9. Long branch attraction and topology testing using weighted least-squares

Borys Wrobel

*Department of Marine Genetics and Biotechnology, Institute of
Oceanology, Polish Academy of Sciences, PL-81712 Sopot, Poland
(bwrobel@iopan.gda.pl)*

WLS tests can be used to tests topologies and branches in trees as long as the distances and some of their associated variances are available. In this work I will compare the performance of the WLS tests with other methods in conditions of long branch attraction, misspecification of the substitution model, substantial rate heterogeneity and/or "topological noise".

The weighted least-squares likelihood ratio test (WLS-LRT) allows for branch and topology testing using only the evolutionary distances and some of their associated variances. WLS can be used when the distances are derived from sequences, but the access to the original sequence data is not necessary; the data may also be, for example, serological or DNA hybridization distances, or distances averaged across different datasets.

The method (implemented in a program WeightLESS; http://www.iopan.gda.pl/~wrobel) involves reducing the T(T-1)/2 variances (where T is the number of taxa) to only two parameters. The logarithm of distance variances is regressed on the logarithm of observed distances. The slope of this regression gives the optimal power to be used in the weighted sum of squares and the y-intercept value is necessary to calculate the likelihood of a given topology. When the likelihood ratio test is used for branch testing, two trees are compared: one is a “contraction” of the other (that is, the length of the branch being tested is set to zero). Under the null hypothesis, both trees are equally likely, and twice the logarithm of the ratio of their likelihoods follows the chi-square distribution with degrees of freedom equal to the difference in the number of branch lengths being estimated (that is, one). In the weighted least-squares context, this statistic can be calculated as the difference between the sum of squares of each tree, divided by the y-intercept parameter.

When constructing confidence sets for trees, the assumption of distance independence, which differentiates the weighted least-squares from the general least-squares approach, is expected to result in a more conservative test. In simulations and analysis of biological data the WLS method gives similar results to those of the Shimodaira-Hasegawa test.

For branch testing, Felsenstein’s (1985) bootstrap selection probability test is perhaps the most commonly used method. However, the statistical basis of this method is unclear. A reliable bootstrap threshold for clade selection is notoriously difficult to establish, and depends on the particular dataset. Our previous results show that WLS-LRT does not suffer from these problems and gives results similar to those of Dopazo’s (1994) internal branch test both in simulations and analysis of real data.

In this work I will compare the performance of the weighted least-squares test in conditions of long branch attraction, when in artifactually produced trees rapidly evolving sequences (represented by long branches on unrooted phylogenetic trees) are placed with other rapidly evolving sequences, even if they are only distantly related. Long branch attraction is the main cause of the inconsistency of the methods of phylogenetic reconstruction and topology testing: when longer sequences are used, the incorrect solution is even more strongly supported. Preliminary results of simulations show that the WLS test is consistent in branch testing when neither the Dopazo test nor the bootstrap probability is, in conditions of long branch attraction, misspecification of the substitution model, substantial rate heterogeneity and/or "topological noise" (when the relations between the taxa is better described by a network and not a tree).