The Evaluation of Dimensionality Reduction Methods to Characterize Phylogenetic Tree-Space

James C. Wilgenbusch[1], Wen Huang[2], Kyle 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 systematically evaluate the performance of several nonlinear dimensionality reduction (NLDR) methods on tree-to-tree distances obtained from independent nonparametric bootstrap analyses of genes from three mid- to large-sized mitochondrial genome alignments.

Study Goals

1.Evaluate the performance and goodness of fit of several popular NLDR methods
2.Estimate the intrinsic dimensionality of tree-to-tree distances
3.Evaluate 2D and 3D projects
4.Compare the tree projects of different mtDNA data sets

Methods

Data

Taxa	Number of Sequences	Reference
Fishes	90	[2] Setiamarga et al., 2008
Mammals	89	[3] Kjer and Honeycutt, 2007
Salamanders	42	[4] 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 Nucleotides			
Gene	Fishes	Mammals	Salamanders	Fishes	Mammals	Salamanders	Color
12S	256	219	119	693	787	809	RED
16S	205	146	106	922	1199	1260	
ATP6	415	540	156	657	708	681	
ATP8	939	362	783	156	164	162	
COI	386	228	106	1539	1542	1548	
COII	444	433	196	690	682	681	GREEN
COIII	643	554	149	783	786	783	
CytB	235	195	122	1164	1140	1131	
ND1	507	170	111	933	969	957	
ND2	371	129	111	990	1048	1014	
ND3	690	1559	355	339	347	330	BLUE
ND4	219	150	108	1371	1384	1332	
ND4L	1362	1056	378	285	290	279	
ND5	188	114	103	1632	1801	1734	
tRNAs	162	146	108	1152	1339	1274	
TOTALS	7022	6001	3011	13306	14186	13975	

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 and calculated the unweighted Robinson-Foulds (RF) distance [5].

Compare NLDR Methods

Visual Inspection

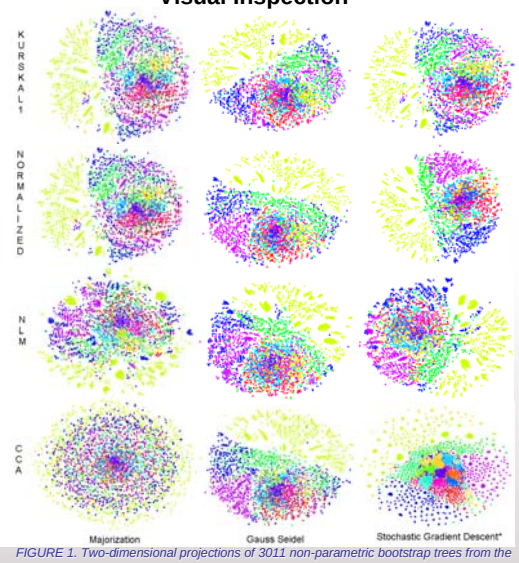

*FIGURE 1. Two-dimensional projections of 3011 non-parametric bootstrap trees from the salamander data set using three cost functions (y-axis) and three optimization algorithms (x-axis). The colors represent the underlying genes used to generate the trees (see color column in Table 2). * Kruskal-1 uses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

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 [6], CON = Continuity [7] and TRU = Trustworthiness [7].

Efficiency of Cost Function and Optimization

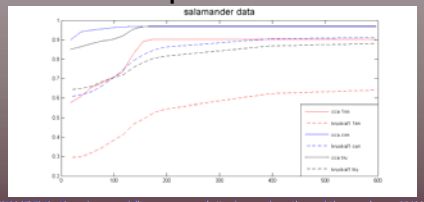

FIGURE 2. Goodness of fit measure plotted as a function of time, where 1NN = 1 Nearest Neighbour [6], CON = Continuity [7] and TRU = Trustworthiness [7].

Intrinsic Dimensionality

Visual Inspection

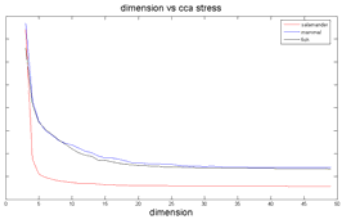

FIGURE 3. Raw stress of CCA plotted as a function of the projection dimensionality. Intrinsic dimensions of Salamander, Mammal and Fishes data are about 7, 15 and 15, respectively.

Analytical Measures

	Salamanders	Mammals	Fishes
NN	3.94	3.41	3.37
COR	5.27	11.77	14.35
ML	7.33	6.21	6.61
VIS	7	15	15

TABLE 3. Several analytical measures of intrinsic dimensionality of the three tree-to-tree distance matrices, where NN = Nearest Neighbour estimator [8,9], COR = Correlation Dimension [10,11], ML = Maximum Likelihood estimator [12]. and VIS result from figure 3 [13].

2D Versus 3D Projections

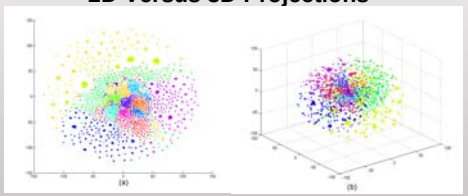

FIGURE 4. (a) Two- and (b) Three-dimensional projections of 3011 non-parametric bootstrap trees from the Salamander data sets using CCA with stochastic gradient descent.

Landscapes of mtDNA Gene Trees

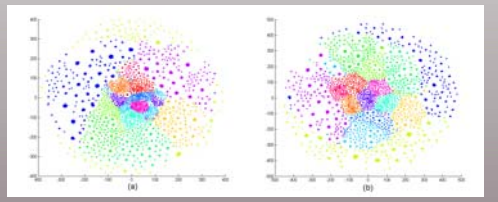

FIGURE 5. Two-dimensional projections of 6001 Mammals (a) and 7022 Fishes (b) non-parametric bootstrap trees using CCA with stochastic gradient descent.

References

1.Hillis, D., Heath, T. & St John, K. Analysis and visualization of tree space. *SYSTEMATIC BIOLOGY* **54**, 471-482 (2005).
2.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).
3.Kjer, K.M. & Honeycutt, R.L. Site specific rates of mitochondrial genomes and the phylogeny of eutheria. *BMC Evol Biol.* **7**, 8 (2007).
4.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).
5.Robinson, D.F. & Foulds, L.R. Comparison of phylogenetic trees. *Math. Biosci* **53**, 131--147 (1981).
6.Van Der Maaten, L.J.P., Postma, E.O. & Van Den Herik, H.J. Dimensionality reduction: A comparative review. *Preprint* (2007).
7.Kaski, S. et al. Trustworthiness and metrics in visualizing similarity of gene expression. *BMC Bioinformatics* **4**, 48 (2003).
8.Pettis, K.W., Bailey, T.A., Jain, A.K. & Dubes, R.C. An intrinsic dimensionality estimator from near-neighbor information. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1**, 25-37 (1979).
9.Verveer, P.J. & Duin, R.P.W. An evaluation of intrinsic dimensionality estimators. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **17**, 81--86 (1995).
10.Grassberger, P. & Procaccia, I. Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena* **9**, 189-208 (1983).
11.Camastra, F. & Vinciarelli, A. Estimating the intrinsic dimension of data with a fractal-based method. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**, 1404--1407 (2002).
12.Levina, E. & Bickel, P.J. Maximum likelihood estimation of intrinsic dimension. *Ann Arbor MI* **48109**, 1052 (2004).
13.Lee, J.A. & Verleysen, M. *Nonlinear dimensionality reduction*. (Springer Verlag: 2007).

Acknowledgements

This work is supported in part by a grant (EF-0849861) from the National Science Foundation. Discussions with Geoffrey Fox regarding the effects of different cost functions and optimization algorithms provided some of the motivation for this investigation.