

Tehrani, Jamshid J.

Drápela, Martin (editor)

## Methods

In: Drápela, Martin. *Sbírka textů k předmětu Úvod do syntaxe*. 1. vyd. Brno: Masarykova univerzita, 2014, pp. 186-191

ISBN 978-80-210-7104-9; ISBN 978-80-210-7107-0 (online : Mobipocket)

Stable URL (handle): <https://hdl.handle.net/11222.digilib/131723>

Access Date: 19. 02. 2024

Version: 20220831

Terms of use: Digital Library of the Faculty of Arts, Masaryk University provides access to digitized documents strictly for personal use, unless otherwise specified.

# Methods

## 1 Phylogenetic Reconstruction

Cladistic analysis employs a branching model of evolution that clusters taxa on the basis of shared derived (evolutionarily novel) traits. Using the principle of parsimony, it involves finding the tree that minimises the total number of character state changes required to explain the distribution of character states among the taxa, known as the "shortest length tree" or "most parsimonious tree". To search for the most parsimonious tree (MPT), the present analysis employed an efficient tree-bisection-reconnection algorithm implemented by the heuristic search option in PAUP 4 [57], carrying out 1,000 replications to ensure a thorough exploration of tree-space. The fit between the data and the MPTs was assessed using the Retention Index (RI) and maximum parsimony bootstrapping. The RI is a measure of how well similarities among a group of taxa can be explained by the retention of shared derived traits on a given tree [34]. A maximum RI of 1 indicates that all similarities can be interpreted as shared derived traits, without requiring additional explanations, such as losses, independent evolution or borrowing. As the contribution of these latter processes increase, generating

similarities that conflict with the tree, the RI will approach 0.

Maximum parsimony bootstrapping is a technique for measuring support for individual clades [58]. It involves carrying out cladistic analyses of pseudoreplicate datasets generated by randomly resampling characters with replacement from the original matrix. Support for the clades returned by the original analysis is then estimated by calculating the frequency with which they occur in the most parsimonious trees obtained from the pseudoreplicates. The bootstrap analyses reported here were carried out in PAUP 4 [57] using heuristic searches of 1,000 replicates.

Bayesian inference proceeds by calculating the likelihood of the data given an initially random tree topology, set of branch lengths and model of character evolution, and iteratively modifies each of these parameters in a Markov Chain Monte Carlo (MCMC) simulation. Moves that improve the likelihood of the data are always accepted, while those that do not are usually rejected (although some may occasionally be accepted within a certain threshold so as to avoid getting trapped in local optima). Following an initial "burn in" period, the likelihood scores will plane out and parameters will fluctuate between similar values, at which point trees are sampled at regular intervals to generate the "posterior distribution". Unlike the trees output by a cladistic analysis, which are based on a single optimality criterion (i.e. parsimony),

the posterior distribution of trees represents a set of phylogenetic  
45 hypotheses that explain the distribution of character states among  
the taxa under a range of plausible evolutionary assumptions.  
The posterior distribution of trees can be summarised by a consensus  
tree or "maximum clade credibility tree", while posterior  
probabilities for individual clades are calculated based on their  
50 frequency in the tree sample. The Bayesian approach has been  
found to be particularly effective when there is wide variance  
in the amount of evolution that has occurred in different regions  
of the character data or tree, since it explicitly incorporates  
these parameters (i.e. branch lengths and substitution model)  
55 into the analysis [59]. The Bayesian analyses reported here  
were carried out in MrBayes 3.2 [60] using the model settings  
for "standard" (morphological) data, with the character coding  
set to "variable" and variance in rates of character evolution  
estimated under a gamma distribution. Two analyses were carried out  
60 simultaneously, each using four MCMC chains that were run for 1  
million generations. Trees were sampled every 1000 generations  
to avoid autocorrelation, with the first 25% of the sample discarded  
as burnin. Log likelihood values for the remaining trees in each  
sample were then graphed as a scatterplot to check that the two runs  
65 had converged.

As with the other two methods, NeighbourNet clusters taxa into hierarchically nested sets. However, unlike cladistics and Bayesian inference, it does not employ a strict branching model of descent with modification, and as such these sets can overlap and intersect with one another. Accordingly, it is claimed that NeighbourNet is better able to capture conflicting signal in a dataset resulting from borrowing and blending among evolutionary lineages [35]. The method involves calculating pairwise distances between the taxa based on the character data, and generating a series of weighted splits that are successively combined using an agglomerative clustering algorithm. Relationships among the taxa are represented by a network diagram, or "splits graph", which shows groupings in the data and distances separating them. Where the splits are highly consistent, the diagram will resemble a branching tree-like structure. Incompatible splits, on the other hand, produce box-like structures that lend a more latticed appearance to the network. The extent of reticulation in the folktale network was quantified using the delta-score and Q-residual score [36], [61]. Both measures calculate conflicting signal by comparing path lengths among pairs of taxa on "quartets" (subsets of four taxa) selected from the network. Quartets are scored from 0 to 1 according to how resolved the splits between each pair of taxa are, with values closer to 0 being more tree-like and values closer to 1 more reticulate.

The estimation of the delta score includes a normalisation constant, whereas Q-residuals had to be normalised by rescaling all between-taxa distances in the network so that they average 1. The NeighbourNet analysis and calculation of d-scores and Q-residuals were carried out in SplitsTree v4.13 [35].

### 95 **Ancestral state reconstruction**

Character states were reconstructed in the putative last common ancestor of ATU 123 and ATU 333 tales through parsimony analysis and Bayesian inference. In the parsimony analysis, the most parsimonious trees (MPTs) from the cladistic analysis were re-rooted so as to make ATU 123 and ATU 333 monophyletic, with the East Asian group forming a sister clade. Next, the evolutionary history of each character was reconstructed on the MPTs by minimising the total number of changes required by each tree. The ancestral state inferred for the last common ancestor of ATU 123/333 tales was then recorded for each tree. The parsimony analyses were carried out in the software program Mesquite, using the Character Trace module [62]. In the Bayesian analysis, phylogenetic relationships among the taxa were reconstructed using a topological prior that forced ATU 333 and ATU 123 to be monophyletic (making the clade present in 100%

of the posterior distribution of trees). The analysis was carried out in MrBayes 3.2 [60], with the other model settings being the same as those used in the original analysis, in which the evolutionary rate across characters was allowed to vary under a gamma distribution.

115 Estimated ancestral states in the last common ancestor of ATU  
123/333 were sampled every 1000 generations to avoid autocorrelation,  
with the first 25% of the sample discarded as burnin. The average  
probabilities for each state were summarised using the Report  
Ancestral State command (report ancstates = yes), integrating  
120 uncertainty in the topological structure of the rest of the tree as  
well as other model parameters.