

Maximum likelihood methods

Try to reconstruct a phylogeny using an explicit model of evolution

To keep the model simple (tractable) often some assumptions are made.

- ▶ the independence of evolution between sites
- ▶ generations are also assumed to be discrete (non overlapping)

The log likelihood of having a given nucleotide at each site for a given topology is computed

Basic conditional probabilities

$$Pr(A \text{ and } B) = Pr(A)Pr(B|A)$$

The term on the right is called a conditional probability and is read as the probability of B given A.

In words, this definition states that the probability that A and B both occur is the probability that A occurs times the probability that B occurs given that A occurred.

Events are said to be independent if and only if

$$Pr(B|A) = Pr(B)$$

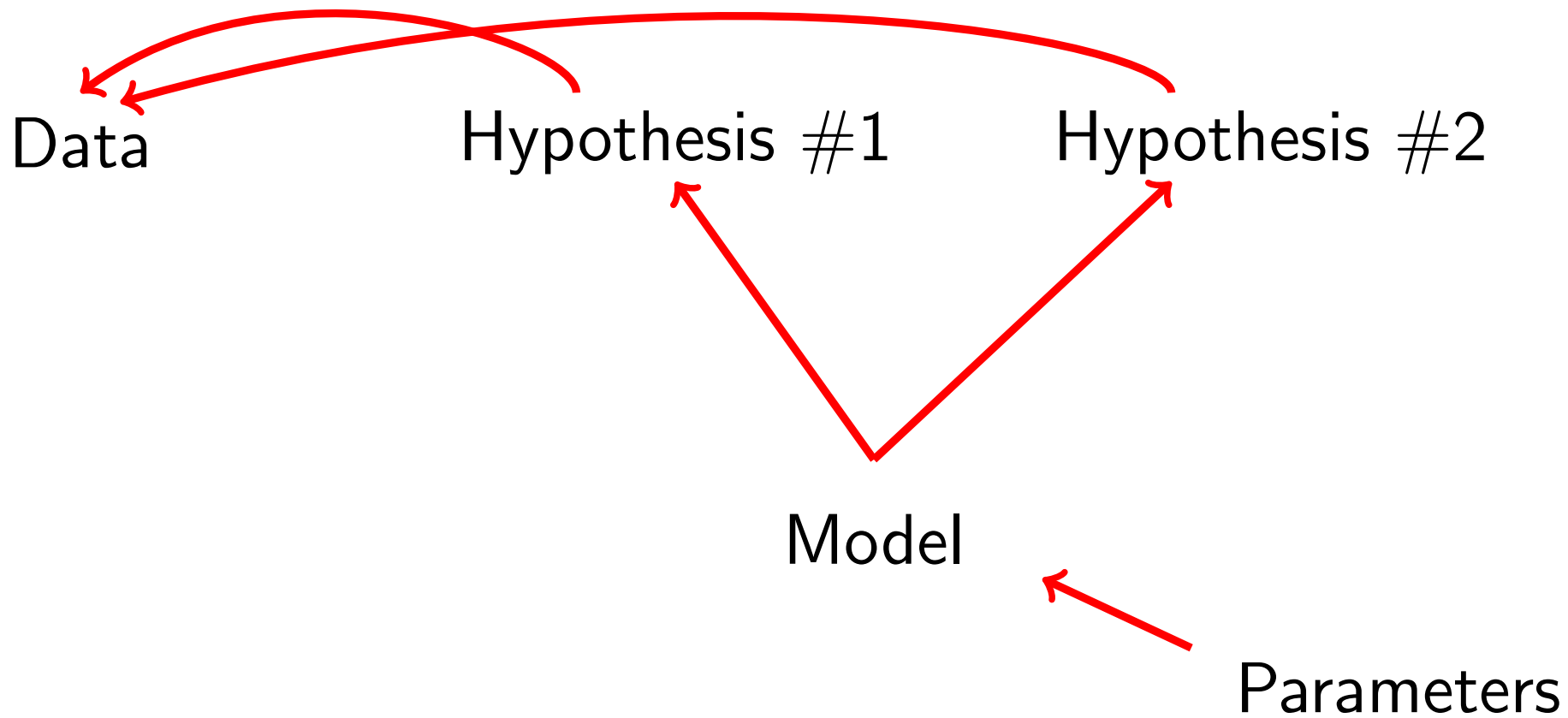
in this case, then

$$Pr(A \text{ and } B) = Pr(A)Pr(B)$$

If you have a collection of data, ... then the probability of the data might be conditional on a hypothesis. This probability is called a likelihood.

$$Pr(\text{Data}|\text{Hypothesis})$$

In our case “data” are sequences and “hypothesis” is a tree.



If you have data D , and some trees T_1, T_2, \dots, T_n then $Pr(D|T_1)$ is called the likelihood of the data given the tree T_1 .

Why the change in terminology? Why not “the probability of”?

Two reasons

1. The sum of likelihoods is not one – $\sum_i Pr(D|T_i) \neq 1$

If you have data D , and some trees T_1, T_2, \dots, T_n then $Pr(D|T_1)$ is called the likelihood of the data given the tree T_1 .

Why the change in terminology? Why not “the probability of”?

Two reasons

1. The sum of likelihoods is not one – $\sum_i Pr(D|T_i) \neq 1$
e.g. throw a single die and say it landed on ‘4’.
 - ▶ Hypothesis 1: The die is fair. Then $Pr('4'| \text{fair die}) = \frac{1}{6}$
 - ▶ Hypothesis 2: The die is loaded 10 to 1 in favour of ‘4’. Then $Pr('4'| 10 \text{ to } 1) = \frac{10}{11}$.

Sum is $\frac{1}{6} + \frac{10}{11} = 1.076$ and larger than one and hence, not a probability.

If you have data D , and some trees T_1, T_2, \dots, T_n then $Pr(D|T_1)$ is called the likelihood of the data given the tree T_1 .

Why the change in terminology? Why not “the probability of”?

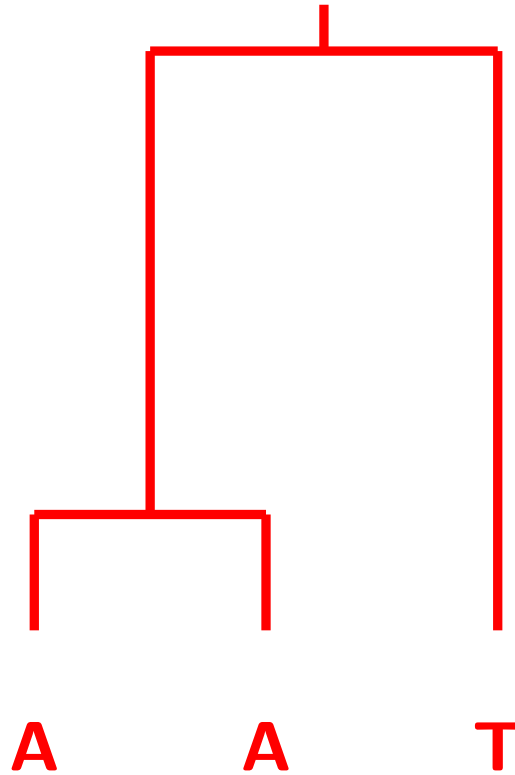
Two reasons

1. The sum of likelihoods is not one – $\sum_i Pr(D|T_i) \neq 1$
2. The second reason is more subtle. It emphasizes that the data is now to be seen as a function of the hypothesis/parameters (a likelihood function).

$$Pr(\text{Data} | H_i, \theta_1, \theta_2, \theta_3, \dots)$$

The values θ_i are parameters of the model, while H_i is the i^{th} hypothesis. Although sometimes the θ_i are of interest in themselves, they are often termed “nuisance parameters”.

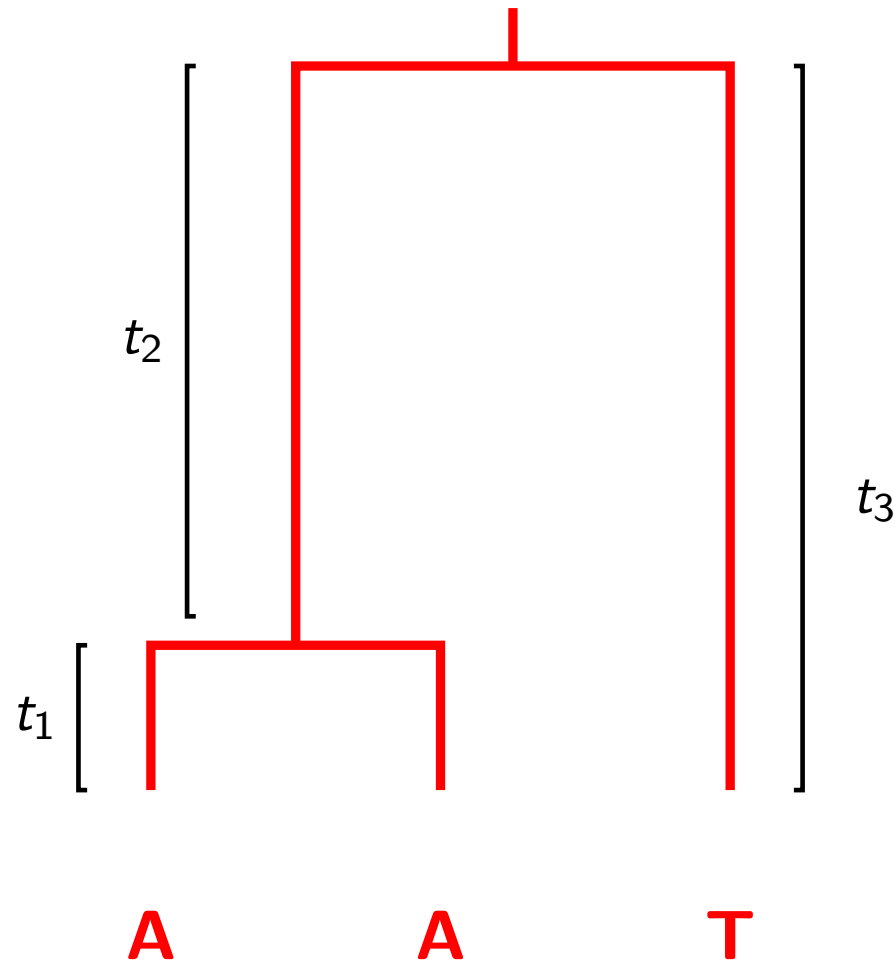
For example, θ might be the rate of evolution. This has to be considered, maximized and changed for each of the different hypotheses but if it is the tree that you are searching for, estimating the rate of evolution is necessary but time-consuming.



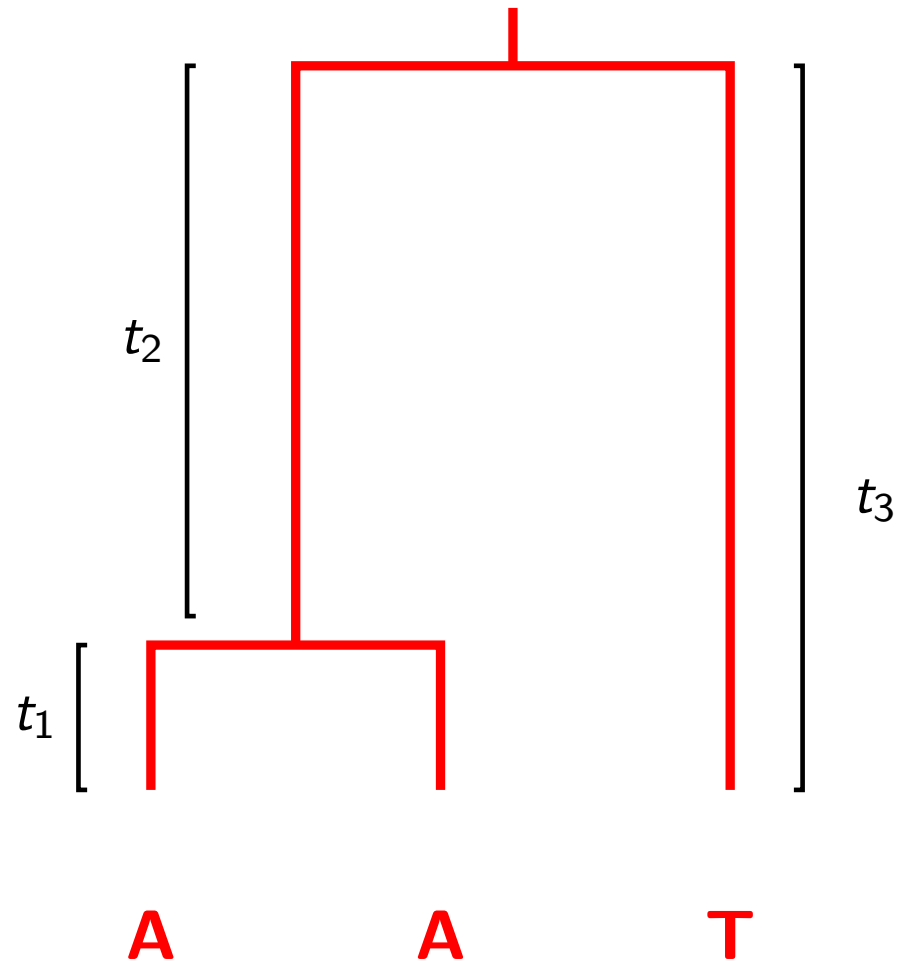
From previous we know

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

$$P_{ij} = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t}$$



Assume $P_{ii}^{t_1} = 0.9$ (that is the probability that the nucleotide stays the same is 0.9 over a length of time equal to t_1) then $P_{ij}^{t_1} = 0.033$ (that is the probability that the nucleotide changes is 0.03 over a length of time equal to t_1).

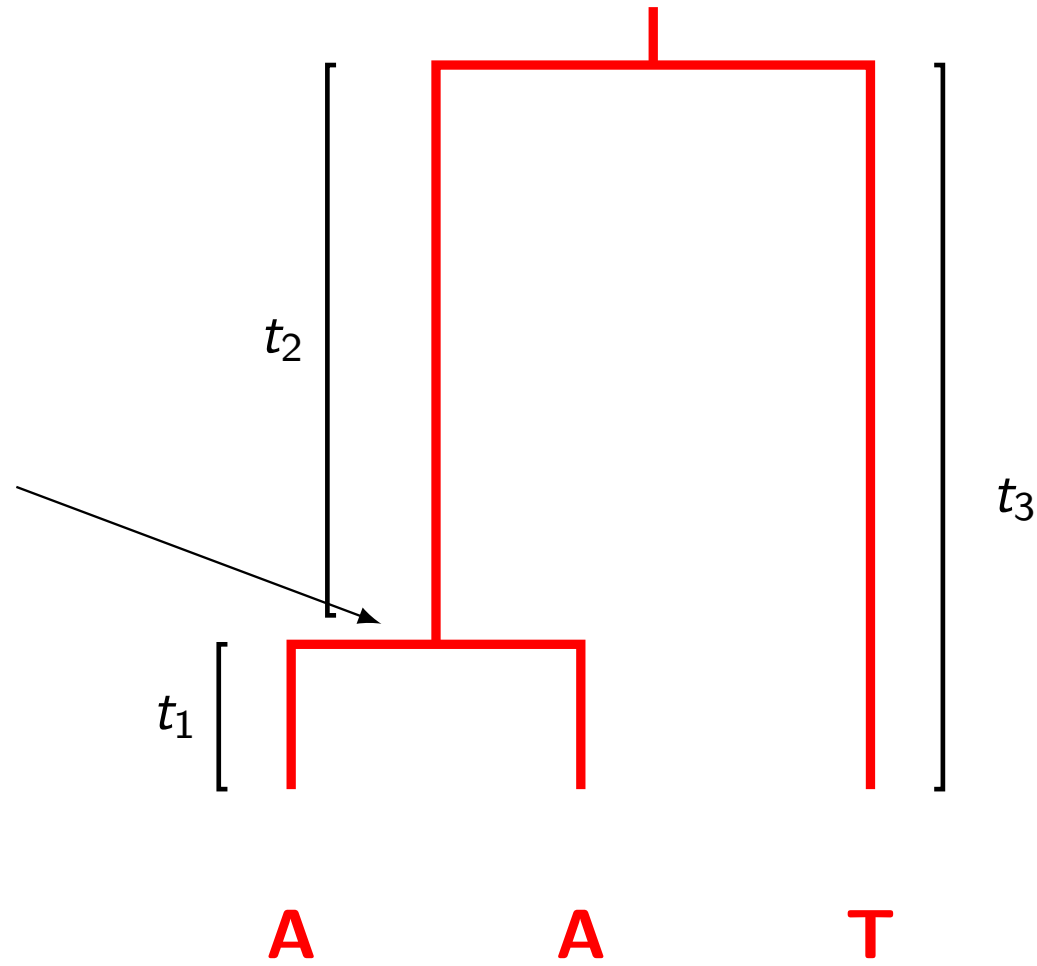


Assume $P_{ii}^{t_1} = 0.9$, $P_{ij}^{t_1} = 0.033$,

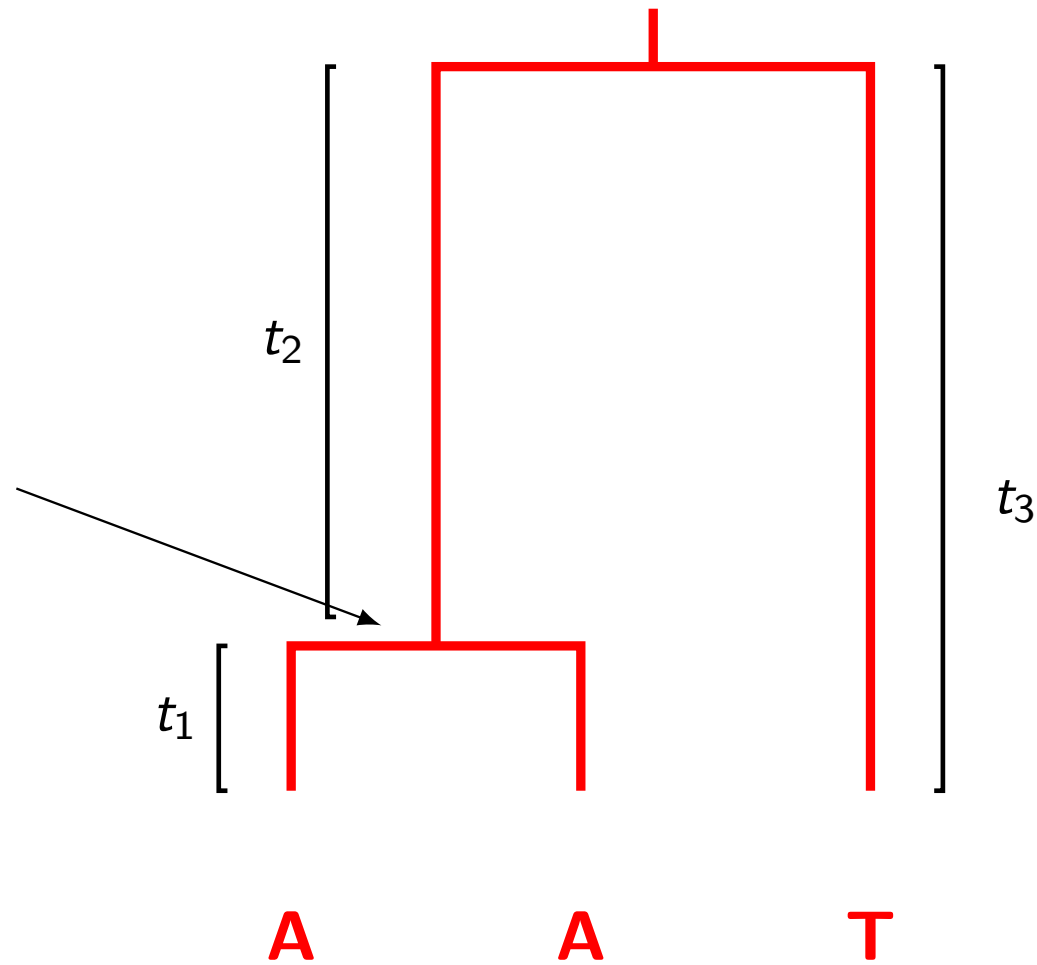
$$P_{ii}^{t_2} = 0.70, P_{ij}^{t_2} = 0.10,$$

$$P_{ii}^{t_3} = 0.55, P_{ij}^{t_3} = 0.15.$$

L(A)?
L(T)?
L(C)?
L(G)?

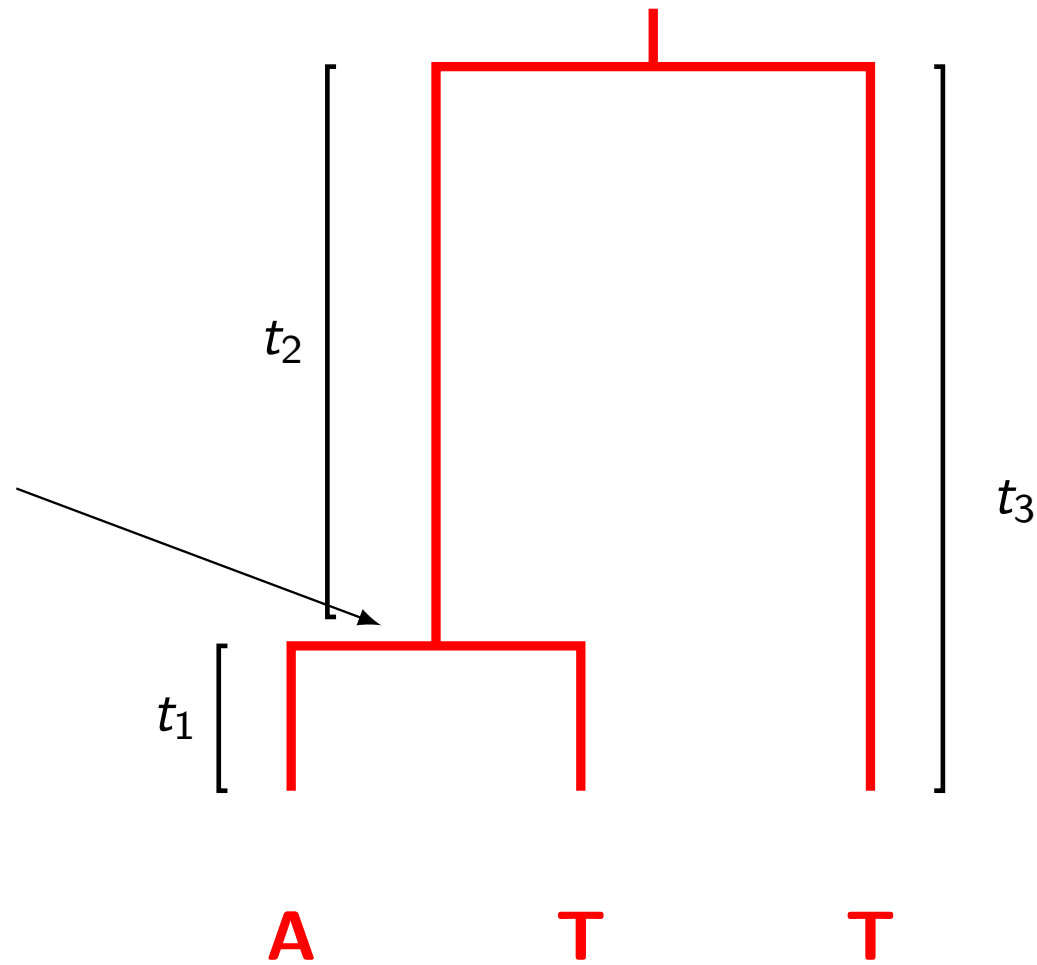


L(A)=0.81
L(T)=0.001
L(C)=0.001
L(G)=0.001



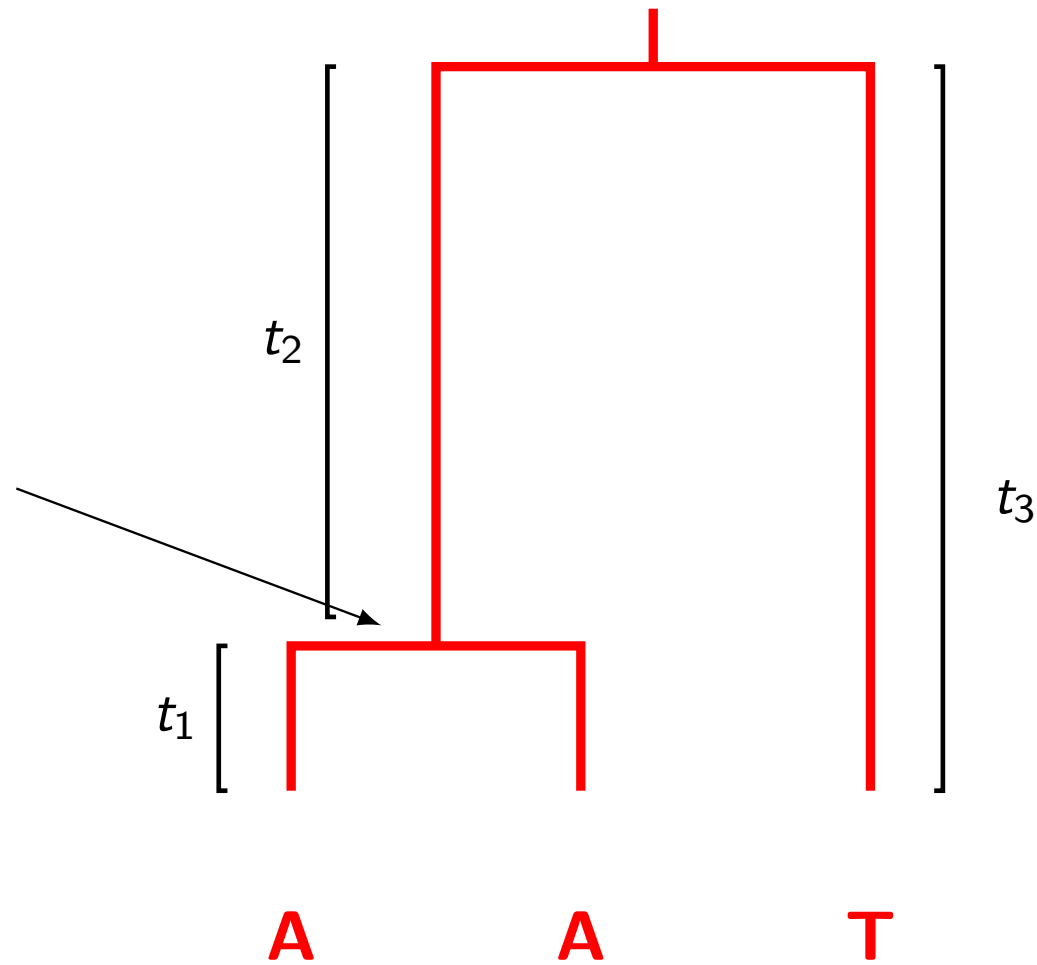
$$P_{ii}^{t_1} = 0.9, P_{ij}^{t_1} = 0.033$$

L(A)=0.03
L(T)=0.03
L(C)=0.001
L(G)=0.001

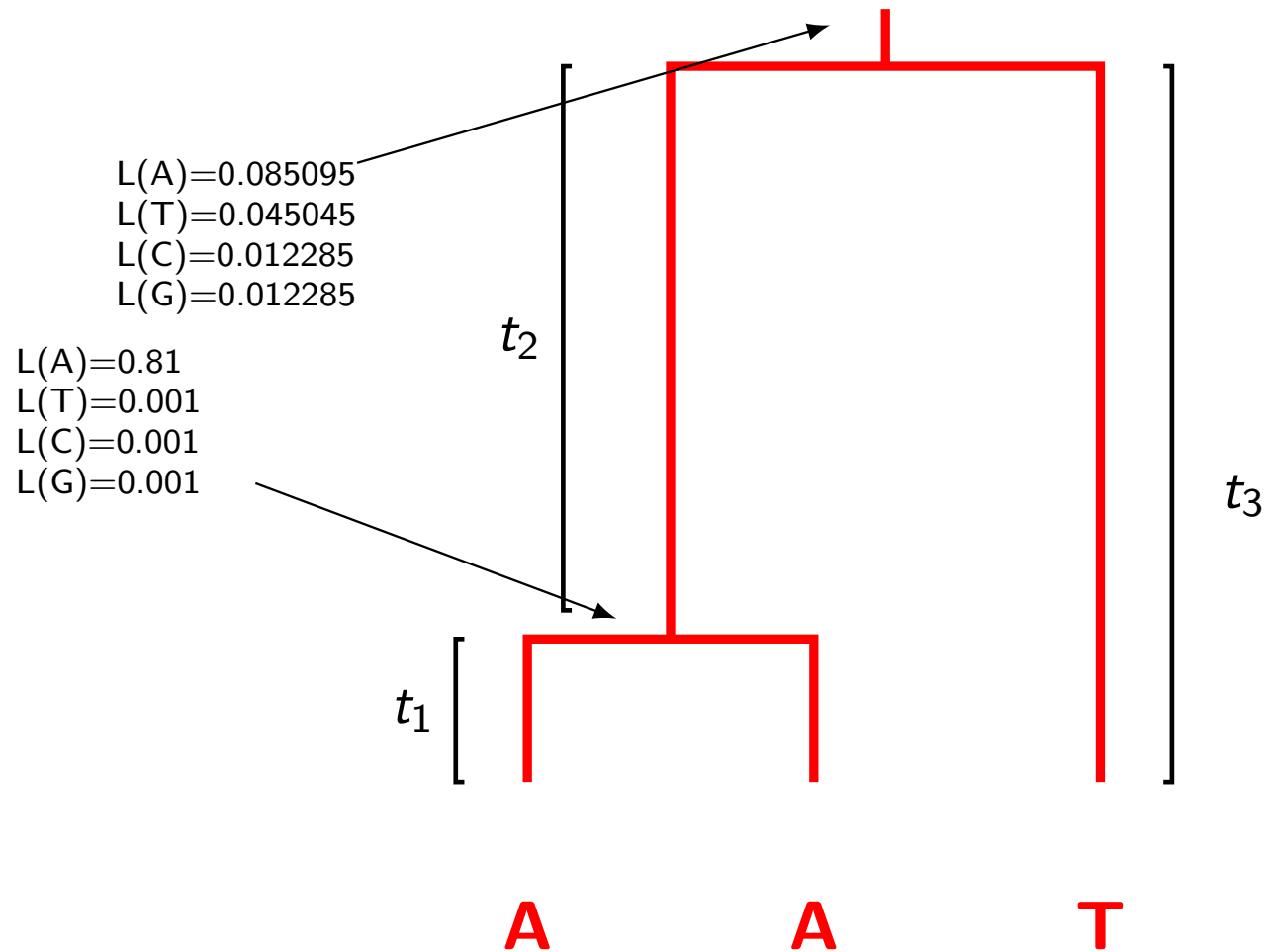


$$P_{ii}^{t_1} = 0.9, P_{ij}^{t_1} = 0.033$$

L(A)=0.81
L(T)=0.001
L(C)=0.001
L(G)=0.001



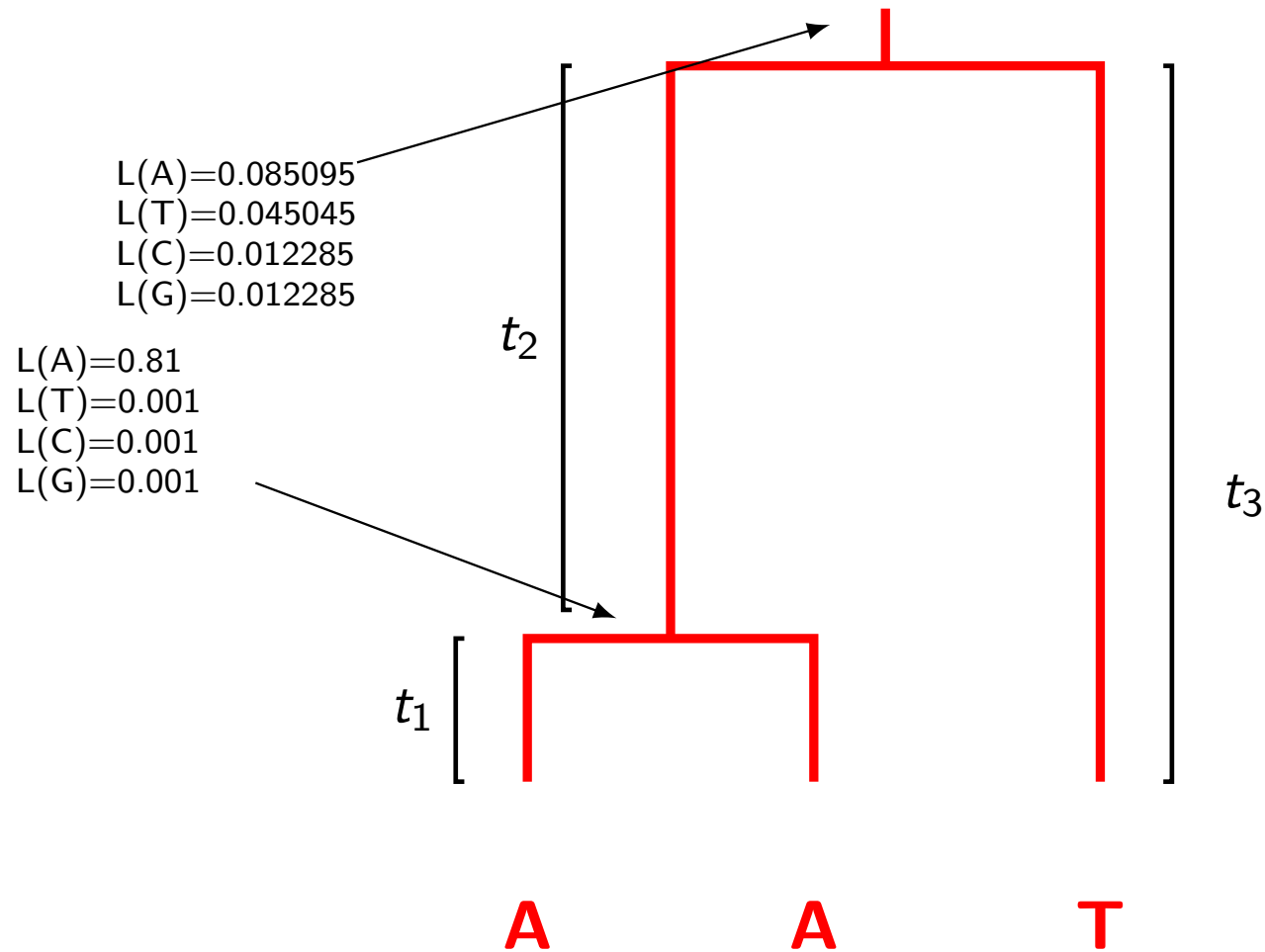
$$P_{ii}^{t_1} = 0.9, P_{ij}^{t_1} = 0.033$$



$$P_{ii}^{t_1} = 0.9, P_{ij}^{t_1} = 0.033,$$

$$P_{ii}^{t_2} = 0.70, P_{ij}^{t_2} = 0.10,$$

$$P_{ii}^{t_3} = 0.55, P_{ij}^{t_3} = 0.15.$$

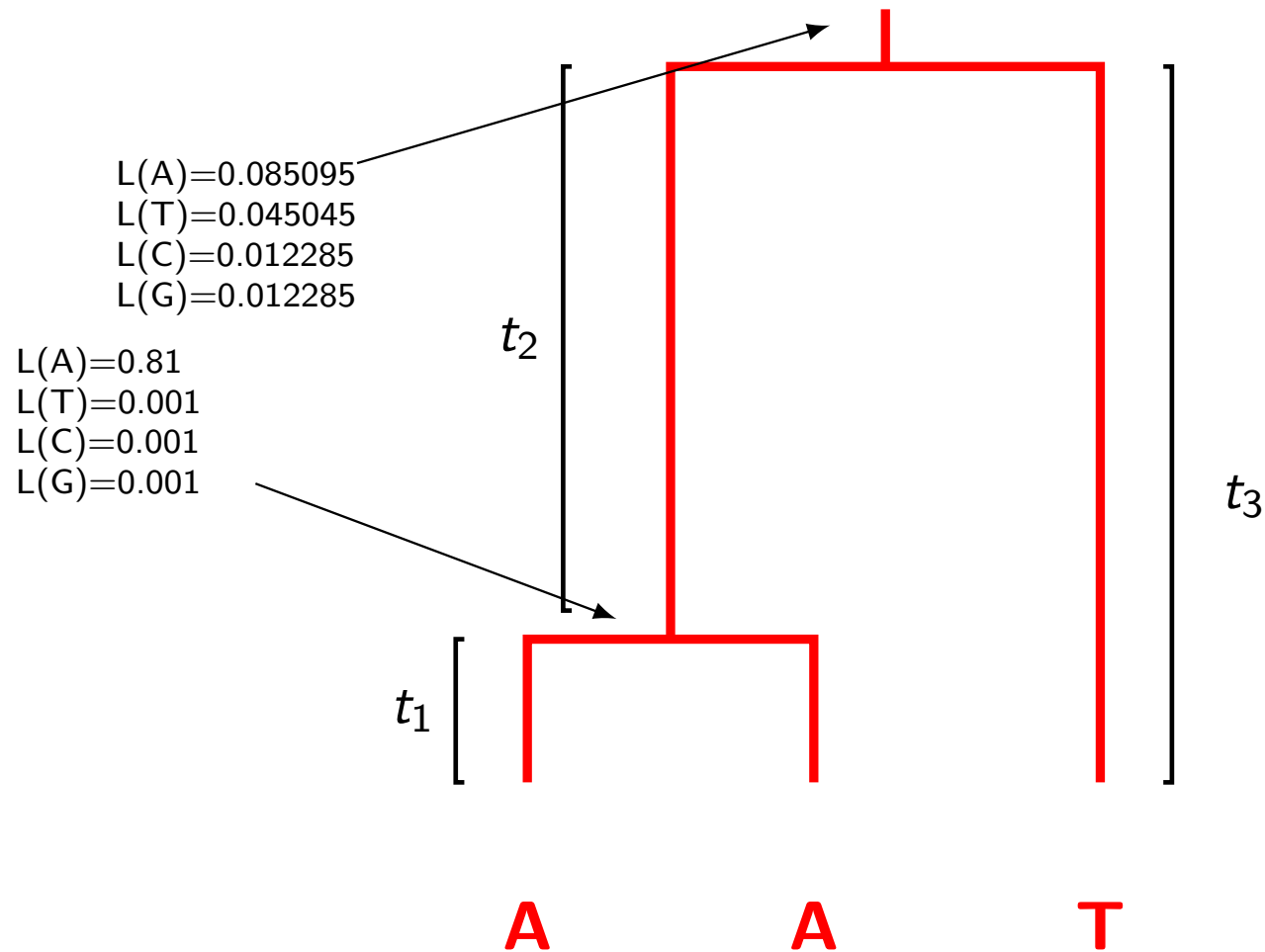


$$P_{ii}^{t_1} = 0.9, P_{ij}^{t_1} = 0.033,$$

$$P_{ii}^{t_2} = 0.70, P_{ij}^{t_2} = 0.10,$$

$$P_{ii}^{t_3} = 0.55, P_{ij}^{t_3} = 0.15.$$

$$\text{Overall likl} = L(A)*0.25 + L(T)*0.25 + \dots$$

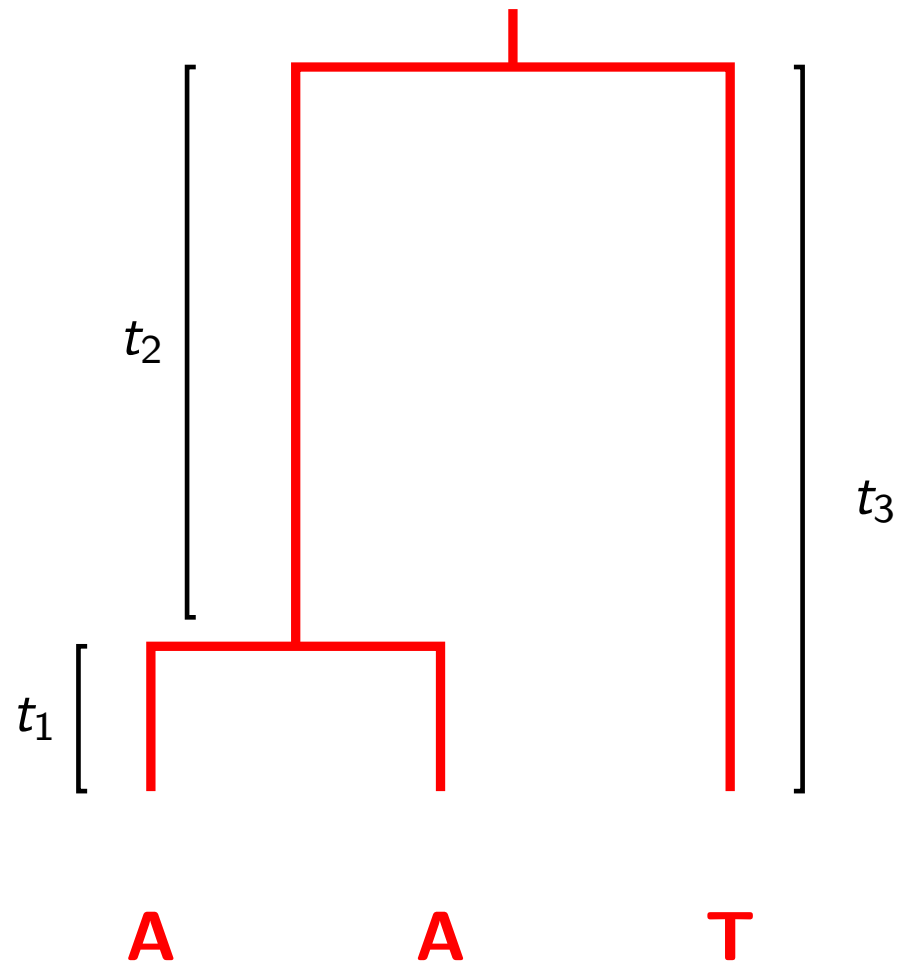


$$P_{ii}^{t_1} = 0.9, P_{ij}^{t_1} = 0.033,$$

$$P_{ii}^{t_2} = 0.70, P_{ij}^{t_2} = 0.10,$$

$$P_{ii}^{t_3} = 0.55, P_{ij}^{t_3} = 0.15.$$

$$\text{Overall likl} = 0.0386775$$



Overall likl = 0.0386775

Maximum likelihood methods

For each site along the sequence a likelihood will be calculated in this fashion.

Hence given a tree topology, T_j , given a rate of change (multiple rates of change), α , given the branch lengths in the tree, a likelihood is calculated.

$$L_i = Pr(\text{site}_i | T_j, \alpha, \text{branch}_1, \text{branch}_2, \dots)$$

Maximum likelihood methods

This is repeated for each site along the sequence.

$$L_1 = Pr(\text{site}_1 | T_i, \alpha, \text{branch}_1, \text{branch}_2, \dots)$$

$$L_2 = Pr(\text{site}_2 | T_i, \alpha, \text{branch}_1, \text{branch}_2, \dots)$$

$$L_3 = Pr(\text{site}_3 | T_i, \alpha, \text{branch}_1, \text{branch}_2, \dots)$$

— — —

$$L_n = Pr(\text{site}_n | T_i, \alpha, \text{branch}_1, \text{branch}_2, \dots)$$

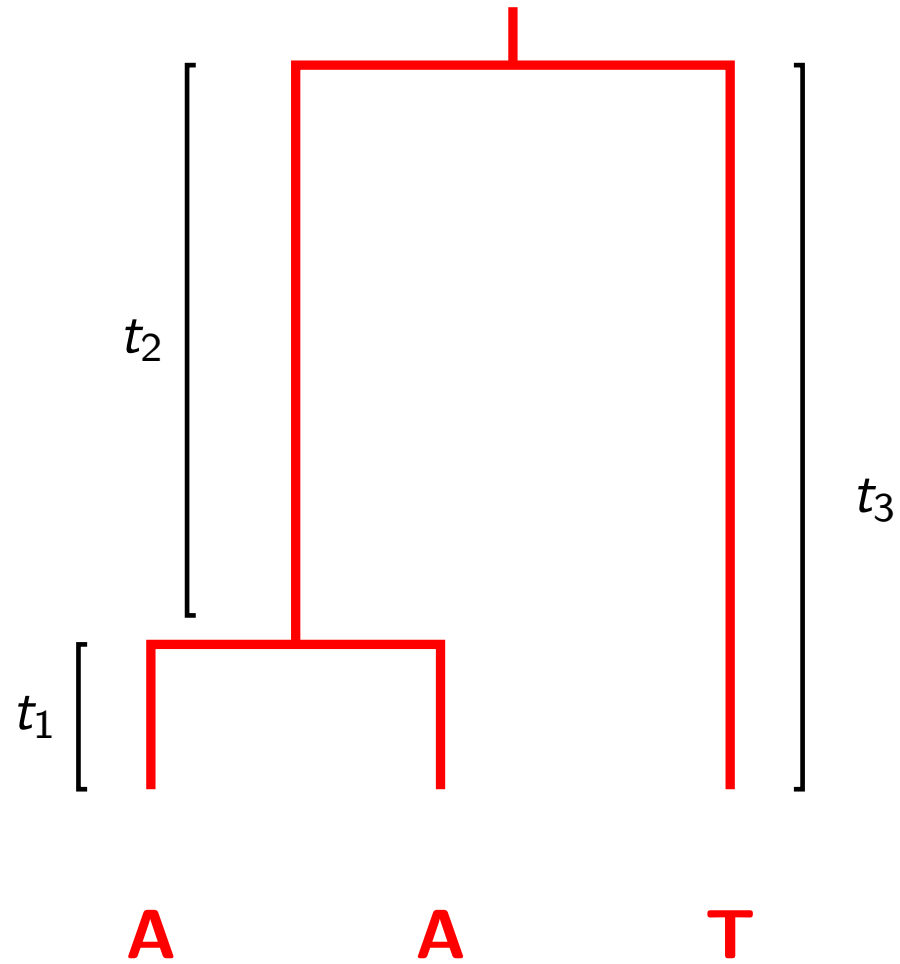
Maximum likelihood methods

To get the overall likelihood these likelihoods (probabilities) are considered independent so ...

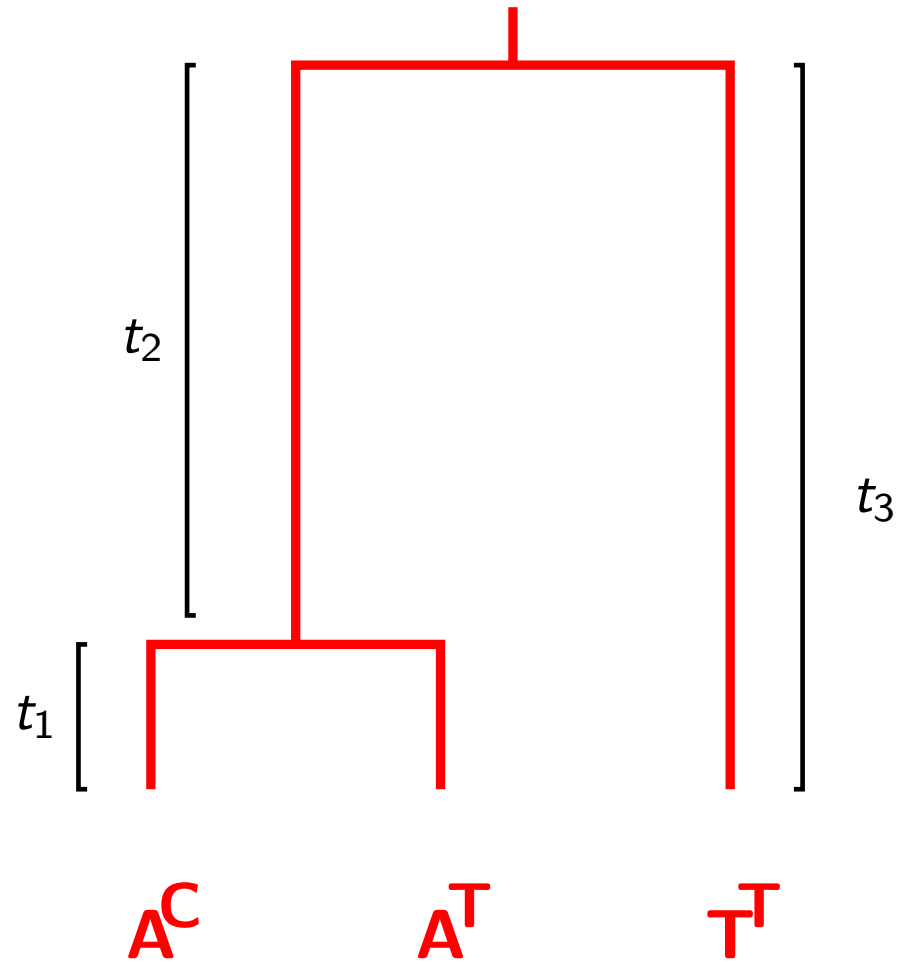
$$L = \prod_{i=1}^n L_i$$

or if log's are considered ...

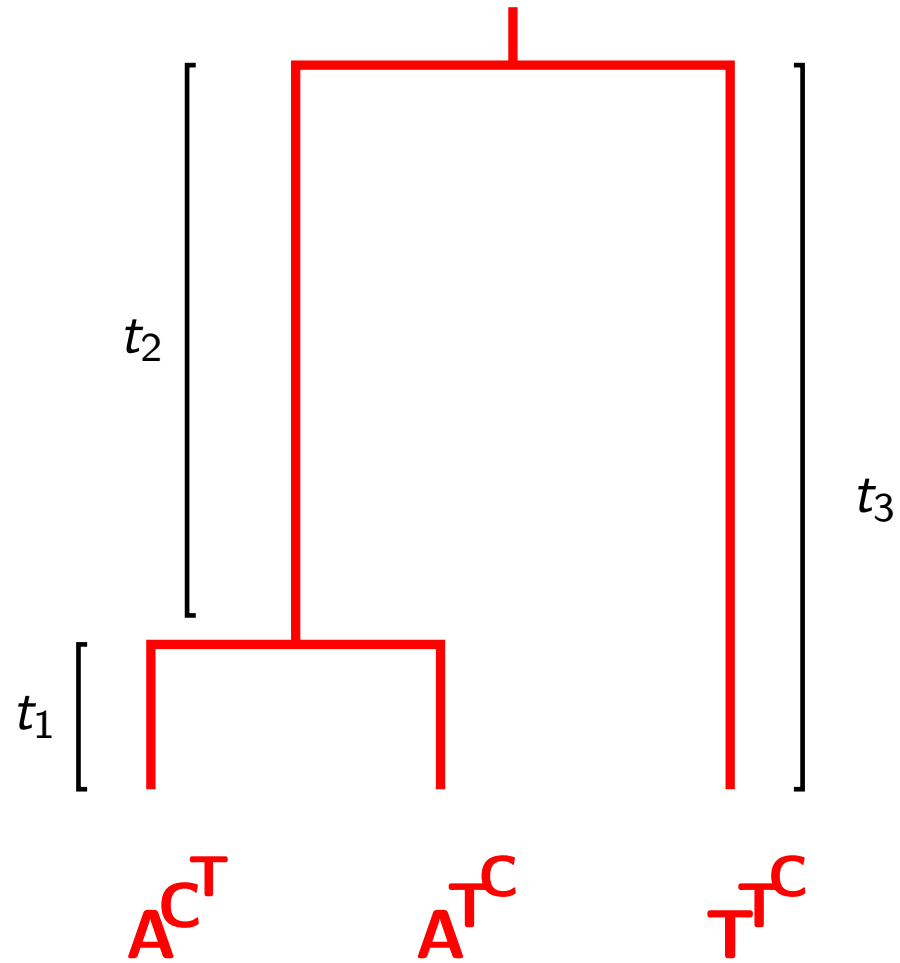
$$\ln(L) = \sum_{i=1}^n \ln(L_i)$$



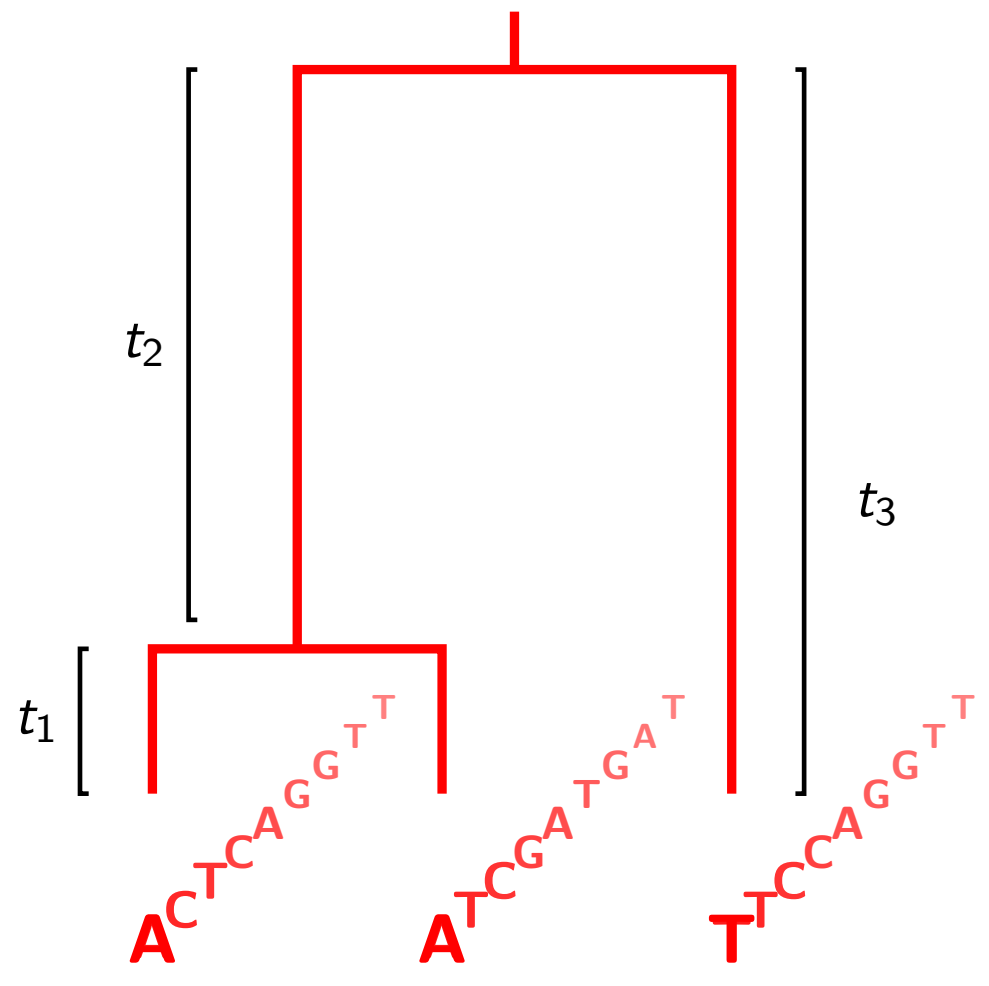
$$L = L_1$$



$$L = L_1 * L_2$$



$$L = L_1 * L_2 * L_3$$



$$L = \prod_i L_i$$

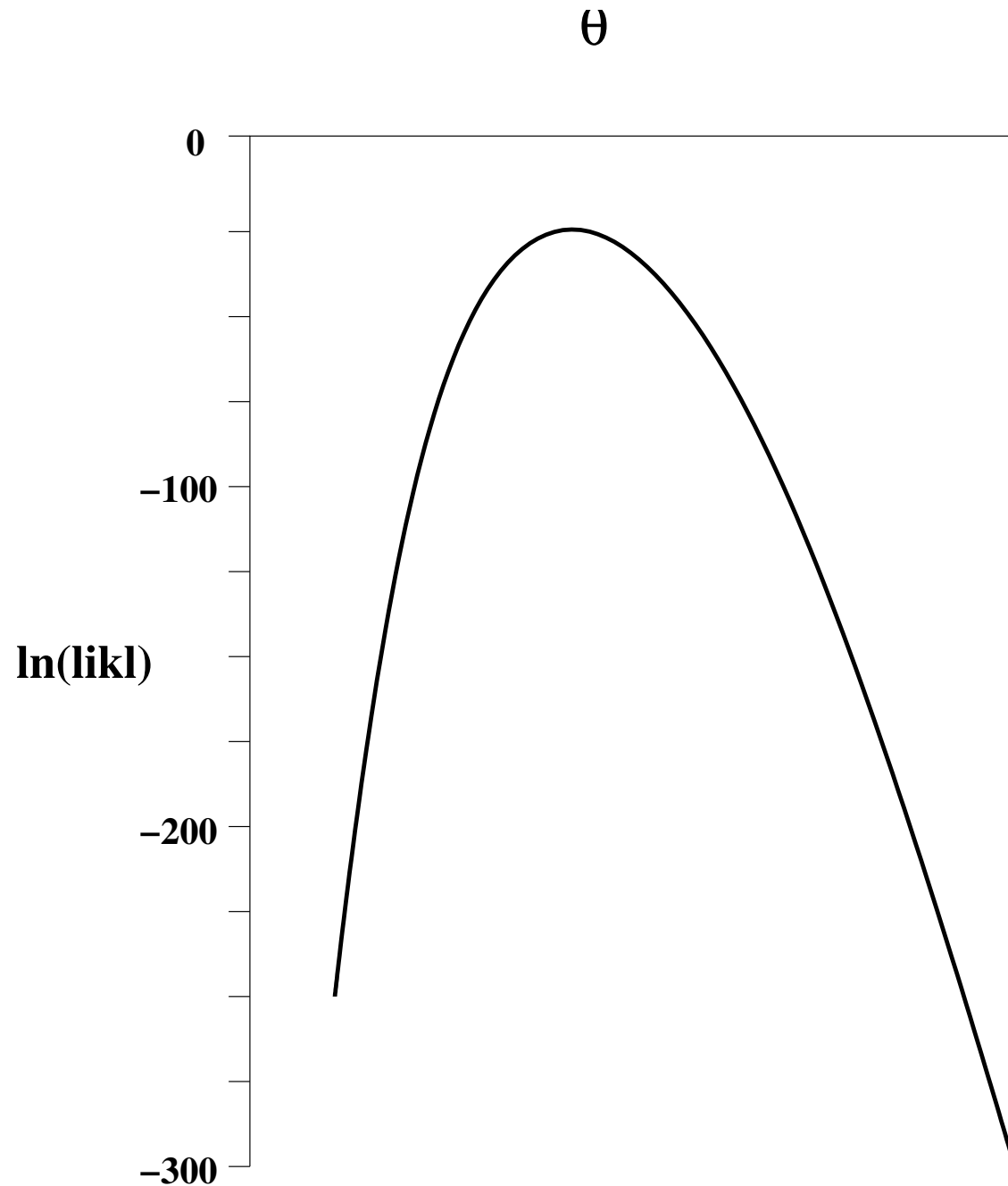
Maximum likelihood methods

Remember that the length of all branches in the tree will be computed. Each individual branch length must be altered to find the one branch length that maximizes the likelihood.

Remember that all of the mutation rates, α (for every type of mutation considered) has to be calculated and the value that maximizes the likelihood found.

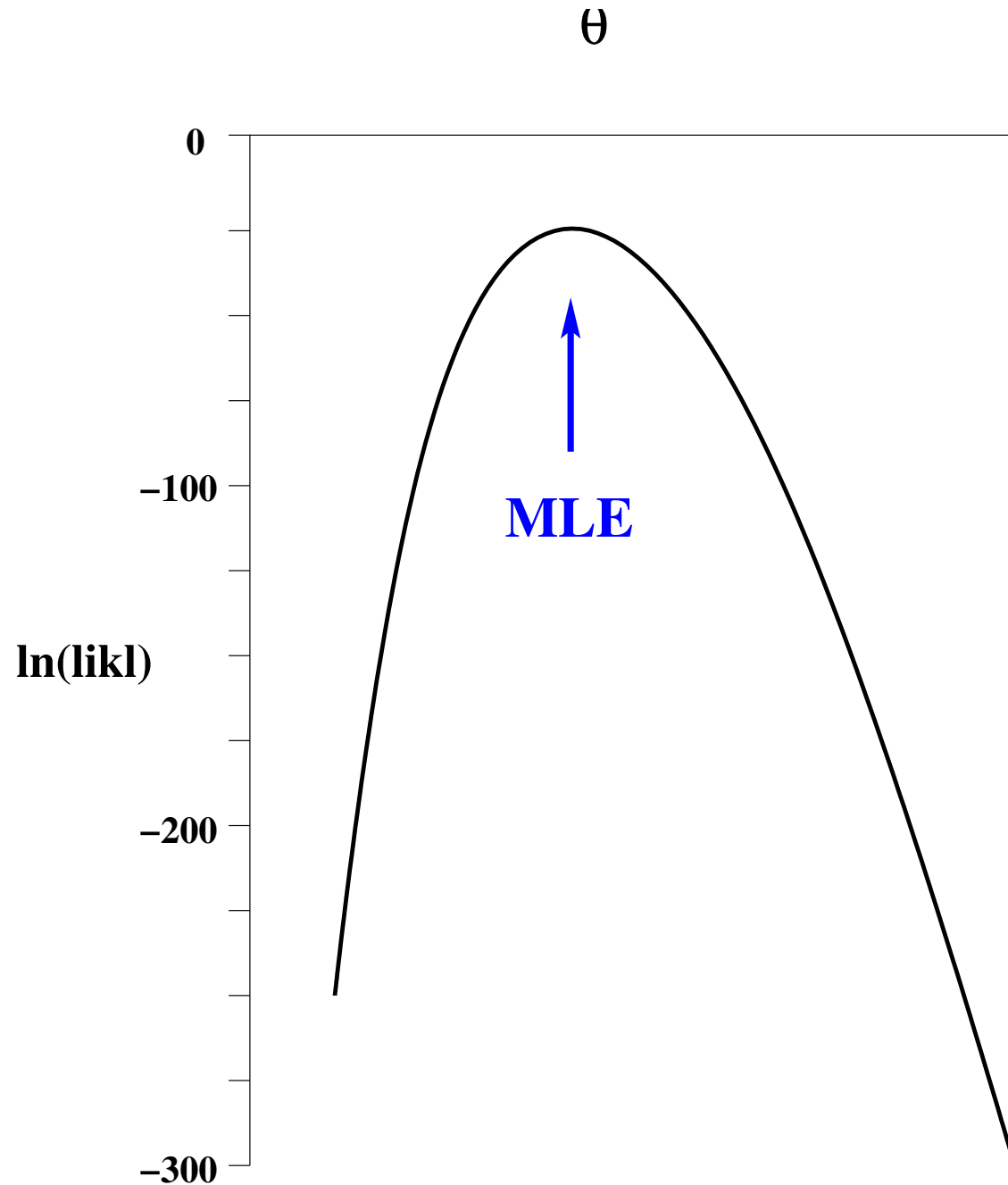
This procedure is repeated on all the possible topologies, the tree with the maximal likelihood is found.

Properties



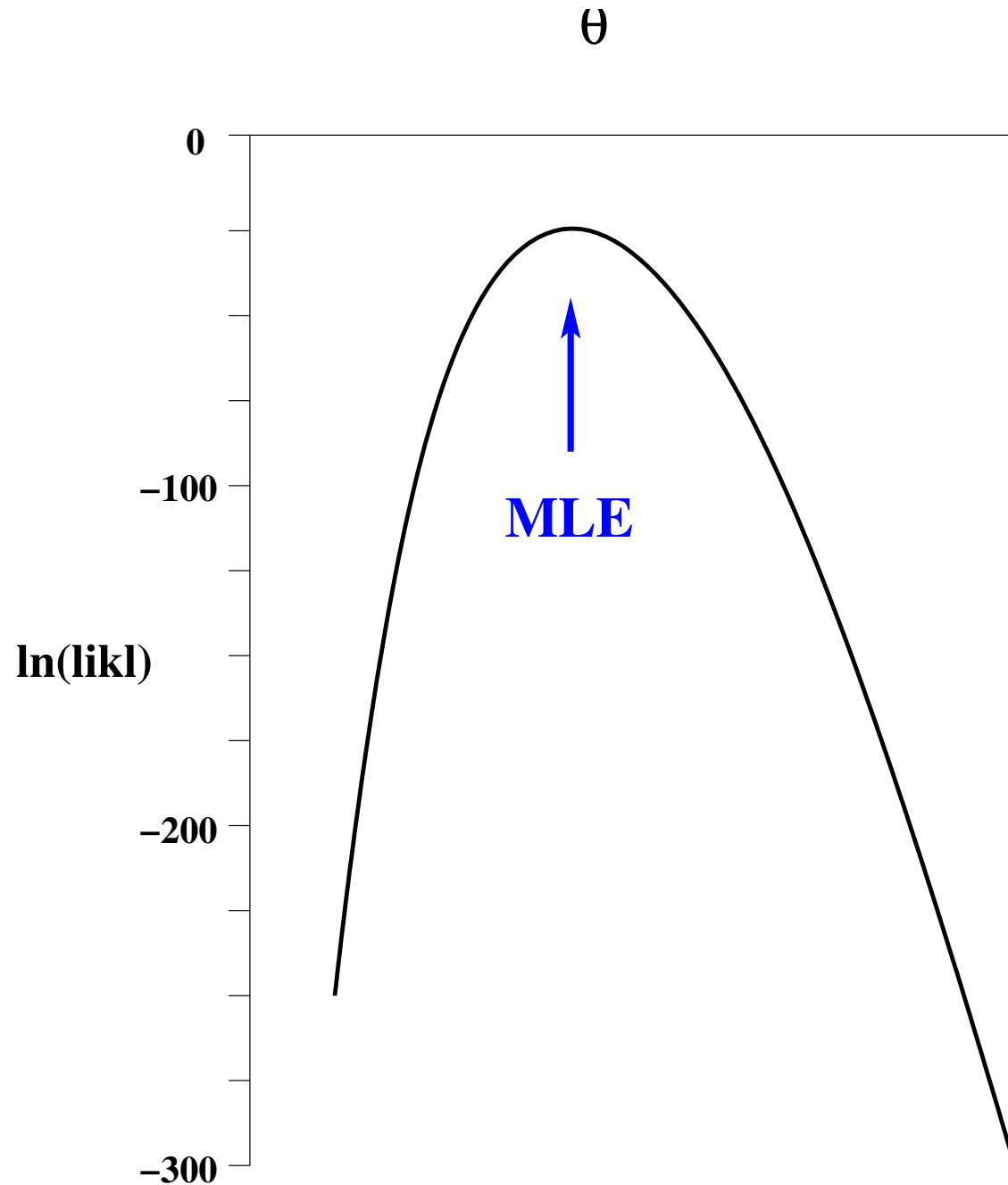
- ▶ The likelihood as a function of the parameter θ .
- ▶ The log likelihood will be negative (and often large).

Properties



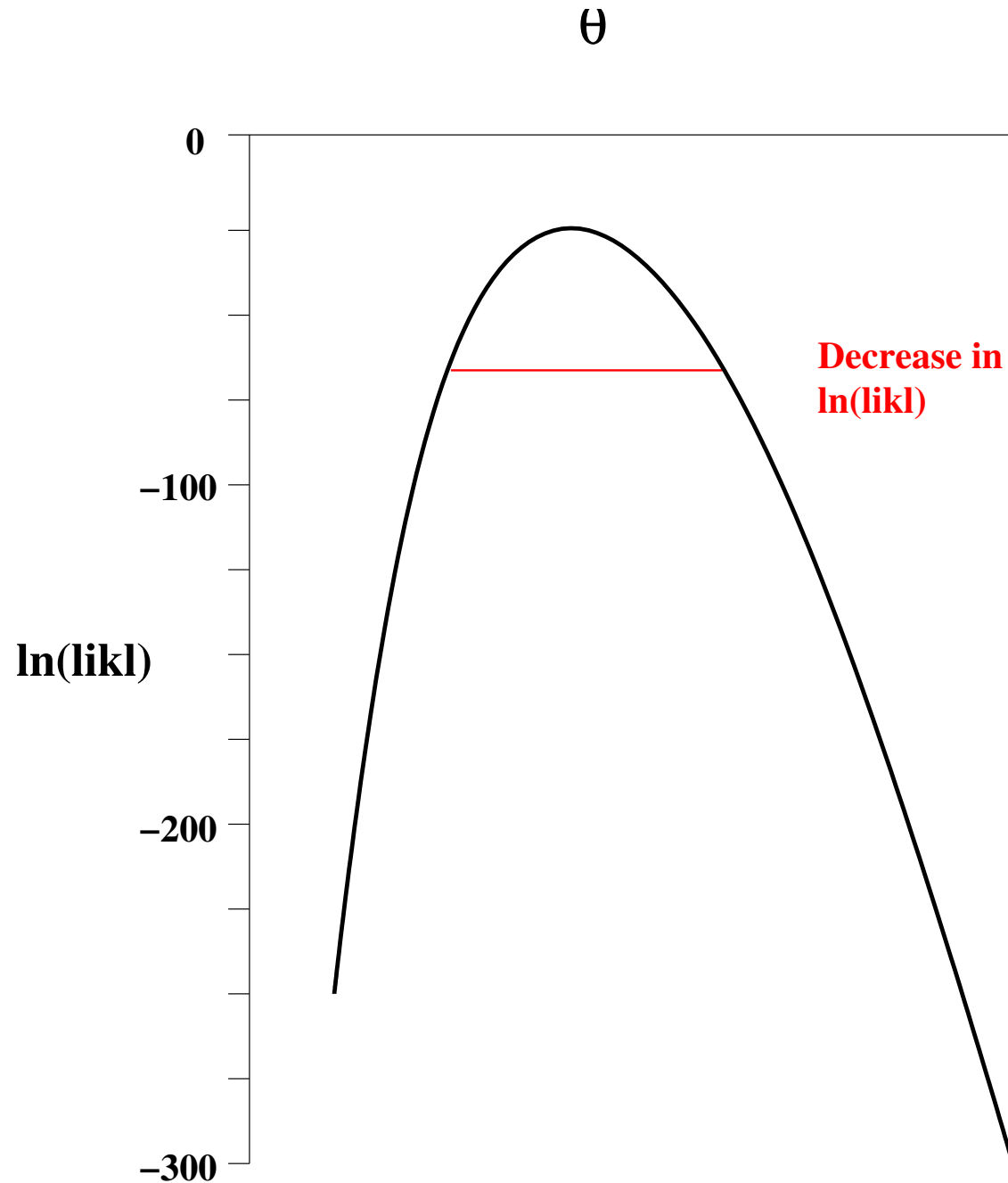
- ▶ The highest probability, the maximum likelihood is at the peak of the curve.
- ▶ The value at this point is called the MLE; the Maximum Likelihood Estimate of the parameter.

Properties



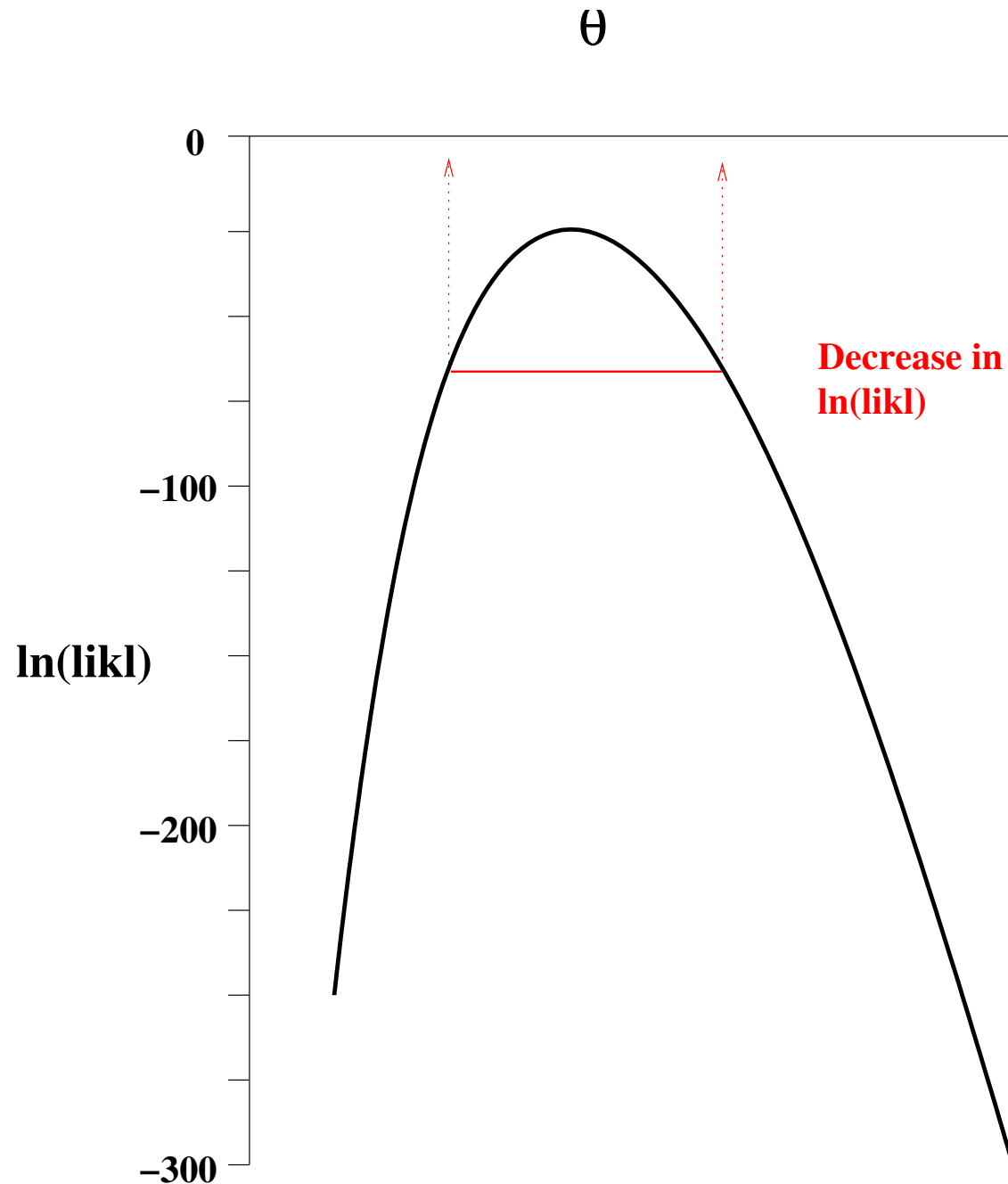
- ▶ The curvature gives an estimate of the variability
- ▶ The variance of the estimate at the point of the MLE is given by the inverse of the second derivative (curvature).

Properties



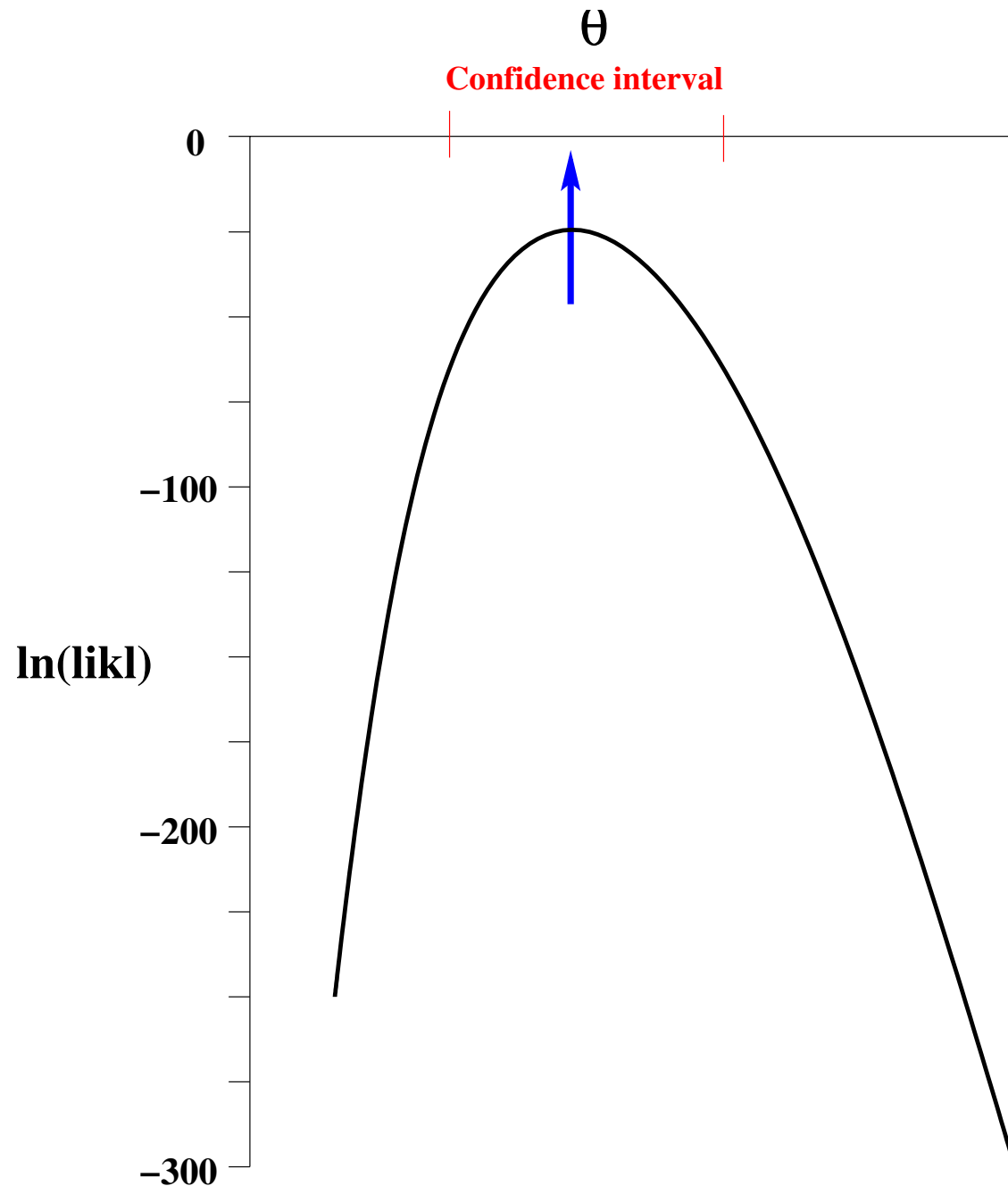
- ▶ A decrease in the likelihood measures a decrease in the assurance of obtaining the data given a hypothesis.

Properties



- ▶ A decrease in the likelihood measures a decrease in the assurance of obtaining the data given a hypothesis.
- ▶ The width of this space gives confidence intervals for the parameters.

Properties



- ▶ The width of this space gives confidence intervals for the parameters.
- ▶ The confidence intervals are not necessarily symmetrical.

Maximum likelihood methods

Now do this for dozens (perhaps hundreds?) of parameters all at once.

Find the maximum in this multi-dimensional space of parameters.

That one spot, where the maximum in all dimensions occurs, is the MLE and provides an estimate for all parameter values.

Properties of likelihoods

- ▶ Maximum likelihood estimators of parameters (MLE's) are those values of the parameters that maximize the likelihood function

Properties of likelihoods

- ▶ Maximum likelihood estimators of parameters (MLE's) are those values of the parameters that maximize the likelihood function
- ▶ In general, MLE's are consistent MVUE's (consistent minimum variance, unbiased estimators)

Properties of likelihoods

- ▶ Maximum likelihood estimators of parameters (MLE's) are those values of the parameters that maximize the likelihood function
- ▶ In general, MLE's are consistent MVUE's (consistent minimum variance, unbiased estimators)
- ▶ Variance of estimates can be obtained from the curvature of the likelihood function around the MLE (i.e. second derivative)

Properties of likelihoods

- ▶ Maximum likelihood estimators of parameters (MLE's) are those values of the parameters that maximize the likelihood function
- ▶ In general, MLE's are consistent MVUE's (consistent minimum variance, unbiased estimators)
- ▶ Variance of estimates can be obtained from the curvature of the likelihood function around the MLE (i.e. second derivative)
- ▶ Tests of nested hypotheses: Under general conditions the ratio of two likelihoods is Chi-square (χ^2) distributed with degrees of freedom equal to the difference in the number of parameters. Thus the calculation of the likelihood provides instant variance and hypothesis testing.

Properties of likelihoods

- ▶ Maximum likelihood estimators of parameters (MLE's) are those values of the parameters that maximize the likelihood function
- ▶ In general, MLE's are consistent MVUE's (consistent minimum variance, unbiased estimators)
- ▶ Variance of estimates can be obtained from the curvature of the likelihood function around the MLE (i.e. second derivative)
- ▶ Tests of nested hypotheses: Under general conditions the ratio of two likelihoods is Chi-square (χ^2) distributed with degrees of freedom equal to the difference in the number of parameters. Thus the calculation of the likelihood provides instant variance and hypothesis testing. (Unfortunately, the general conditions are sometimes not met in phylogenetics and an information criteria is often used instead)