

The Shape and Dimensionality of Phylogenetic Tree-Space Based on Mitochondrial Genomes

James C. Wilgenbusch, Department of Scientific Computing, Florida State University

Wen Huang, Department of Mathematics, Florida State University

Kyle A. Gallivan, Department of Mathematics, Florida State University

Visually representing phylogenetic trees supported by different genes or by other *a priori* defined data partitions in 2- or 3-Dimensional (D) space can be a useful way for investigators to gain a better perspective on potential problems sometimes associated with the analysis of large multi-source data sets (Hillis et al., 2005). The practice of visually representing sets of competing phylogenetic trees in a geometric space can be separated into three major and sometimes computationally intensive components: 1) the selection of a set of phylogenetic trees to be compared; 2) the calculation of pairwise distances (p_{ij}) between all members of the set of selected phylogenetic trees; and 3) the calculation of coordinates in 2D or 3D space such that the Euclidean distance between the projected points (d_{ij}) closely corresponds to the original tree-to-tree distances (p_{ij}). We focus our evaluation on the last of these components by systematically comparing the performance of a single linear dimensionality reduction method commonly referred to as “classical MDS” and several non-linear dimensionality reduction (NLDR) methods. Nonparametric bootstrap analyses were performed on 15 mitochondrial gene partitions found in three published mitochondrial genome alignments (Kjer and Honeycutt, 2007; Setiamarga et al., 2008; Zhang et al., 2008). The bootstrap analyses resulted 3011, 6001, and 7022 trees for which tree-to-tree distances (p_{ij}) were calculated using the Robinson-Foulds distance metric. Nonlinear dimensionality reduction methods differed from one another by the algorithm (e.g., Linear Iteration, Majorization, Gauss-Seidel, and Stochastic Gradient) used to optimize several cost or stress functions [e.g., Kruskal-1 Stress, Normalized Stress, Nonlinear Mapping (NLM), Curvilinear Components Analysis (CCA)]. The resulting 2D and 3D coordinates were compared using several criteria including: 1) length of time required to reach convergence, 2) dependence of projections on initial conditions, 3) trustworthiness and discontinuity (Venna and Kaski, 2005) of the reduced data when compared to the original distances, and 4) the coherence of the data when visualized in 2D and 3D space. Dimensionality reduction methods and evaluation criteria were performed using software developed by the authors (GPL: <http://bpd.sc.fsu.edu/index.php/component/content/article/64>), while bootstrap analyses and tree-to-tree distances were calculated using PAUP* (Swofford, 2002).

Correctly characterizing phylogenetic tree-space by dimensionality reduction methods is critical if this approach is to be of value as an interpretive or a diagnostic tool. We demonstrate that different dimensionality reduction methods can significantly influence the appearance and interpretation of 2D and 3D projections of tree-to-tree distances. In particular, among the cost functions and optimization algorithms that we evaluated, we found that CCA and the stochastic gradient decent method gave the best representation of the original tree-to-tree distances as indicated by the trustworthiness and continuity metrics. Somewhere between three and 15 dimensions were needed to fit each of the tree-to-tree distance matrices according to several intrinsic dimensionality estimators. Therefore, visualizing the phylogenetic tree-to-tree distances in 3D generally facilitates the interpretation of these data. We plan to use this analysis framework to evaluate other tree-to-tree distance metrics, dimensionality reduction costs functions, and optimization algorithms. We also plan to apply our findings and the software developed as a part of this project to help refine evolutionary models used to infer phylogenetic trees, to alert practitioners to convergence problems where MCMC is used to infer phylogenies, and to improve heuristic search strategies.

Hillis, D. et al. (2005) Analysis and visualization of tree space. *Systematic Biology*, **54**, 471-482.

Kjer, K.M. and Honeycutt, R.L. (2007) Site specific rates of mitochondrial genomes and the phylogeny of eutheria. *BMC Evolution Biology*, **7**, 8.

Setiamarga, D. et al. (2008) Interrelationships of Atherinomorpha (medakas, flyingfishes, killifishes, silversides, and their relatives): The first evidence based on whole mitogenome sequences. *Molecular Phylogenetics and Evolution*, **49**, 598-605.

- Swofford, D.L. (2002) PAUP*: phylogenetic analysis using parsimony (* and other methods). Version 4:Beta 10.
- Venna, J. and Kaski, S. (2005) Local multidimensional scaling with controlled tradeoff between trustworthiness and continuity. In, *Proceedings of WSOM*, pp. 695–702.
- Zhang, P. et al. (2008) Phylogeny and biogeography of the family Salamandridae (Amphibia: Caudata) inferred from complete mitochondrial genomes. *Molecular Phylogenetics and Evolution*, **49**, 586-597.