

LIKELIHOOD IN MOLECULAR PHYLOGENETICS

Peter G. Foster
The Natural History Museum, London

July, 2001

Likelihood in molecular phylogenetics

- Why use likelihood?
- Simple likelihood calculations
- Choosing a model
- Hypothesis testing
- New research
- Practical likelihood

Why use likelihood?

- It takes into account branch lengths
 - Accurate branch lengths even if there are superimposed hits (*ie* more than one mutation at the same site)
 - Using Markov models does not so much “correct” for superimposed hits, rather it incorporates superimposed hits into the core of the process
- It is explicit
 - assumptions are stated, not hidden
- You can make the model fit the data
- It is efficient and powerful
 - it uses all the data

You can make the model fit the data

If the data ...

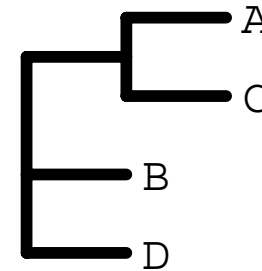
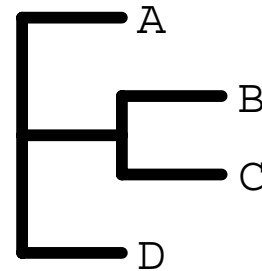
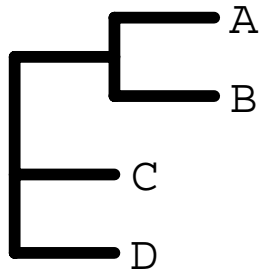
- have an unusual composition
- have a transition/transversion ratio different from 1
- have both quickly and slowly evolving sites

...you can use that information in your model.

It uses all the data

A acgcaa
B acataa
C atgtca
D gcgtta

- There are no parsimony informative sites in these data
- Under un-weighted parsimony, all three possible trees have equal length



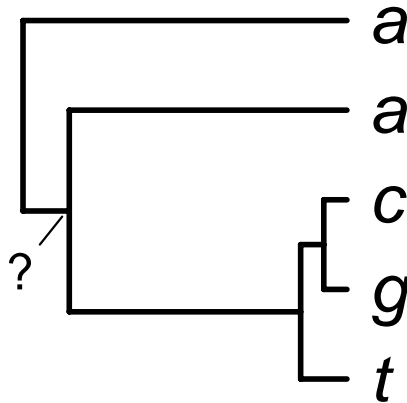
It uses all the data

```
A  acgcaa  
B  acataa  
C  atgtca  
D  gcgtta
```

- Although there are no parsimony informative sites, there appears to have been several evolutionary events, which should provide useful phylogenetic information.
- It appears that transitions are more common than transversions:
- The constant site provides useful information regarding the tendency of the **a** to stay the same.
- If we use this information, then one tree is more optimal than the other two.

What is the ancestral state?

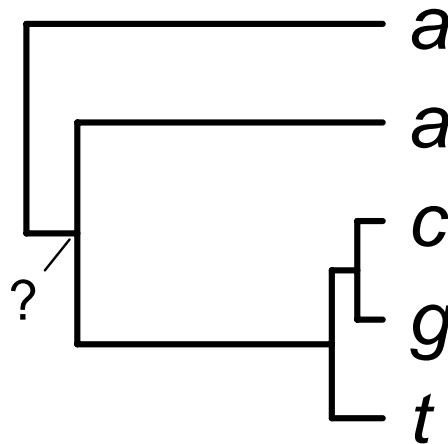
Consider one site on this tree, with these character states—



- Ancestral state **a** is most parsimonious
- Wide character state variation together with short branches tells us that this is a fast site
—so we should expect a large amount of change over the long branches.
- Likelihood is equivocal about the ancestral state (it could be anything)

Branch lengths under parsimony and likelihood

- Parsimony considers that you would have the same expectation that a character would change along both long and short branches.
- Likelihood and distance methods, using models, consider that change is more probable along long branches than along short branches.



Long branch attraction by parsimony

Since outgroup taxa often have long branches...

- perhaps ingroup taxa go basal
- perhaps outgroup taxa go in the ingroup

Model-based methods do not suffer from LBA, but only if the model fits the data well.

Molecules do not evolve like morphological characters

- Molecular sequences appear to evolve mostly by random change, with a small amount of selection.
- This behavior can be described well by stochastic models which incorporate among-site rate variation.
- This allows us to use probabilistic methods in our analyses
 - and puts our analyses on a sound statistical footing.

Likelihood is appropriate for data generated by a random process

These data were probably *not* generated by a random process

000000000000

010301001000

222022100100

131130010011

- There are no constant sites.
- There is an obvious ancestral taxon.
- Some characters are binary, some are multi-state

Likelihood

In general...

The likelihood is the probability of the data given the model.

In phylogenetics, we can say (loosely, \pm) that the tree is part of the model

The likelihood is the probability of the data given the tree and the model.

The likelihood supplies a natural order of preferences among the possibilities under consideration.

-Fisher, 1956

Flip a coin— get a “head”

What is the likelihood of that data?

- The likelihood depends on the model
- If you think its a fair coin, the likelihood of the data is 0.5
- If you think it is a two-headed coin, the likelihood of the data is 1.0
- ...So the model that you use can have a big effect on the likelihood

Likelihood calculations

- In molecular phylogenetics, the data are an alignment of sequences
- Each site has a likelihood
 - this differs depending on the model and tree
- The total likelihood is the product of the site likelihoods
 - or the sum of the log of the site likelihoods
- The maximum likelihood tree is the tree topology that gives the highest likelihood under the given model.
- We use reversible models, so the position of the root does not matter.

Reference 1

P. G. Foster 2001. “The Idiot’s Guide to the Zen of Likelihood in a Nutshell in Seven Days for Dummies, Unleashed” *Unpublished manuscript*

- Elementary likelihood calculations and definitions
- Probability and rate matrices
- Finding the maximum likelihood branch length
- Calculating likelihood values on a tree
- Checking that PAUP* gets the correct likelihood values

Choosing a model

- Don't “assume” a model
- Rather, find a model that fits your data.

Models are described in terms of...

- tendency of one base to change to another
- composition
- site-to-site rate variation

Models often have “free” parameters. These can be fixed to a reasonable value, or estimated by ML.

Tendency of one base to change to another

- This can be described by a rate matrix
- The most complex in paup is the GTR, general time-reversible
- Other models are simplifications of this
 - HKY, F81, K2P, JC, etc ...

GTR: General time-reversible model

$$\mathbf{R} = \begin{bmatrix} - & a & b & c \\ a & - & d & e \\ b & d & - & f \\ c & e & f & - \end{bmatrix}$$

- Symmetrical, so time-reversible
- There are 6 substitution types (`lset nst=6`), so 5 free parameters
- You can restrict these using the `rclass` subcommand in `lset`
 - eg* `lset rclass=(a b c c b a)` to make a subset with only 3 substitution types
 - The program `modeltest` uses `rclass` a lot, see the file `modelblock3`

Base frequencies (composition)

- equal
- specified
- empirical
 - often a good approximation to ML-estimated, and much faster
- estimated by ML
- For DNA, there are 4 compositions, so 3 free parameters

Among-site rate heterogeneity

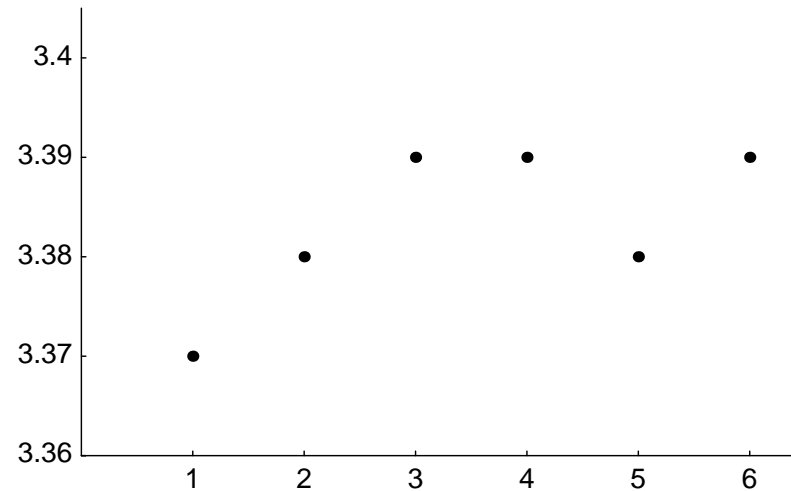
- pInvar
- gamma-distributed variable sites
 - has an average rate of 1.0
 - shape can change greatly with only one parameter (α , `shape` in `paup`)
 - approximated with a discrete gamma distribution with `nCat` divisions
- pInvar + gamma
- site-specific
 - good for codons

Parameters

- Models differ in their free, *ie* adjustable, parameters
- More parameters are often necessary to better approximate the reality of evolution
- The more free parameters, the better the fit (higher the likelihood) of the model to the data. (Good!)
- The more free parameters, the higher the variance, and the less power to discriminate among competing hypotheses. (Bad!)
- We do not want to “over-fit” the model to the data

Diminishing returns?

What is the best way to fit a line (a model) through these points?

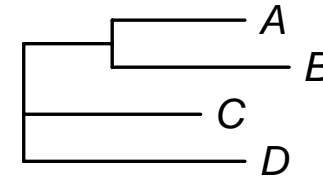


- A linear fit would be fair
- A quadratic fit would be better
- It would be possible for a higher-order fit to go thru every point
 - but you would only be fitting noise, *ie* “over-fitting” the data

How to tell if adding (or removing) a certain parameter is a good idea?

- Use statistics!
- The null hypothesis is that the presence or absence of the parameter makes no difference
- In order to assess significance you need a null distribution

Is it worth adding a parameter? —An example



We have some DNA data, and this tree.

- Evaluate with JC (Jukes-Cantor) model: log likelihood is -1008.587
- Evaluate with K2P (Kimura 2-parameter) model: log likelihood is -1008.268
 - The K2P model has one more parameter than the JC model, the **tRatio**.
- We got a better log likelihood with the extra parameter, by 0.319
- Is the extra parameter worth adding?

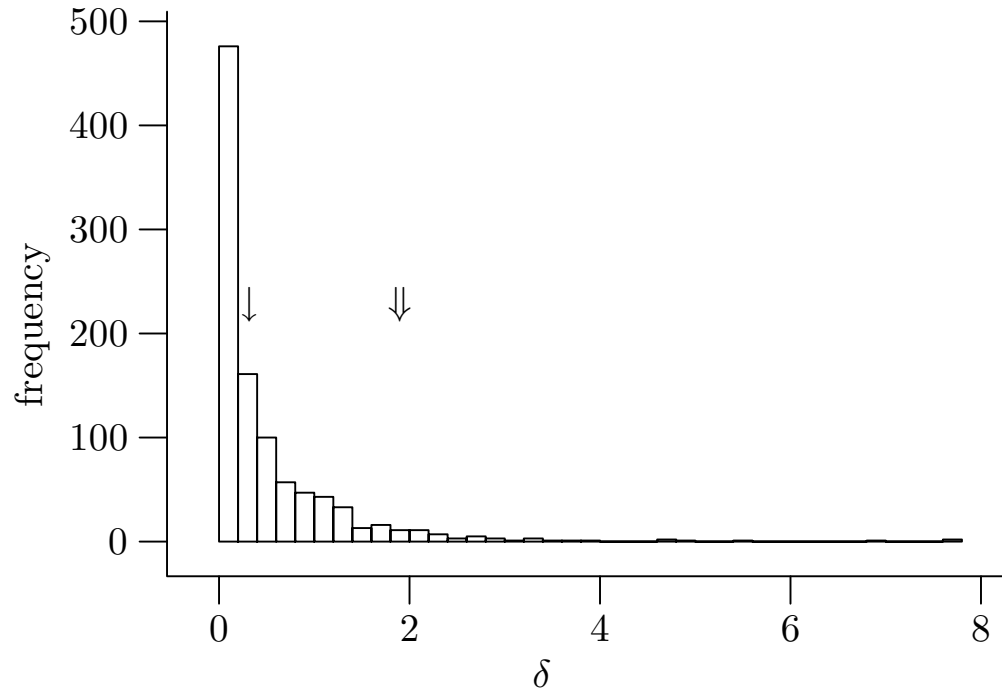
Stats test

- Null hypothesis (generally): the extra parameter does not make any difference
- Null hypothesis (specifically): the tree and the JC model
- We need to know how much of an improvement in likelihood we can expect *due to noise alone* when we add the parameter

Stats test

- Null hypothesis: the tree and the JC model
- We need a null distribution, which we can get by simulating fake data many times under the null hypothesis
- Evaluate the likelihood of each simulated data set with both the JC and the K2P models
- Keep the log likelihood differences—they are the null distribution

Stats test



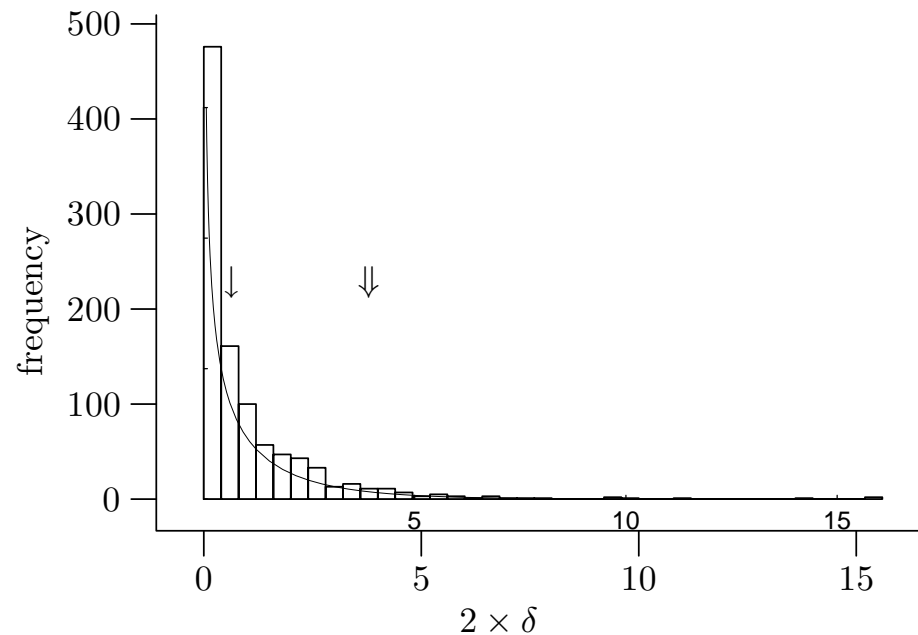
We have generated many true null hypothesis data sets and evaluated them under the JC model and the K2P model. 95% of the differences are under 2. The statistic for our original data set was 0.319, and so it is *not* significant. In this case it is *not* worthwhile to add the extra parameter ([tRatio](#)).

You can use χ^2 approximation to assess significance of the effect of parameters

- Double the difference in log likelihoods is approximately χ^2 distributed
- The 95% point
 - by the previous simulation: 3.79
 - by $\chi^2_{df=1}$: 3.84

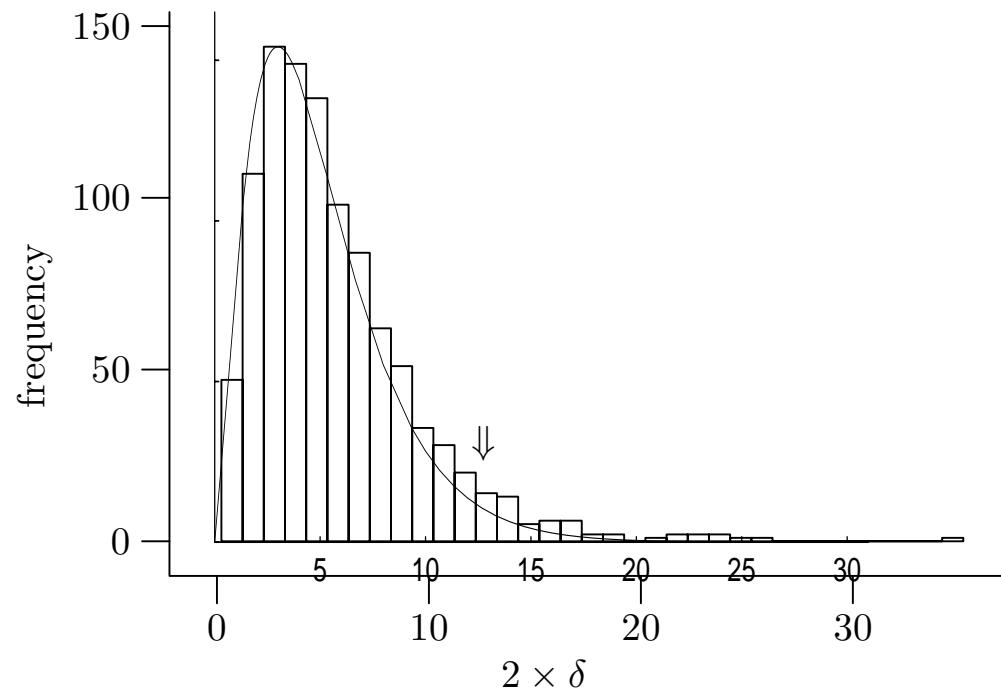
You can use χ^2 approximation to assess significance of the effect of parameters

The curve is a χ^2 distribution with 1 degree of freedom.



Now JC vs GTR, 5 degrees of freedom

The curve is a χ^2 distribution with 5 degrees of freedom. The histogram is generated from simulations under the null hypothesis.



Reference 2

D. Posada and K. A. Crandall 1998. “MODELTEST: testing the model of DNA substitution” *Bioinformatics* 14: 817-818.

- Automates the process of choosing a model
- Uses PAUP to do the likelihood calculations
- 56 different models are tested
- site-specific rate variation is *not* covered

Choosing a model— “maxing out”

- Often you will need the most complex model, GTR+I+G, or GTR+SS.
- It makes you think that you could improve the model by adding other parameters
- ...and you might be right.

When the model does not fit the data...

- A badly fitting model may over- or under-estimate the true evolutionary distance
- An underestimate (*eg* the assumed “model” in parsimony) will show long branch attraction
- An overestimate (rare) will show long branch repulsion

Among-site rate variation

- It is usually important that among-site rate variation be modelled
- Failure to do so will lead to underestimated distances and long-branch attraction
- Use **pInvar** or **gamma**-distributed variable rates, or a combination of the two
- ...or site-specific (**ss**) rates

Comparing tree topologies using likelihood

- When we compare trees, we can't use the same strategy that we used to compare nested models
- The Kishino-Hasegawa test to compare trees has been in use for 10 years.
- However, the KH test has problems
- A similar but better test is the recent Shimodaira-Hasegawa test

Reference 3

Goldman, Anderson, and Rodrigo 2000. Likelihood-based tests of topologies in phylogenetics. *Syst. Biol* 49: 652–670.

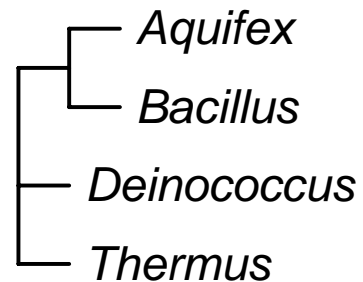
- A critique of the Kishino-Hasegawa test
- An explanation of various ways of comparing tree topologies with likelihood

Comparing tree topologies using the SH test

- The Shimodaira-Hasegawa test can tell you whether sub-optimal trees are significantly worse than the ML tree.
- If sub-optimal trees are *not* significantly worse than the ML tree (often the case!) then the ML tree is not a strong hypothesis, perhaps because the data are weak.
- You often can get reasonable sub-optimal trees using topologically-constrained searches.

Convergent composition

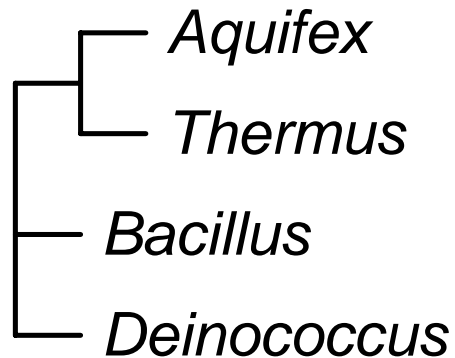
- *Deinococcus* are radiation resistant bacteria.
- *Thermus* are thermophilic
 - There is good evidence for a close phylogenetic relationship between *Deinococcus* and *Thermus*
- *Aquifex* is another thermophile, and *Bacillus* is a mesophile
 - Neither is closely related to either *Deinococcus* or *Thermus*



- We can take this as the “true” topology.

Convergent composition

- However, the two thermophiles share a compositional bias, and group together, giving the wrong tree with many phylogenetic methods.



The shared compositional bias of *Aquifex* and *Thermus* is so strong that the true phylogenetic signal is masked, and the two taxa “attract” each other in the tree

SH test of the 3 possible trees with these 4 taxa

Shimodaira-Hasegawa test:

SH test using RELL bootstrap (one-tailed test)

Number of bootstrap replicates = 1000

Tree	-ln L	Diff	-ln L	P
attract	3983.00041		(best)	
true	3985.30568	2.30526		0.465
other	3995.26719	12.26677		0.027*

* P < 0.05

- Maximum likelihood with the GTR+G model erroneously finds the “attract” tree as the best tree
- However, the true tree cannot be rejected under this model

Heuristic search strategies

- It is too time-consuming to estimate parameters (other than branch lengths) while searching.
- Parameters should be fixed to reasonable values before searching.

Don't do this...

```
lset pinvar = estimate  
      shape = estimate;  
hsearch;
```

Do this...

```
lset pinvar = 0.213  
      shape = 0.679;  
hsearch;
```

Or this...

```
lset pinvar = previous  
      shape = previous;  
hsearch;
```

General strategy: *Successive iteration*

- Optimize parameters to reasonable values on a single tree.
- Fix parameters, and search.
- Re-optimize parameters based on the best tree from the search.
- Repeat until things don't improve anymore.

How to get the initial parameters?

- Rule of thumb: Parameters for reasonably good trees do not differ much.
- Start with a reasonably good tree, eg a MP tree, and use that to get valid initial parameters.

Branch swapping

- NNI is fastest, but least complete
- SPR is intermediate
- TBR is best, most complete, but slowest

A fast heuristic search strategy

- Successive iteration
- Quickly get the model parameters close to their final values using inexpensive NNI and SPR branch swapping
- Then one round of TBR branch swapping, with `lset approxlim=2`
- Finally, all that is needed is to finish with one round of expensive TBR branch swapping, with `approxlim=5`.

This strategy is much faster than successive iteration using only TBR branch swapping.

New research in likelihood

- Bayesian methods, MCMC
- codon models
- heterogeneous models
 - heterogeneous over the data
 - heterogeneous over the tree
- likelihood of morphological data
- covarion model
- modelling indels

Reference 4

Paul Lewis 2001. Phylogenetic systematics turns over a new leaf. *TREE* 16: 30–37.

- Very short history of models
- codon and secondary structure models
- likelihood of morphological data
- Bayesian methods, MCMC

Reference 0

Swofford, Olsen, Waddell, and Hillis, 1996. *in Hillis et al, Molecular Systematics.*