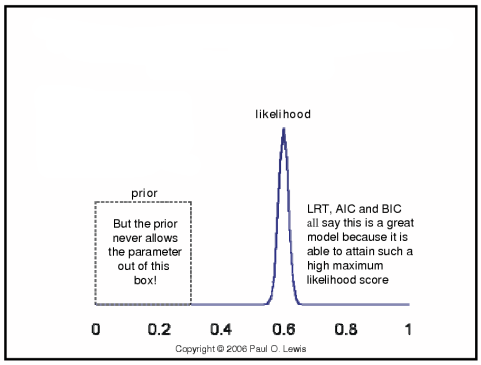
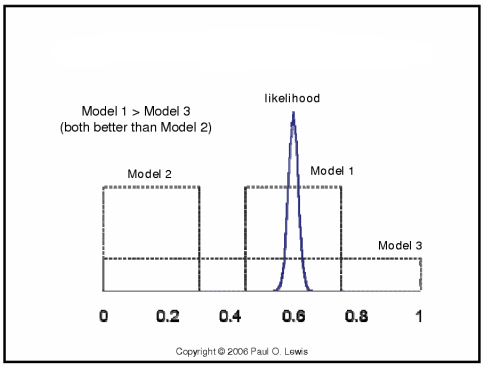
Minimum posterior predictive loss model selection (Paul Lewis)

Bayesian model = likelihood model + prior model (in contrast to LRT, AIC, BIC, etc. which do not take into account the priors).

Choice of prior can turn a good model into a bad one.





Bayes factors punish vague priors (model 1 > model 3 > model 2).

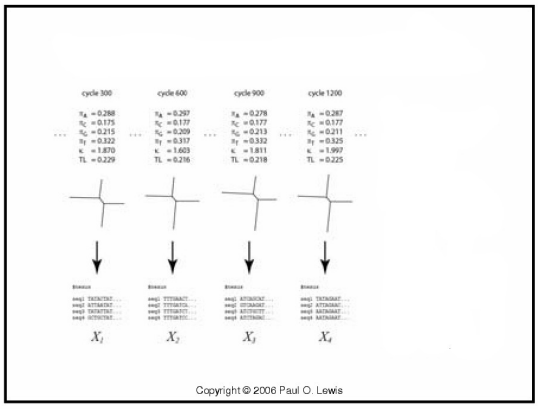
Contrast between model fit and predictive ability:

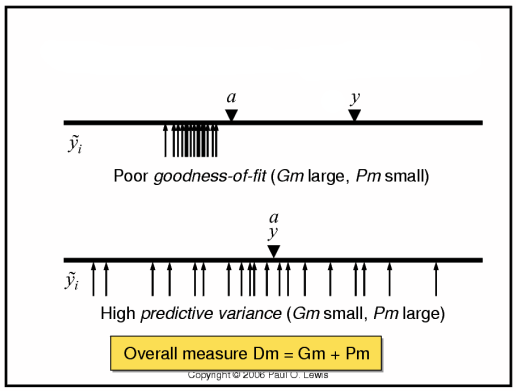
complex model -> better fit -> less predictive ability

simple model -> worse fit -> more predictive ability

more complex models make more variable predictions on unseen data

Posterior predictive approach – does a model predict the kind of data that you've actually seen?





To test models, use parameters+tree to simulate data during MCMCMC and compare simulated dataset to original using some metric (loss function).

Deviance loss function – combines difference in frequency of each site pattern between two datasets, weighted by frequency.

A poor model will predict datasets with very different stats to the real dataset; hence you choose the model with the minimum loss function (LOSS FUNCTION == difference between real dataset and average of posterior predicted dataset).

Overall measure takes into account goodness of fit (Gm) and variance (Pm).

Estimating sequence divergence times (Bruce Rannala)

Using molecular clocks, calibrated with multiple fossils

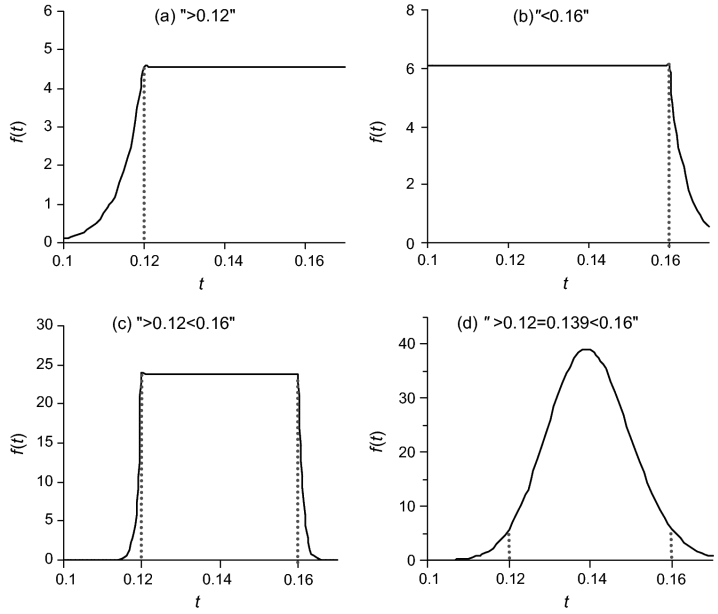
Averaging over uncertainty in fossil dates -> Bayesian approach

previous attempts use uniform prior over some time interval – probability of date outside the interval is 0

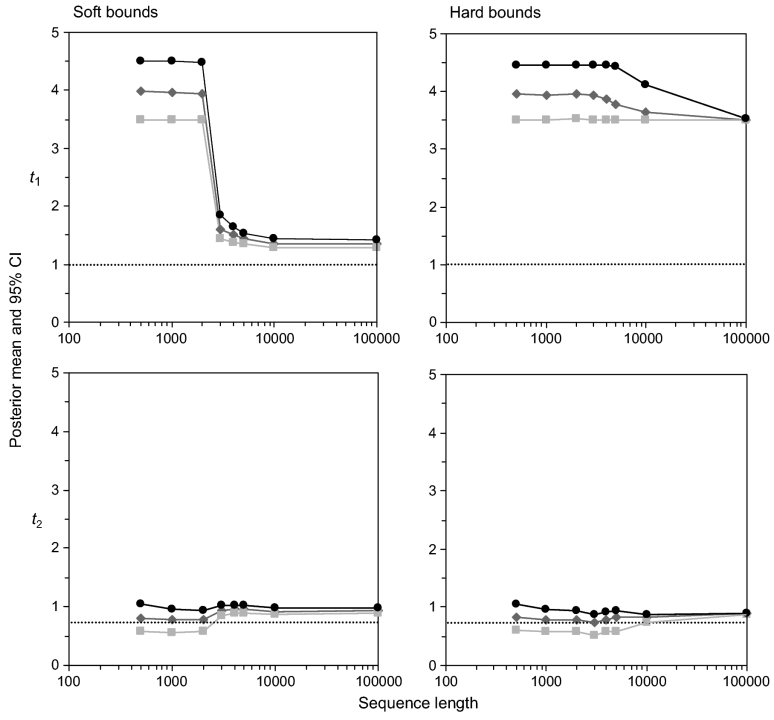
soft bounds necessary so that incorrect date estimates can be overcome by likelihood – small but positive probability outside the bounds

Known tree topology; calculate posterior distribution of divergence times using birth-death model to specify prior for divergence times without fossil estimates

Prior probability densities implemented:



uncertainty in prior dates means that uncertainty will remain even with infinite dna data (i.e. all *relative* branch lengths and hence *relative* divergence times known perfectly).



Species trees from gene trees (Liang Liu)

Causally, species tree -> gene tree ->DNA

but analytically DNA -> gene tree -> species tree

How to calculate P(Gene tree | species trees) ?

2 step mcmc

1. estimate posterior distribution of gene trees
2. use that to estimate posterior distribution of species trees

Use birth/death process to model generation of gene trees from species trees.

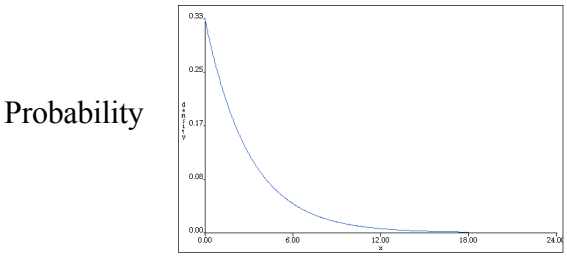
gene trees are are correlated due to their shared species tree, so they should be estimated jointly.

distribution of gene trees given species trees is derived from coalescent theory.

Hyperpriors

Parameter of a prior distribution expressed in terms of another distribution

e.g. Normal prior: branch length = exponential(3)



Hyperprior: branch length = exponential(x); x=gamma(2)