An Evaluation of Tree-to-Tree Distance Metrics used to Visualize Phylogenetic Tree Landscapes

James C. Wilgenbusch[1], Wen Huang[2], Kyle A. Gallivan[2]

[1]Department of Scientific Computing, [2]Department of Mathematics, 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 2D or 3D space [1]. In this study, we evaluate several tree-to-tree distance metrics using trees obtained from independent nonparametric bootstrap analyses of genes from a mitochondrial genome alignment.

Study Goals

1.Visually and analytically evaluate projections of four commonly used tree-to-tree distance metrics.
2.Estimate the intrinsic dimensionality of tree-to-tree distance metrics.

Methods

Aligned whole salamander mitochondrial DNA (mtDNA) genomes were obtained from Zhang et al. [2]. The software package PAUP* 4.0b10 [3] was used to perform 100-replicate nonparametric bootstrap analyses [4] on each of 15-gene partitions contained within the mtDNA alignment. The maximum likelihood (ML) criterion and a heuristic search [neighbor joining starting tree, Sub-tree Pruning and Regrafting (SPR) branch swapping with a reconstruction limit of 10] were used to select optimal phylogenetic trees for each bootstrap replicate. Parameters of the ML model (i.e., nucleotide substitution rates, base frequencies [5] and an among site rate heterogeneity parameter [6]) were independently optimized for each gene partition on a neighbor joining tree constructed for each gene partition. A special purpose script by JCW (available upon request) was used to distribute phylogenetic analyses in parallel on FSU's shared HPC system.

The program TreeScaper [7] was used to evaluate several dimensionality reduction cost functions and optimization algorithms. The Curvilinear Components Analysis (CCA) cost function and the stochastic gradient decent optimization algorithm provided the best fit to the original tree-to-tree distances according to several goodness of fit measures [8, 9] (Fig.1).

Efficiency of Cost Function and Optimization

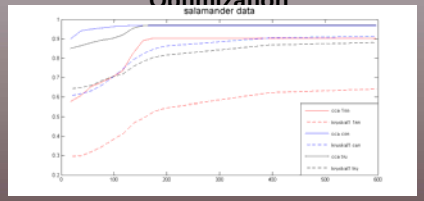

FIGURE 1. Goodness of fit measure plotted as a function of iteration, where 1NN = 1 Nearest Neighbour [8], CON = Continuity [9] and TRU = Trustworthiness [9].

Plots of Tree-to-Tree distances

Visual Inspection

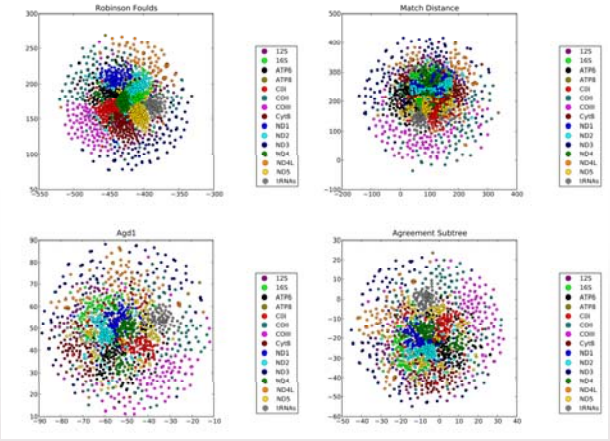

FIGURE 2. Two-dimensional projections of 1921 non-parametric bootstrap trees from the salamander data set using four tree-to-tree distance metrics (Robinson Foulds [10], Match Distance [11, 12], Agd1 [13], and Agreement Subtree [13]). The colors represent the underlying genes used to generate the trees. Projections were made using TreeScaper [7] with the cost function set to CCA and the optimization algorithm set to Stochastic Gradient Decent.

Distribution of Tree-to-Tree Distances

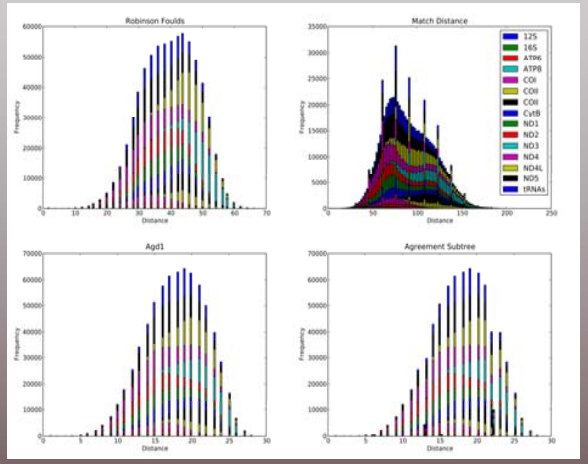

Figure 3. Distributions of raw tree-to-tree distances for four distance metrics evaluated under this study.

Intrinsic Dimensionality

Visual Inspection

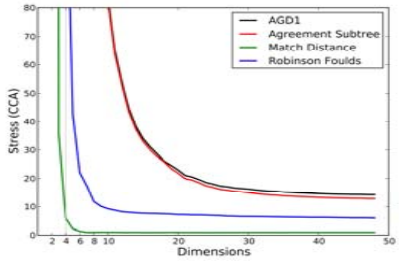

FIGURE 4. Raw stress of CCA plotted as a function of the projection dimensionality. More than 4 dimensions does not greatly improve the raw stress for the Match Distance metric.

Relationship Among Distance Metrics

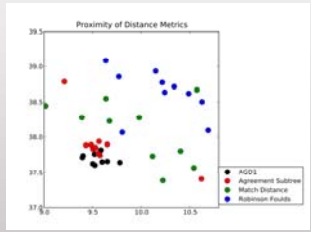

Figure 5. A projection of the relationship among 2D matrices, similar to those displayed in Figure 2. Ten projections for each of the four tree-to-tree distance metrics were generated based on different initialization conditions. The 40 2D projections were compared using a Procrustes analysis and the program TreeScaper [7] with the cost function set to CCA and the optimization algorithm set to the stochastic gradient decent method was used to display the result of the Procrustes.

Observations

1.Tree-to-tree distance metrics can qualitatively and quantitatively influence projections of "tree landscapes."
2.The projection of the RF-distances shows groups of related mtDNA gene trees best.
3.The Match Distance metric discriminates among bootstrap trees better than other metrics.
4.Fewer dimensions are required to optimally project the "Match Distance" tree-to-tree distance metric.

References

1. Hillis, D., Heath, T. & St John, K. 2005. Analysis and visualization of tree space. Systematic Biology 54, 471-482.
2. Zhang, P., Papenfuss, T., Wake, M., Qu, L. & Wake, D. Phylogeny and biogeography of the family Salamandridae (Amphibia: Caudata) inferred from complete mitochondrial genomes. Molecular Phylogenetics and Evolution 49, 586-597 (2008).
3. Swofford, D. L. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods) (version, b10).
4. Felsenstein, J. 1985. Confidence limits on phylogenies: an approach Using the Bootstrap. Evolution 39: 783-791.
5. Yang, Z. 1994. Estimating the pattern of nucleotide substitution. Journal of Molecular Evolution 39: 105-111.
6. Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. Journal of Molecular Evolution 39: 306-314.
7. Huang, W., Wilgenbusch, J. C., Gallivan, K. A. 2010. TreeScaper. available online at: http://bpd.sc.fsu.edu/index.php/diagnostic-software.
8. Van Der Maaten, L.J.P., Postma, E.O. & Van Den Herik, H.J. 2007. Dimensionality reduction: A comparative review. Preprint.
9. Kaski, S, et al. 2003. Trustworthiness and metrics in visualizing similarity of gene expression. BMC Bioinformatics 4, 48.
10. Robinson, D.F. & Foulds, L.R. 1981. Comparison of phylogenetic trees. Math. Biosci 53, 131–147.
11. Bogdanowicz, D. 2008. Comparing phylogenetic trees using a minimum weight perfect matching. Proceedings of the 2008 1st International Conference on IT, Gdansk, Poland
12. Bogdanowicz, D. and Giaro, K. 2010. Comparing arbitrary unrooted phylogenetic trees using generalized matching split distance. Information Technology (ICIT), 2nd International Conference, 259–262.
13. Goddard, W., Kubicka, E., Kulicki, G. and McMorris, F. R. 1994. The agreement metric for labeled binary trees. Mathematical Bioscience 123:215-226.

Acknowledgements

This work is supported in part by a grant (EF-0849861) from the National Science Foundation. Discussions with David Swofford provided some of the motivation for this investigation. Special thanks to Yu Lin, Vaibhav Rajan, and Bernard Moret for supplying a copy of their software for calculating Match Distances.