Evaluation and Application of Nonlinear Dimensionality Reduction Methods for Phylogenetic Inference

Wen Huang1, James C. Wilgenbusch2, Kyle A. Gallivan1
1Department of Mathematics, 2Department of Scientific Computing, Florida State University, Tallahassee, FL

INTRODUCTION

Phylogenetic analyses of large and diverse data sets generally result in large sets of competing phylogenetic trees. Consensus tree methods used to summarize sets of competing trees discard important information regarding the similarity and distribution of competing trees. A more fine grain approach is to use a dimensionality reduction method to project tree-to-tree distances in low dimension Euclidean space [1]. Such an approach gives us a way to better understand the processes and patterns of evolution and well as how well suited our models and methods are performing. For example, analyses of different data partitions may support different phylogenies because reconstruction methods sometimes fail to adequately accommodate process heterogeneity underlying data partitions found within an alignment [2, 3, 4, 5] or because some data partitions simply do not share the same evolutionary history [6]. Furthermore, large data sets are typically more computationally challenging to analyze and often call for more extreme heuristic shortcuts, which may fail to converge to a global optimum [7].

In this study, first, we systematically evaluate the performance of several nonlinear dimensionality reduction (NLDR) methods on several tree-to-tree distances obtained from independent nonparametric bootstrap analyses of genes from three mid-to-large-sized mitochondrial genome alignments. Second, we apply the most reliable NLDR method to visualize the consequences of removing potentially misleading characters from an alignment of 169 Elassobranchus protein coding sequences comprised of 1 mtDNA and 7 nuclear loci. Characters were removed from the alignment based on how well they fit a model of stationarity using a program called DRUIDS [8]. We expect that sets of trees favored by individual loci will be more difficult to distinguish in projections (i.e., landscapes) of phylogenetic trees obtained from analyses of an alignment after the DRUIDS filter is applied.

Study Goals

1. Evaluate the performance and goodness of fit of several popular distance-based NLDR methods
2. Compare the tree projects of different mtDNA data sets
3. Evaluate different tree-to-tree metrics
4. Evaluate the effect of nonstationary characters on tree inference.

Methods of NLDR

Data


Plots of Tree-to-Tree distances


Quantitative Comparisons


References


Acknowledgements


This work is supported in part by a grant (DB-000080) from the National Science Foundation.