Statiscal Properties of Phylogenetic Tree Construction Methods

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 with parameters $M_0 = (T, N)$ 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.

Focusing only on the tree (easy extension to (T, N) for notational ease), 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

$$\lim_{n \to \infty} \text{Prob}\left(\underbrace{\| \hat{T}_n - T \|}_{} < \varepsilon\right) = \underline{\qquad}$$

measurement of how close $\hat{T}_n$ is to the true value T

probabilistic quantification of when $\hat{T}_n$ will be with 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

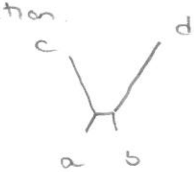
If $\hat{T}_{method}$ is not consistent, in practice no amount of data collection will help you to estimate T with "method."

Which methods are consistent?

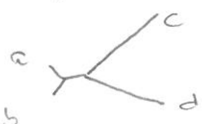
⋮

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

≡ 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 Cavendar-Ferris-Neyman CFN model. "Felsenstein Zone"

Details: Method:
Parsimony on a 4-taxon tree

Model: Explained below CFN 2-state model

Data: Pattern frequencies xxyy xyxy xyyx

$\uparrow$ $\uparrow$ $\uparrow$

Counts from data n_1 n_2 n_3

(order reversed in book.)

3-trees to choose from

T_1: ab|cd $\quad ps(T_1) = n_1 + 2n_2 + 2n_3 = 2n - n_1$

T_2: ac|bd $\quad ps(T_2) = 2n_1 + n_2 + 2n_3 = 2n - n_2$

T_3: ad|bc $\quad ps(T_3) = 2n_1 + 2n_2 + n_3 = 2n - n_3$

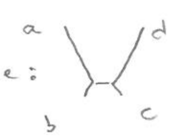
where n = # of informative sites

Parsimony Criterion: Choose T_i with n_i largest (so $2n - n_i$ smallest)

End data analysis.

Begin: Generate sequences assuming the CFN model

Tree:



root $\longrightarrow$

Book: Here root at a

2-states

Root Distribution is $p_\rho = (.5\ \ .5)$

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

$$M_{long} = \begin{pmatrix} 1-q, & q \\ q & 1-q \end{pmatrix} \qquad M_{short} = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix} \qquad p, q \in (0, .5)$$

With this model

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

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

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

↑

Work (including HW)

:)

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

$\lim_{n\to\infty} \frac{n_i}{n} = p_i$ it must be that $p_1 > p_2, p_3$