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

Wen Huang[1], James Wilgenbusch[2], Kyle Gallivan[1]

[1]Department of Mathematics, [2]Department of Scientific Computing, Florida State University, Tallahassee, FL, USA

Introduction

Recent advances in molecular techniques have resulted in a rapid increase in the availability of genomic data sets. These data sets represent a rich foundation for researchers seeking to better understand the evolutionary history of extant organisms or genes. At the same time, this wealth of new data presents some significant computational challenges; e.g., adding more sequences to an analysis exponentially increases the number of candidate trees that must be evaluated and adding new genes can increase process heterogeneity among data partitions and confound the substitution models used to score trees. These challenges are driving the development of new tools and methods designed to alert biologists to possible analysis problems. One such tool involves the use of dimensionality reduction methods to characterize the relationship among competing evolutionary hypothesis, also known as phylogenetic tree space.

In this study, we systematically evaluate the performance of several linear and non-linear dimensionality reduction methods on tree-to-tree distances obtained from independent nonparametric bootstrap analyses of genes from three mid- to large-sized mitochondrial genome alignments. The practice of visually representing sets of competing phylogenetic trees in a geometric space can be separated into three major and sometime computationally intensive components: 1) the selection of a set of phylogenies to be compared (T); 2) the calculation of pairwise distances ($d_t(i,j)$) between all members of the set of selected trees (T); and 3) the calculation of coordinates in two or three dimensional space such that the Euclidean distance between the points ($d_x(i,j)$) closely corresponds to the original tree-to-tree distances ($d_t(i,j)$). We focus our comparisons on the last of these components.

Methods

Data

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

Taxa	Number of Sequences	Reference
Fishes	90	Setiamarga et al., 2008
Mammals	89	Kjer and Honeycutt, 2007
Salamanders	42	Zhang et al., 2008

Table 2. Phylogenetic trees were obtained for each of the three mtDNA data sets by performing a maximum likelihood (GTR+ Γ) nonparametric bootstrap (Felsenstein, 1985) analysis (100 replicates) on each of the 15-mtDNA genes.

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

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 (Robinson and Foulds, 1981).

Cost Functions, Algorithms, and Evaluation Metrics

We consider three cost functions in order to generate a low-dimensional Euclidean representation of the set of bootstrap trees:

$$E_{SMACOF} = \sum_{i=1,j<i}^{N} (d_t(i,j) - d_x(i,j))^2$$

$$E_{NLM} = \frac{1}{c} \sum_{i=1,j<i}^{N} \frac{(d_t(i,j) - d_x(i,j))^2}{d_t(i,j)}$$

$$E_{CCA} = \sum_{i=1,j<i}^{N} (d_t(i,j) - d_x(i,j))^2 F_\lambda(d_x(i,j)),$$

where $c = \sum_{i=1,j<i}^{N} d_t(i,j)$, $F_\lambda(d_x(i,j)) = \exp(-d_x(i,j)/\lambda)$, $d_t(i,j)$ is the distance between T_i and T_j and $d_x(i,j) = \|x_i - x_j\|_2$. The cost functions were optimized using the following algorithms; majorization, a deterministic Gauss-Seidel-Newton method, and a stochastic gradient descent method.

Two metrics were used to evaluate the quality of the non-linear dimensionality reduction. These are trustworthiness ($M_1(k)$) and continuity ($M_2(k)$) (Kaski et al., 2003):

$$M_1(k) = 1 - \frac{2}{Nk(2N - 3k - 1)} \sum_{i=1}^{N} \sum_{j \in U_k(i)} (r(i,j) - k)$$

$$M_2(k) = 1 - \frac{2}{Nk(2N - 3k - 1)} \sum_{i=1}^{N} \sum_{j \in V_k(i)} (\hat{r}(i,j) - k)$$

where $U_k(i)$: is the k-neighborhood of x_i, $r(i,j)$: is the rank of x_j in the neighborhood $U_k(i)$, $V_k(i)$: is the k-neighborhood of T_i, and $\hat{r}(i,j)$: is the rank of T_j in the neighborhood $V_k(i)$.

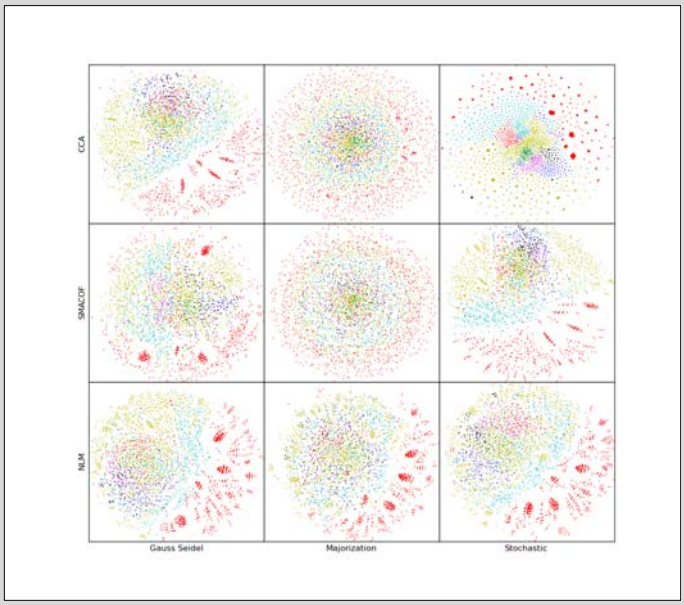

Figure 1. Two-dimensional projections of 3011 non-parametric bootstrap trees from the salamander data set using three cost functions (y-axis) and three optimization algorithms (x-axis). The colors represent the underlying genes used to generate the trees (see color column in Table).

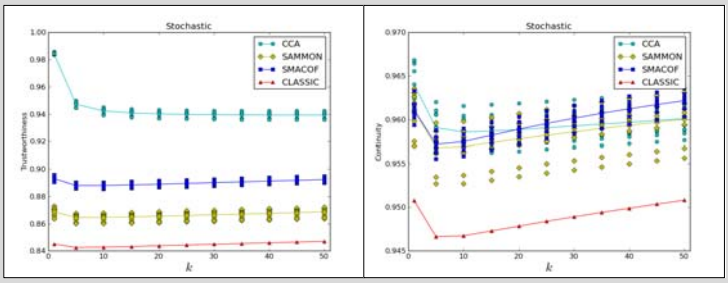

Figure 2. The trustworthiness (left) and continuity (right) for each cost function (with ten random staring point initializations) were plotted as a function of k, where k is defined as the size of the neighborhood around each point in the projection. The stochastic gradient decent method was used to optimize each cost function.

Conclusions

Visually representing phylogenic trees supported by different genes or by other *a priori* defined data partitions in 2D or 3D space is 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). For example, we demonstrate using several nonlinear dimensionality reduction (NDR) methods that different mtDNA genes favor different phylogenetic trees. This result likely indicates that the substitution model used to score trees is failing to adequately accommodate underlying process heterogeneity because mtDNA genes are non-recombining and should share a common history.

Furthermore, we reveal that different NDR methods significantly influence the interpretation of tree-to-tree distances when projected into 2D space. In particular, we found that the CCA cost function 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. Correctly characterizing phylogenetic tree-space by NDR methods is critical if this approach is to be of value as an interpretive or a diagnostic tool. For example, this method might also be used to alert practitioners to convergence problems where MCMC is used to infer phylogenies, conditions when heuristic search methods are inadequate, and partitions in combined phylogenetic data sets that do not share a common evolutionary history.

In the future we plan to use the analysis framework presented here to evaluate additional tree-to-tree distance metrics, dimensionality reduction costs functions, and optimization algorithms. We have also generated results not presented here, which indicate that visualizing the phylogenetic tree-space in 3D allows for better interpretation of these data. We plan to apply our findings and the software developed as a part of this project to help refine evolutionary models used to infer to more promising parts of the tree landscape.

Acknowledgements

We acknowledge the Florida State University shared High-Performance Computing facility and staff for contributions to results presented in this poster. This work was in part supported by a grant from the National Science Foundation (EF-0849861).

References

Felsenstein,J. (1985) Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution*, **39**, 783-791.

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 Evol Biol.*, **7**, 8.

Robinson,D.F. and Foulds,L.R. (1981) Comparison of phylogenetic trees. *Math. Biosci*, **53**, 131–147.

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.

Kaski,S. et al. (2003) Trustworthiness and metrics in visualizing similarity of gene expression. *BMC Bioinformatics*, **4**, 48.

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