# Evaluating Tree Topologies

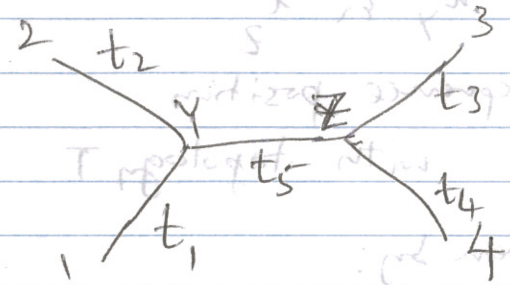
2 Feb 2015



Evaluate Trees with a maximum likelihood method

1) you need a model for the evolutionary processes that have occurred

2) model is time-reversible $\Rightarrow$ give the same likelihood regardless of which time direction is assumed for a branch $\Rightarrow$ location of root unimportant

3) 


branch lengths are evolutionary time as opposed to distance in the previous examples.

4 leaves - 1 to 4
unknown ancestral seq - at Y & Z

probabilities depend on topology (T)

consider each branch separately

$P(j \mid i, t)$ : prob of base i mutating into base j in time t.

- analytical form depends on evolutionary model

<u>JC:</u>

$$P(j \mid i, t) = \begin{cases} \frac{1}{4}(1 + 3e^{-4\alpha t}) & j = i \\ \frac{1}{4}(1 - e^{-4\alpha t}) & j \neq i \end{cases}$$

not $e^{-8\alpha t}$ because that referred to simultaneous evolution of two organisms.

④ What is common with parsimony?
- consider sequence data
- look at each alignment ptn in turn
- ignore sites with gaps
- arbitrary & temporary use of a tree node as the root during the calculation

⑤ 1, 2, 3, 4 are sequences, what we are looking for likelihood of specific bases $x_Y$ & $x_Z$ occurring at a particular sequence position at internal nodes Y and Z with topology T & branch lengths $t_i$ is given by:

$$P(x_1, x_2, x_3, x_4, x_Y, x_Z \mid T, t_1, t_2, t_3, t_4, t_5) =$$

$$q_{x_Y} P(x_1 \mid x_Y, t_1) P(x_2 \mid x_Y, t_2) P(x_Z \mid x_Y, t_5) P(x_3 \mid x_Z, t_3) P(x_4 \mid x_Z, t_4)$$

prob of base $x_Y$ at internal node Y (taken as root)

each branch evolved independently, so the probabilities are multiplied together.

$x_1, x_2 \ldots$ | at the location under consideration these are the bases in sequences 1, 2, ...

(5)

(6) internal nodes Y & Z — are sequences

they could have any possible base for the $i^{th}$ position

$\mathcal{L}_i$ = likelihood of observed sequences $x_1$ to $x_4$

= consider all possible bases at $x_Y$ and $x_Z$

$$= \sum_{x_Y = A,C,G,T} q_{x_Y} P(x_1 | x_Y, t_1) P(x_2 | x_Y, t_2) \left[ \sum_{x_Z = A,C,G,T} P(x_Z | x_Y, t_5) \times P(x_3 | x_Z, t_3) \times P(x_4 | x_Z, t_4) \right]$$

$\ln \mathcal{L} = \sum_i \ln \mathcal{L}_i$ ] over all positions, total likelihood

$i$ → sum over all sequence positions.

node Z would be the root, same result would have been obtained.

(7) for a given topology T, we need to find branch lengths, + possible parameters of specific to evolutionary model (

$\alpha$ in JC)

— take partial derivatives of $\ln \mathcal{L}$ w.r.t. individual parameters & do function minimization

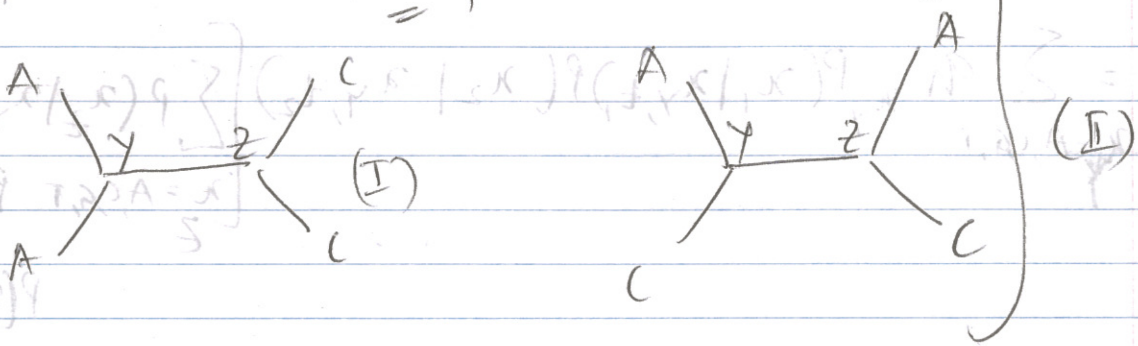
⑧ a search of tree topologies → Calculate $\alpha$ for each →

e.g.

4 seq. / each has one base; A, A, C, C

branch length same, JC model

2 possible trees | one parameter = branch length




$q_A = q_C = q_G = q_T = 1/4$

$E = \frac{1}{4}(1 + 3e^{-4\alpha t})$ prob of a branch of length $3\alpha t$ between nodes with identical bases

$D = \frac{1}{4}(1 - e^{-4\alpha t})$ with different bases

$(Y, Z) = (A, C)$ ⇒ 4 branches have the same at each end ⇒ E

1 branch different D

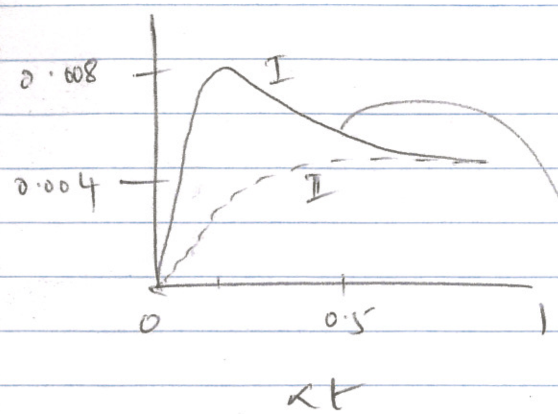
$\alpha_1 = \frac{1}{4} E^4 D$

$\alpha_I = \frac{1}{4}(E^4 D + 2E^3 D^2 + 4E^2 D^3 + 2ED^4 + 7D^5)$

(A, C) (A, A | C, C)

$$L_{II} = \frac{1}{4}\left( 2E^3D^2 + 2E^2D^3 + 10ED^4 + 2D^5 \right)$$

(A, A) | C, C.  A, C | C, A.




$\alpha t = 0.01$

$3\alpha t = 0.31$

most contribution from (A, C) first term

$\alpha t \to 1$ sufficiently long time no difference between I & II

all base choices for nodes Y & Z become equally likely