

Consistency, Long Branch Attraction, Robustness to Model Violations, ...

Of the 3 methods of tree construction we have learned about --

MP, distance methods, ML -- which is best?

...

For sure, one wants a statistical estimator to be CONSISTENT

(informally, if your data is perfectly in accord with the model

[or method], the method X reconstructs the true tree.)

For the formal definition, suppose you have chosen a ^{specific} model $M_0 = (\mathcal{T}, \mathcal{N})$ with parameters \mathcal{N} = numerical parameters and your data consists of site pattern frequencies computed from n independent trials of the experiment. i.e. n sites were generated under M_0 independently.

for notational ease

Focusing only on the tree (easy extension to $(\mathcal{T}, \mathcal{N})$),

then let \hat{T}_n denote the estimator from your method based on a sample of size n . Suppose $\varepsilon > 0$ is arbitrary. Then if

$$\lim_{n \rightarrow \infty} \text{Prob} \left(\underbrace{\|\hat{T}_n - T\|}_{\text{measurement of how close } \hat{T}_n \text{ is to the true value } T} < \varepsilon \right) =$$

↑
probabilistic quantification of when \hat{T}_n will be within ε of T

↑
let the sample size go to ∞

then \hat{T}_n is a CONSISTENT ESTIMATOR of $T = (T, N)$

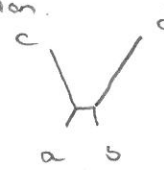
Formalization of a very basic requirement for inference.

If \hat{T}_{method} is not consistent, in practice no amount of data collection will help you to estimate T with "method."

Which methods are consistent?

⋮

Parsimony and Long Branch Attraction.

We know that a metric tree  can be hard to infer since a, b are the closest, i.e. their DNA sequences should look the most similar, but a, c are sister. Indeed, NJ was introduced to address this issue. We might expect parsimony to struggle for such trees. For such a tree, a "method" might infer



a tree in which the long branches are attracted.
 \equiv Long Branch Attraction

This phenomenon (LBA) extends to larger trees $n > 4$ and can throw off inference if too remote an outgroup is contained in data. Another way to view this is taxa c, d are essentially independent if those terminal branch lengths are long enough.

We will show that parsimony (MP) can be inconsistent

under a "2-state JC" model called the Cavender-Farris-Neyman CFN model. "Felsenstein Zone"

Method:

Details: Parsimony on a 4-taxon tree

Model: Explained below CFN 2-state model

Data: Pattern frequencies

xxyy

xyxy

xyyx

↑

↑

↑

Counts from data

 n_1 n_2 n_3

(Order reversed in book.)

3-trees to choose from

 $T_1: ab|cd$

$$PS(T_1) = n_1 + 2n_2 + 2n_3 = 2n - n_1$$

 $T_2: ac|bd$

$$PS(T_2) = 2n_1 + n_2 + 2n_3 = 2n - n_2$$

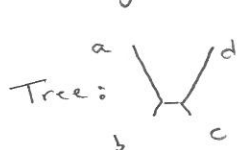
 $T_3: ad|bc$

$$PS(T_3) = 2n_1 + 2n_2 + n_3 = 2n - n_3$$

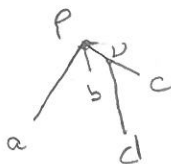
where $n = \#$ of informative sitesParsimony Criterion: Choose T_i with n_i largest (so $2n - n_i$ smallest)

End data analysis.

Begin: Generate sequences assuming the CFN model



root →



Book: Here root at a

2-states.

Root Distribution is $p_i = (.5, .5)$

2 Markov matrices one for short edges, one for long edges

$$M_{\text{long}} = \begin{pmatrix} 1-q & q \\ q & 1-q \end{pmatrix}$$

$$M_{\text{short}} = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$$

$$p, q \in (0, .5)$$

$$xxyy \quad p_1 = (1-q)^2 p(1-p)^2 + 2q(1-q)p(1-p)^2 + q^2 p^3$$

$$xyxy \quad p_2 = (1-q)^2 p^2(1-p) + 2q(1-q)p^2(1-p) + q^2(1-p)^3$$

p. 168 book

$$xyyx \quad p_3 = (1-q)^2 p^3 + 2q(1-q)p(1-p)^2 + q^2 p(1-p)^2$$

↑
Work (including HW)

(1)

If MP were to choose the true tree T_1 , then since

$$\lim_{n \rightarrow \infty} \frac{n_i}{n} = p_i$$

it must be that $p_1 > p_2, p_3$

To test this, we compute

Assuming $p, q \in (0, 1/2)$ Why?

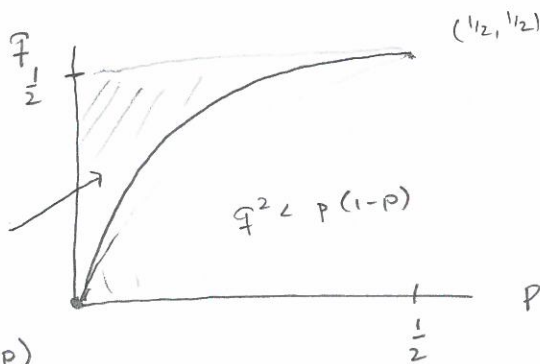
$$p_1 - p_3 = (1-2p)(p(1-p) - q^2)$$

$$p_1 - p_2 = p(1-2p)(1-2q)$$

$p_1 - p_3 > 0$ always, but

$$p_1 - p_2 > 0 \text{ iff } p(1-p) - q^2 > 0 \quad \text{i.e.} \quad p(1-p) > q^2$$

$$\text{i.e.} \quad q^2 < -p^2 + p$$



$$q^2 > p(1-p)$$

Felsenstein Zone



with parameters

Theorem: If sequences evolve under the CFN model on

given as above, then for parameters chosen with $q^2 > p(1-p)$ Maximum

Parsimony gives an inconsistent estimator of T.

Summary:

Parsimony, a non-model-based method yields inconsistent estimators for sequence data perfectly in accord with a CFN model in some parts of parameter space

ML yields a consistent estimator \hat{T}_{MLE} under many models. (list in text.)

NJ yield a consistent estimator provided the distance used is the appropriate one for the model chosen ($GTR \leftrightarrow d_{GTR}$, etc.)

Math Students: Identifiability of parameters key to consistency proofs.

Though consistency of \hat{T}_n is essential for sound inference of T , there are other practical considerations

1) Since consistency is concerned with $\lim_{n \rightarrow \infty}$, for empirical datasets, how big should n be so that \hat{T}_n is a "good" estimate?

2) Phylogenetic models are of course a poor description of the true evolutionary process of DNA so how robust our methods to violations of model assumptions?

ROBUSTNESS of MLE?

Bootstrapping:

Way to assess support for branches in \hat{T}

"The bootstrap"

Suppose \hat{T} is estimated from an alignment of length n

s_1 :	A		n
			T
			T
			C
			T
s_ℓ :	A		T
			↑
			n-th site

And that ML (or any other method) was used to construct $\hat{T} = \hat{T}_n$

Then

- 1) Sample with replacement from the columns in the alignment to create a new dataset of length n , say R_1 (= replicate 1)
 - 2) Construct a tree $\hat{T}_{n,1}$ from alignment R_1 and keep track of the number of edges it has that are on \hat{T}

↑
bootstrap
 - 3) Repeat as many times as desired, typically $K=1000$ times ($K=100$)
 - 4) Embellish \hat{T} by adding bootstrap support numbers to its branches
- The bootstrap support number is proportion of bootstrap trees with that edge