

Maximum Likelihood for phylogenetics

MEP: 14 Oct 2004

foundation

- ◆ likelihood methods have been around for about 85 yrs, invented by RA Fisher
- ◆ revolutionized statistics
- ◆ suggested for molecular data to inferring phylogenies in the 1960's, but too computationally difficult at that time

background...

- ◆ $\text{Prob}(D|H)$ is the likelihood of hypothesis H (our model)
- ◆ However, it is NOT the probability of hypothesis H - that would be $\text{Prob}(H|D)$ and will visit that in a couple weeks when we talk about Bayesian estimates of phylogenies - we make no assumptions about the probability of hypothesis H , our model
- ◆ $\text{Prob}(D|H)$ is the probability our our data, given the hypothesis
- ◆ we can test a variety of H 's (models) and find the one that maximizes the probability of observing our data
- ◆ for phylogenetics, we can think of D as our sequence data and H as the tree & model of evolution - this allows us to find the tree (& model) that maximizes the probability of observing those sequences, the ML tree

an example:

- ◆ imagine you are at home and hear a "roll... CRASH" in the room above you
- ◆ think of a model: " H : gremlins bowling"; given this H , your observation (roll, crash) has a high likelihood
- ◆ think of an alternative model: " H : Bush and Condi playing cards"; given this H , your observation his a lower likelihood
- ◆ in neither case do we draw a conclusion about the probability of the model - in both cases, they are improbable

reminder

- ◆ likelihoods do not sum to 1.0 while probabilities do
- ◆ given a tree & a model, we could calculate the prob of obtaining all possible data sets; these probs would sum to 1.0
- ◆ however, we are only interested in one of these: the observed data set
- ◆ the likelihood is not the prob that the tree is the true tree; it is the prob that the tree has given rise to the observed data

assumptions, requirements:

- ◆ likelihood calculations assume that:
 1. evolution in different sites (on the given tree) is independent
 2. evolution in different lineages is independent
- ◆ maximum likelihood request 3 elements:
 1. a model of sequence evolution
 2. a tree (specifying a topology & branch lengths)
 3. the observed data
- ◆ ML solves 2 problems:
 1. for a given topology, what set of branch lengths makes the observed data most likely (what is the max likelihood value for that tree)?
 2. which tree of all possible trees has the greatest likelihood?

how?

- ◆ for a given tree, we can calculate $P_{ij}(t)$, the probability that nucleotide i at time 0 becomes j at time t at a given site
- ◆ we express this as $P_{ij}(v)$, with $v=rt$ varying among branches; in effect, we are estimating v
- ◆ $nLL = \sum L_k$, for k sites where L_k is maximizes for each site across the whole tree
- ◆ if j is known and i (the ancestor) is unknown:
 - ◆ we must compute the likelihood for all possible i (A, T, G, C)
 - ◆ sometimes, both i & j are internal nodes and unknown
 - ◆ the likelihood is the sum across all possible nucleotides at these nodes