen Fehler ist

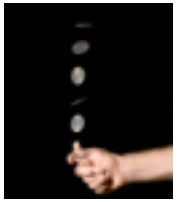
$$p \leq 1$$

Daraus folgt die Wahrscheinlichkeit für den korrekten Buchstaben:

$$q = 1 - p$$

Für eine faulen oder rechtschreib-schwachen Mönch wäre $p = q = 0.5$

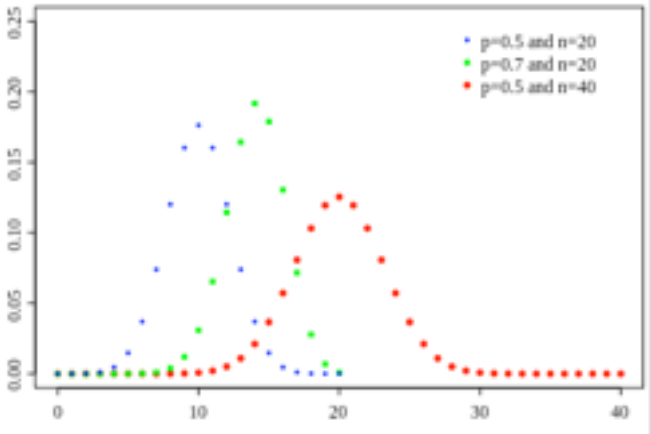
Evolution wirft eine Münze mehrfach hintereinander…



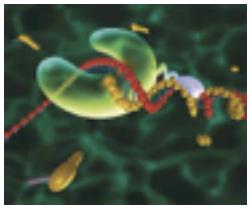
Da der Mönch mehr als einen Buchstaben kopieren muss, muss er seine Münze mehrfach hintereinander werfen. Tatsächlich genauso häufig wie es Buchstaben zu kopieren gibt.

Das Ergebnis von ***n*** aufeinander folgenden Bernoulli Experimenten wird durch eine **Binomialverteilung** modelliert. Die Zufallsvariable ***X*** stellt die Anzahl der falsch kopierten Buchstaben in *n* Experimenten da. Die Wahrscheinlichkeit $X = \boldsymbol{k}$ Fehler zu beobachten ist dann:

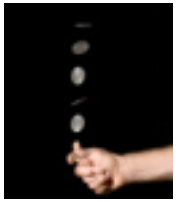
$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$



Probability density function of a binomial distribution



Evolution wirft eine Münze mehrfach hintereinander…



- Glücklicherweise kann die DNA Polymerase einen Text viel exakter kopieren als jeder Mönch. Die Fehlerwahrscheinlichkeit ***p*** ist also sehr klein (Größenordnung 10^{-9} pro Position und Generation in Eukaryoten).

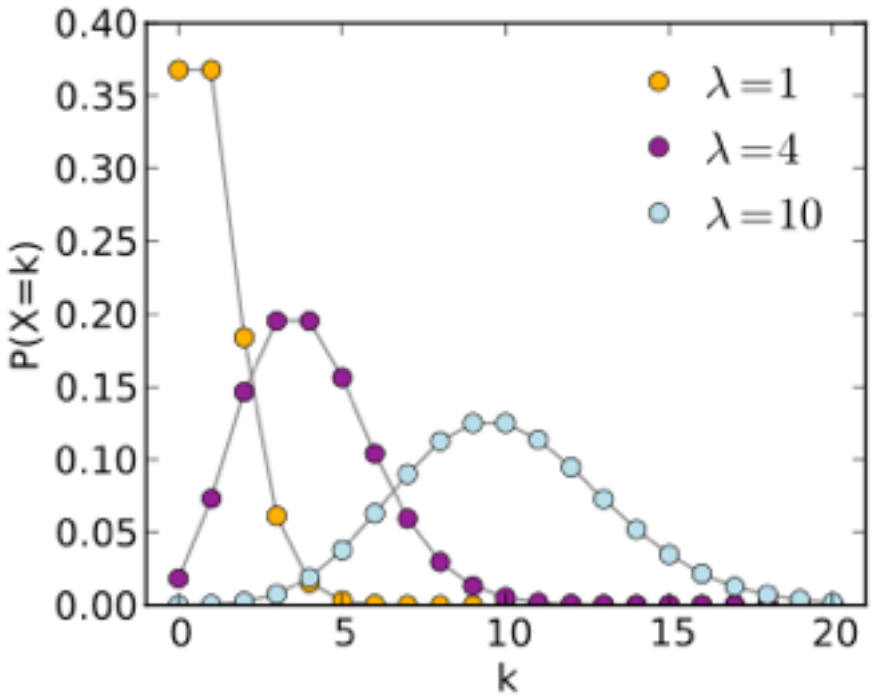
- Weiterhin ist die Anzahl an aufeinanderfolgenden Experimenten ***n*** sehr gross (denken sie and die 3.2 Milliarden Basenpaare des menschlichen Genoms).

- Unter diesen Umständen konvergiert die Binomialverteilung zur **Poisson-Verteilung** mit $np = \lambda > 0$. Die Wahrscheinlichkeit ***k*** Substitutionen zu beobachten ist dann

$$\Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Die Poisson-Verteilung hat den Erwartungswert n*p = λ

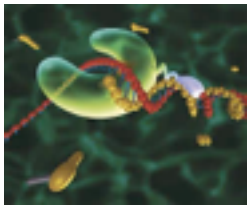
Die Poisson Verteilung modelliert die Anzahl von Substitutionen in einem bestimmten Zeitinterval *t*



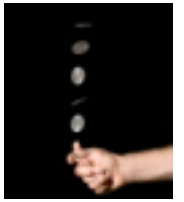
Probability density function

Die Poisson Verteilung ist eine diskrete Wahrscheinlichkeitsfunktion die die Wahrscheinlichkeit einer bestimmten Anzahl von Ereignissen in einem fixen Zeitinterval bestimmt.

Voraussetzung: Diese Ereignisse finden mit einer bekannten mittleren Rate λ statt und das Eintreten ist unabhängig vom Zeitpunkt des letzten Ereignisses.



Wie lange müssen wir bis zum ersten Ereignis warten, i.e. *k=1*?



Schaut schwierig aus, lässt sich aber aus dem bisher gehörten recht einfach herleiten:

Die Wahrscheinlichkeit dass das erste Ereignis T_1 nach Zeit ***t*** eintritt ist gleich der Wahrscheinlichkeit, dass bis einschließlich ***t*** nichts passiert ist.

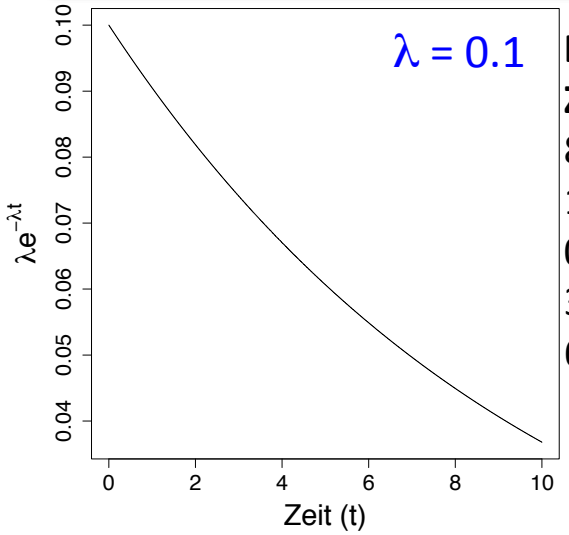
$$\Pr(T_1 > t) = \Pr(N(t) = 0)^*$$

Wir können also die **Wartezeit *t*** bis zum Eintritt des ersten Events wie folgt modellieren:

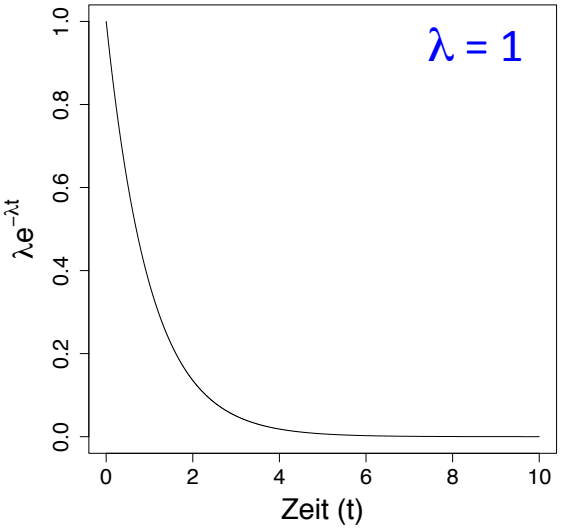
$$\Pr(N(t) = 0) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t}$$

*N(t) is our Poisson process

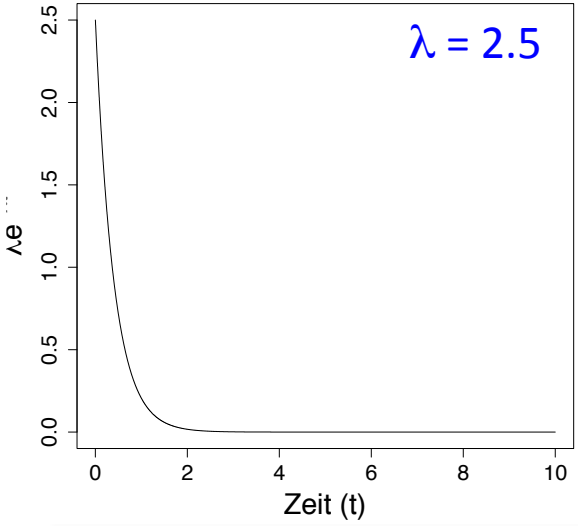
Die Wahrscheinlichkeitsdichtefunktion für verschiedene Raten



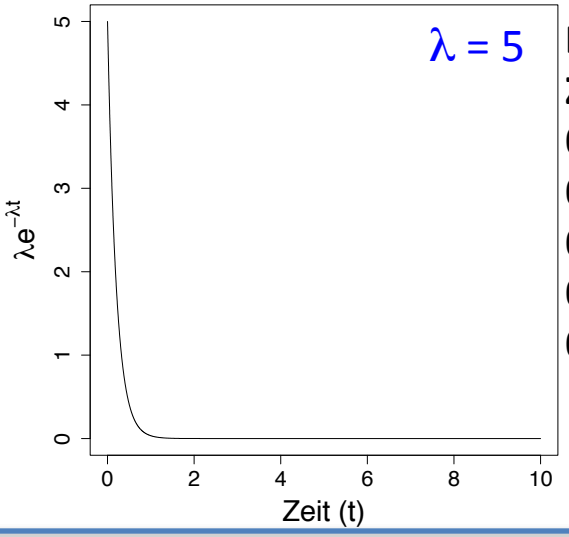
Ergebnis von 5 Ziehungen
8.1
15.8
0.7
3.6
6.0



Ergebnis von 5 Ziehungen
2.5
2.3
0.7
2.2
0.4

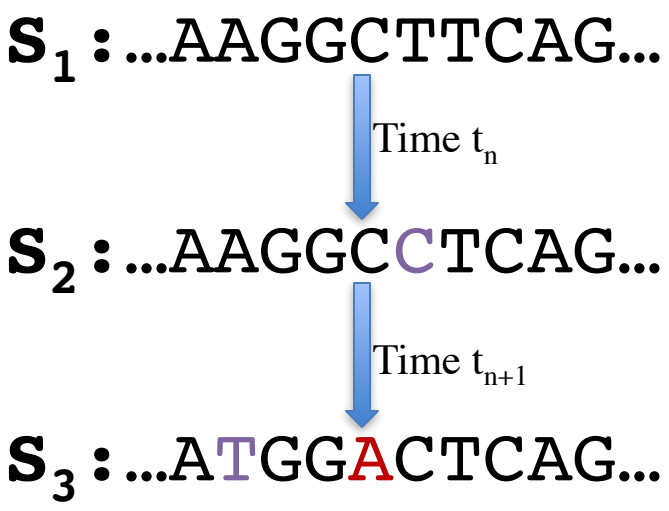


Ergebnis von 5 Ziehungen
0.3
0.4
0.2
0.1
1.3



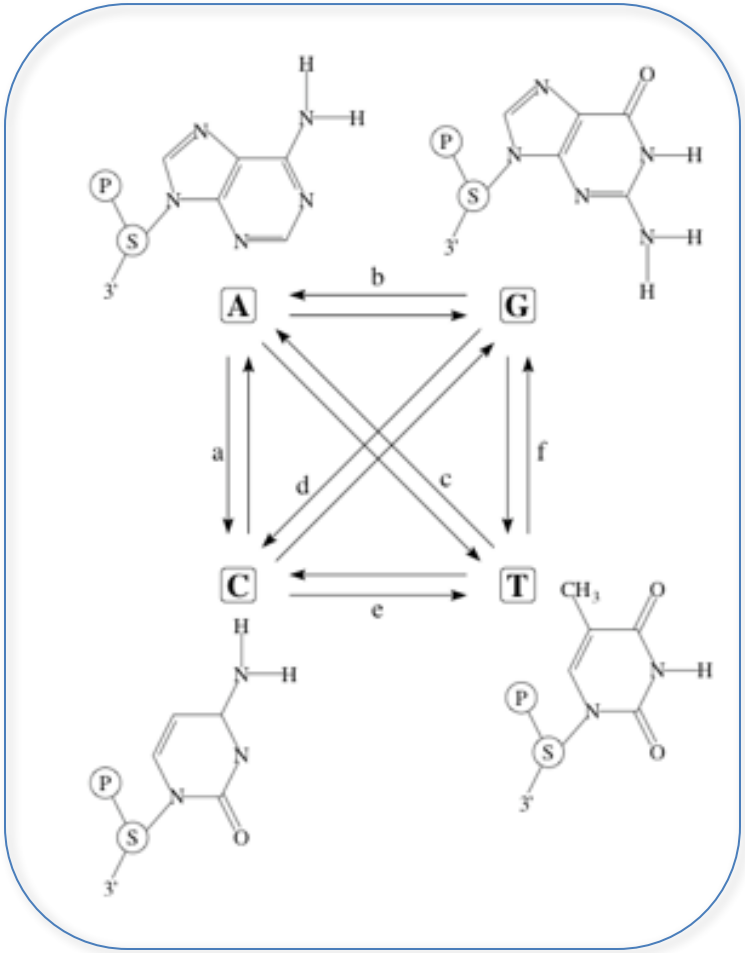
Ergebnis von 5 Ziehungen
0.30
0.01
0.27
0.16
0.14

Zurück zur Modellierung von Sequenzevolution als eine Zeit-kontinuierliche Markov-Kette



1. **Markov-Kette erster Ordnung** Der Evolutionsprozess hat keine Erinnerung, i.e. Sequenz S_2 mutiert zu S_3 in Zeit t_{n+1} unabhängig von S_1
2. **Stationar**
   Die Häufigkeiten π_j der Nukleotide oder Aminosäure verändern sich nicht.
3. **Zeit-reversibel**
   $\pi_i \cdot P_{ij}(t) = P_{ji}(t) \cdot \pi_j$

Die Ratenmatrix *Q* enthält die Raten für die 12 möglichen Übergänge zwischen den Basen*. Nehmen wir Zeit-Reversibilität an reduziert sich die Anzahl der Raten** auf 6

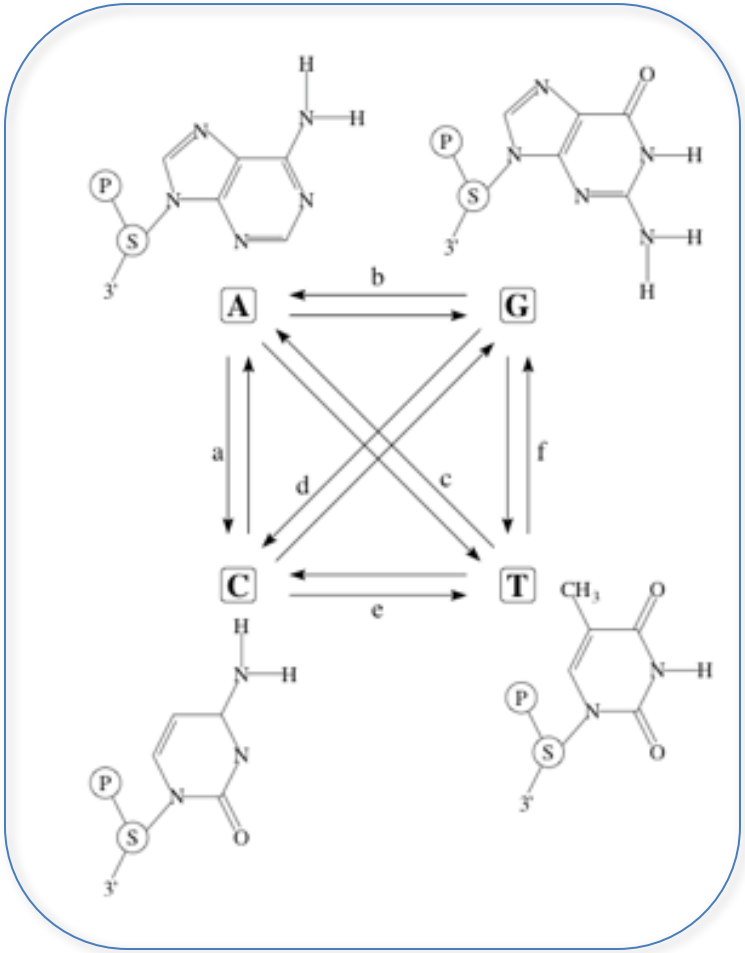


$$Q = \begin{array}{c} \begin{array}{cccc} \mathbf{A} & \mathbf{C} & \mathbf{G} & \mathbf{T} \end{array} \\ \begin{pmatrix} - & a & b & c \\ a & - & d & e \\ b & d & - & f \\ c & e & f & - \end{pmatrix} \end{array}$$

Die Rate der Substitution von Base *i* zu Base *j* wird durch den entsprechenden Eintrag q_{ij} in der Ratenmatrix Q gegeben

*Die vier Basen kann man auch als 'Zustände' ansehen. **Raten sind **KEINE WAHRSCHEINLICHKEITEN**, da sie Zeit-unabhängig sind!!

Die Ratenmatrix Q
Konventionen



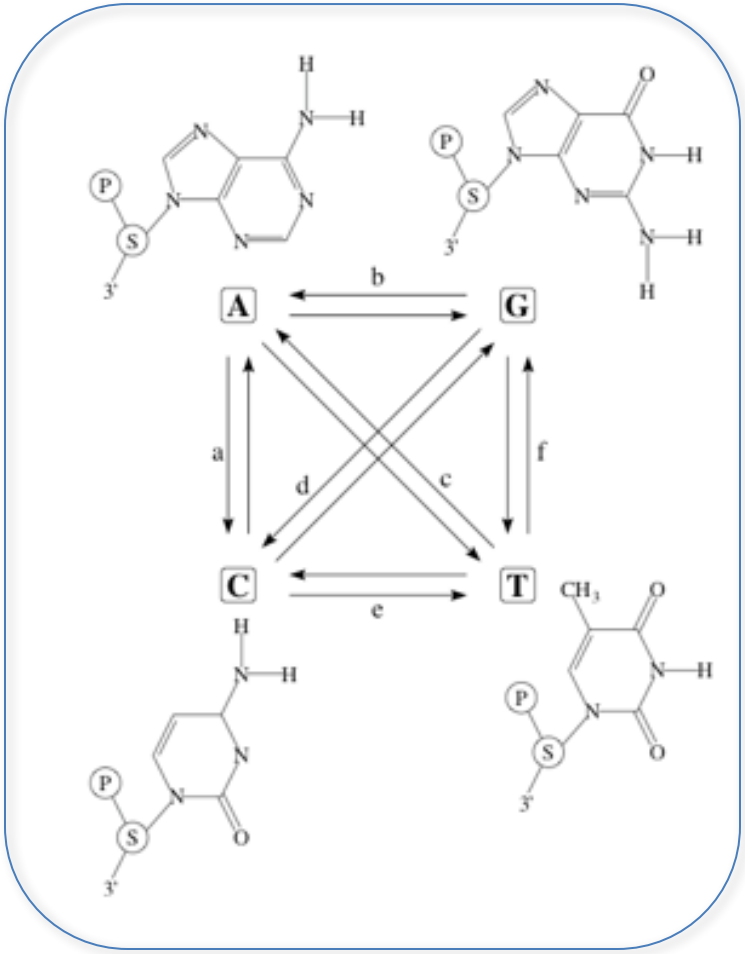
$$Q = \begin{array}{c} \begin{array}{cccc} \mathbf{A} & \mathbf{C} & \mathbf{G} & \mathbf{T} \end{array} \\ \begin{pmatrix} q_{AA} & a & b & c \\ a & q_{CC} & d & e \\ b & d & q_{GG} & f \\ c & e & f & q_{TT} \end{pmatrix} \end{array}$$

Die diagonalen Einträge q_{ii} werden so gewählt, dass die Zeilensumme 0 ergibt.

Also,

$$q_{ii} = -\sum_{j \neq i} q_{ij}$$

Der Häufigkeitsvektor Π enthält die Gleichgewichtshäufigkeiten der 4 Basen*.

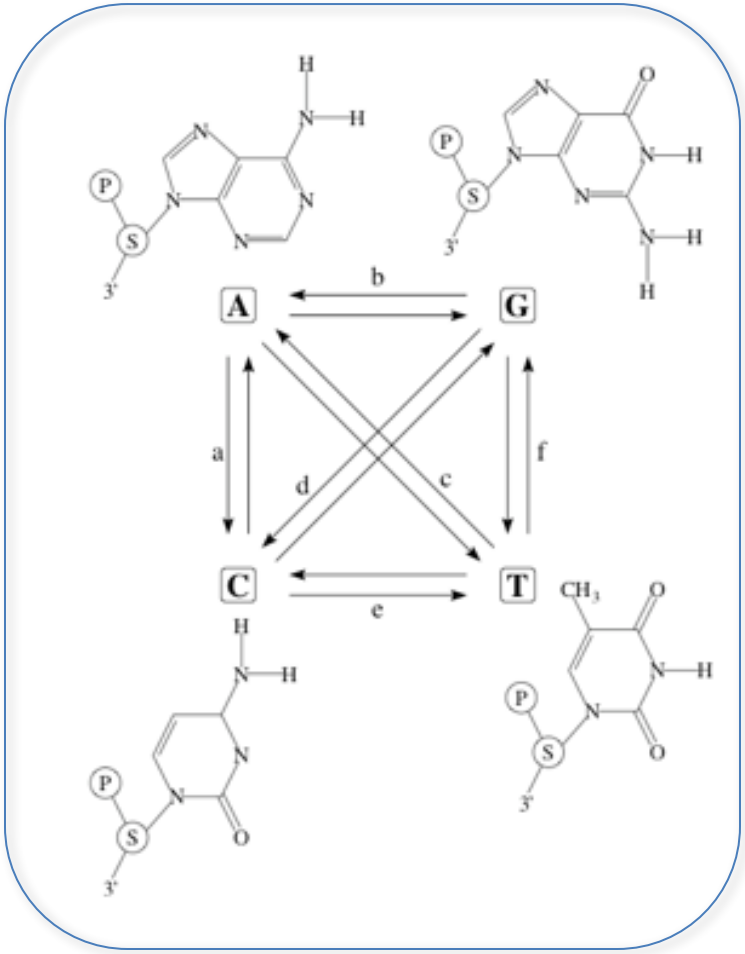


$$Q = \begin{array}{c} \begin{array}{cccc} \mathbf{A} & \mathbf{C} & \mathbf{G} & \mathbf{T} \end{array} \\ \begin{pmatrix} q_{AA} & a & b & c \\ a & q_{CC} & d & e \\ b & d & q_{GG} & f \\ c & e & f & q_{TT} \end{pmatrix} \end{array}$$

$$\Pi = (\pi_A, \pi_C, \pi_G, \pi_T)$$

*Unsere Annahme ist, dass der Substitutionsprozess die Basenhäufigkeit nicht verändert, wir müssen sie also im Modell spezifizieren

Die Gesamtsubstitutionsrate von Base *i* zu Base *j* ist das Produkt der Rate q_{ij} und der Basenhäufigkeit π_j.*



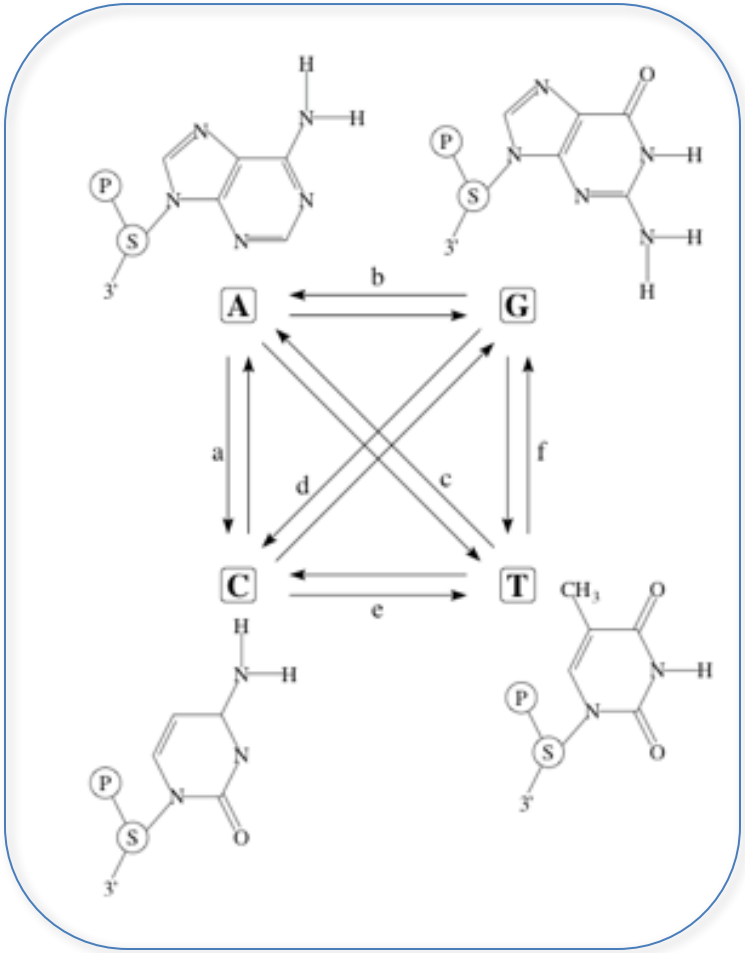
$$Q = \begin{array}{c} \begin{array}{cccc} \mathbf{A} & \mathbf{C} & \mathbf{G} & \mathbf{T} \end{array} \\ \begin{pmatrix} q_{AA} & a & b & c \\ a & q_{CC} & d & e \\ b & d & q_{GG} & f \\ c & e & f & q_{TT} \end{pmatrix} \end{array}$$

$$\Pi = (\pi_A, \pi_C, \pi_G, \pi_T)$$

$$r_{ij} = \pi_j \times q_{ij}; \forall i \neq j \in \{A, C, G, T\}$$

*"It is unclear whether it is biologically reasonable to consider these two sets of parameters as representing different forces that affect nucleotide substitution, but this distinction is mathematically convenient." Z. Yang (1994) J Mol Evol 39:105-111

Die Ratenmatrix Q
Konventionen



$$Q = \begin{array}{c} \begin{array}{cccc} \mathbf{A} & \mathbf{C} & \mathbf{G} & \mathbf{T} \end{array} \\ \begin{pmatrix} q_{AA} & a & b & c \\ a & q_{CC} & d & e \\ b & d & q_{GG} & f \\ c & e & f & q_{TT} \end{pmatrix} \end{array}$$

$$\Pi = (\pi_A, \pi_C, \pi_G, \pi_T)$$

Die Ratenmatrix wird so skaliert, dass sie zu 1 aufaddiert.

$$-\sum_i \pi_i q_{ii} = 1$$

Typischerweise re-skaliert man die Substitutionsratenmatrix so, dass die Gesamtrate 1 wird. Wenn man dies tut, kann man Zeit in Substitutionen pro Position messen.

$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} - & a & b & c \\ a & - & d & e \\ b & d & - & f \\ c & e & f & - \end{pmatrix} \end{array}$$

$$\Pi = (\pi_A, \pi_C, \pi_G, \pi_T)$$

$$\pi_C a + \pi_G b + \pi_T c +$$

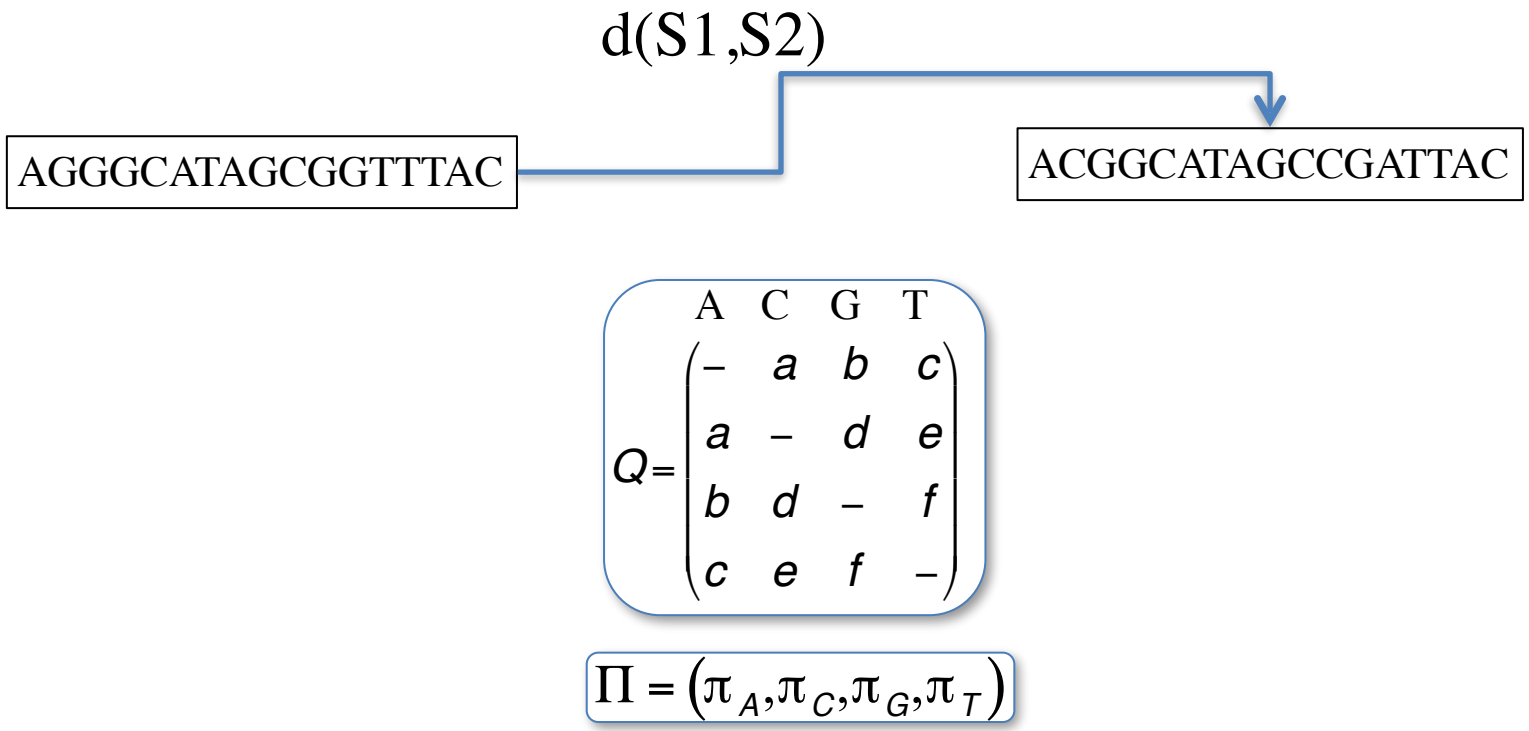
$$\pi_A a + \pi_G d + \pi_T e +$$

$$\pi_A b + \pi_C d + \pi_T f +$$

$$\pi_A c + \pi_C e + \pi_G f = 1$$

Es ist einfach zu sehen, dass für $\pi_i = \frac{1}{4}, \forall i \in \{A, C, G, T\}$ und $q_{ij} = \alpha, \forall i \neq j \in \{A, C, G, T\}$ $\alpha = \frac{1}{3}$

Wunderbar, aber wir haben die Geschichte mit der Frage begonnen wieviele Substitutionen stattgefunden haben seit sich zwei Sequenzen zuletzt einen gemeinsamen Vorfahren geteilt haben.



Wir müssen die Modellierung also weiterführen und irgendwie die Zeit zurück ins Spiel bringen um damit Raten in Wahrscheinlichkeiten zu überführen

Wie überführt man Raten in Wahrscheinlichkeiten ?

$$Q = \begin{pmatrix} q_{AA} & q_{AC} & q_{AG} & q_{AT} \\ q_{CA} & q_{CC} & q_{CG} & q_{CT} \\ q_{GA} & q_{GC} & q_{GG} & q_{GT} \\ q_{TA} & q_{TC} & q_{TG} & q_{TT} \end{pmatrix}$$

$$P(t) = \begin{pmatrix} p_{AA}(t) & p_{AC}(t) & p_{AG}(t) & p_{AT}(t) \\ p_{CA}(t) & p_{CC}(t) & p_{CG}(t) & p_{CT}(t) \\ p_{GA}(t) & p_{GC}(t) & p_{GG}(t) & p_{GT}(t) \\ p_{TA}(t) & p_{TC}(t) & p_{TG}(t) & p_{TT}(t) \end{pmatrix}$$

?

Betrachten wir die Veränderungswahrscheinlichkeit an einer Position in einem Zeitraum Δt

Übergangsratenmatrix *Q*

$$Q = \begin{pmatrix} q_{AA} & q_{AC} & q_{AG} & q_{AT} \\ q_{CA} & q_{CC} & q_{CG} & q_{CT} \\ q_{GA} & q_{GC} & q_{GG} & q_{GT} \\ q_{TA} & q_{TC} & q_{TG} & q_{TT} \end{pmatrix}$$

Gesamtrate den Zustand *i* zu verlassen

$$q_i = \sum_{j \neq i} q_{ij}$$

Was ist die Wahrscheinlichkeit ein A, C, G oder T an einer Position nach Zeit ***t*** zu beobachten?

$$P(t) = (p_A(t), p_C(t), p_G(t), p_T(t))^T$$

Spaltenvektor

Was ist die Wahrscheinlichkeit nach $t+\Delta t$ immer noch ein **A** zu beobachten?

$$p_A(t+\Delta t) = p_A(t) - p_A(t) q_A \Delta t + \sum_{i \neq A} p_i(t) q_{iA} \Delta t$$

Wahrscheinlichkeit zu ***t*** ein **A** gehabt zu haben

Wahrscheinlichkeit dass sich ein **A** in *delta t* in einen anderen Zustand verändert hat.

Wahrscheinlichkeit dass sich ein anderer Zustand innerhalb *delta t* in ein **A** verändert hat

Betrachten wir die Veränderungswahrscheinlichkeit an einer Position in einem Zeitraum Δt

Was ist die Wahrscheinlichkeit nach $t+\Delta t$ immer noch ein **A** zu beobachten?

$$p_A(t+\Delta t) = p_A(t) - p_A(t)q_A\Delta t + \sum_{i \neq A} p_i(t)q_{iA}\Delta t$$

Wahrscheinlichkeit zu ***t*** ein **A** gehabt zu haben

Wahrscheinlichkeit dass sich ein **A** in einen anderen Zustand verändert hat

Wahrscheinlichkeit dass sich ein anderer Zustand in ein **A** verändert hat

Nun generalisieren wir das für einen beliebigen Zustand, also eine beliebige Base

$$P(t+\Delta t) = P(t) + P(t)Q\Delta t$$

Wir formen ein wenig um*

$$\frac{P(t+\Delta t) - P(t)}{\Delta t} = P(t)Q$$

Für Δt -> *0* folgt

$$P'(t) = P(t)Q$$

***und erhalten etwas das wie eine Sekantensteigung aussieht, und das schreit förmlich nach Ableitung für Δt -> 0**

Die Übergangswahrscheinlichkeit im Zeitraum *t* ist exponentialverteilt

$$P'(t) = P(t)Q$$

- Wir erinnern uns an die Annahme: unser Markov-Prozess ist stationär. Damit ist Q unabhängig von *t* und somit eine Konstante!
- Entsprechend handelt es sich um eine lineare Differenzialgleichung der allgemeinen Form f'(x)=Af(x)
- Wir suchen also eine Funktion deren Ableitung die Funktion selbst multipliziert mit einer Konstanten ist. Die **Exponentialfunktion** erfüllt diese Bedingung, also folgt

$$P(t) = e^{Qt}P(0)$$

Einheitsmatrix (neutrales Element)

Wie überführt man Raten in Wahrscheinlichkeiten am Beispiel des Jukes Cantor Modells (JC69)

$$Q = \begin{array}{c} \begin{array}{cccc} \mathbf{A} & \mathbf{C} & \mathbf{G} & \mathbf{T} \end{array} \\ \begin{pmatrix} -\frac{3}{4}\alpha & \frac{1}{4}\alpha & \frac{1}{4}\alpha & \frac{1}{4}\alpha \\ \frac{1}{4}\alpha & -\frac{3}{4}\alpha & \frac{1}{4}\alpha & \frac{1}{4}\alpha \\ \frac{1}{4}\alpha & \frac{1}{4}\alpha & -\frac{3}{4}\alpha & \frac{1}{4}\alpha \\ \frac{1}{4}\alpha & \frac{1}{4}\alpha & \frac{1}{4}\alpha & -\frac{3}{4}\alpha \end{pmatrix} \end{array}$$

Aus Q und Π berechnen wir nun die Übergangswahrscheinlichkeitsmatrix* $P(t)$ *als*

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

*The substitution probability matrix is a stochastic matrix describing a Markov chain over a finite state space *S*

Wie überführt man Raten in Wahrscheinlichkeiten am Beispiel des Jukes Cantor Modells (JC69)

$$Q = \begin{array}{c} \begin{array}{cccc} \mathbf{A} & \mathbf{C} & \mathbf{G} & \mathbf{T} \end{array} \\ \begin{pmatrix} -3\alpha & a & a & a \\ a & -3\alpha & a & a \\ a & a & -3\alpha & a \\ a & a & a & -3\alpha \end{pmatrix} \end{array} *$$

$$**$$

$$P(t) = \begin{array}{c} \begin{array}{cccc} \mathbf{A} & \mathbf{C} & \mathbf{G} & \mathbf{T} \end{array} \\ \begin{pmatrix} \frac{1}{4}+\frac{3}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} \\ \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}+\frac{3}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} \\ \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}+\frac{3}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} \\ \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}+\frac{3}{4}e^{-4\alpha t} \end{pmatrix} \end{array}$$

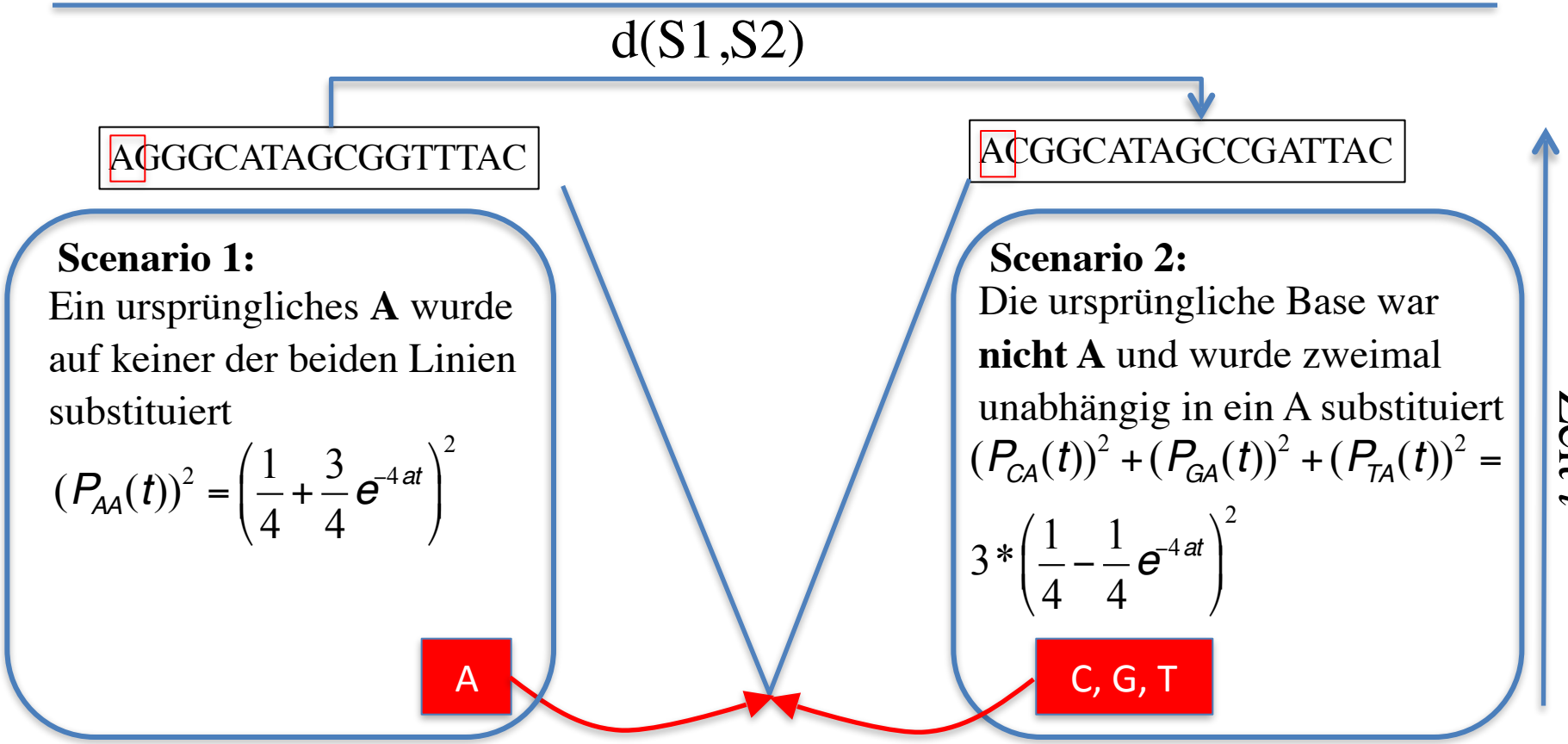
Daraus ergibt sich die Wahrscheinlichkeit, dass sich eine Base im Zeitraum *t* nicht verändert hat als

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-4at}$$

*Wir haben das ein wenig vereinfacht, indem wir die konstante Basenhäufigkeit ¼ in α integriert haben

**Konsultieren Sie die Regeln zur Berechnung einens Matrixexponentials

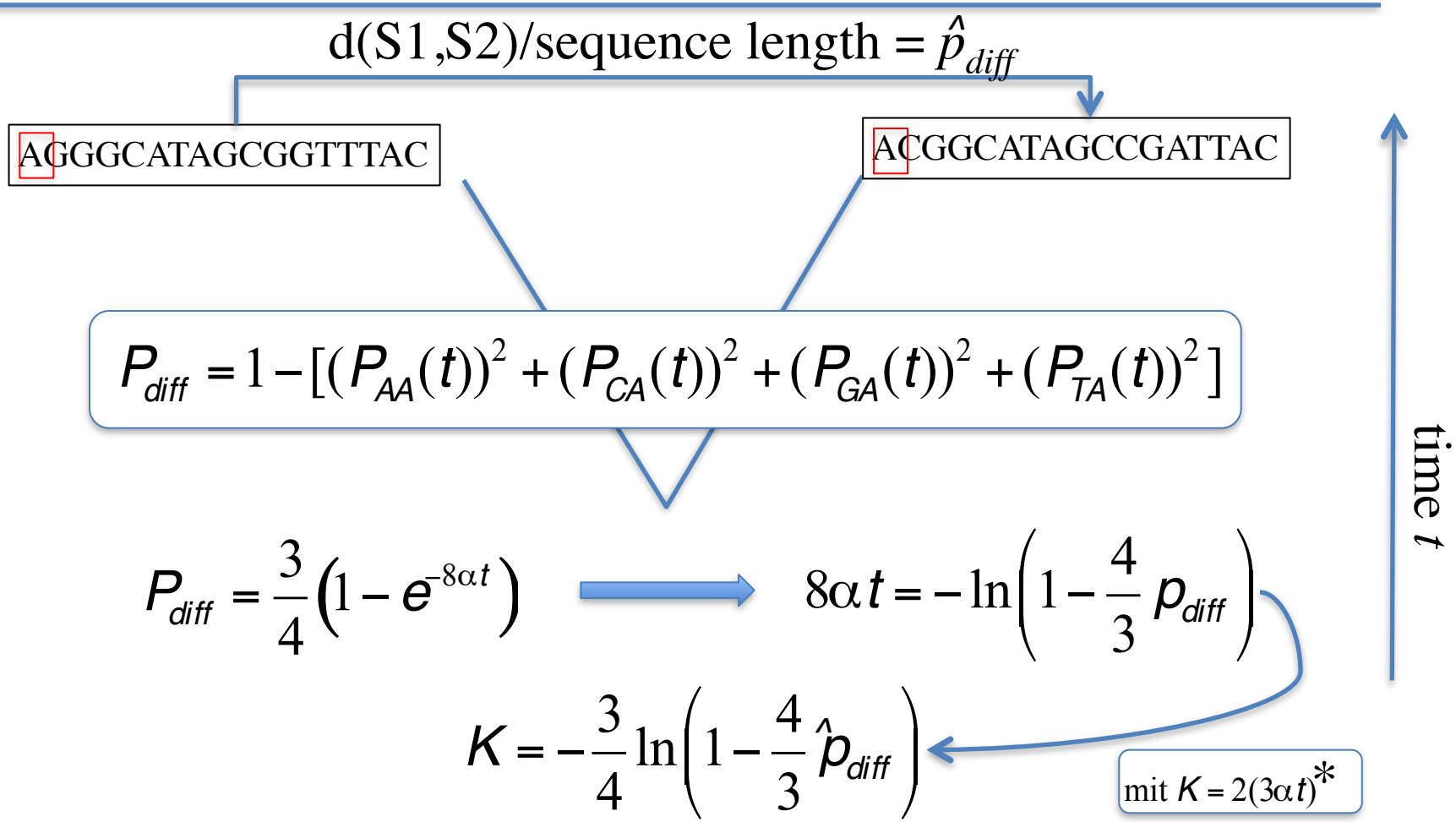
Die Wahrscheinlichkeit den gleichen Buchstaben in zwei Sequenzen zu beobachten kann zwei Gründe haben



Daraus folgt nun die Gegenwahrscheinlichkeit eine **unterschiedliche** Base zu beobachten als

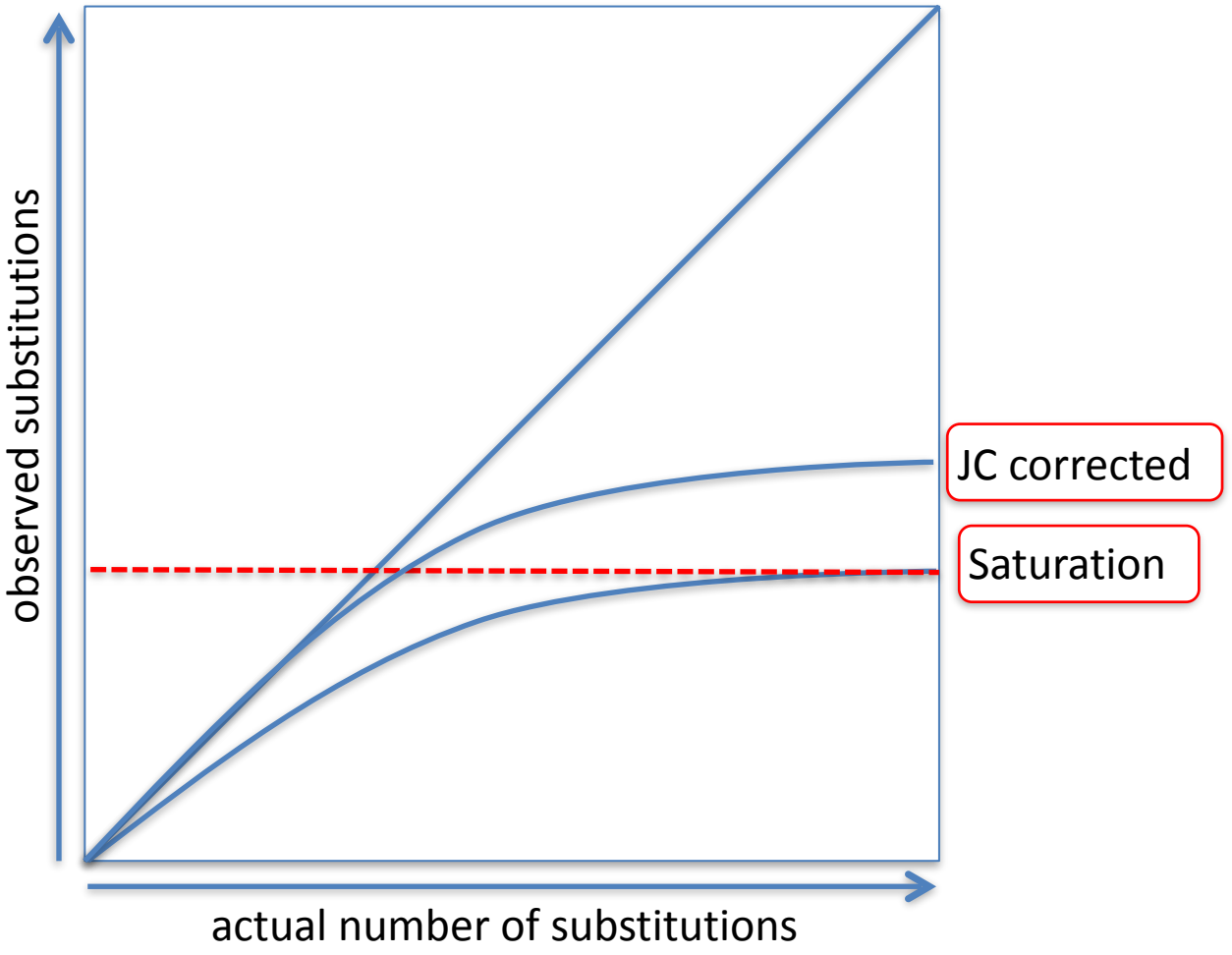
$$P_{diff} = 1 - [(P_{AA}(t))^2 + (P_{CA}(t))^2 + (P_{GA}(t))^2 + (P_{TA}(t))^2]$$

Mit P_{diff} können wir nun die Jukes Cantor Korrekturformel bestimmen ohne α oder ***t*** *zu kennen*!

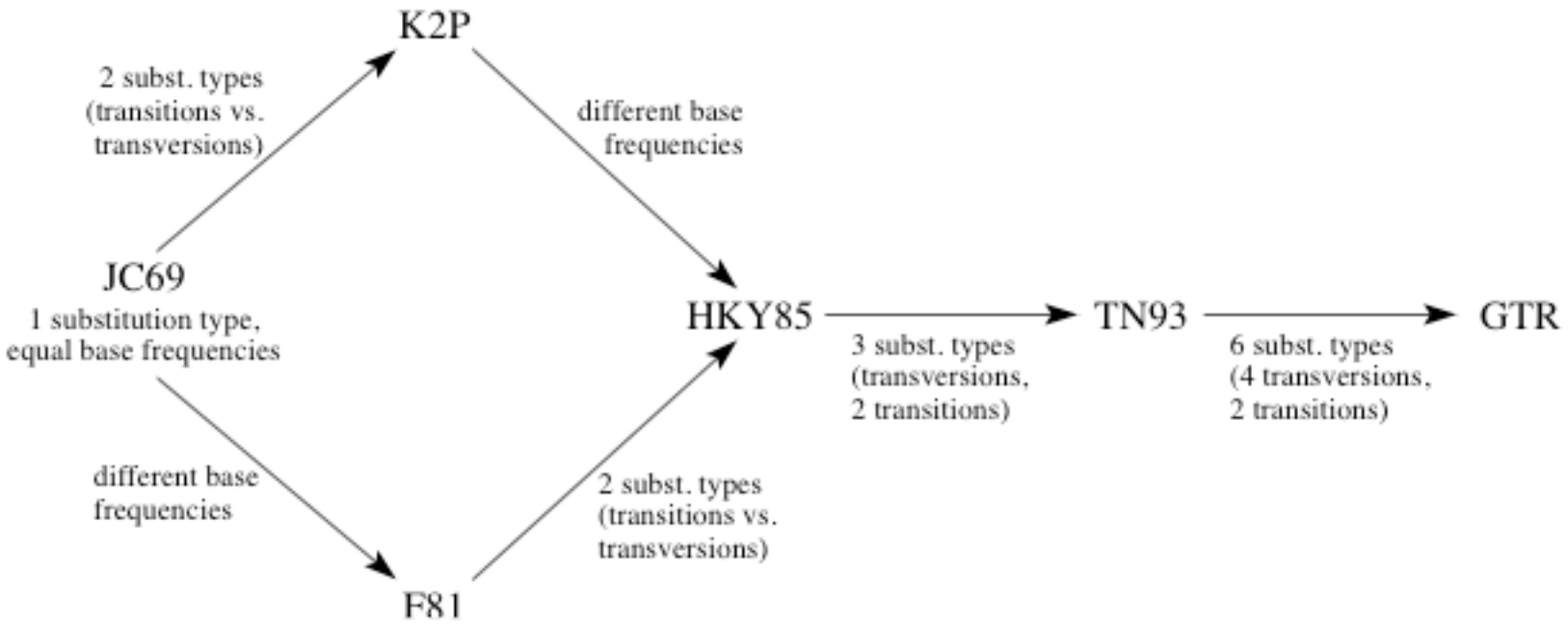


*Gegeben unser Modell erwarten wir 3αt Substitutionen pro Base und Ast, also 6αt auf beiden Ästen. Erinnern Sie sich an die JC Ratenmatrix und den Erwartungswert einer Poisson-Verteilung

Mit Hilfe unseres Substitutionsmodells können wir nun die beobachteten Distanzen korrigieren



Es gibt die verschiedensten Substitutionsmodelle für DNA Sequenzevolution



Further modification:
rate heterogeneity: invariant sites, Γ-distributed rates, mixed.

…und gleiches gilt auch für Proteinsequenzen

- Dayhoff matrix (Dayhoff et al., 1978)
- JTT (Jones et al., 1992)
- WAG (Whelan and Goldman, 2000; distantly related sequences)
- VT (Mueller and Vingron, 2000; distantly related sequences)
- mtREV (Adachi and Hasegawa, 1996; mirochondrial sequences)
- mtMAM (Yang et al., 1998; mammalian mitochondria)
- …