1 Appendix B: Tree Comparisons

1.1 Normalised Tree Similarity

For any tree with n taxa compared using a tree difference metric m, Normalized Tree Similarity, NTS_m (Bogdanowicz et al., 2012), represents the similarity score for the two trees given the expected difference between two random Yule trees (Bogdanowicz et al., 2012) with n taxa. If $\bar{d}_{m,n}(rand)$ is the average difference between two random Yule trees with n taxa and $d_{m,n}(x,y)$ the difference between the two trees x and y each containing n taxa, then:

$$NTS_{m,n}(x,y) = \frac{\bar{d}_{m,n}(rand) - d_{m,n}(x,y)}{\bar{d}_{m,n}(rand)} \tag{1}$$

NTS ranges from one to -∞. For any m, n, when NTS = 1, the trees are identical, when NTS = 0 the trees are no more different than expected by chance, and when $NTS < 0$, the trees are more different than expected when comparing two random trees.

We used the NTS method to scale all the Robinson Foulds and Triplets metrics calculated in our analyses, using the TreeCmp JavaScript (Bogdanowicz et al., 2012).

1.2 Bhattacharyya Coefficient

The Bhattacharyya Coefficient calculates the probability of overlap of two distributions (Bhattacharyya, 1943). When it is equal to zero, the probability of overlap of the distributions is also zero, and when it is equal to one, the two distributions are entirely overlapping. It forms an elegant and easy to compute continuous measurement of the probability of similarity between two distributions. The coefficient is calculated as the sum of the square root of the relative counts shared in n bins among two distributions.

$$\text{Bhattacharyya Coefficient} = \sum_{i=1}^{n} \sqrt{\sum a_i \times \sum b_i} \tag{2}$$

where

$$a_i = \frac{\text{Number of counts in bin } i \text{ for the distribution } a}{\text{Total number of counts for the distribution } a} \tag{3}$$

and

$$b_i = \frac{\text{Number of counts in bin } i \text{ for the distribution } b}{\text{Total number of counts for the distribution } b} \tag{4}$$

The precision of the Bhattacharyya Coefficient is directly related to the number of bins, n. If n is low, the overlap will be overestimated and if n is too high, the overlap will be underestimated. In this analysis, we determined the number of bins using Silverman's rule of thumb which states that n should be 0.9 times the minimum of the standard deviation and the interquartile range of the distribution, divided by 1.34 times the sample size of the distribution to the negative one-fifth power (`bw.nrd0()` function in R (Silverman, 1986)).

References

Bhattacharyya, A. 1943. On a measure of divergence between two statistical populations defined by their probability distributions. Bulletin of the Calcutta Mathematical Society 35:99–109.

Bogdanowicz, D., K. Giaro, and B. Wróbel. 2012. TreeCmp: Comparison of trees in polynomial time. Evolutionary Bioinformatics 8:475–487.

Silverman, B. 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall/CRC Monographs on Statistics & Applied Probability Taylor & Francis.