Modeling Molecular Evolution: Hypothesis testing

Maria Anisimova

Institute of Computational Life Sciences
Zurich University of Applied Sciences - ZHAW

Life Sciences and Facility Management

Institute of Computational Life Sciences

10.05.2022, Lund

Swiss Institute of Bioinformatics

Many substitution models exist.
Which model to use? The best!

But what does this mean?
Need a criterion

Sir R. A. FIsher

Likelihood

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

L(M; θ I D) = Pr(D I M; θ)

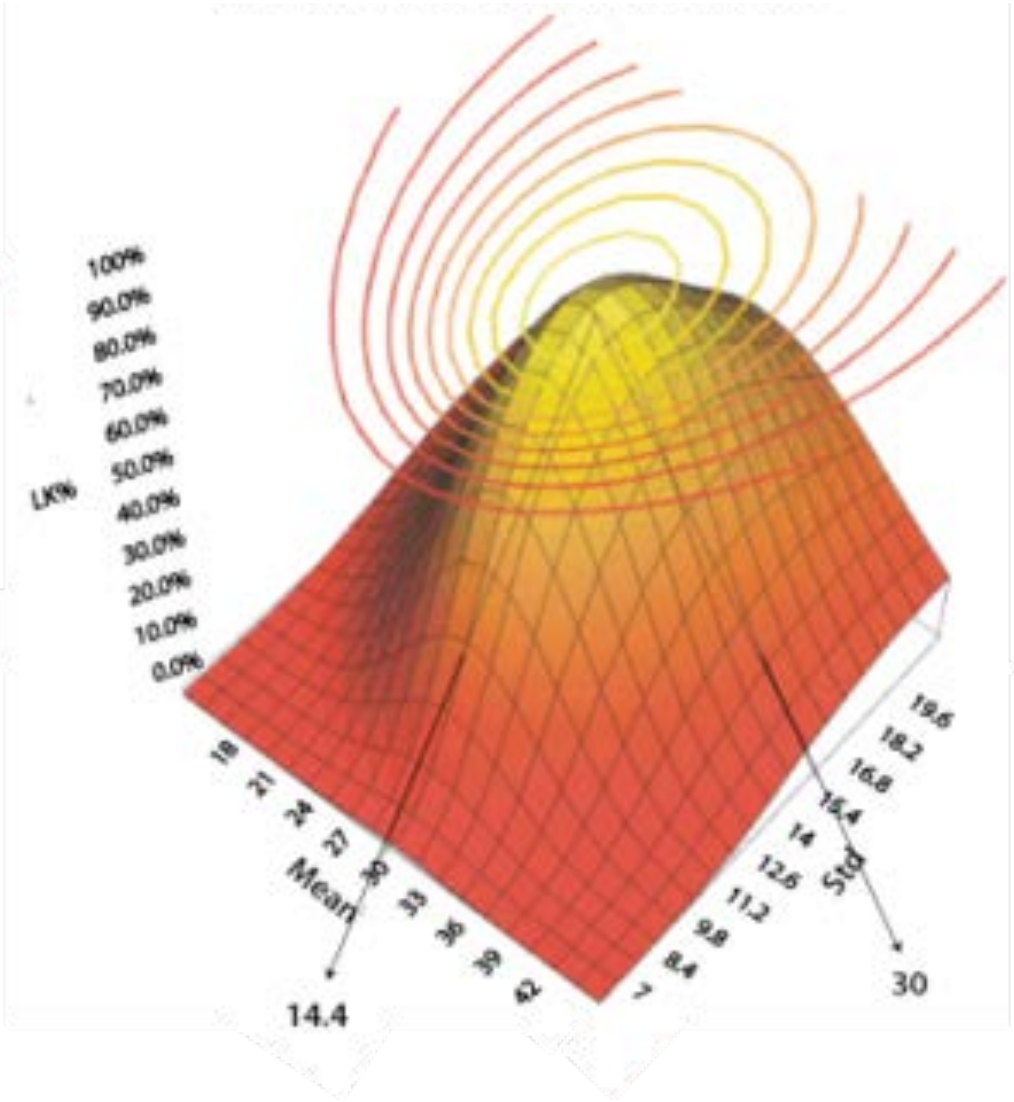
Maximum likelihood (ML) inference finds $\hat{\theta}$, the best-supported value of parameters θ:

such that L(M; $\hat{\theta}$ I D) ≥ L(M; $\hat{\theta}$ I 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



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

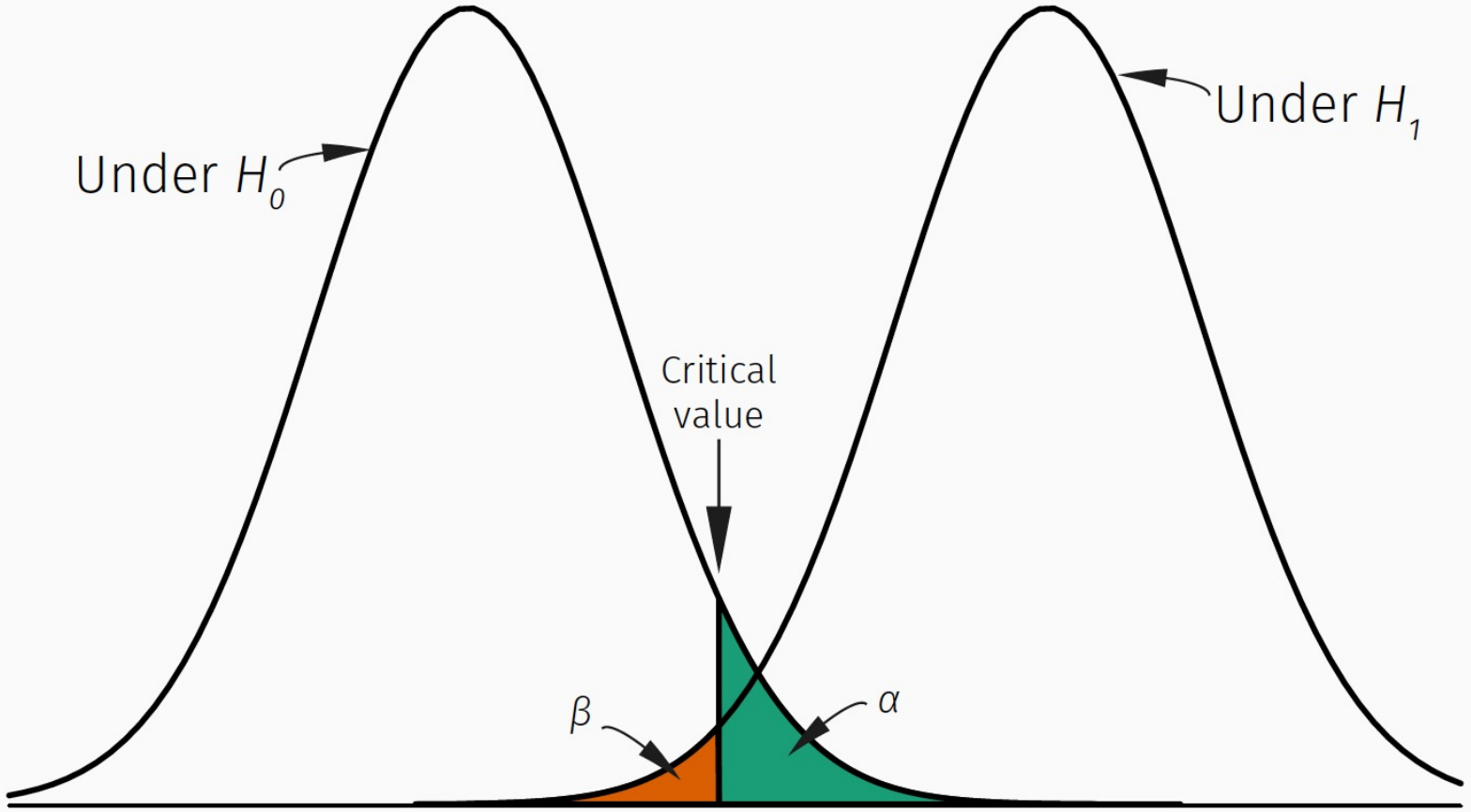
- A *test statistic T* is used as a reduction of the data
- The range of values for rejecting H_0 being
- tested is called the *critical region*
- There are good and bad tests, leading to the wrong inference or statistical errors:

  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

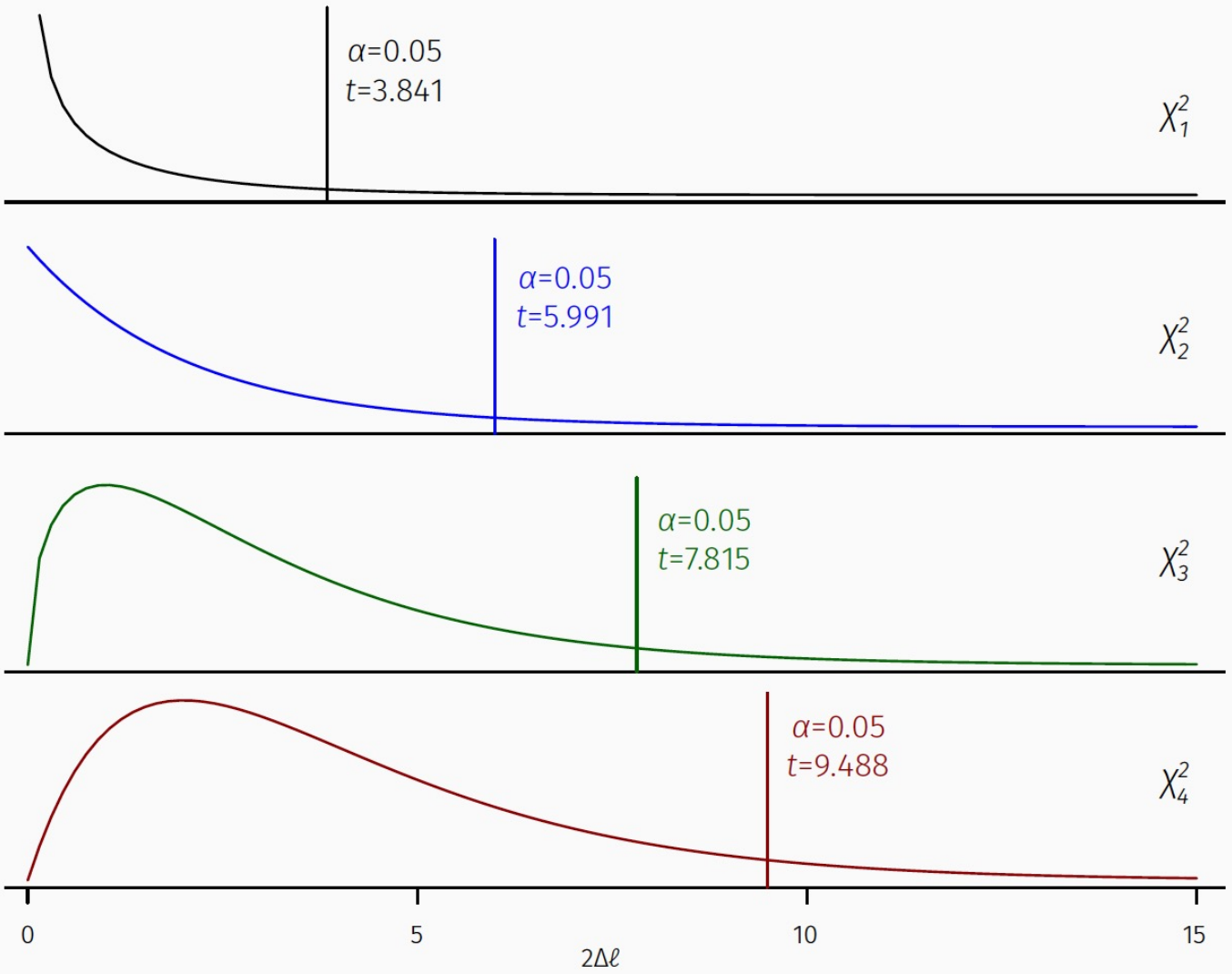
Choose a small value of α (e.g. 0.05 or 0.01)

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



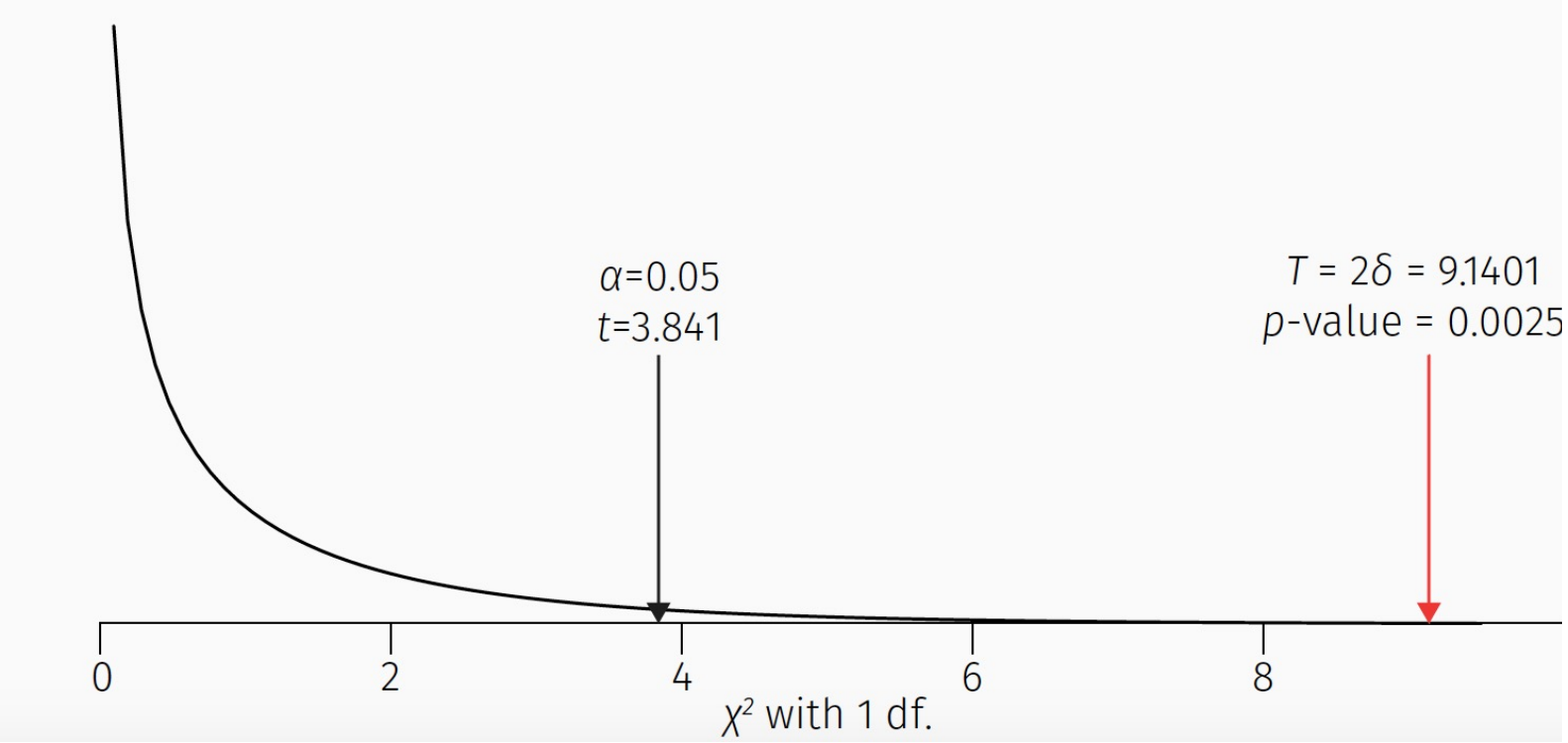
Exampl LRT: Fair vs biased coin

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

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

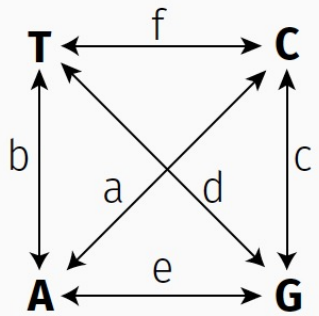
p-value = 0.0025 < 0.05

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

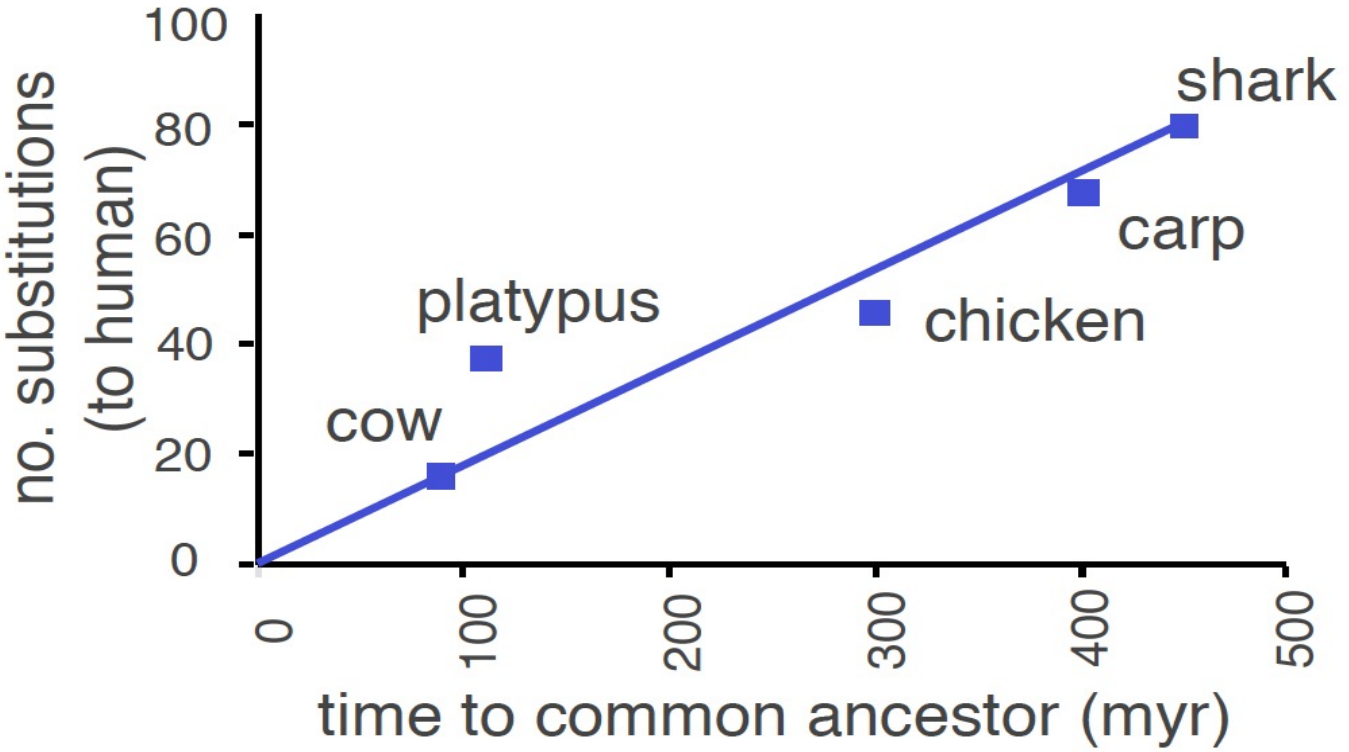
H_1: GTR+Γ (GTR parameters + α parameter)

GTR is nested within GTR+Γ, as $\alpha \to \infty$ recovers GTR

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

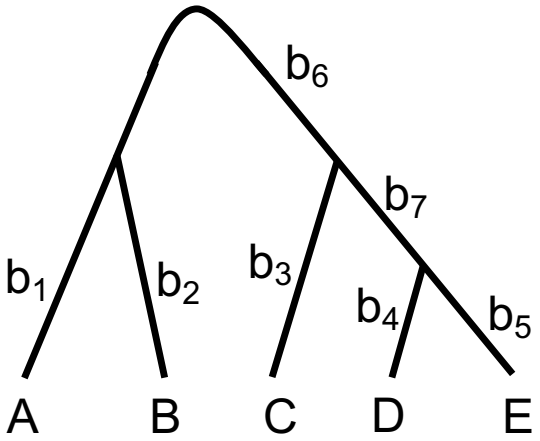
- Test 2δ with 50:50 mixture of point mass 0 and χ^2 with d.f. = 1
- Critical values are 2.71 at 5% and 5.41 at 1%
- See Goldman & Whelan (2000) for further details and table of critical values.

LRT: constant rate over time?



LRT: constant rate over time?

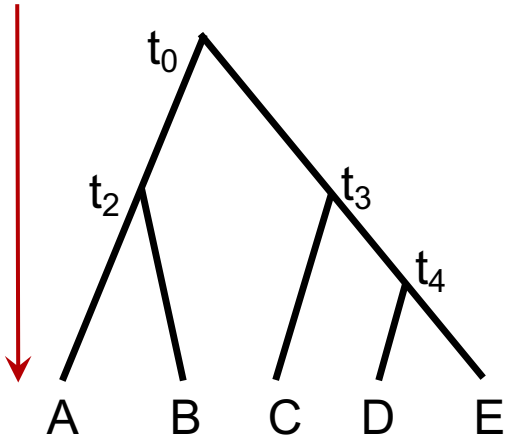
H_1: no clock
Parameters:
$2T - 3 = 7$ for T taxa



$$b_1 = b_2$$
$$b_4 = b_5$$
$$b_3 = b_4 + b_7$$

$T - 2 = 3$
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.

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