

“Distance Method Adjustments and a Test for General Heterotachy in Phylogenetic Estimation”

Jihua (Jerry) Wu
PhD Candidate in Statistics
Dalhousie University

ABSTRACT

This presentation is about using distance methods for phylogenetic tree estimation and inference in the presence of heterotachy. It is composed of four parts; the first two parts are about phylogenetic tree reconstruction and the last two about heterotachy detection.

In the first part, we give several identifiable results for distance methods, which provides a solid theoretical foundation for the following rest of the presentation. We prove that for a fixed and given rate matrix Q , as long as the rate matrix has at least two non-zero distinct eigenvalues, two distributions of pairwise site patterns coincide only if the pairwise distances and shape parameters are the same. However, when the rate matrix Q is completely unspecified, a single distance is not sufficient for an identifiable model; we will give illustrative examples of two different sets of parameters giving the same pairwise distributions. We also prove that when two distinct pairwise distributions are available, the rate matrix Q is completely unknown and has at least two non-zero distinct eigenvalues, the rate matrix Q , pairwise distances and the shape parameter α can be identified simultaneously. Thus distance-based methods alone can be used for rates-across-sites and substitution model parameter estimation. Extending this result we show that, if the rate variation is described by an arbitrary rate distribution and pairwise distributions are available for all distances, the rate matrix Q , the pairwise distance and the rates-across-sites distribution can be identified.

In the second part of the presentation, we give a general definition of heterotachy as multivariate rates-across-sites variation, where the rates or branch lengths are modeled by an arbitrary distribution. We also show that all of the commonly considered heterotachy models can be considered special cases of this general definition. We use this result to establish that, for pairs of taxa, heterotachy is usual, univariate rates-across-sites variation. Motivated by this characterization of heterotachy, we assume that rates between two taxa follow different distributions. For convenience, gamma distributions with different shape parameters for different pairs are used. Through maximum likelihood estimation, we find the best distance and the best (shape) parameter for the corresponding distribution for each pair of taxa. Then the neighbor-joining method is used to find the best tree topology. Using a famous example of heterotachy where uncorrected ML performed poorly, we simulated replicate DNA sequence alignments with two symmetrical rate partitions along a four-taxon tree. Our pairwise alpha heterotachy adjustment (PAHA) consistently showed better phylogenetic accuracy (the fraction of replicates from which the true tree was recovered) on simulated data than either uncorrected ML or MP methods. We also applied PAHA method to Chloroplast Data, with interesting results.

Motivated by the results from the second part of the thesis, we extract information from α 's calculated through PAHA method and constructs a test statistic to decide whether a RAS model or general heterotachy model is appropriate for a group of sequences.

We use similar ideas to construct a heterogeneity test which can be used to distinguish ER model from RAS model. The tests are shown to perform well under a wide range of simulation conditions. Finally, we used these two tests to analyze three real data sets, and found that while all of these data sets showed significant evidence of heterotachy, there were subtrees for which the data was consistent with an equal rates or rates-across-sites model.

In the last part of the thesis, we propose another test to find the within gene heterotachy. Modifying the PAHA method proposed in the second part, we adjust for between-gene heterotachy by assuming different gene trees for different genes. Any additional heterotachy must be a consequence of within-gene heterotachy. The results of simulations show that the within gene test performs well under a wide range of conditions. We implement the test to chloroplast genome sequences data, and found some special genes.