# Evaluation and Application of Nonlinear Dimensionality Reduction Methods for Phylogenetic Inference

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

[1]Department of Mathematics, [2]Department 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 Elasmobranch 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

| Taxa | Number of Sequences | Reference |
|---|---|---|
| Fishes | 90 | [9] Setiamarga et al., 2008 |
| Mammals | 89 | [10] Kjer and Honeycutt, 2007 |
| Salamanders | 42 | [11] Zhang et al., 2008 |

TABLE 1. Aligned whole mitochondrial DNA (mtDNA) genomes were obtained from three published studies representing a diverse set of animal taxa.

| | Number of Trees | | | | Number of Trees | | |
|---|---|---|---|---|---|---|---|
| Gene | Fishes | Mammals | Salamanders | Gene | Fishes | Mammals | Salamanders |
| 12S | 256 | 219 | 119 | ND1 | 507 | 170 | 111 |
| 16S | 205 | 146 | 106 | ND2 | 371 | 129 | 111 |
| ATP6 | 415 | 540 | 156 | ND3 | 690 | 1559 | 355 |
| ATP8 | 939 | 362 | 783 | ND4 | 219 | 150 | 108 |
| COI | 386 | 228 | 106 | ND4 | 1362 | 1056 | 378 |
| COII | 444 | 433 | 196 | ND4L | 188 | 114 | 103 |
| COIII | 643 | 554 | 149 | ND5 | 162 | 146 | 108 |
| CytB | 235 | 195 | 122 | **TOTALS** | **7022** | **6001** | **3011** |

TABLE 2. Phylogenetic trees were obtained for each of the three mtDNA data (GTR+ Γ) nonparametric bootstrap analysis (100 replicates) on each of the 15-mtDNA genes

A tree-to-tree distance matrix was created for the Fish, Mammal, and Salamander data set by concatenating the bootstrap trees found for gene. First of all, let us concentrate on the unweighted Robinson-Foulds (RF) distance [12].

## Compare NLDR Methods

### Visual Inspection

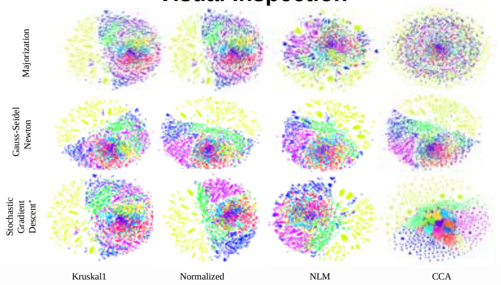

FIGURE 1. Two-dimensional projections of 3011 non-parametric bootstrap trees from the salamander data set using four cost functions (x-axis) and three optimization algorithms (y-axis). The colors represent the underlying genes used to generate the trees (see Table 2). * Kruskal-1 uses the linear iteration method instead of the stochastic gradient descent method used by the other cost functions in this row.

### Goodness of Fit Measures

| Salamander | | Majorization | Gauss Seidel | Stochastic | Linear Iteration |
|---|---|---|---|---|---|
| KRUSKAL-1 | 1NN | 0.631518 | 0.636533 | - | 0.656958 |
| | CON | 0. 867292 | 0.868435 | - | 0.883922 |
| | TRU | 0. 889536 | 0.890508 | - | 0.904859 |
| NORMALIZED | 1NN | 0. 631518 | 0.643607 | 0.692826 | - |
| | CON | 0. 867292 | 0.872708 | 0.898833 | - |
| | TRU | 0. 889536 | 0.892184 | 0.96152 | - |
| NLM | 1NN | 0.585785 | 0.62461 | 0.618765 | - |
| | CON | 0.852738 | 0.875596 | 0.871919 | - |
| | TRU | 0.952199 | 0.96244 | 0.961883 | - |
| CCA | 1NN | 0.629326 | 0.650017 | 0.897077 | - |
| | CON | 0.847438 | 0.8747 | 0.972035 | - |
| | TRU | 0.819831 | 0.897908 | 0.965572 | - |

TABLE 3. Three goodness of fit measures used to evaluate each combination of cost function and optimization algorithm: 1NN = 1 Nearest Neighbour [13], CON = Continuity [14] and TRU = Trustworthiness [14].

### Landscapes of mtDNA Gene Trees

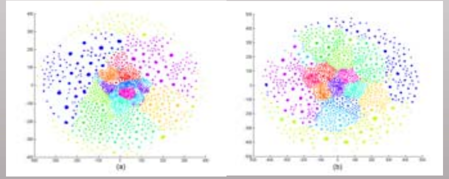

FIGURE 2. Two-dimensional projections of 6001 Mammals (a) and 7022 Fishes (b) non-parametric bootstrap trees using CCA with stochastic gradient descent.

## Plots of Tree-to-Tree distances

### Visual Inspection

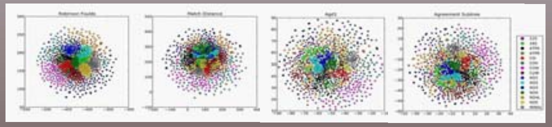

FIGURE 3. Two-dimensional projections of 1921 non-parametric bootstrap trees from the salamander data set using four tree-to-tree distance metrics (Robinson Foulds [12], Match Distance [15, 16], Agd1 [17], and Agreement Subtree [17]). The colors represent the underlying genes used to generate the trees. Projections were made using TreeScaper [18] with the cost function set to CCA and the optimization algorithm set to Stochastic Gradient Decent.

## Method of testing

### Data

| | Number of ML Bootstrap Trees | | | Number of ML Bootstrap Trees | |
|---|---|---|---|---|---|
| Gene | Unfiltered | Filtered | Gene | Unfiltered | Filtered |
| RAG1 | 120 | 116 | ND2 | 116 | 139 |
| ACT | 137 | 133 | PROX1 | 112 | 110 |
| KBTBD2 | 111 | 106 | SCFD2 | 113 | 113 |
| TOB101 | 161 | 145 | RAG2 | 116 | 121 |
| | | | **TOTALS** | **986** | **983** |

TABLE 4. The number of ML (GTR+ Γ+Pinvar) nonparametric bootstrap (100 replicates) trees and the number of characters in each gene partition before and after the DRUIDS filter.

## Results

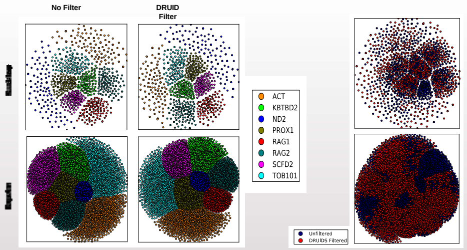

FIGURE 4. Projections of bootstrap and Bayesian trees obtained from the analysis of unfiltered and DRUIDS filtered alignments. Each locus was analyzed independently. RF-distances were calculated on concatenated sets of trees obtained from each analysis and RF-distances were projected using CCA and Stochastic Gradient Decent (i.e., a dimensionality reduction method). The colored points in the left projections represent trees favored by different loci. The colors in the right plots represent trees obtained from unfiltered and DRUIDS filtered alignments. No characters were removed by the DRUIDS filter for the SCFD2 locus.

## Quantitative Comparisons

| | 1NN | | | Random Index Method | | |
|---|---|---|---|---|---|---|
| Measure | Original | 2D | 3D | Original | 2D | 3D |
| Unfiltered | 0.997972 | 0.998986 | 0.998986 | 0.1397 | 0.1482 | 0.1453 |
| DRUID Filtered | 0.997965 | 0.997965 | 0.997965 | 0.1397 | 0.1456 | 0.1442 |

TABLE 5. Two cluster-based methods were used to quantify whether the DRUID filtered data lessened the distinction among sets of trees favored by different loci. Both the 1NN [13] and Random Index Methods suggest that filtering the data does not lessen the distinction, which is consistent with our visualizations.

## References

1.Hillis, D., Heath, T. & St John, K. Analysis and visualization of tree space. *SYSTEMATIC BIOLOGY* **54**, 471-482 (2005).
2. Alfaro, M.E and Huelsenbeck, J.P. (2006) Comparative performance of bayesian and AIC-based measures of phylogenetic model uncertainty. *Syst. Biol.*, 41, 89–96.
3. Bull, J.J. *et al.* (1993) Partitioning and Combining Data in Phylogenetic Analysis. *Syst. Biol.*, 42, 384–397.
4. Nylander *et al.* (2004) Bayesian phylogenetic analysis of combined data. *Syst. Biol.*, 53, 47-67.
5. Pagel, M. and Meade, A. (2004) A Phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol.*, 53, 571-581.
6. Maddison, W.P. (1997) Gene Trees in Species Trees. *Syst. Biol.*, 46, 523–536.
7. Sanderson, M.J. and Kim, J. (2000) Parametric phylogenetics?, *Syst. Biol.* 49, 817–829.
8. Fedrigo, O., et al. (2005). DRUIDS--detection of regions with unexpected internal deviation from stationarity. *Journal of experimental zoology. Part B, Molecular and developmental evolution*, *304*(2), 119-28.
9.Setiamarga, D. et al. 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 (2008).
10.Kjer, K.M. & Honeycutt, R.L. Site specific rates of mitochondrial genomes and the phylogeny of eutheria. *BMC Evol Biol.* **7**, 8 (2007).
11.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).
12.Robinson, D.F. & Foulds, L.R. Comparison of phylogenetic trees. *Math. Biosci* **53**, 131–147 (1981).
13.Van Der Maaten, L.J.P., Postma, E.O. & Van Den Herik, H.J. Dimensionality reduction: A comparative review. *Preprint* (2007).
14.Kaski, S. et al. Trustworthiness and metrics in visualizing similarity of gene expression. *BMC Bioinformatics* **4**, 48 (2003).
15. Bogdanowicz, D. 2008. Comparing phylogenetic trees using a minimum weight perfect matching. Proceedings of the 2008 1st International Conference on IT, Gdansk, Poland
16. Bogdanowicz, D. and Giaro, K. 2010. Comparing arbitrary unrooted phylogenetic trees using generalized matching split distance. Information Technology (ICIT), 2nd International Conference, 259–262.
17. Goddard, W., Kubicka, E., Kubicki, G. and McMorris, F. R. 1994. The agreement metric for labeled binary trees. Mathematical Bioscience 123:215-226.
18. Huang, W., et al. (2010). TreeScaper: software to visualize tree landscapes. *http://bpd.sc.fsu.edu/index.php/diagnostic-software*

## Acknowledgements

This work is supported in part by a grant (EF-0849861) from the National Science Foundation.