Taming the Beast Workshop

Molecular Evolution Models

David Rasmussen & Carsten Magnus

June 27, 2016

Outline

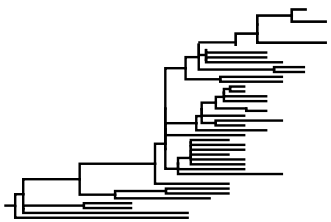
- Models of sequence evolution:
  - rate matrices
  - Markov chain model
- Variable rates amongst different sites: "+Γ"
- Codons and data partitions
- Implementation in BEAST2

Levels of evolution

genotype

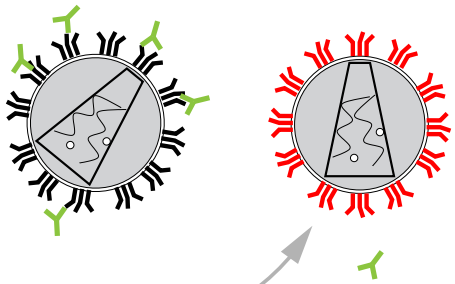
sequence level

ACUGAACGUGACUACUG
ACUGAACGUAACUACUG



phenotype

e.g. antigenic level: Antibody binding to HIV



codon: three nucleotides encode for one amino acid

one nucleotide change can already change the phenotype

alphabet:

4 nucleotides: DNA: TCAG
RNA: UCAG

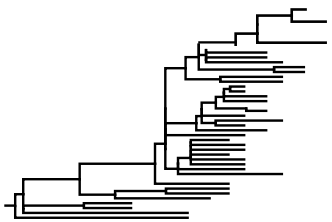
20 amino acids

Levels of evolution

genotype

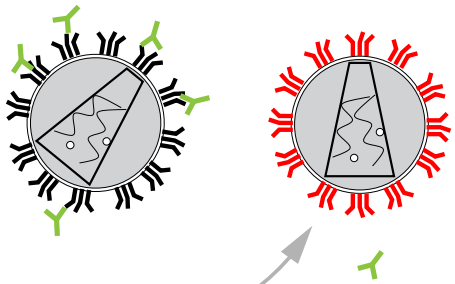
sequence level

ACUGAACGUGACUACUG
ACUGAACGUAACUACUG



phenotype

e.g. antigenic level: Antibody binding to HIV



codon: three nucleotides encode for one amino acid

one nucleotide change can already change the phenotype

alphabet:

4 nucleotides: DNA: TCAG
RNA: UCAG

20 amino acids

When comparing two nucleotide sequences we have to keep in mind that they are the result of mutation during replication (genotypic level) and selection (phenotypic level).

Sequence alignment

```
ATTACGAC
TCTACGAC
```

way of arranging sequences to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences

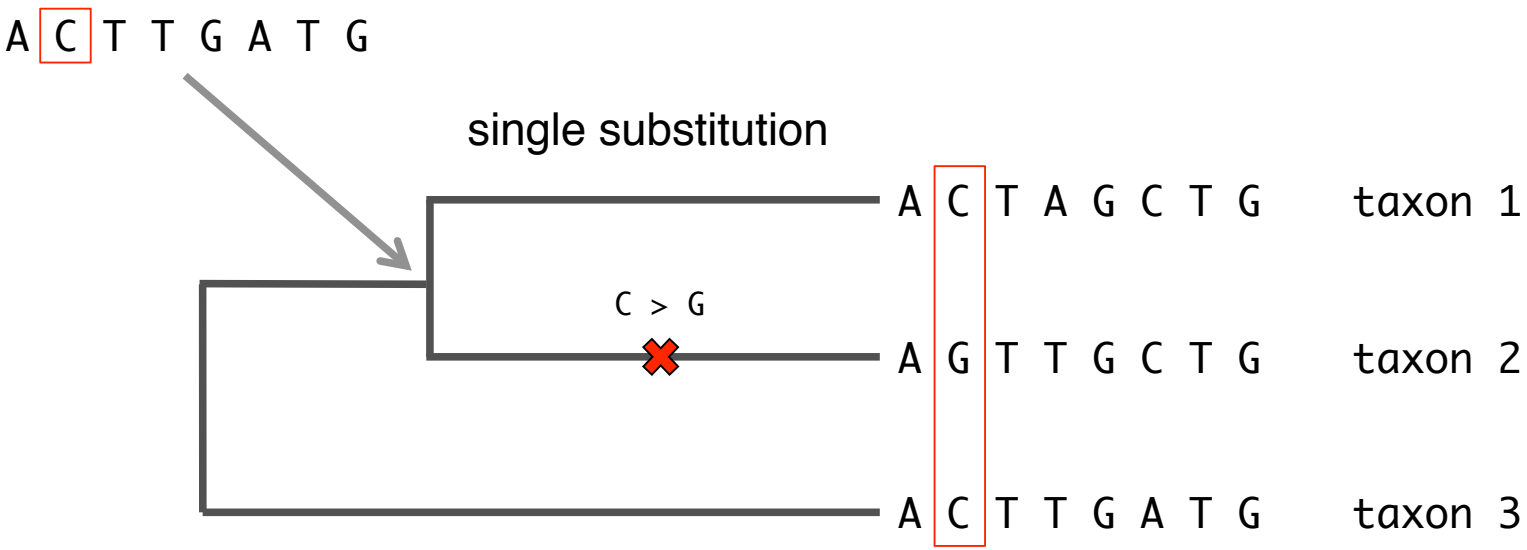
- To find an alignment: concept of **positional homology**: nucleotides (or amino acids) show positional homology if they exist at equivalent positions in the respective sequence.
- Programs for alignment MUSCLE, CLUSTAL which can be called from e.g. AliView, MegAlign,...

BEAST analysis starts with aligned sequences!!!
→ file format `.fas`, `.fasta`, `.nexus`

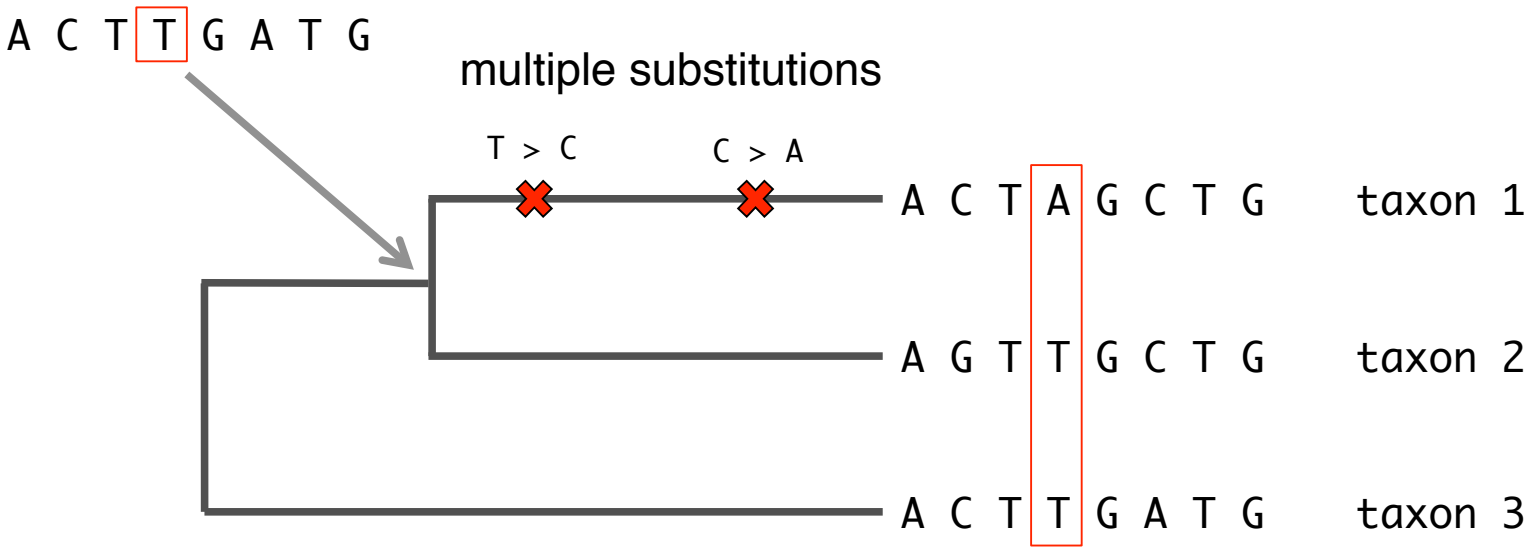
Models for nucleotide substitions

The fundamental problem

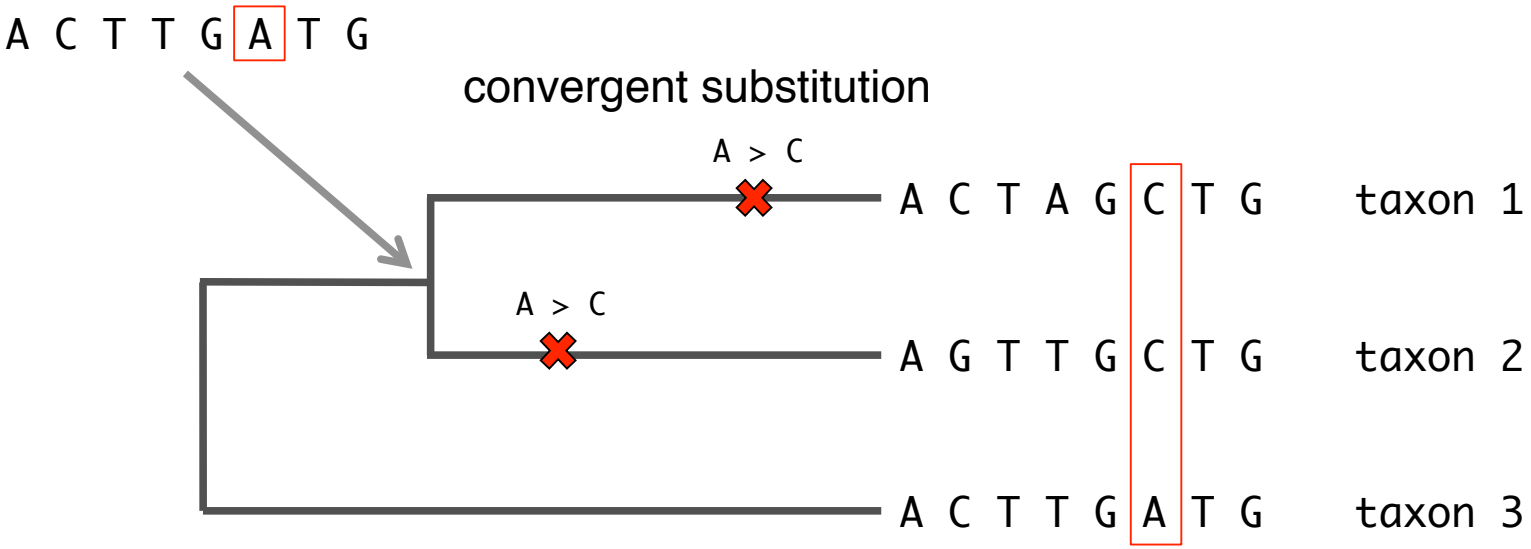
The fundamental problem



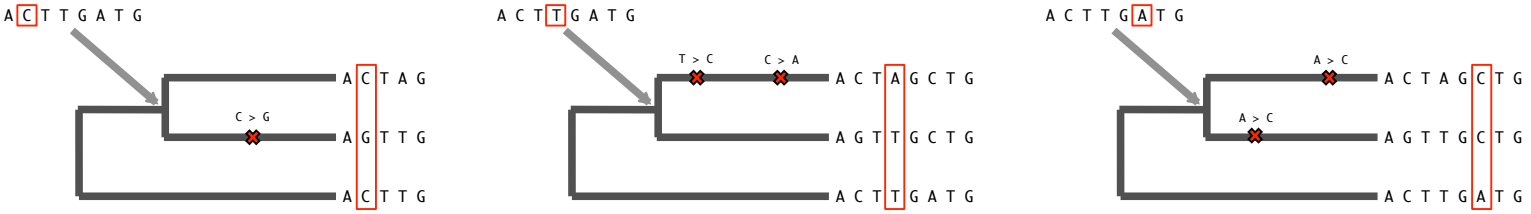
The fundamental problem



The fundamental problem



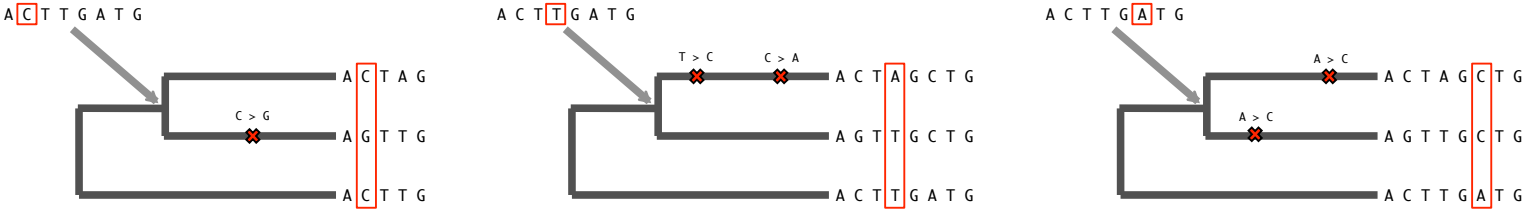
The fundamental problem



Problem of phylogenetics:

We observe sequences but not their evolutionary history. Thus we have to take **all** possible evolutionary trajectories into account.

The fundamental problem



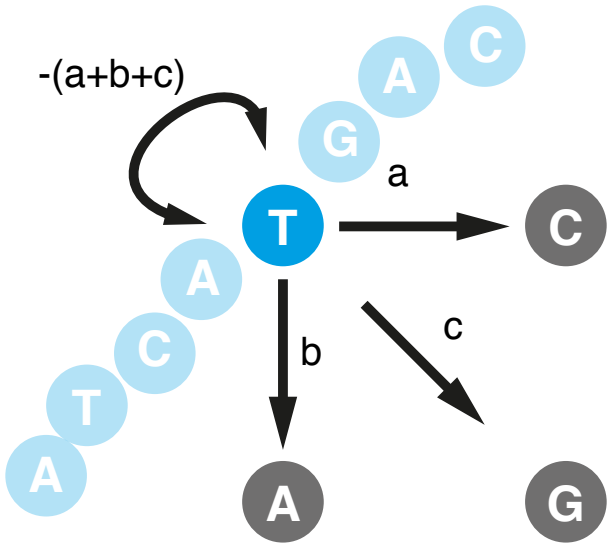
Problem of phylogenetics:
We observe sequences but not their evolutionary history. Thus we have to take **all** possible evolutionary trajectories into account.

☞ The sequence evolution model appears in the posterior:

A model for nucleotide substitutions

State space of each nucleotide position: $\mathcal{S} = \{\mathrm{T}, \mathrm{C}, \mathrm{A}, \mathrm{G}\}$

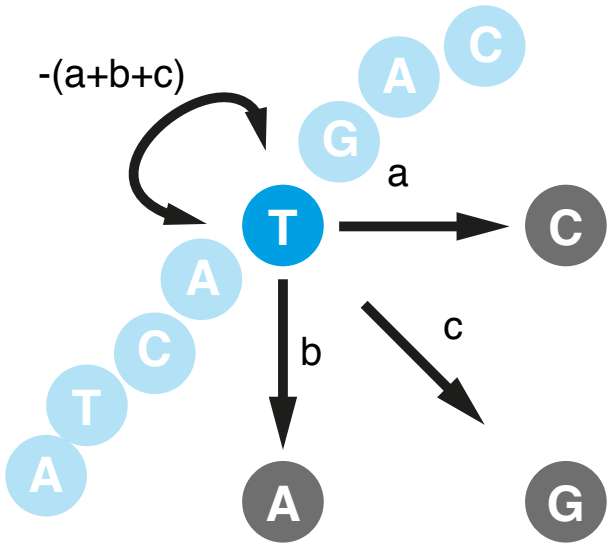
Example: Assume the process is at state T



A model for nucleotide substitutions

State space of each nucleotide position: $\mathcal{S} = \{\mathrm{T}, \mathrm{C}, \mathrm{A}, \mathrm{G}\}$

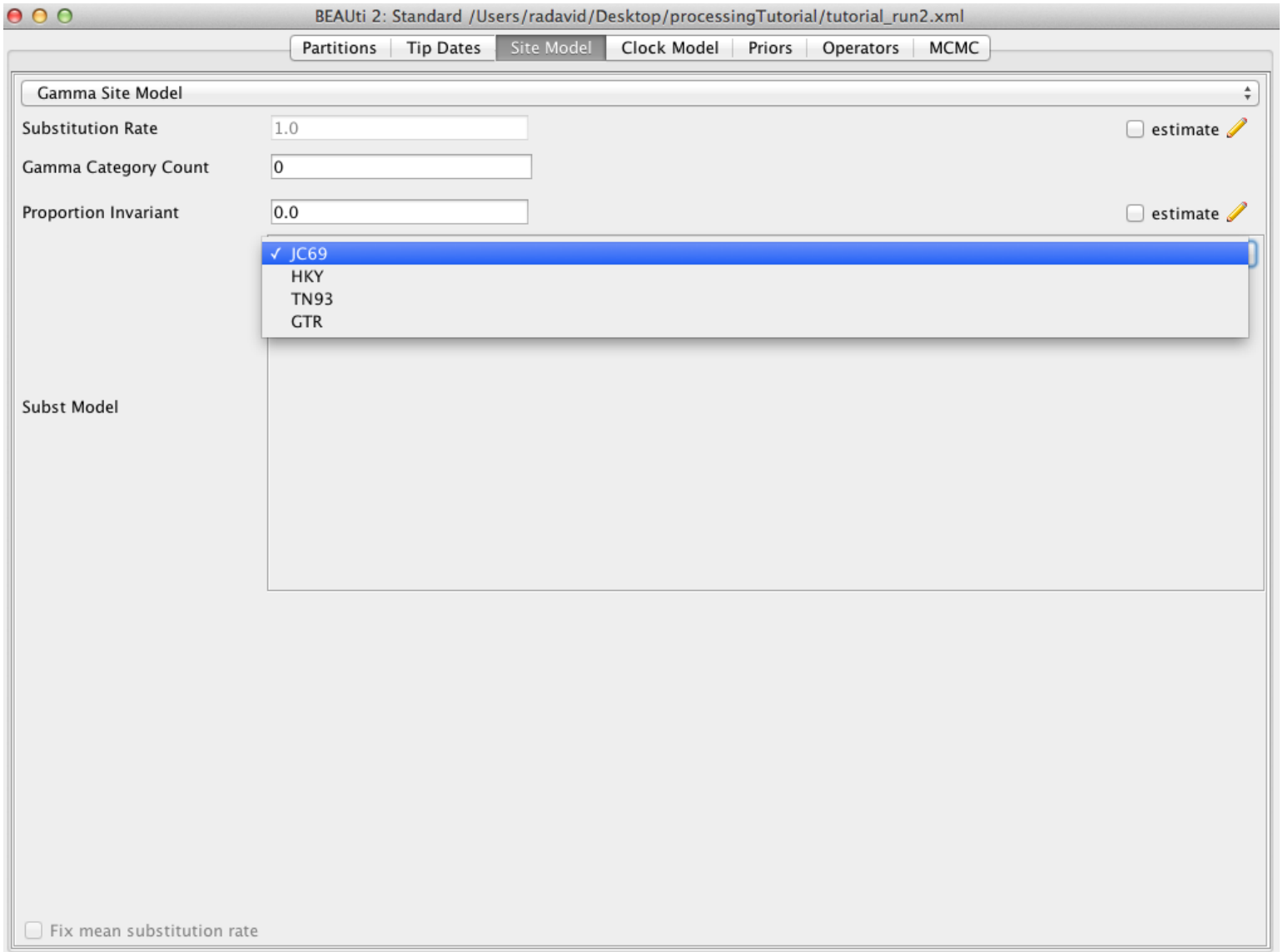
Example: Assume the process is at state T



Substitution rate matrix:

$$\begin{array}{c} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\ \begin{pmatrix} \text{-(a+b+c)} & \mathrm{a} & \mathrm{b} & \mathrm{c} \\ \mathrm{d} & \text{-(d+e+f)} & \mathrm{e} & \mathrm{f} \\ \mathrm{g} & \mathrm{h} & \text{-(g+h+i)} & \mathrm{i} \\ \mathrm{j} & \mathrm{k} & \mathrm{l} & \text{-(j+k+l)} \end{pmatrix} \end{array}$$

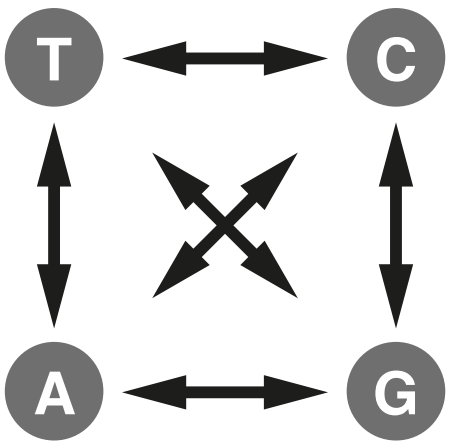
Site models in BEAST2



The easiest substitution model: JC69

JC69:

- named after TH Jukes, CR Cantor: Evolution of protein molecules. 1969 [Jukes and Cantor, 1969].
- all substitution have the same rate, λ



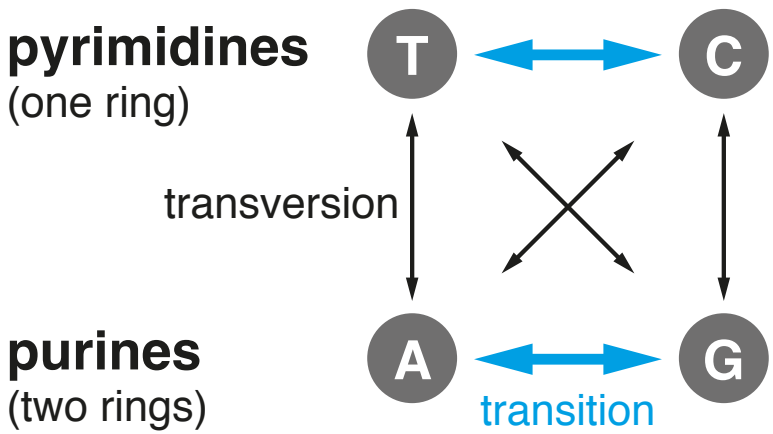
Substitution rates:

$$\begin{array}{c} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\ \begin{pmatrix} \cdot & \lambda & \lambda & \lambda \\ \lambda & \cdot & \lambda & \lambda \\ \lambda & \lambda & \cdot & \lambda \\ \lambda & \lambda & \lambda & \cdot \end{pmatrix} \end{array}$$

Accounting for transition/transversion: K80

K80:

- named after M Kimura: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. 1980. [Kimura, 1980]
- transitions happen at rate α, transversions at rate β



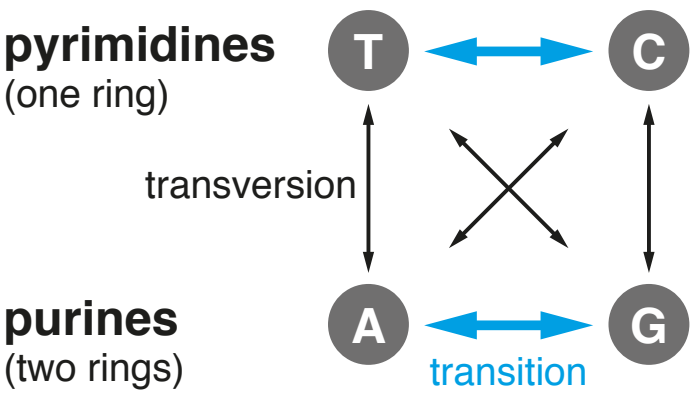
Substitution rates:

$$\begin{array}{c} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\ \begin{pmatrix} \cdot & \alpha & \beta & \beta \\ \alpha & \cdot & \beta & \beta \\ \beta & \beta & \cdot & \alpha \\ \beta & \beta & \alpha & \cdot \end{pmatrix} \end{array}$$

Accounting for transition/transversion: HKY

HKY:

- named after [Hasegawa et al., 1984, Hasegawa et al., 1985]
- accounting for transitions (rate α), transversions (rate β)
- after a long period of evolution, equilibrium frequencies are reached



Substitution rates:

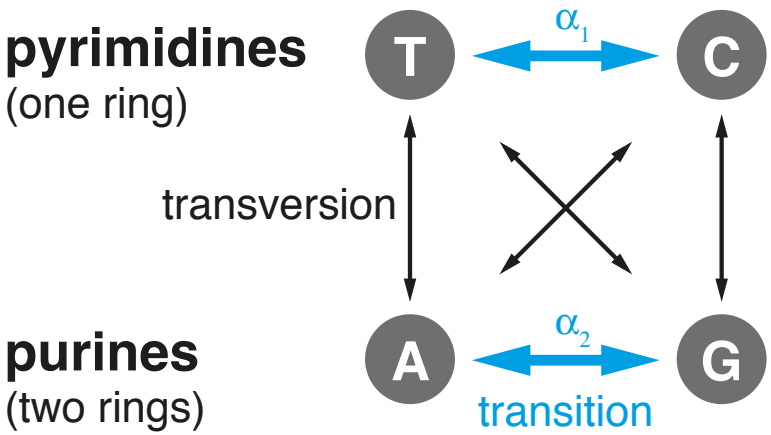
$$\begin{array}{c} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\ \begin{pmatrix} \cdot & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha\pi_T & \cdot & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & \cdot & \alpha\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha\pi_A & \cdot \end{pmatrix} \end{array}$$

$$= \begin{pmatrix} \cdot & \alpha & \beta & \beta \\ \alpha & \cdot & \beta & \beta \\ \beta & \beta & \cdot & \alpha \\ \beta & \beta & \alpha & \cdot \end{pmatrix} \cdot \begin{pmatrix} \pi_T & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_A & 0 \\ 0 & 0 & 0 & \pi_G \end{pmatrix}$$

Accounting for transition/transversion: TN93

TN93:

- named after [Tamura and Nei, 1993]
- accounting for different transition rates between T and C as well as A and G
- after a long period of evolution, equilibrium frequencies are reached



Substitution rates:

$$\begin{array}{c} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\ \begin{pmatrix} \cdot & \alpha_1\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha_1\pi_T & \cdot & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & \cdot & \alpha_2\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha_2\pi_A & \cdot \end{pmatrix} \end{array}$$

A more general substitution model: GTR

GTR (REV):

- generalised time-reversible model
- based on three papers:
  [Tavaré, 1986, Yang, 1994, Zharkikh, 1994]

Substitution rates:

$$\begin{array}{c} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\ \begin{pmatrix} \cdot & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & \cdot & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & \cdot & f\pi_G \\ c\pi_T & e\pi_C & f\pi_A & \cdot \end{pmatrix} \end{array}$$

\+ quite flexible

\+ time-reversible

\- not completely general

The most general substitution model – implemented in BEAST2 but not in BEAUti

UNREST:

- unrestricted model first described in [Yang, 1994]
- each substitution has a (different) rate

Substitution rates:

$$\begin{array}{c} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\ \begin{pmatrix} \cdot & a & b & c \\ d & \cdot & e & f \\ g & h & \cdot & i \\ j & k & l & \cdot \end{pmatrix} \end{array}$$

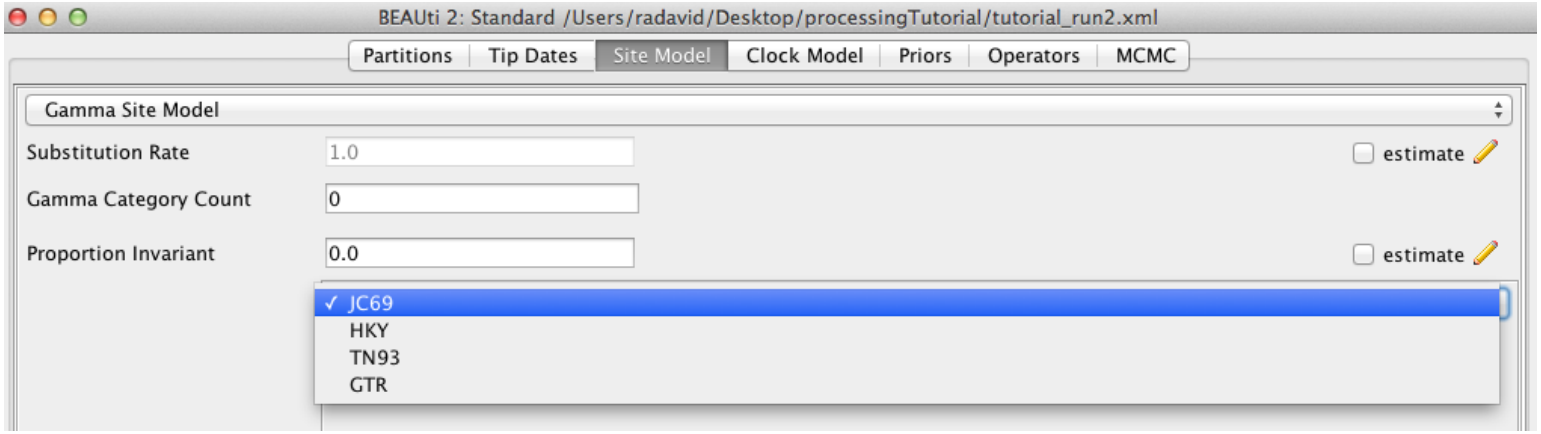
\+ most general case

\+ all other models are special cases of UNREST

\- mathematical very complicated and not handy to use

\- not time-reversible

Substitution models in BEAUti



model	parameters	description
JC69	1	all substitutions have the same rate
K80	2+3*	accounts for transition and transversions, not in BEAUti
HKY	2+3*	distinction between transition and transversions, including equilibrium frequencies
TN93	3+3*	different rates for transitions
GTR	6+3*	general, but still time-reversible
UNREST	12	most general, not time-reversible, not in BEAUti

* Can be empirically estimated from the alignment or inferred alongside the substitution rates.

The fundamental problem - again

Problem of phylogenetics:

We observe sequences but not their evolutionary history. Thus we have to take all possible evolutionary trajectories into account.

The fundamental problem - again

Problem of phylogenetics:
We observe sequences but not their evolutionary history. Thus we have to take all possible evolutionary trajectories into account.

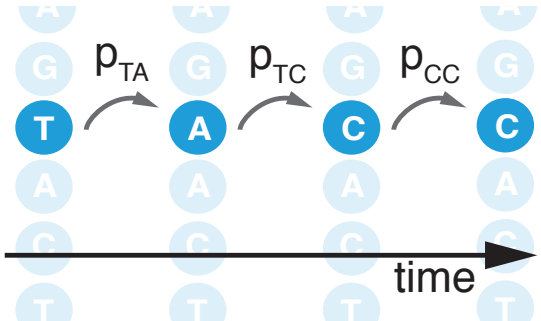
So far we determined **rates** of nucleotide substitutions. But we need **probabilities**.

Nucleotide substitutions as Markov chains (MC)

Definition of a Markov chain (see also [Ross, 1996])

stochastic process, i.e. a series of random experiments through time

Nucleotide substitutions as MC



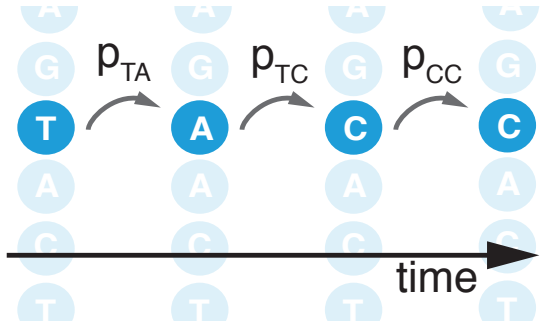
Nucleotide substitutions as Markov chains (MC)

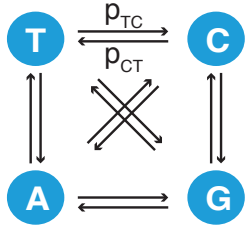
Definition of a Markov chain (see also [Ross, 1996])

stochastic process, i.e. a series of random experiments through time

lives on a **state space** and jumps to the different states

Nucleotide substitutions as MC





Nucleotide substitutions as Markov chains (MC)

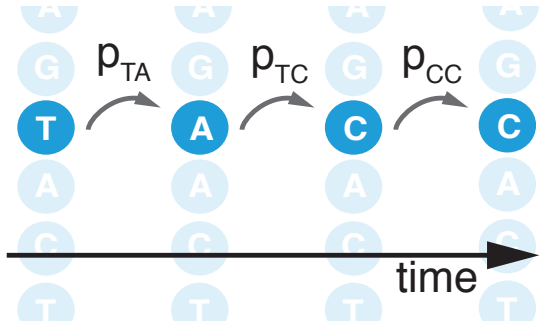
Definition of a Markov chain (see also [Ross, 1996])

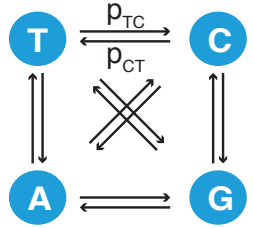
stochastic process, i.e. a series of random experiments through time

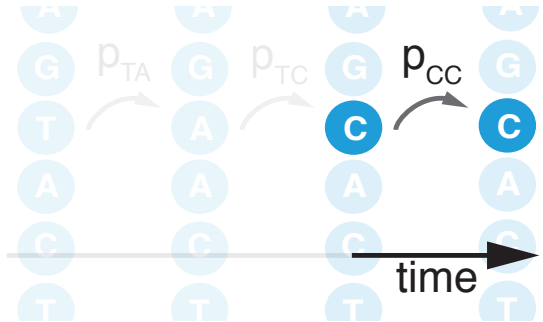
lives on a **state space** and jumps to the different states

memorylessness: the probability of jumping to a state only depends on the actual state

Nucleotide substitutions as MC







Why Markov chains are a great model for nucleotide substitutions

- memorylessness: a nucleotides substitution happens independently from the substitution history at this site
- substitution rate matrix defines the transition probabilities
  - applying theories of linear algebra we can calculate the transition probability matrix according to:

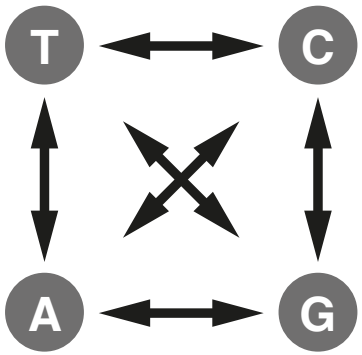
$$P(t) = e^{Qt} = U\,\mathrm{diag}(e^{\epsilon_1 t}, e^{\epsilon_2 t}, e^{\epsilon_3 t}, e^{\epsilon_4 t})U^{-1}$$

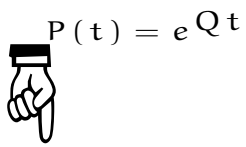
- the **transition probabilities take into account every possible substitution path** (Chapman-Kolmogorov theorem)

Example of transition probabilities: JC69

Substitution rates:

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





transition probability matrix:

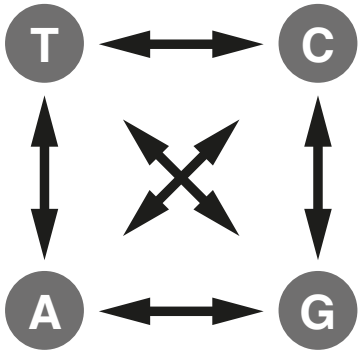
$$P(t) = \begin{pmatrix} p_0(t) & p_1(t) & p_1(t) & p_1(t) \\ p_1(t) & p_0(t) & p_1(t) & p_1(t) \\ p_1(t) & p_1(t) & p_0(t) & p_1(t) \\ p_1(t) & p_1(t) & p_1(t) & p_0(t) \end{pmatrix}$$

with $p_0(t) = \frac{1}{4} + \frac{3}{4}e^{-4\lambda t}$
and $p_1(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}$

Example of transition probabilities: JC69

Substitution rates:

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



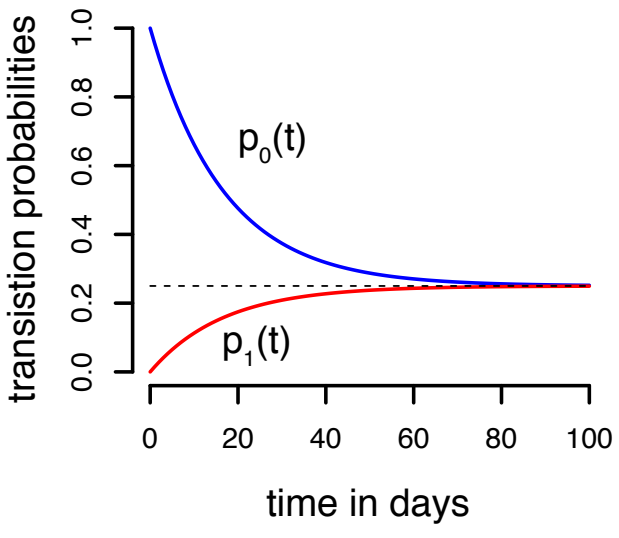
$$P(t) = e^{Qt}$$

$$\lambda = 0.015 \frac{\text{substitutions per site}}{\text{day}}$$

transition probability matrix:

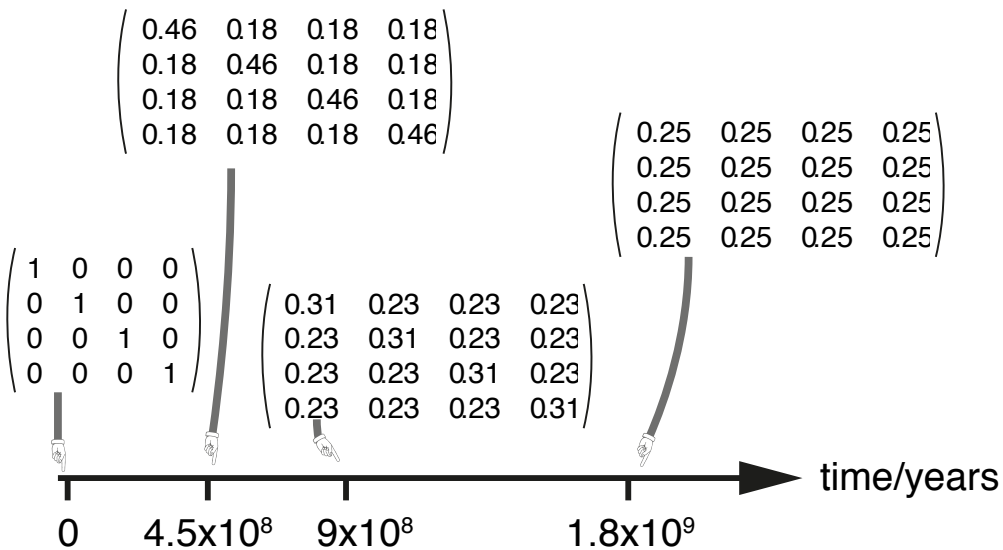
$$P(t) = \begin{pmatrix} p_0(t) & p_1(t) & p_1(t) & p_1(t) \\ p_1(t) & p_0(t) & p_1(t) & p_1(t) \\ p_1(t) & p_1(t) & p_0(t) & p_1(t) \\ p_1(t) & p_1(t) & p_1(t) & p_0(t) \end{pmatrix}$$

with $p_0(t) = \frac{1}{4} + \frac{3}{4}e^{-4\lambda t}$
and $p_1(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}$



JC69: Stationary distribution

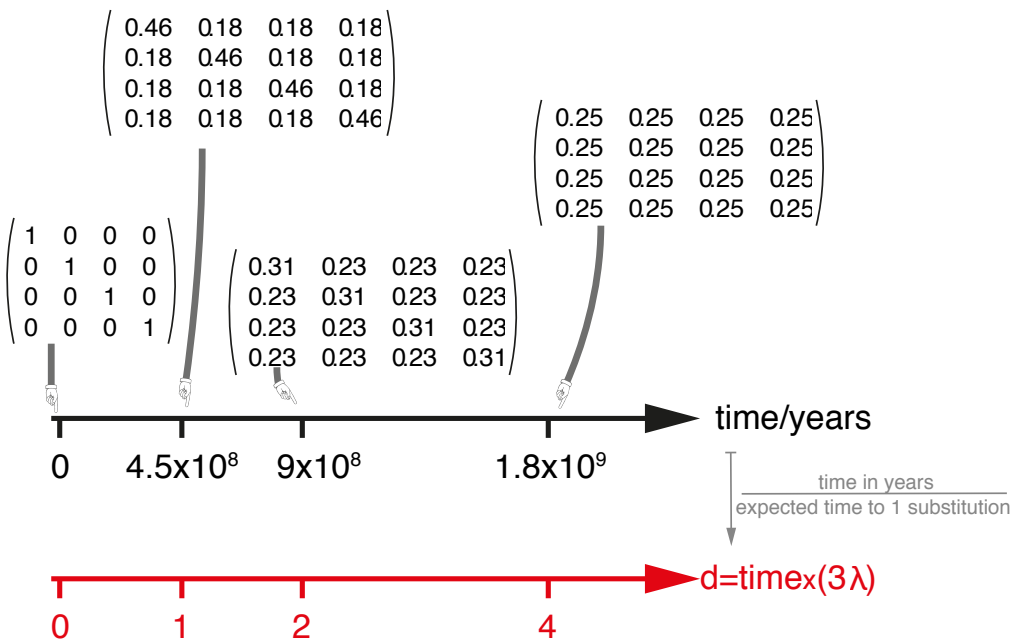
Suppose we have a sequence that evolves with rate $\lambda = 2.2/3 \times 10^{-9} \frac{\text{substitutions per site}}{\text{year}}$. We follow the evolution of 4 different sites with T at site 1, C at site 2, A at site 3 and G at site 4 at time point 0. How likely is it, that after time $\mathbf{t}$ has passed, there is a T,C,A or G at the four different positions? To answer this question, we follow the time evolution of the transition probability matrix $\mathbf{P}(\mathbf{t})$:



- when $\mathbf{t} \to \infty$ stationary distribution is reached
- Any long sequence (e.g. TTTTTT...) at time 0, will be composed of equal amounts of T,C,A,G after time $\mathbf{t} \to \infty$

JC69: Time transformation

The times we look at, e.g. in species evolution, are very often very large. Thus, instead of real time, we display an evolutionary time scale in terms of sequence distances. As one substitution happens at rate 3λ in JC69 (keep in mind that in other models the expected time to substitution is different!), we expect one substitution to happen after time $1/(3\lambda)$. This is due to exponentially distributed waiting times for an event happening at a certain rate. This means, that we expect one substitution after $\frac{1}{2.2\times 10^{-9}} \approx 4.5 \times 10^{-8}$ years in our example.



$\mathrm{t} = \frac{\mathrm{d}}{3\lambda}$ in JC69

Trick from physics: compare units:

$[\mathrm{t}]$ =years

$[\frac{\mathrm{d}}{3\lambda}] = \frac{\#\text{ substitutions}}{\#\text{ substitutions/year}}$

Variable substitution rates across sites

Variable rates

- so far: all sites in the sequence evolve at the same rate
- but: substitution rates might differ over the genome
  - mutation rates might differ over sites
  - selective pressure might be different on the phenotypic level

Variable rates

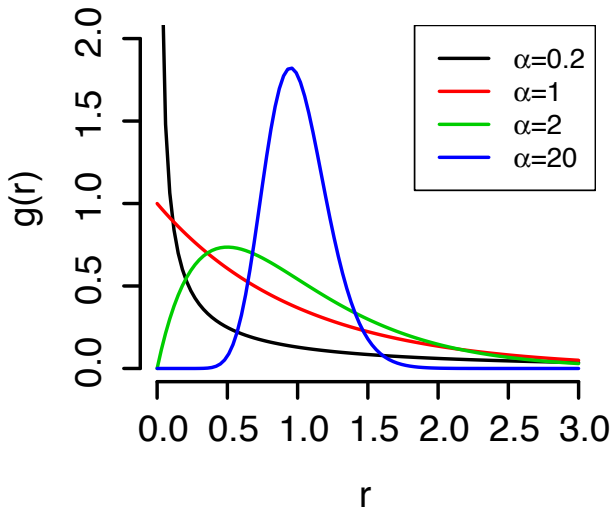
- so far: all sites in the sequence evolve at the same rate
- but: substitution rates might differ over the genome
  - mutation rates might differ over sites
  - selective pressure might be different on the phenotypic level

We extend the existing models, by replacing the constant rates by Γ-distributed random variables (notation: JC69+Γ, HKY+Γ, ...)

Example: JC69+Γ

$\lambda \mapsto \lambda R$

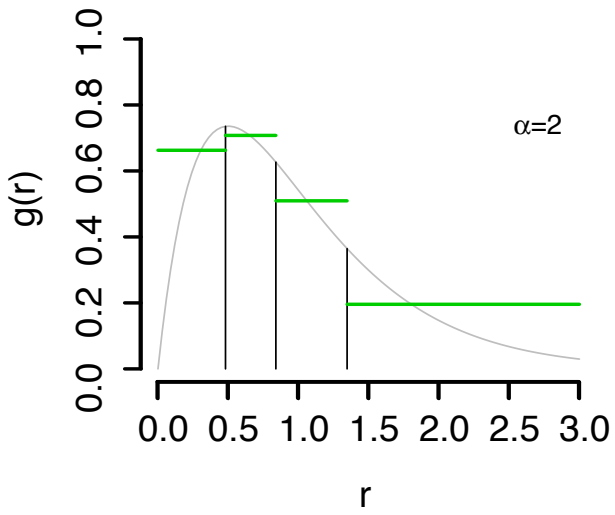
we replace the substitution rate λ by λR, where R is a Γ-distributed random variable with shape parameter α and mean 1.



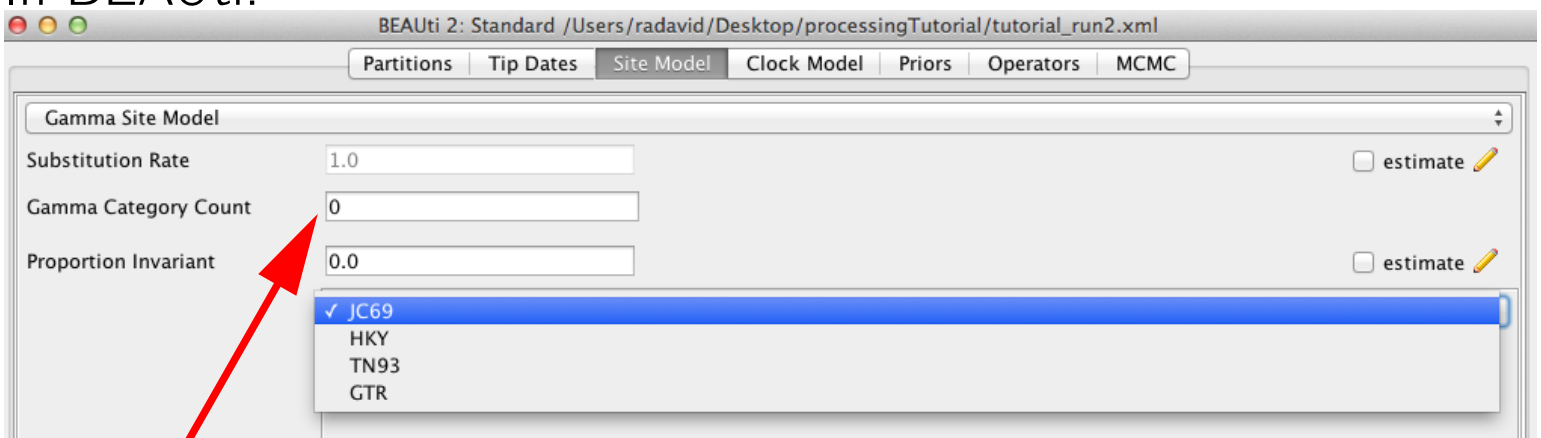
Example: JC69+Γ

$\lambda \mapsto \lambda R$

we replace the substitution rate λ by λR, where R is a Γ-distributed random variable with shape parameter α and mean 1.



In BEAUti:

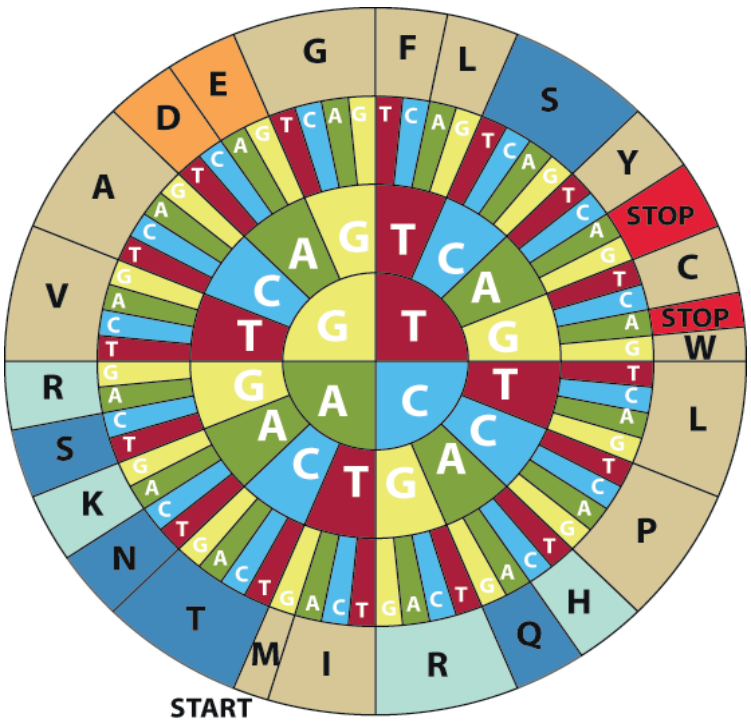


Change number of Gamma Category Count to allow for rate variation. 4 to 6 categories work normally well.

Codons and data partitions

The codon sun

A codon consists of three nucleotides, translating to one of the 20 amino acids:



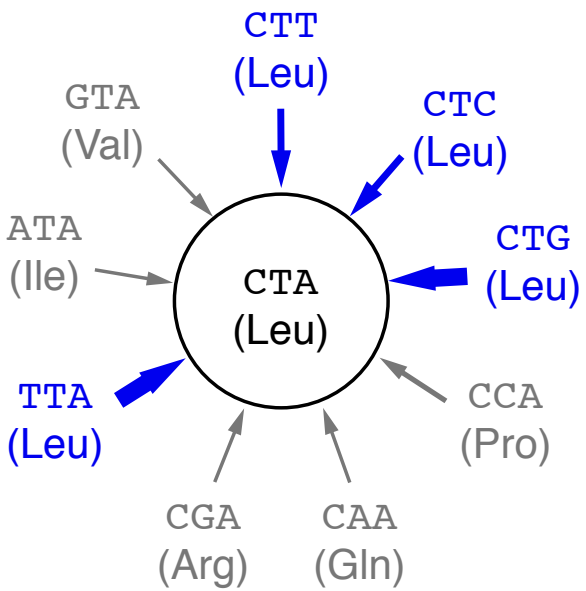
[Sanger, 2015]

Amino Acid	Three-Letter Abbreviation	One-Letter Symbol	Molecular Weight
Alanine	Ala	A	89Da
Arginine	Arg	R	174Da
Asparagine	Asn	N	132Da
Aspartic acid	Asp	D	133Da
Asparagine or aspartic acid	Asx	B	133Da
Cysteine	Cys	C	121Da
Glutamine	Gln	Q	146Da
Glutamic acid	Glu	E	147Da
Glutamine or glutamic acid	Glx	Z	147Da
Glycine	Gly	G	75Da
Histidine	His	H	155Da
Isoleucine	Ile	I	131Da
Leucine	Leu	L	131Da
Lysine	Lys	K	146Da
Methionine	Met	M	149Da
Phenylalanine	Phe	F	165Da
Proline	Pro	P	115Da
Serine	Ser	S	105Da
Threonine	Thr	T	119Da
Tryptophan	Trp	W	204Da
Tyrosine	Tyr	Y	181Da
Valine	Val	V	117Da

[Promega, 2015]

Example: Codon CTA

Overview over substitution rates to the same codon CTA, the thickness of arrows represent different rates:



- **synonymous substitutions:** AA does not change
- **nonsynonymous substitutions:** AA does change
- bigger arrows: transition
- smaller arrows: transversion

adapted from [Yang, 2014]

Varying substitution rates amongst the codon positions

[Bofkin and Goldman, 2007] have shown that in protein encoding regions

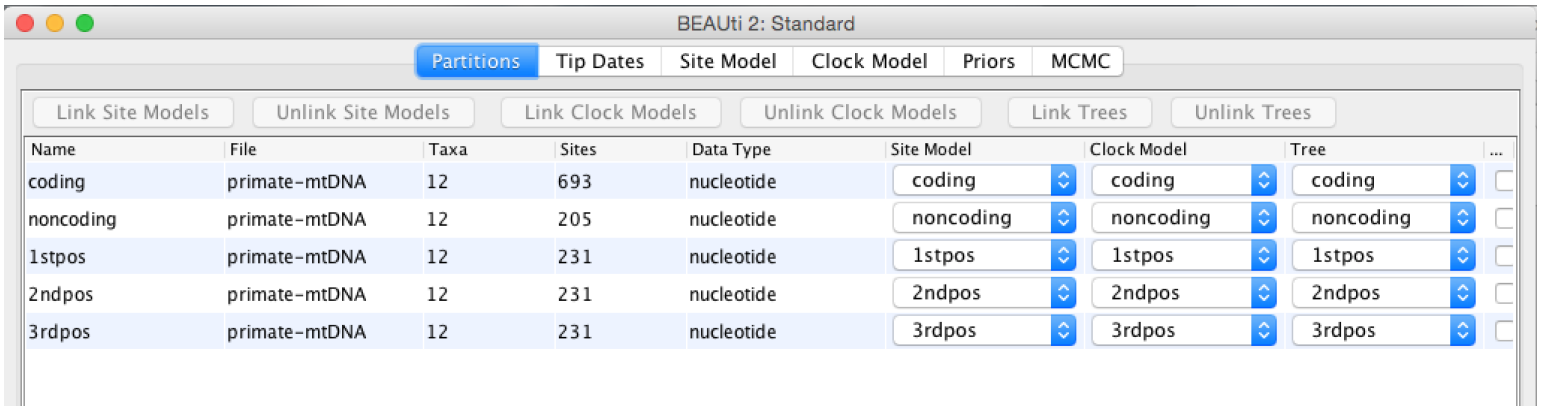
- second codon positions evolve more slowly than first codon positions
- third codon positions evolve faster than first codon positions

Varying substitution rates amongst the codon positions

[Bofkin and Goldman, 2007] have shown that in protein encoding regions

- second codon positions evolve more slowly than first codon positions
- third codon positions evolve faster than first codon positions

⇒ Different codon positions can have different evolutionary rates. BEAST2 allows for estimating these rates separately.



file `BEAST2.4.x/examples/nexus/primate-mtDNA.nex`

Including the choice of substitution rate model into your BEAST analysis

Rate models in BEAST2

- BEAST2 allows for including different site models into your analysis (☞ `Site Model` tab in BEAUti)
- Which site model is the best for your data?

Rate models in BEAST2

- BEAST2 allows for including different site models into your analysis (☞ `Site Model` tab in BEAUti)
- Which site model is the best for your data?

`T:` package bModelTest: Bayesian site model selection for nucleotide data

Rate models in BEAST2

- BEAST2 allows for including different site models into your analysis (☞ `Site Model` tab in BEAUti)
- Which site model is the best for your data?

`T:` package bModelTest: Bayesian site model selection for nucleotide data

`T:` package SubstBMA: modelling across-site variation in the nucleotide

References I

- Bofkin, L. and Goldman, N. (2007). Variation in Evolutionary Processes at Different Codon Positions. *Molecular Biology and Evolution*, 24(2):513–521.
- Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the Human Ape Splitting by a Molecular Clock of Mitochondrial-Dna. *Journal of Molecular Evolution*, 22(2):160–174.
- Hasegawa, M., Yano, T., and Kishino, H. (1984). A New Molecular Clock of Mitochondrial-Dna and the Evolution of Hominoids. *Proceedings of the Japan Academy Series B-Physical and Biological Sciences*, 60(4):95–98.
- Jukes, T. and Cantor, C. (1969). Evolution of protein molecules. *Mammalian Protein Metabolism.*, pages 21–123.
- Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, 16(2):111–120.
- Promega (2015). The amino acids: https://www.promega.com/ /media/files/resources/technical references/amino acid abbreviations and molecular weights.pdf.
- Ross, S. M. (1996). *Stochastic Processes. Second edition.* Wiley.
- Sanger (2015). The codon sun: ftp://ftp.sanger.ac.uk/pub/yourgenome/downloads/activities/kras-cancer-mutation/krascodonwheel.pdf.
- Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10(3):512–526.
- Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. In *Some mathematical questions in biology—DNA sequence analysis (New York, 1984)*, pages 57–86. Amer. Math. Soc., Providence, RI.
- Yang, Z. (1994). Estimating the pattern of nucleotide substitution. *Journal of molecular evolution*, 39(1):105–111.
- Yang, Z. (2014). *Molecular Evolution – A Statistical Approach*. Oxford University Press.
- Zharkikh, A. (1994). Estimation of evolutionary distances between nucleotide sequences. *Journal of molecular evolution*, 39(3):315–329.