Modeling Molecular Evolution

Maria Anisimova
Applied Computational genomics Team
Institute of Applied Simulation
ZHAW and SIB

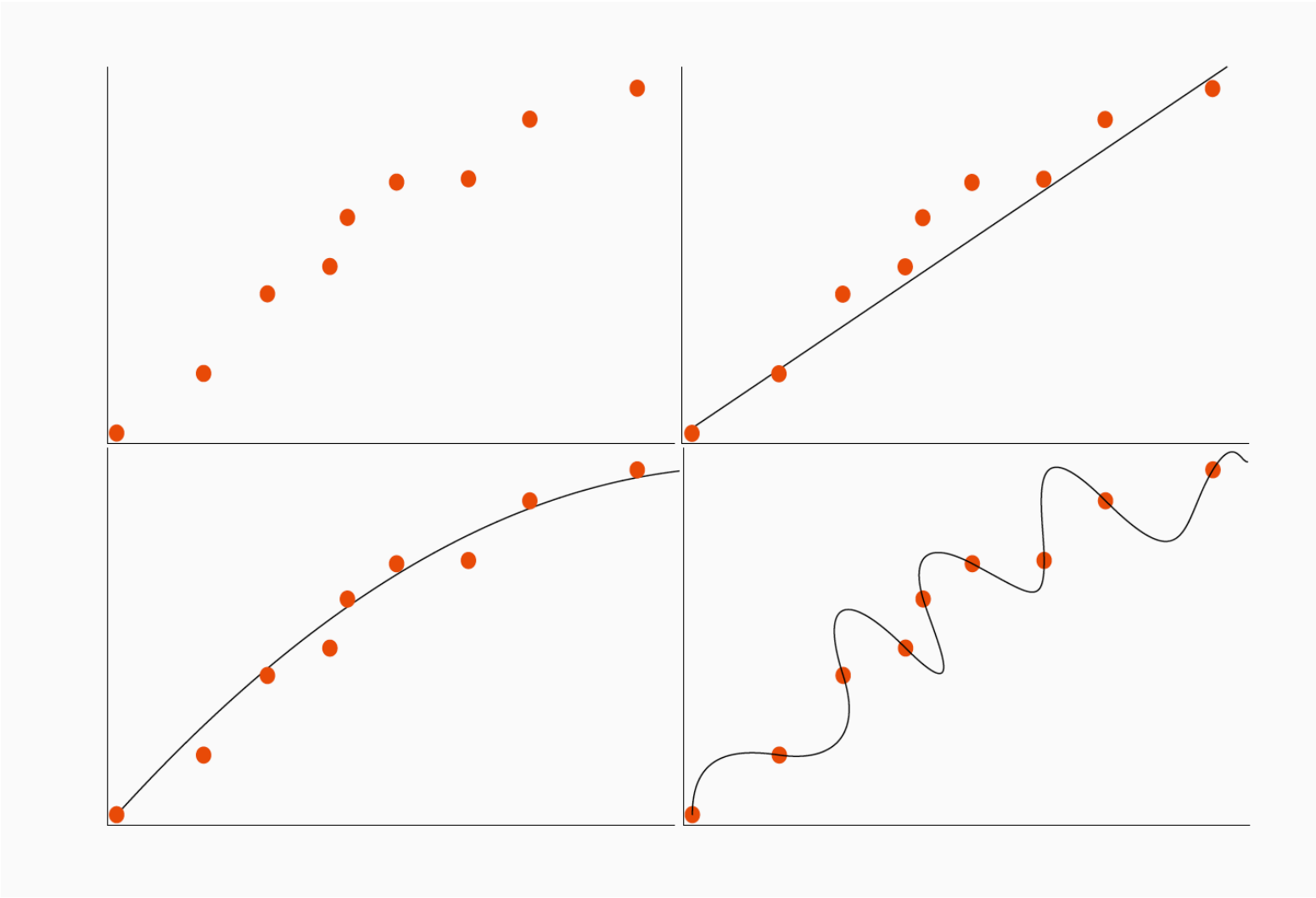
Zürcher Hochschule
für Angewandte Wissenschaften

zh
aw

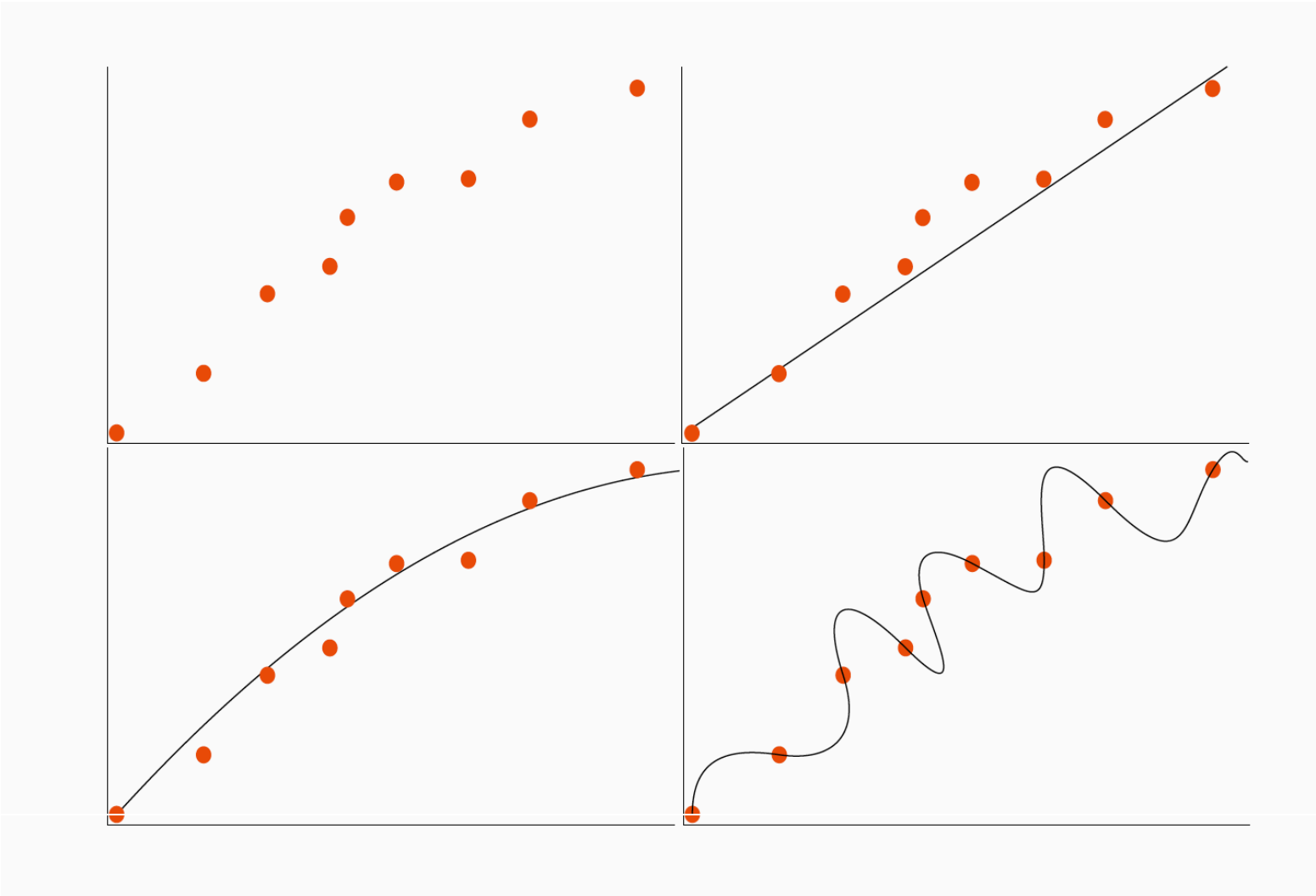
2016

SIB 15 YEARS
Swiss Institute of
Bioinformatics

All models are wrong…



All models are wrong…



… but some are useful. (*Box 1976*)

Models of molecular evolution allow to:

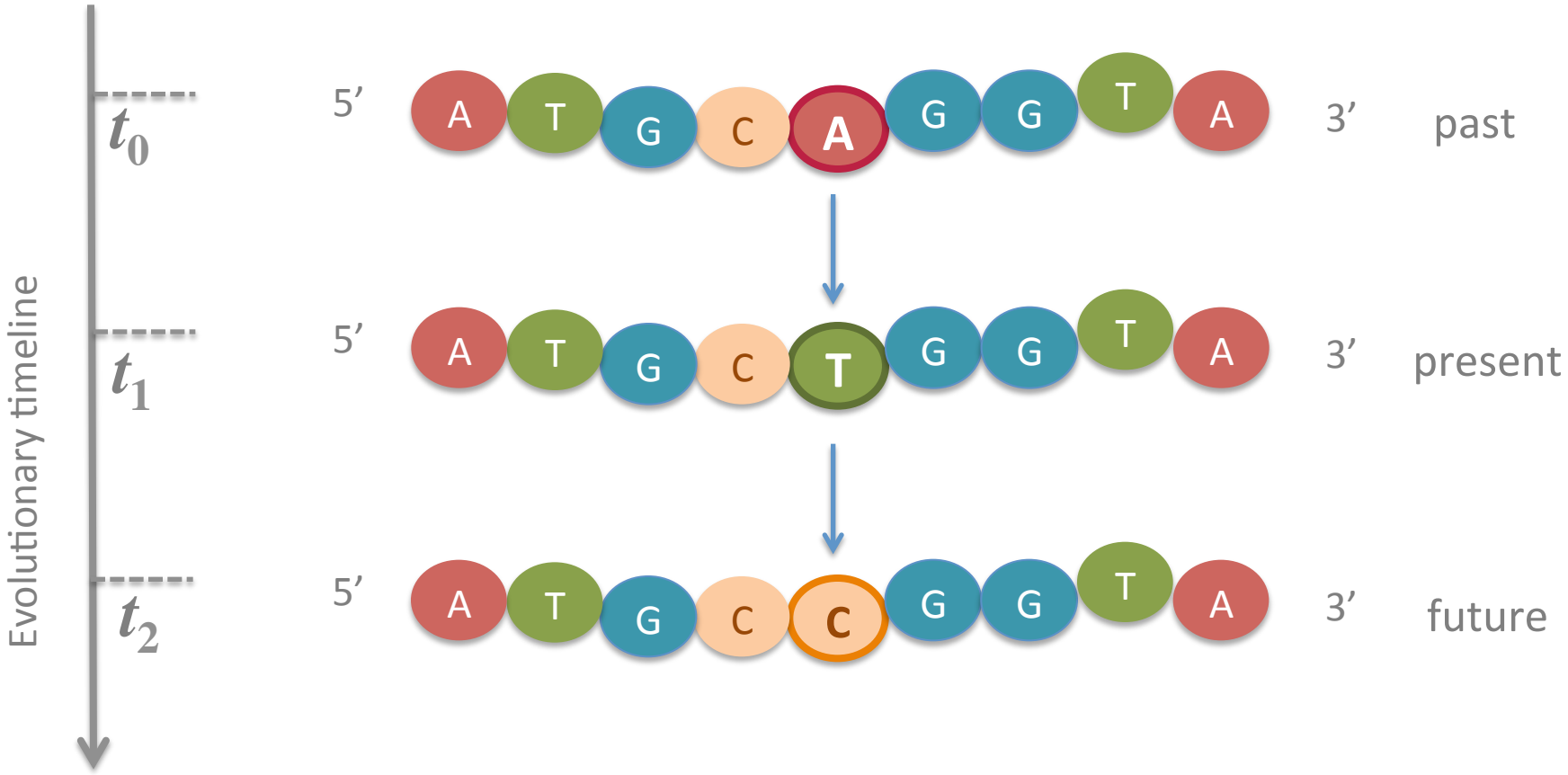
- do hypothesis testing
- study molecular evolution patterns
- infer homologs conservation: what sites are preserved? Which are under positive selection? Function?
- infer sites involved in evasion from immune response and used in vaccine design
- infer mutation rates, biases and date speciation events
- study evolution of gene families using phylogenetics
- how does environment/ecology affect genomes?
- connection between genotype and phenotype?

Andrey Markov

1856 - 1922

Russian mathematician
Described the rules of a process:
inspired by Eugene Onegin of Pushkin

Markov model of substitution



Memoriless property:

$$\Pr(C_{future} \mid T_{present} \,\&\, A_{past}) = \Pr(C_{future} \mid T_{present})$$

Markov model of substitution: summary

The future depends only on the current state

States $X(t)$: discrete or continuous

Time t: discrete (eg, # generations) or continuous (exponential waiting times)

Simple/convenient mathematically

Typical assumptions:

- Independence of evolution at sites
- Stationarity
- Homogeneity
- Time reversibility

More formally…

A **discrete** Markov process $X(t)$ in time t is a family of R.V. such that

for any (continuous or discrete) states $x_0, x_1, \ldots, x_t, x_{t+1}$ and any discrete t :

$$\Pr\{X(t+1)=x_{t+1} \mid X(t)=x_t,\ X(t-1)=x_{t-1}, \ldots,\ X(1)=x_1,\ X(0)=x_0\}$$

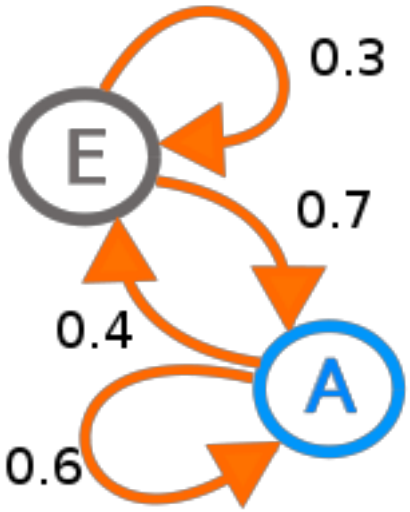
$$= \Pr\{X(t+1)=x_{t+1} \mid X(t)=x_t\}$$

A **continuous** Markov process has continuous index, defined for a family of R.V. $\{X(t),\ 0 \le t < \infty\}$

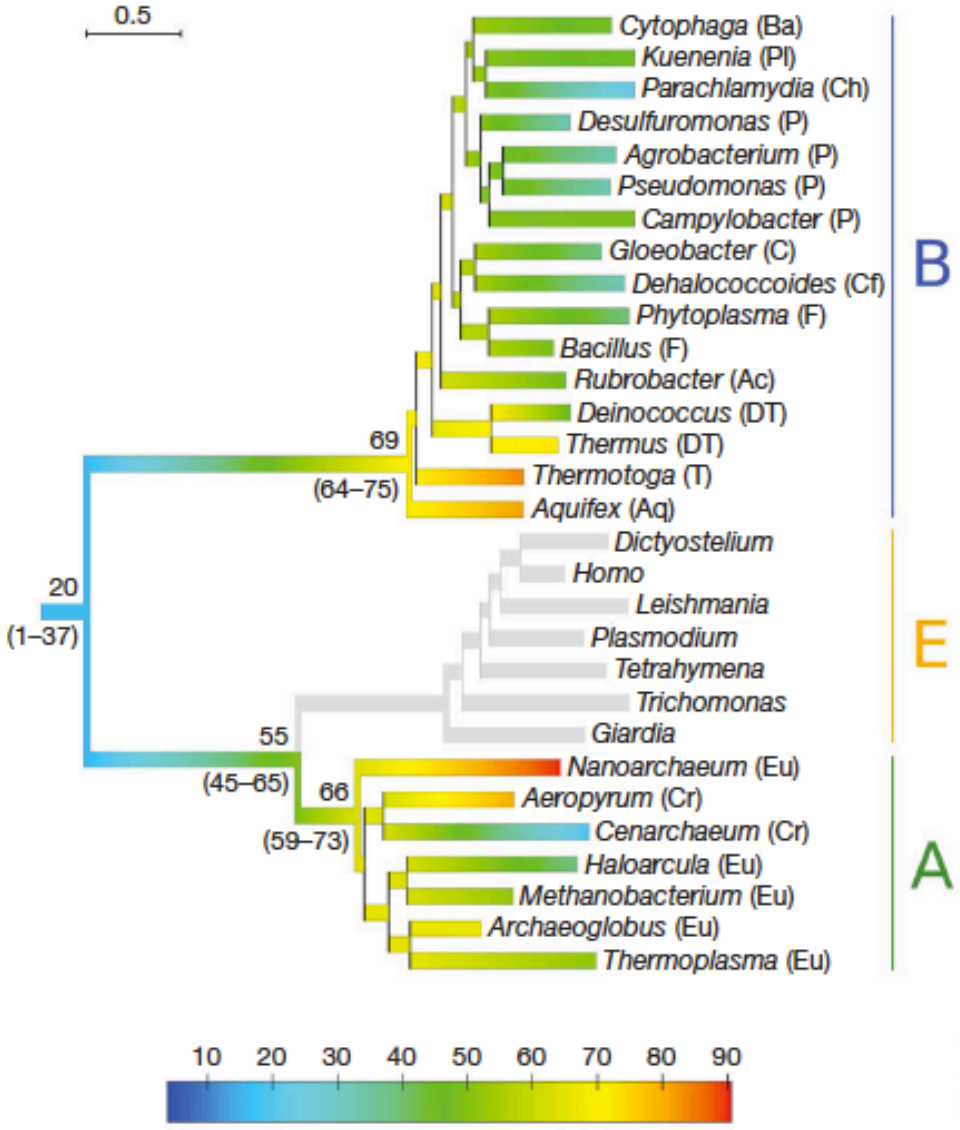
Generating matrix is needed!

For a **homogeneous** Markov process:

$$\Pr\{X(t+1)=x \mid X(t)=y\} = \Pr\{X(t)=x \mid X(t-1)=y\} \text{ for any } t$$



Example of a simple 2-state process:
http://en.wikipedia.org/wiki/Markov_chain



From Boussau et al. 2008, Nature

Nonthermophilic LUCA?

Figure 2 | Evolution of thermophily over the tree of life. Protein-derived nhPhyloBayes OGT estimates (and their 95% confidence intervals for key ancestors) for prokaryotic organisms are colour-coded from blue to red for low to high temperatures. Colours were interpolated between temperatures estimated at nodes. The eukaryotic domain, in which OGT cannot be estimated, has been shaded. The colour scale is in °C; the branch length scale is in substitutions per site. A, archaeal; B, bacterial; E, eukaryotic domains. Ac, Actinobacteria; Aq, Aquificae; Ba, Bacteroidetes; C, Cyanobacteria; Cf, Chloroflexi; Ch, Chlamydiae; Cr, Crenarchaeota; DT, Deinococcus/Thermus; Eu, Euryarchaeota; F, Firmicutes; P, Proteobacteria; Pl, Planctomycetes; T, Thermotogae.

4-state Markov chain for DNA

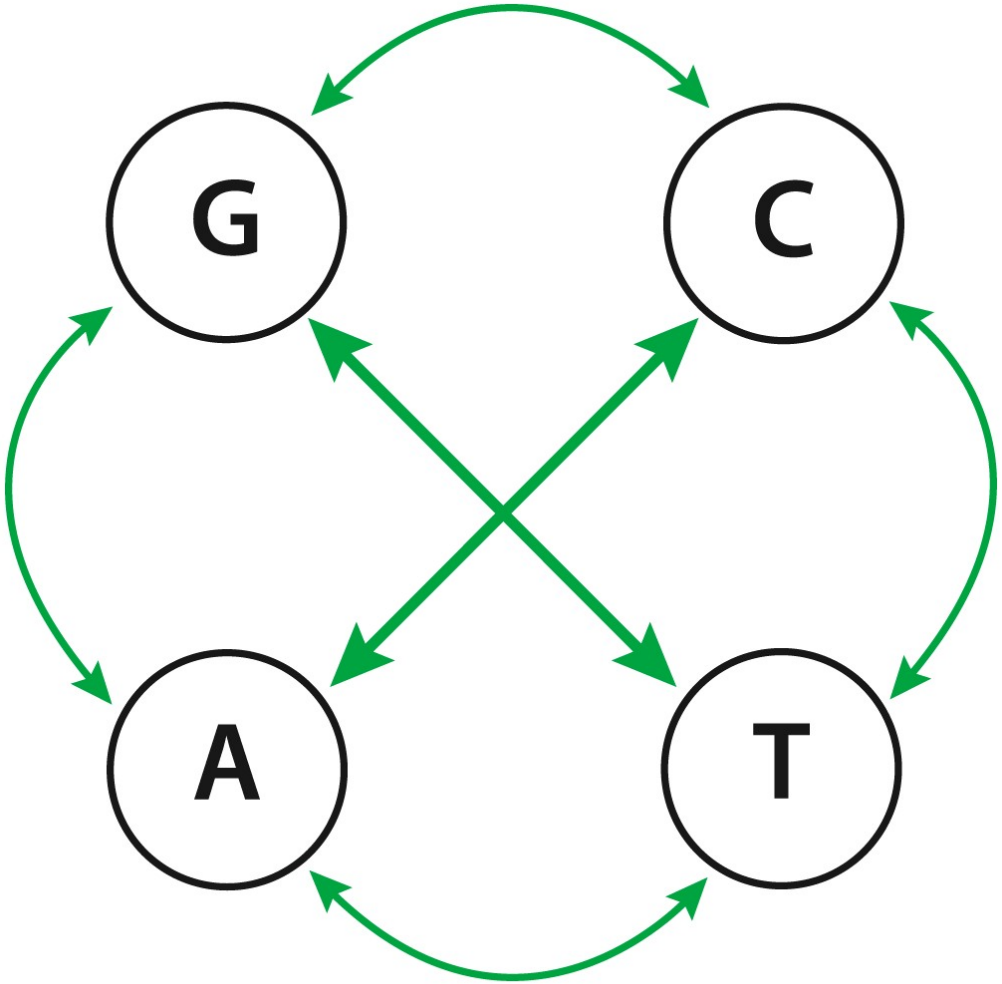


Figure 3.13 Phylogenomics: A Primer (© Garland Science 2013)

Markov model of DNA substitution

Sites evolve independently (i.i.d.)

Continuous-time Markov process describes substitutions at any site

Character at time *t* is R.V. X(t) ∈ {A,C,G,T}

Process generating matrix *Q*

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

q_{ij} are instantaneous rates from *i* to *j*

Process leaves state *i* at rate: $-q_{ii} = \sum_{j \neq i} q_{ij}$

$\Pr\{ X(t+\Delta t)=j \mid X(t)=i \}_{i \neq j} = q_{ij} \Delta t$

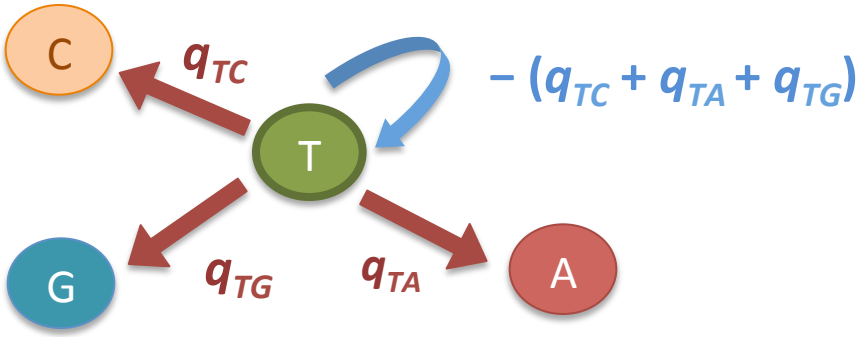
If q_{ij} constant over time the process is homogeneous

Q determines transition matrix $P(t) = \{p_{ij}(t)\} = \{\Pr\{ X(t)=j \mid X(0)=i \}\}$, $t > 0$

$$\frac{dP(t)}{dt} = P(t)Q \text{ and } P(0) = I \quad \Rightarrow \quad \boxed{P(t) = \exp(Qt)}$$

The instantaneous rate matrix of the Markov process

$$Q = \{q_{ij}\} = \begin{pmatrix} -\sum_{j \neq A} q_{Tj} & q_{TC} & q_{TA} & q_{TG} \\ q_{CT} & -\sum_{j \neq C} q_{Cj} & q_{CA} & q_{CG} \\ q_{AT} & q_{AC} & -\sum_{j \neq G} q_{Aj} & q_{AG} \\ q_{GT} & q_{GC} & q_{GA} & -\sum_{j \neq T} q_{Gj} \end{pmatrix}$$



Total rate of change = Rate of staying in the same state

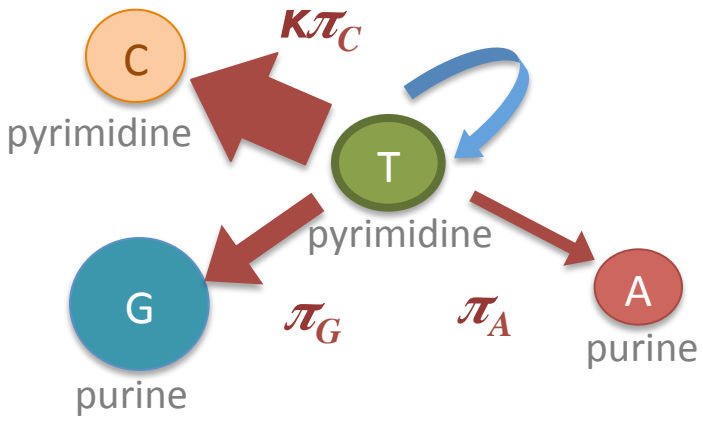
$$q_{TC} + q_{TG} + q_{TA} = -(q_{TC} + q_{TA} + q_{TG})$$

HKY model, Hasegawa-Kishino-Yano (1985)

$$Q_{\text{HKY}} = \begin{pmatrix} \bullet & \kappa\pi_C & \pi_A & \pi_G \\ \kappa\pi_T & \bullet & \pi_A & \pi_G \\ \pi_T & \pi_C & \bullet & \kappa\pi_G \\ \pi_T & \pi_C & \kappa\pi_A & \bullet \end{pmatrix}$$

Unequal frequencies

$\pi_T, \pi_C, \pi_A, \pi_G$



Transition (ts) vs. transversion (tv) rate ratio:

$\kappa = ts/tv$

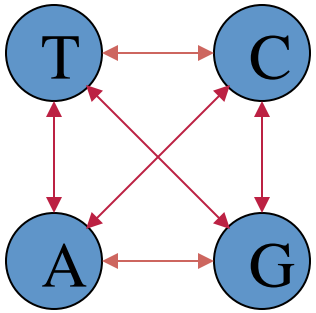
Common models of nucleotide evolution

$$Q_{\mathrm{JC69}} = \begin{pmatrix} \bullet & \lambda & \lambda & \lambda \\ \lambda & \bullet & \lambda & \lambda \\ \lambda & \lambda & \bullet & \lambda \\ \lambda & \lambda & \lambda & \bullet \end{pmatrix}$$

Jukes and Cantor (1969)

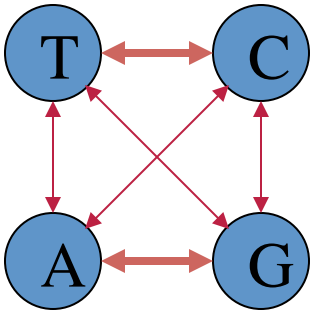
$$Q_{\mathrm{K80}} = \begin{pmatrix} \bullet & \alpha & \beta & \beta \\ \alpha & \bullet & \beta & \beta \\ \beta & \beta & \bullet & \alpha \\ \beta & \beta & \alpha & \bullet \end{pmatrix}$$

Kimura (1980)



pyrimidines

purines



transversions

transitions

Common models of nucleotide evolution

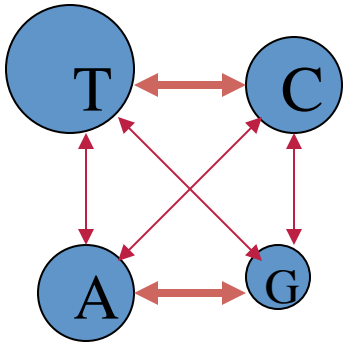
$$Q_{\text{HKY85}} = \begin{pmatrix} \bullet & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha\pi_T & \bullet & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & \bullet & \alpha\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha\pi_A & \bullet \end{pmatrix}$$

$$Q_{\text{TN93}} = \begin{pmatrix} \bullet & \alpha_1\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha_1\pi_T & \bullet & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & \bullet & \alpha_2\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha_2\pi_A & \bullet \end{pmatrix}$$

Hasegawa, Kishino, Yano (1984-85)

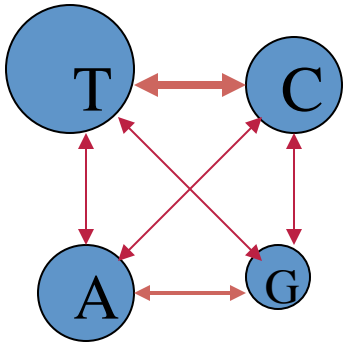
Tamura and Nei (1993)

Similar to F81 (Felsenstein 1981)



pyrimidines

purines



transversions

transitions

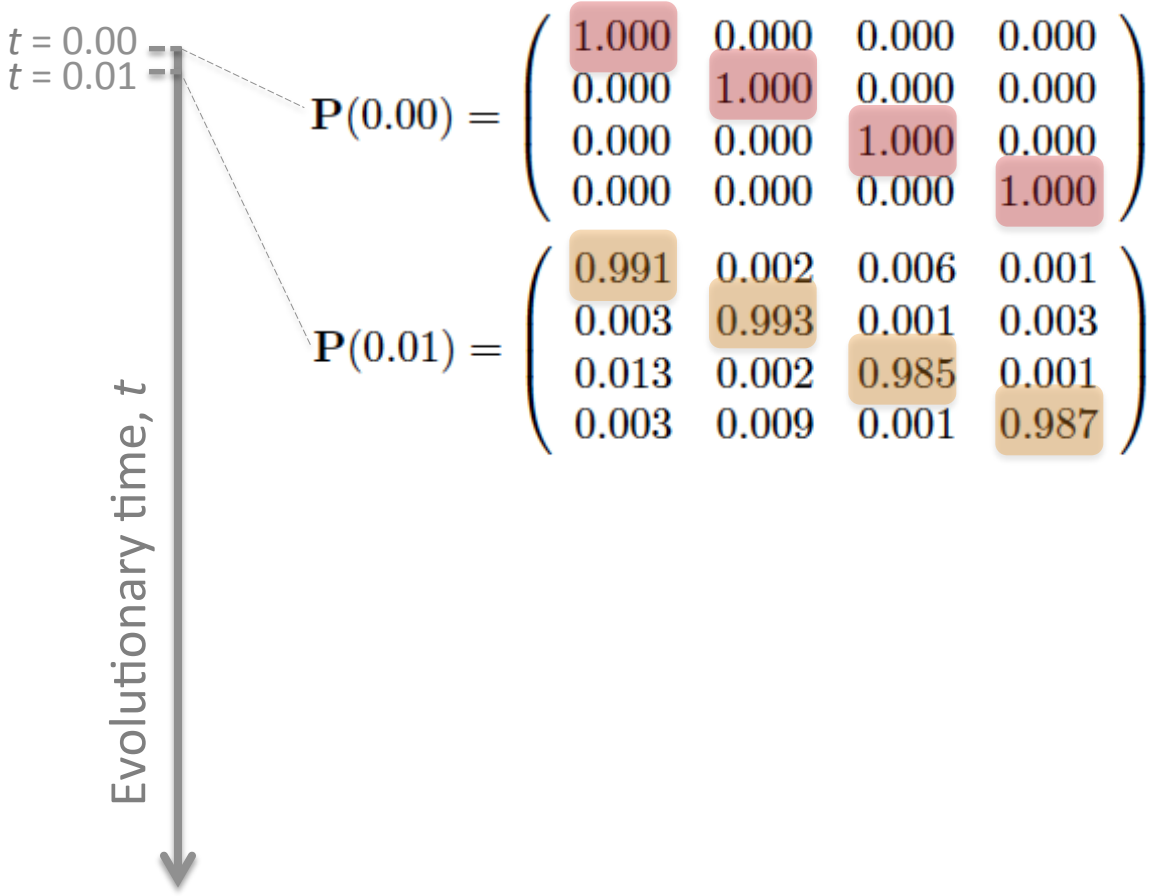
The probability of transition over time

$$\frac{dP(t)}{dt} = P(t)Q \text{ et } P(0)=I \qquad \Rightarrow \qquad P(t) = \exp(Qt)$$

$$\mathbf{P}(0.00) = \begin{pmatrix} 1.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix}$$

The probability of transition over time

$$\frac{dP(t)}{dt} = P(t)Q \text{ et } P(0)=I \quad \Rightarrow \quad P(t) = \exp(Qt)$$



HKY model:

$\kappa = 5$

$\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T) =$

$(0.4, 0.3, 0.2, 0.1)$

The probability of transition over time

$$\frac{dP(t)}{dt} = P(t)Q \text{ et } P(0)=I \quad \Rightarrow \quad P(t) = \exp(Qt)$$

HKY model:

$\kappa = 5$

$\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T) =$

$(0.4, 0.3, 0.2, 0.1)$

$$\mathbf{P}(0.00) = \begin{pmatrix} 1.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix}$$

$$\mathbf{P}(0.01) = \begin{pmatrix} 0.991 & 0.002 & 0.006 & 0.001 \\ 0.003 & 0.993 & 0.001 & 0.003 \\ 0.013 & 0.002 & 0.985 & 0.001 \\ 0.003 & 0.009 & 0.001 & 0.987 \end{pmatrix}$$

$$\mathbf{P}(1.00) = \begin{pmatrix} 0.580 & 0.141 & 0.232 & 0.047 \\ 0.188 & 0.587 & 0.094 & 0.131 \\ 0.464 & 0.141 & 0.348 & 0.047 \\ 0.188 & 0.394 & 0.094 & 0.324 \end{pmatrix}$$

The probability of transition over time

$$\frac{dP(t)}{dt} = P(t)Q \text{ et } P(0)=I \quad \Rightarrow \quad P(t) = \exp(Qt)$$

HKY model:

$\kappa = 5$

$\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T) =$

$(0.4, 0.3, 0.2, 0.1)$

$$\mathbf{P}(0.00) = \begin{pmatrix} 1.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix}$$

$$\mathbf{P}(0.01) = \begin{pmatrix} 0.991 & 0.002 & 0.006 & 0.001 \\ 0.003 & 0.993 & 0.001 & 0.003 \\ 0.013 & 0.002 & 0.985 & 0.001 \\ 0.003 & 0.009 & 0.001 & 0.987 \end{pmatrix}$$

$$\mathbf{P}(1.00) = \begin{pmatrix} 0.580 & 0.141 & 0.232 & 0.047 \\ 0.188 & 0.587 & 0.094 & 0.131 \\ 0.464 & 0.141 & 0.348 & 0.047 \\ 0.188 & 0.394 & 0.094 & 0.324 \end{pmatrix}$$

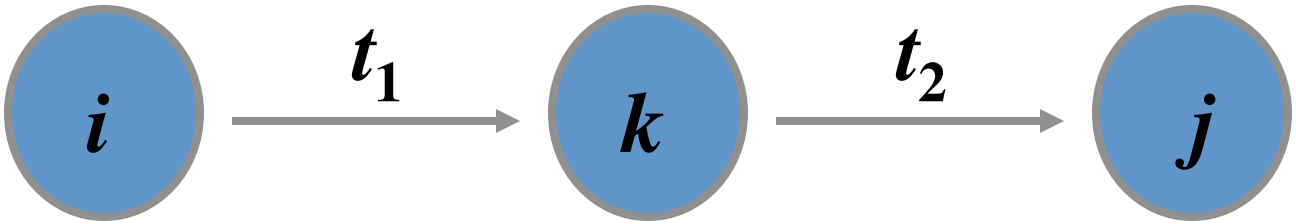
$$\mathbf{P}(100) = \begin{pmatrix} 0.400 & 0.300 & 0.200 & 0.100 \\ 0.400 & 0.300 & 0.200 & 0.100 \\ 0.400 & 0.300 & 0.200 & 0.100 \\ 0.400 & 0.300 & 0.200 & 0.100 \end{pmatrix}$$

Convergence
to stationary frequencies
stationnaires:

Multiple substitutions

Markov process accounts for multiple hits and hidden changes.
By Chapman-Kolmogorov theorem:

$$p_{ij}(t_1+t_2) = \sum_k p_{ik}(t_1)\, p_{kj}(t_2) \quad \text{for } k \in \{\text{T, C, A, G}\}$$



Stationarity

initial distribution of Markov chain $X(t)$:

$\boldsymbol{\pi}(0) = (\pi_T(0), \pi_C(0), \pi_A(0), \pi_G(0))$

At time t: $\boldsymbol{\pi}(t) = \boldsymbol{\pi}(0)\, \boldsymbol{P}(t)$

OR $\pi_i(t) = \pi_T(0)\, p_{Ti}(t) + \pi_C(0)\, p_{Ci}(t) + \pi_A(0)\, p_{Ai}(t) + \pi_G(0)\, p_{Gi}(t)$

The process is stationary if $\forall t > 0 \;\; \boldsymbol{\pi}(t) = \boldsymbol{\pi}(0)$

Stationary distribution: $\boldsymbol{\pi} = \boldsymbol{\pi} \boldsymbol{P}(t) \quad \Rightarrow \quad \boldsymbol{\pi} \boldsymbol{Q} = 0$

($\boldsymbol{\pi}$ is an eigenvector for eigenvalue 0)

OR $\sum_i \pi_i q_{ij} = 0$ (for $\forall\, j$)

$-\pi_j q_{jj} = \sum_{i \neq j} \pi_i q_{ij}$

(Total flow out of j = Total flow into j)

Time reversibility

Markov process is *time-reversible* if and only if

$\forall (i \neq j) \qquad \pi_i q_{ij} = \pi_j q_{ji}$

(In steady state: flow $i \rightarrow j$ = flow $j \rightarrow i$)

OR $\forall (t, j, i \neq j) \qquad \pi_i p_{ij}(t) = \pi_j p_{ji}(t)$

If reversibility assumed:

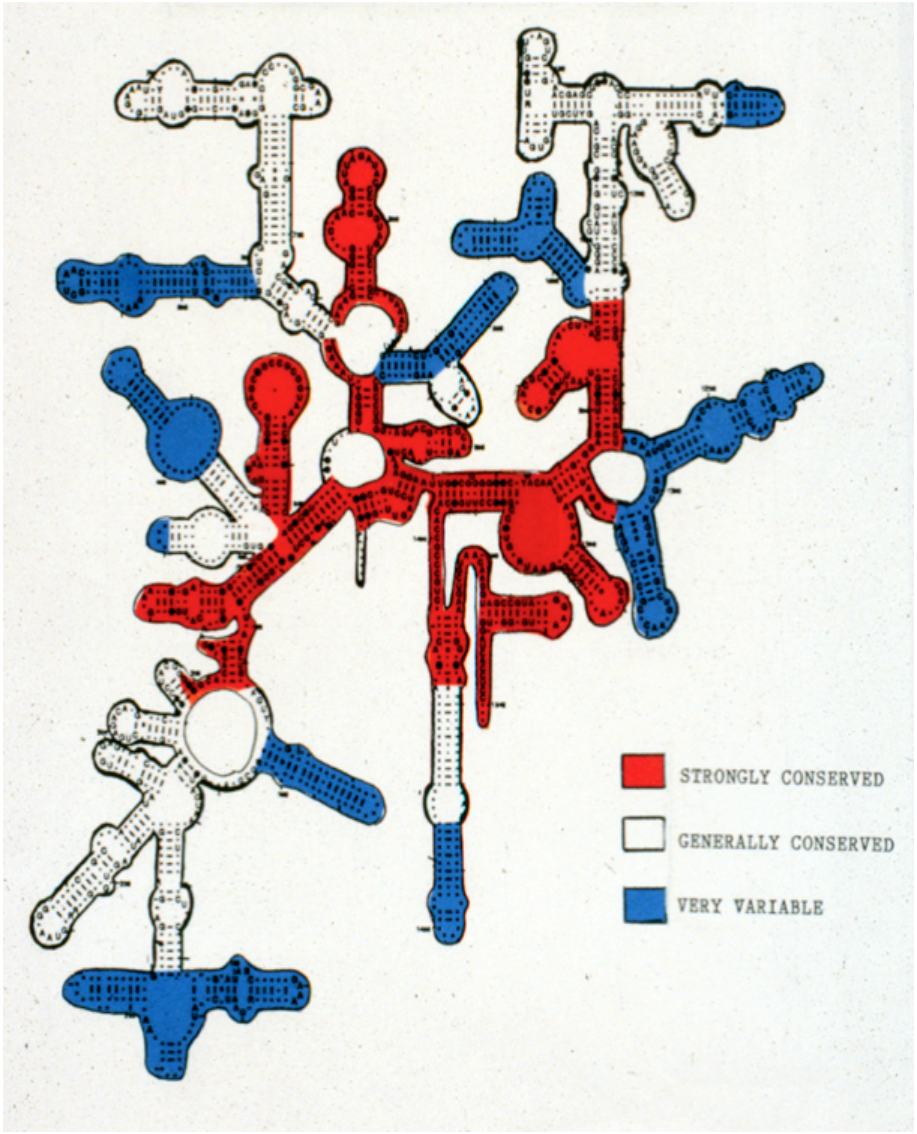
$q_{ij} = s_{ij}\pi_j$, where $s_{ij} = s_{ji}$ is exchangeability between i and j

Q is described by 9 independent parameters (GTR or REV, Tavare 1986):

$$Q = \begin{pmatrix} \bullet & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & \bullet & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & \bullet & f\pi_G \\ c\pi_T & e\pi_C & f\pi_A & \bullet \end{pmatrix} = \begin{pmatrix} \bullet & a & b & c \\ a & \bullet & d & e \\ b & d & \bullet & f \\ c & e & f & \bullet \end{pmatrix} \begin{pmatrix} \pi_T & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_A & 0 \\ 0 & 0 & 0 & \pi_G \end{pmatrix}$$

Model with no reversibility constraint: UNREST (Yang 1994)

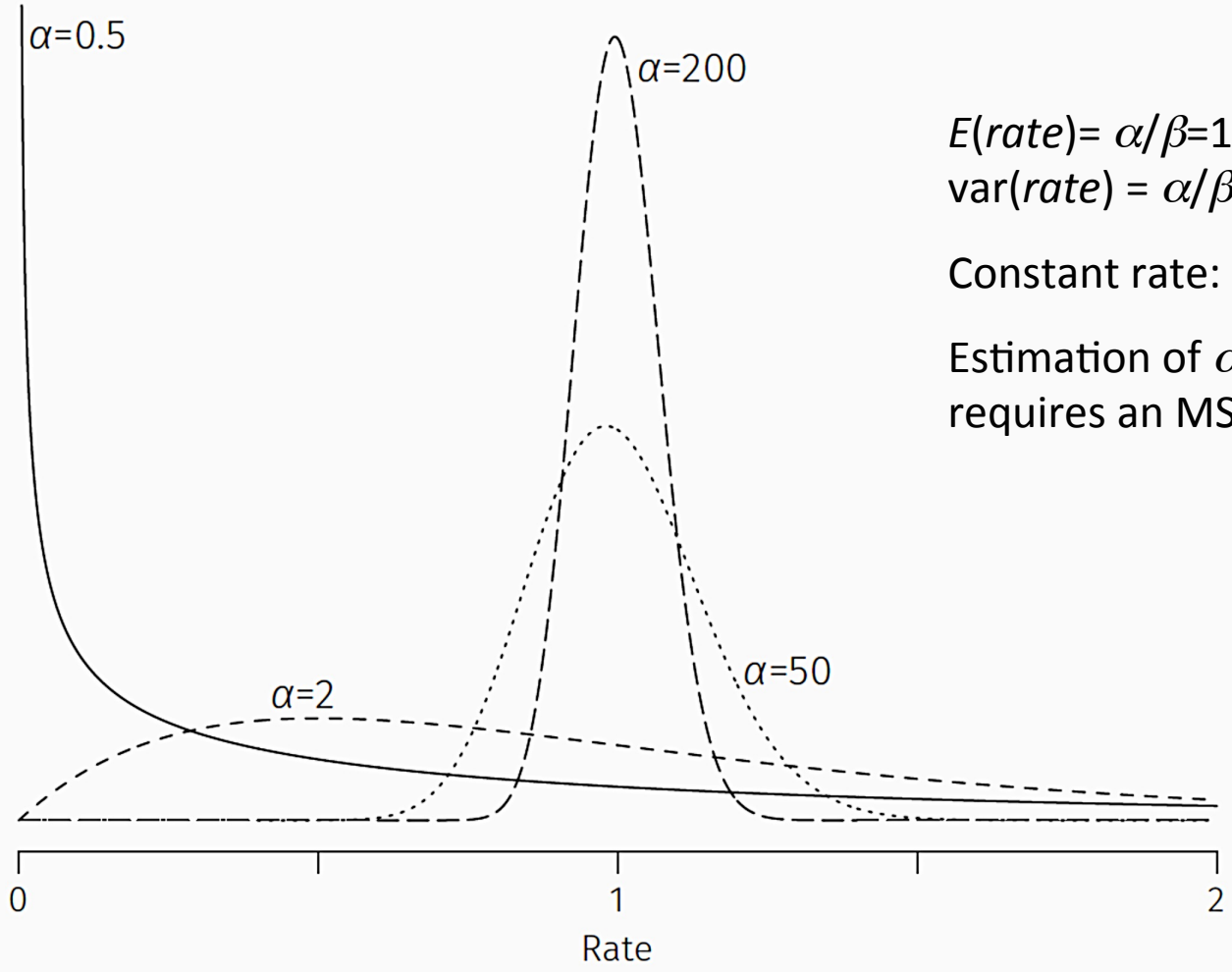
Across-sites rate variability



Small subunit ribosomal RNA (18S or 16S)

Across-sites rate variability

Can be modeled using the Γ-distribution with α = β



$E(rate) = \alpha/\beta = 1$

$var(rate) = \alpha/\beta^2 = 1/\alpha$

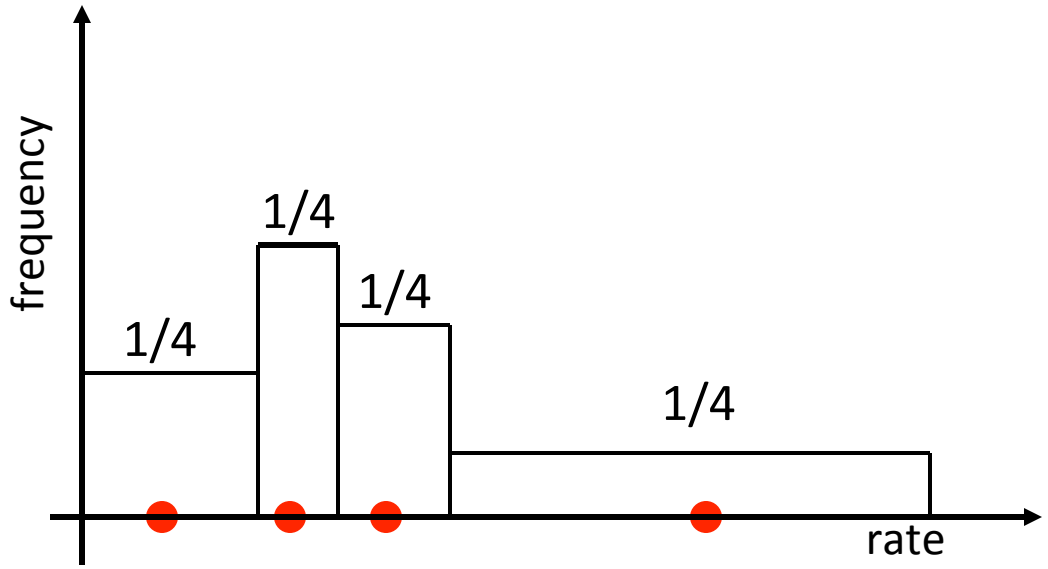
Constant rate: $\alpha \to \infty$

Estimation of α requires an MSA

The gamma distribution has no biological justification, it was chosen for its convenience.

Across-sites rate variability

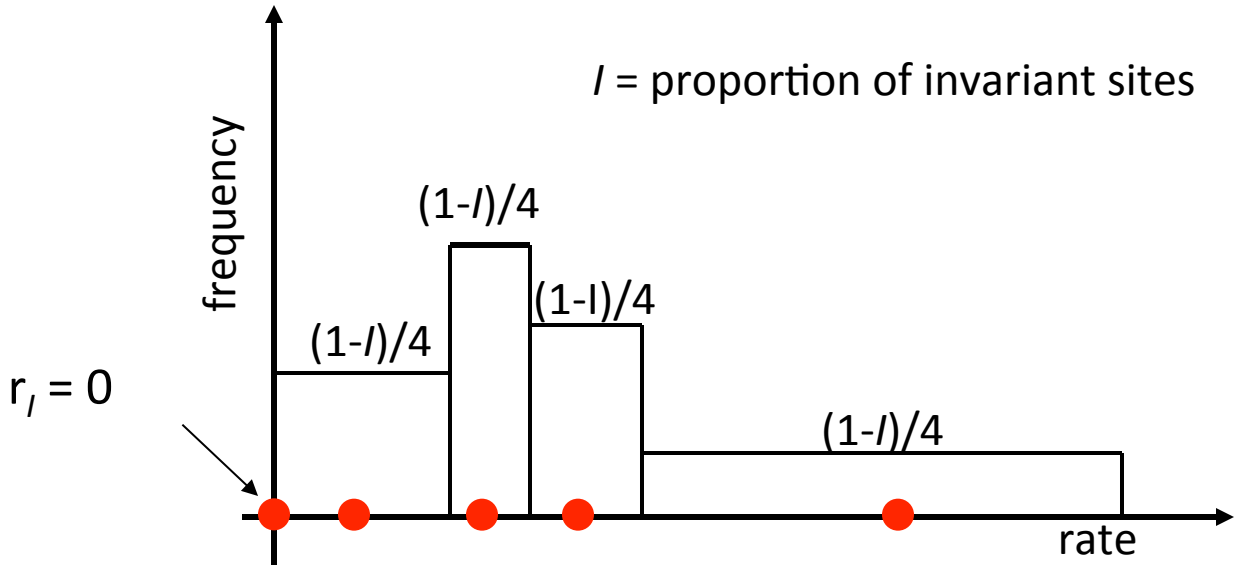
The Γ-distribution is simplified by discretization, for example with 4 classes of equal weight:



Across-sites rate variability

Γ + I model allows a proportion of invariable sites

I should be estimated from the data



How many discrete categories?

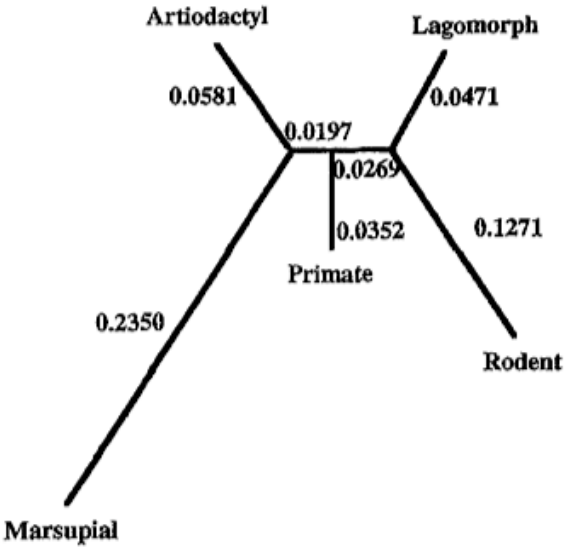


Fig. 3. The maximum likelihood tree for the five orders of mammals from the α and β globin genes (570 bp). The F84 + Γ model was assumed. Branch lengths are measured by the average numbers of nucleotide substitutions per site.

as the F84 + Γ and F84 + dG4 models: the other two

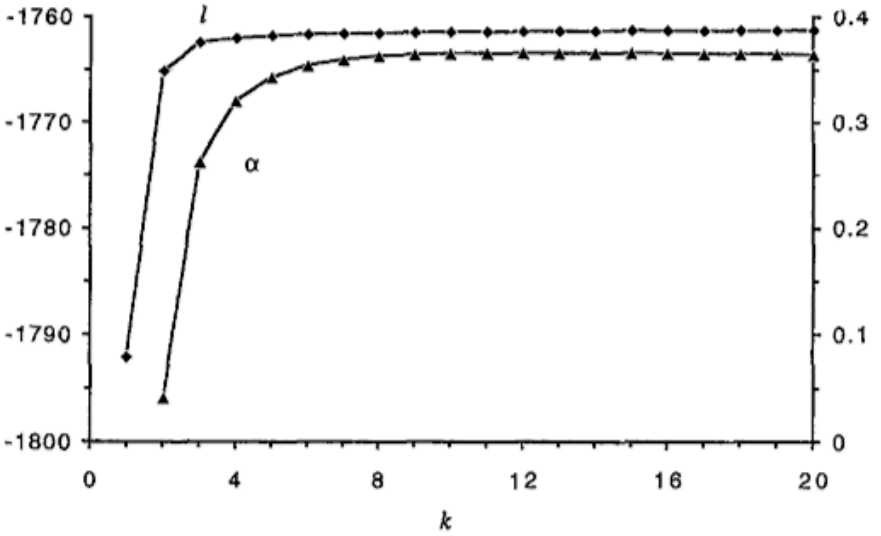


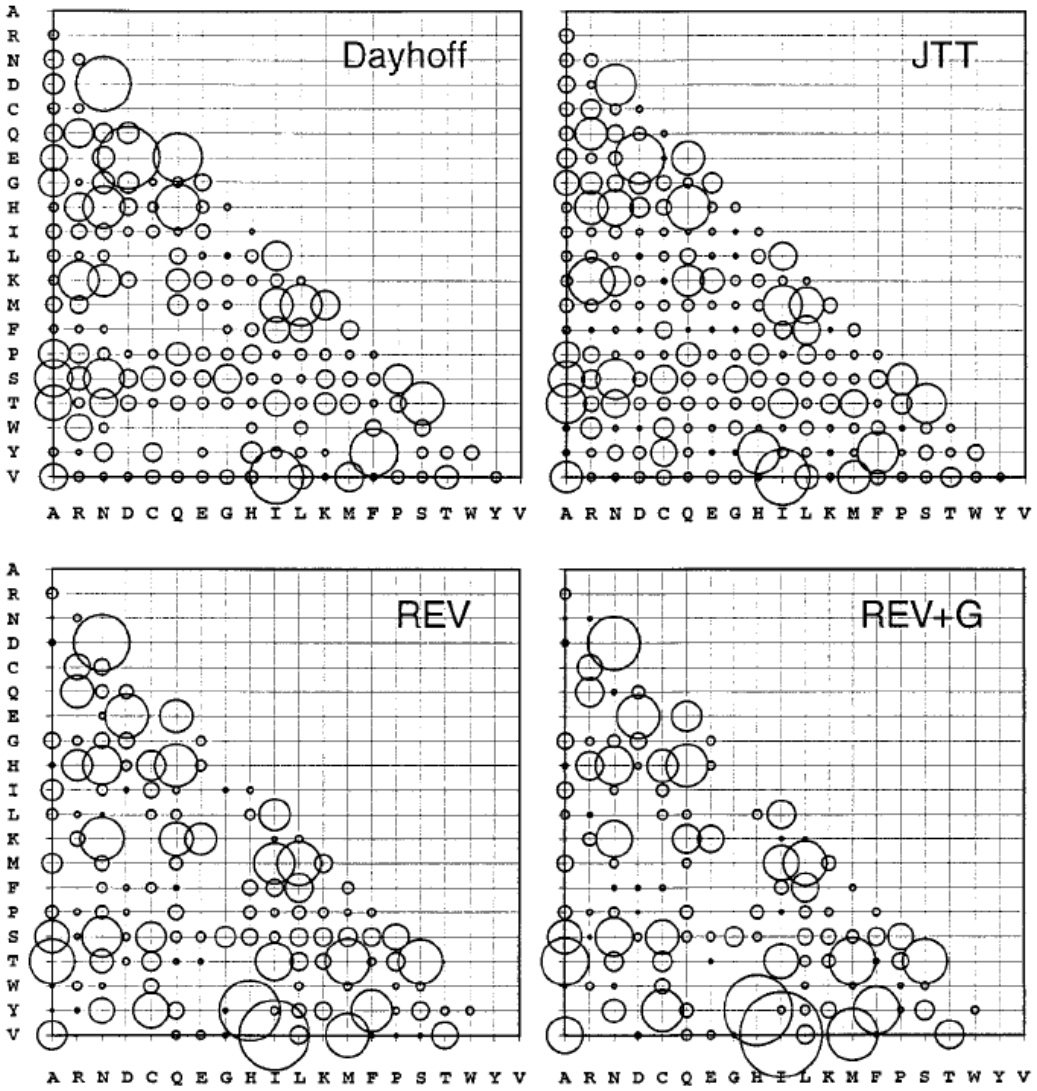
Fig. 4. Likelihood values and estimates of the α parameter as functions of k, the number of categories in the discrete gamma model. The α and β globin genes for the five mammalian orders (570 bp) are analyzed, assuming the best tree (Fig. 3) and the F84 + dG model. The average nucleotide frequencies are $\pi_T = 0.2200$, $\pi_C = 0.2449$, $\pi_A = 0.2761$, and $\pi_G = 0.2590$, with $\ell_{max} = -1{,}579.76$. When $k = \infty$, that is, with the F84 + Γ model, $\ell = -1{,}761.17$ and $\hat{\alpha} = 0.360$.

From Yang 1994

Table 1. Maximum likelihood estimates of the α parameter[a]

Sequences	Species	$\hat{\alpha}$	Refs
Nuclear genes			
α- and β-globin genes, positions 1 and 2	5 mammals	0.36	10,23
Albumin genes, all positions	5 vertebrates	1.05	44
Insulin genes, all positions	5 vertebrates	0.40	44
c-myc genes, all positions	5 vertebrates	0.47	44
Prolactin genes, all positions	5 vertebrates	1.37	44
16S-like rRNAs, stem region	5 species	0.29	45
16S-like rRNAs, loop region	5 species	0.58	45
$\psi\eta$-globin pseudogenes	6 primates	0.66	23
Viral genes			
Hepatitis B virus genomes	13 variants	0.26	46
Mitochondrial genes			
12S rRNAs	9 rodents	0.16	22
895-bp mtDNAs	9 primates	0.43	10
Positions 1 and 2 of 13 genes[b]	11 vertebrates	0.13–0.95	28
Position 1 of four genes	6 primates	0.18	19
Position 2 of four genes	6 primates	0.08	19
Position 3 of four genes	6 primates	1.58	19
D-loop region of mtDNAs[c]	25 humans	0.17	12
Protein sequences			
Mitochondrial cytochrome *b*	16 deuterostomes	0.44	12

Empirical models for proteins



Empirical models for proteins

JTT

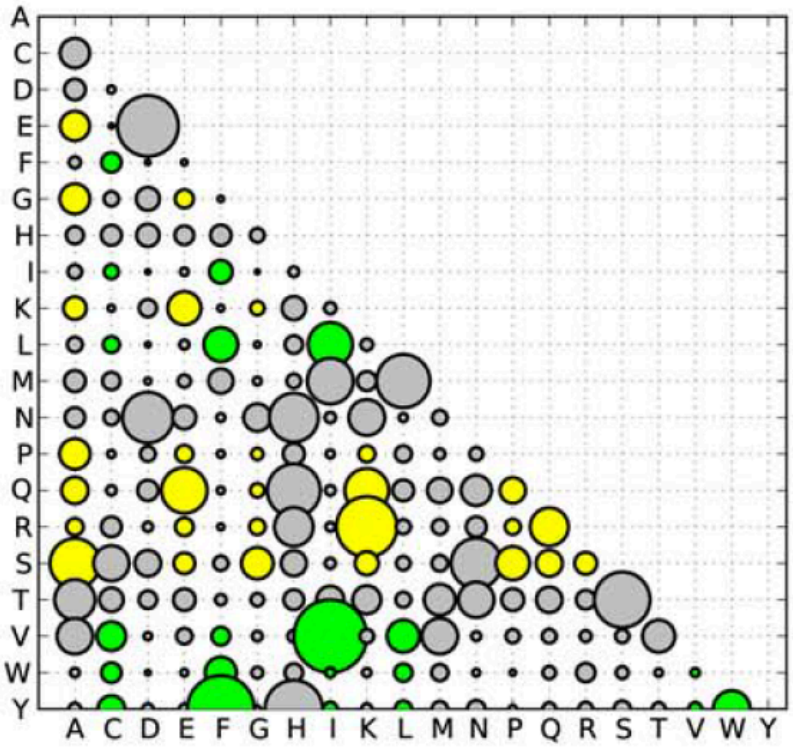
WAG

LG

HIV

Order/IDP

α/β



Q-matrix

AA stationary frequencies: π_i

+ F option: estimate frequencies from data

Which model?

Which model?

The best!

Which model?

The best!

Need a criterion to decide “the best”?

Sir R. A. FIsher

Likelihood

The likelihood of model M, parameters θ,
Given data D is:

$$L(M; \theta \mid D) = Pr(D \mid M; \theta)$$

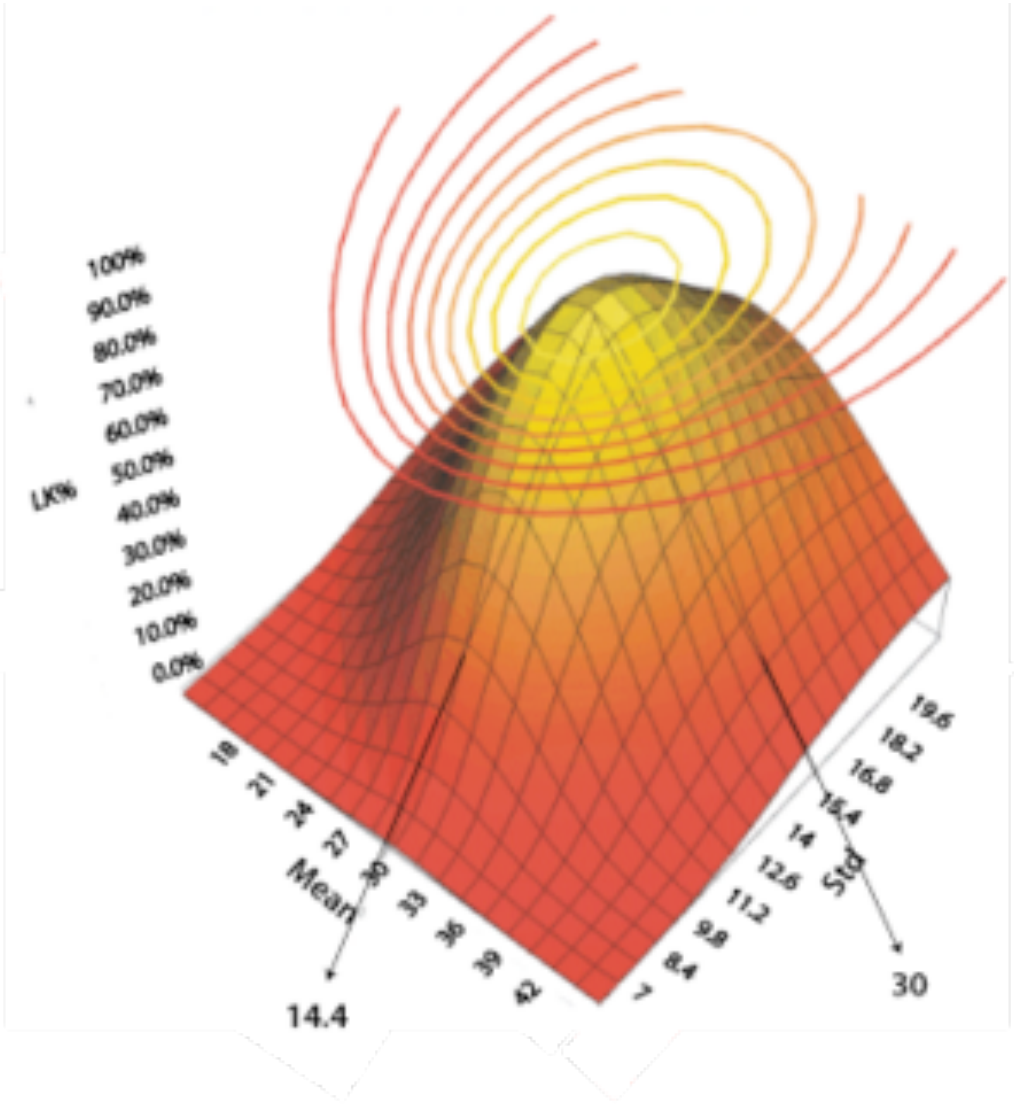
Maximum likelihood (ML) inference finds $\hat{\theta}$, the best-supported value of parameters θ:

such that $L(M; \hat{\theta} \mid D) \geq L(M; \hat{\theta} \mid D)$ for all other θ

M with parameters θ describes your hypothesis.

The ML method was pioneered by Sir R.A. Fisher in 1921-22
Lindgren (1968), Edwards (1984)

Likelihood


100%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
LK%
18
21
24
27
30
33
36
39
42
Mean
14.4
7
8.4
9.8
11.2
12.6
14
15.4
16.8
18.2
19.6
Std
30

Hypothesis tests

A hypothesis is a statement about the state of nature. It may need substantiation, verification or rejection.

A test of a hypothesis assigns one of the inferences:

- 'accept' the hypothesis or
- 'reject' the hypothesis for some result of an experiment

Example: Fair coin

Toss coin 100 times, observe 65 heads and 35 tails.
Null hypothesis H_0: "The coin is fair"
(i.e. probability 0.5 for Heads)

Calculate the likelihood:

$$L(H_0|D) = \binom{100}{65} \times 0.5^{65} \times 0.5^{35} = 0.000864$$

$$\log(L(H_0|D)) = \log(0.000864) = -7.0541$$

Example: Biased coin

Alternative hypothesis H_1:
"The coin is biased with probability p of heads"
The ML estimate of p is 65/100 = 0.65
Optimized the likelihood:

$$L(H_1|D) = \binom{100}{65} \times p^{65} \times (1-p)^{35}$$

$$= \binom{100}{65} \times 0.65^{65} \times 0.35^{35} = 0.08340$$

$$\log(L(H_1|D)) = \log(0.08340) = -2.484$$

H_1 is more likely, but is the result significant?

Hypothesis testing

Test the null hypothesis H_0 against the alternative H_1

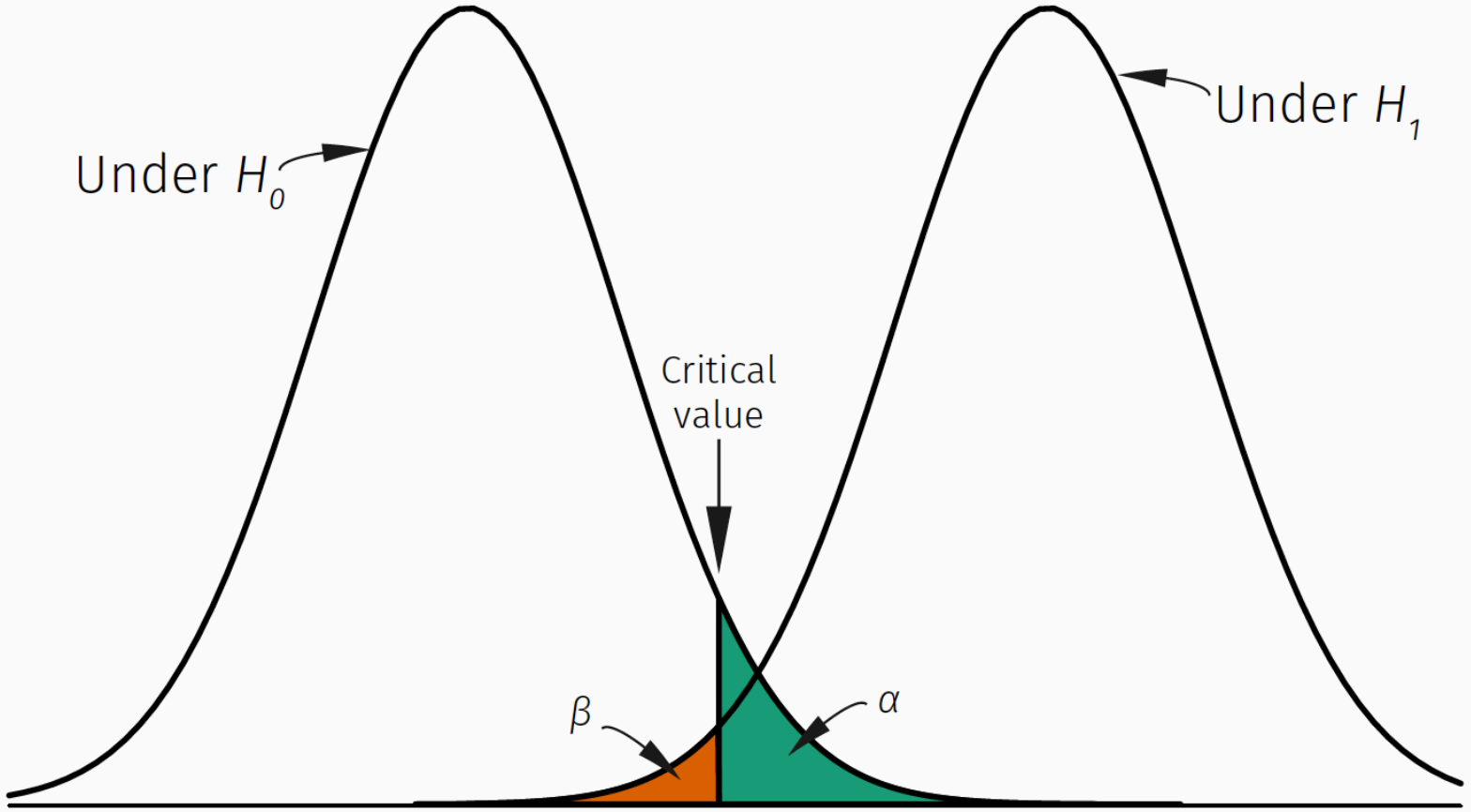
- A *test statistic T* is used as a reduction of the data
- The range of values for rejecting H_0 being
- tested is called the *critical region*
- There are good and bad tests, leading to the wrong inference or statistical errors:

  Type I error: Rejecting H_0 when H_0 is true.

  Type II error: Accepting H_0 when H_0 is false

Type I and II errors

- α = "size of type I error" = P_{H_0}(reject H_0)
- β = "size of type II error" = P_{H_1}(accept H_0)



Nested hypotheses

Two models are *nested* if one model can be reduced to another model by constraining some of its parameters.

In our example: forcing p = 0.5 in H_1 reduces it to H_0

H_1 has one more parameter than H_0

$$P(H_1, p) = \binom{100}{65} \times p^{65} \times (1-p)^{35}$$

Fix p to 0.5

$$P(H_1, p = 0.5) = \binom{100}{65} \times 0.5^{65} \times 0.5^{35} = P(H_0)$$

Likelihood ratio test (LRT)

Test H_0 against H_1, given they are nested

Use likelihood ratio statistic:

$$\ell_0 = \log\{L(H_0)\}$$
$$\ell_1 = \log\{L(H_1)\}$$
$$T = 2\delta = 2\log\left(\frac{L(H_1)}{L(H_0)}\right) = 2(\ell_1 - \ell_0)$$

When H_0 is correct, the LRT statistic is asymptotically distributed as χ^2 distribution with *k* degrees of freedom (equal to the difference in the number of parameters in H_0 and H_1)

Significance level and *p*-value

Choose the rejection region given null is true:

$P(T \geq t \mid H_0) = \alpha$

T is the calculated test statistic from data

t is the chosen cut-off for the critical region

α is the desired significance level

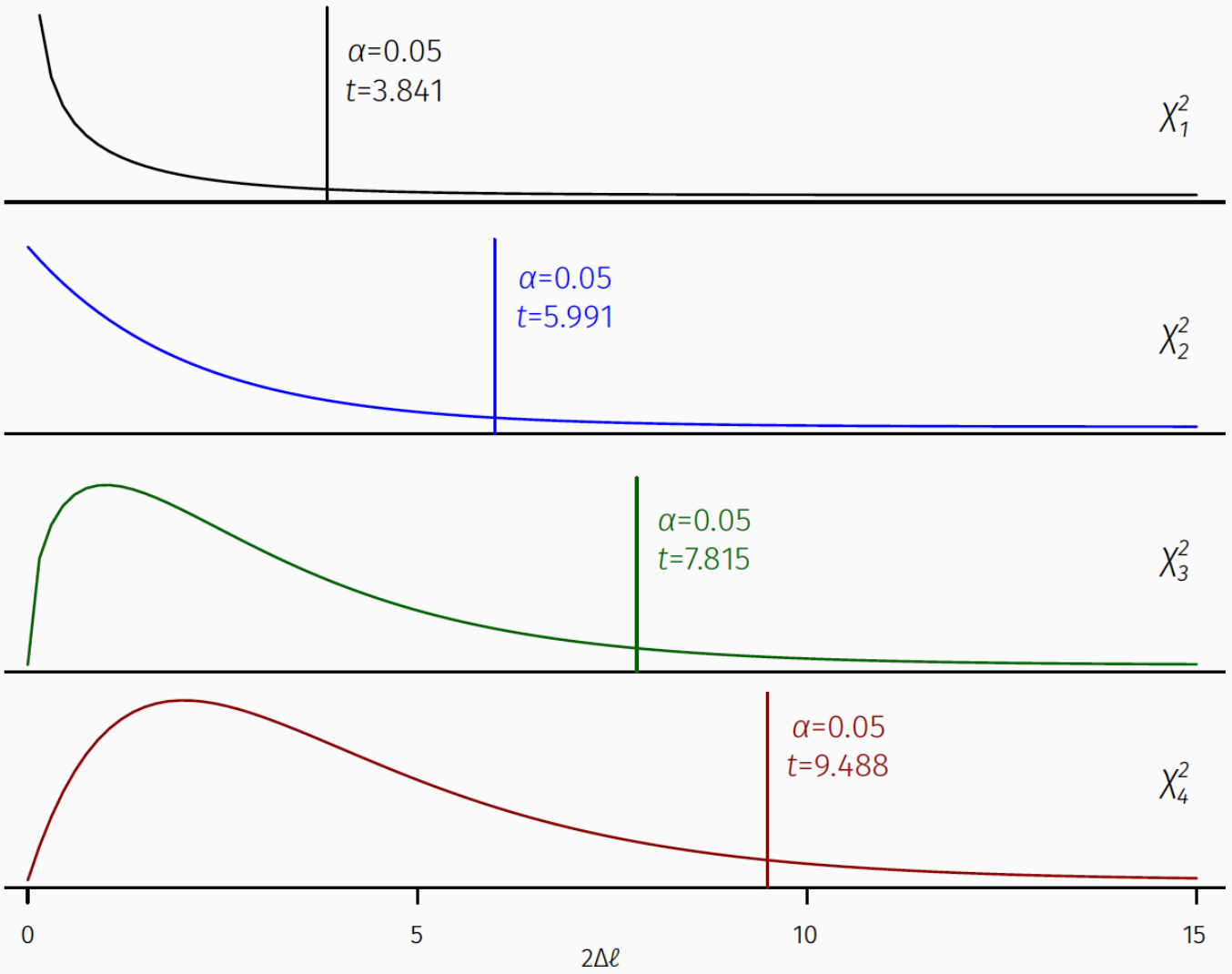
Choose a small value of α (e.g. 0.05 or 0.01)

For example, for χ^2 with d.f. = 1:

$P(T \geq 3.841) = 0.05$ and $P(T \geq 6.634) = 0.01$

p-value is probability of a result at least as extreme as that observed if H_0 were true

X^2 distributions



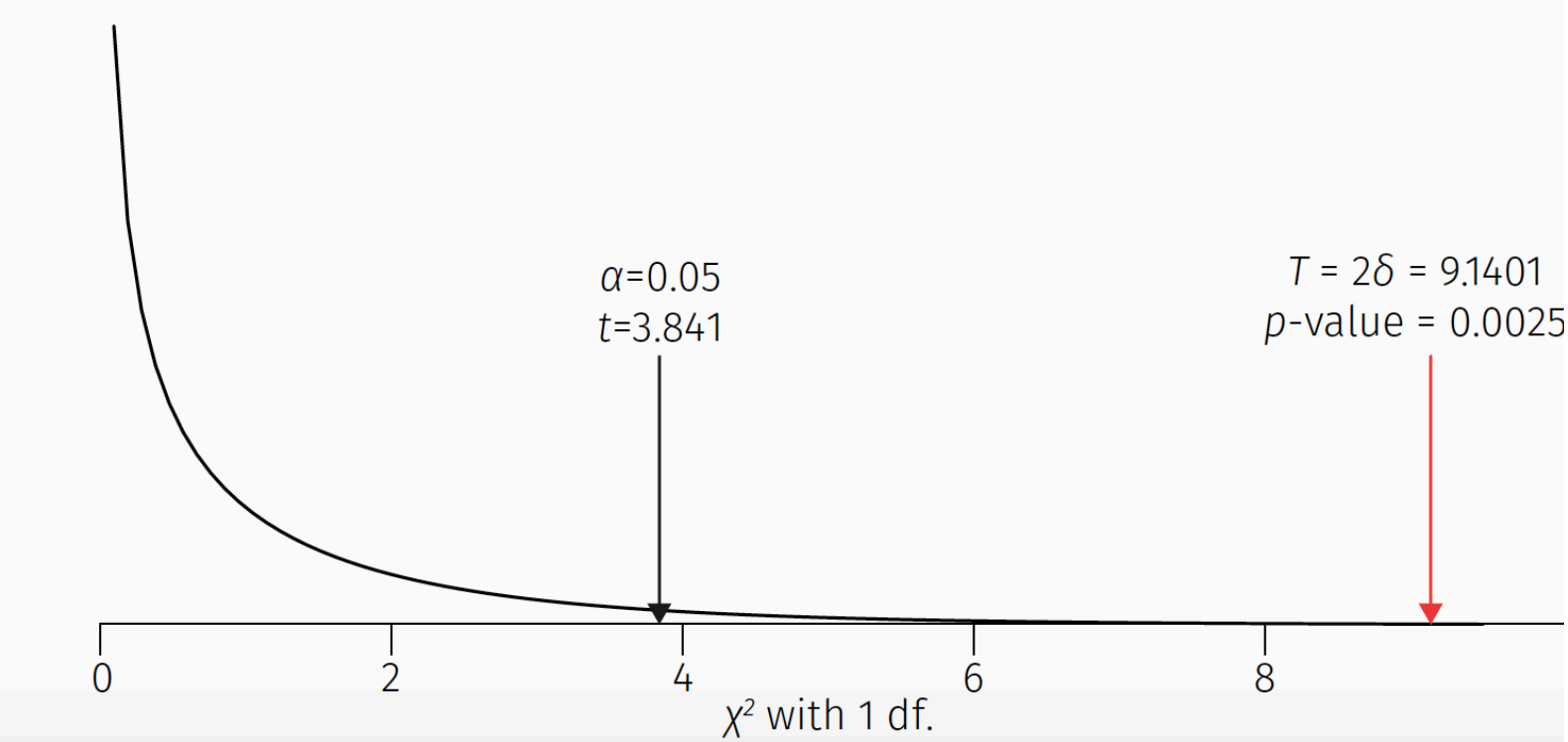
Exampl LRT: Fair vs biased coin

$2\delta = 2(\ell_1 - \ell_0) = 2(-2.484 - -7.0541) = 9.1401$

1 more parameter (p) in H_1, so use χ^2 with 1 d.f.

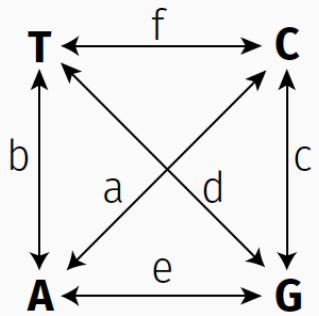
p-value = 0.0025 < 0.05

Reject the null H_0 in favour of the alternative H_1



Nested models

Model	Base frequencies	Substitution rates	Free parameters
JC	$\pi_T = \pi_C = \pi_A = \pi_G$	a = b = c = d = e = f	0
K80	$\pi_T = \pi_C = \pi_A = \pi_G$	a = b = c = d ≠ e = f	1
F81	$\pi_T \neq \pi_C \neq \pi_A \neq \pi_G$	a = b = c = d = e = f	3
HKY	$\pi_T \neq \pi_C \neq \pi_A \neq \pi_G$	a = b = c = d ≠ e = f	4
GTR	$\pi_T \neq \pi_C \neq \pi_A \neq \pi_G$	a ≠ b ≠ c ≠ d ≠ e ≠ f	8



Adapted from Posada & Crandall (2001).

LRT: JC vs K80

H_0: JC model
H_1: K80 model (with κ or ts/tv rate ratio)

- Both hypotheses use the same tree topology and have
- same number of branch length parameters.
- JC is nested within the K80 model.
- Fixing κ = 1 in K80 gives the JC model.
- The difference in number of parameters is 1 (κ).
- Perform the LRT by comparing 2δ with χ^2 d.f. = 1

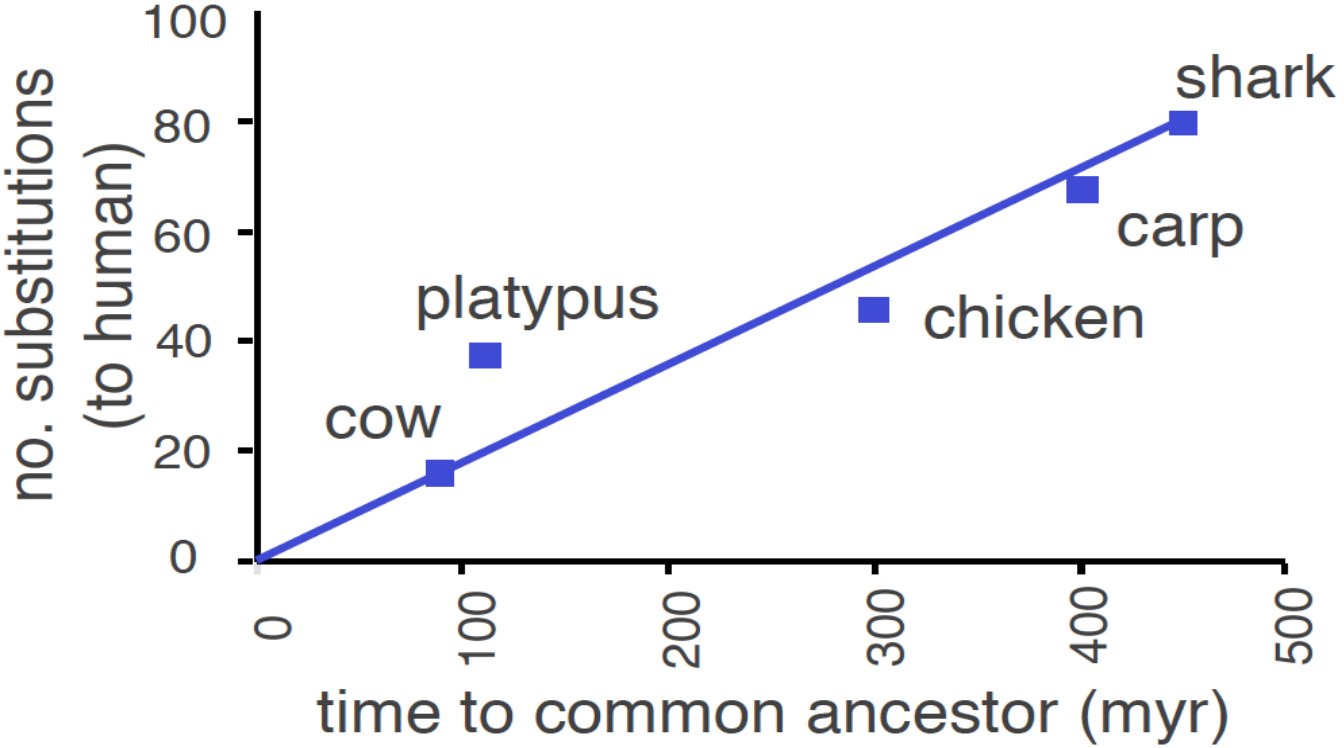
LRT: GTR vs GTR+Γ

H_0: GTR model
H_1: GTR+Γ (GTR parameters + α parameter)
GTR is nested within GTR+Γ, as $\alpha \rightarrow \infty$ recovers GTR

But, this value is on the boundary of the parameter space, so:

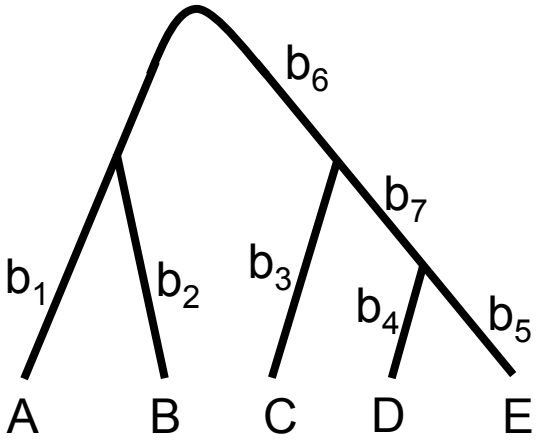
- Test 2δ with 50:50 mixture of point mass 0 and χ^2 with d.f. = 1
- Critical values are 2.71 at 5% and 5.41 at 1%
- See Goldman & Whelan (2000) for further details and table of critical values.

LRT: constant rate over time?



LRT: constant rate over time?

H_1: no clock
Parameters:
$2T - 3 = 7$ for T taxa



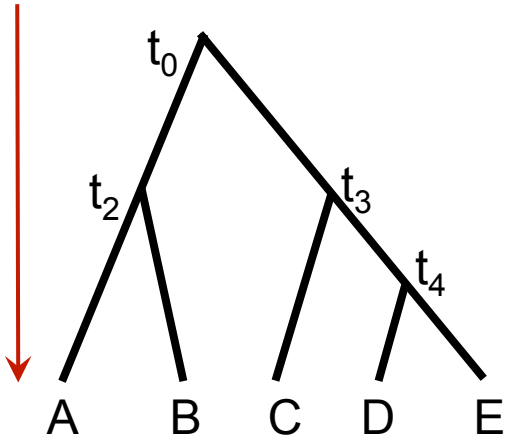
$b_1 = b_2$

$b_4 = b_5$

$b_3 = b_4 + b_7$

$T - 2 = 3$
constraints

H_0: clock
Parameters:
$T - 1 = 4$



Akaike Information Criterion

$$AIC = 2k - 2\log(L)$$

k is number of free model parameters
L is the maximum likelihood

- More parameters lead to a larger penalty
- We choose the model with the lowest AIC value
- Can be used with non-nested models
- Can rank models

AICc and BIC

For small sample size *n* compared to the number of parameters *k* (e.g. $n / k < 40$) use corrected AIC:

$$AIC_c = 2k - 2\log(L) + \frac{2k(k+2)}{n-k-1}$$

Bayesian information criterion is related to AIC.
BIC has a larger penalty for parameters than AIC,
so is more conservative and prefers simpler models.

$$BIC = k\log(n) - 2\log(L)$$