Visualizing the Consequences of Model Mis-specification in Phylogenetic Landscapes

James Wilgenbusch[1], Wen Huang[2], Kyle A. Gallivan[2], Gavin Naylor[3], Shannon Corrigan[3], Nicolas Straube[3], and Chenhong Li[3]

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

[3]College of Charleston, Charleston, SC

INTRODUCTION

Large multilocus data sets are increasingly common and offer new opportunities to better understand the processes and patterns of evolution. These new data sets are not without challenges, however. 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 [1, 2, 3, 4] or because some data partitions simply do not share the same evolutionary history [5]. 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 [6].

We use a dimensionality reduction method (similar to [7]) 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.

Methods

1. The program DRUIDS [8] was used to identify nonstationary sites in each of the eight genes included in the multiple sequence alignment.
2. Maximum likelihood non-parametric bootstrap analyses using PAUP* [9] and Bayesian MCMC analyses using MrBayes [10] were conducted on the original alignment and on the DRUIDS filtered alignment for each locus.
3. The program PAUP* was used to calculate the unweighted Robinson-Foulds [11] tree-to-tree distance among all bootstrap trees and among the last 1000 MCMC trees. The last 1000 trees represents 1 mil. generations.
4. The program TreeScaper [12] was used to project in 2D and 3D different sets RF-distances obtained from different sets of concatenated phylogenetic trees.

	Number of ML Bootstrap Trees		Number of Nucleotides		
Gene	Unfiltered	Filtered	Unfiltered	Filtered	Color
RAG1	120	116	921	835	Red
ACT	137	133	456	444	Dark Orange
KBTBD2	111	106	1074	1004	Lime
TOB101	161	145	696	630	Aqua
ND2	116	139	999	905	Blue
PROX1	112	110	1041	960	Olive
SCFD2	113	113	510	510	Fuchsia
RAG2	116	121	699	679	Teal
TOTALS	**986**	**983**	**6396**	**5967**	

TABLE 1. 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.

Dimensionality Test	Unfiltered	DRUID Filtered
NN	6.91808	6.63336
COR	4.53759	4.57536
ML	17.2628	18.6626

TABLE 2. The intrinsic dimensionality of each tree-to-tree distance matrix was measured using three tests; NN = Nearest Neighbour estimator [13,14], COR = Correlation Dimension [15,16], and ML = Maximum Likelihood estimator [17].

Results

Visualizing Multi-Gene Landscapes

To test if clustering of related trees was caused by an artifact of the dimensionality reduction method, we plotted RF-distances of trees inferred from random sets of characters collected over the entire alignment. The size of the character sets corresponded with the size of the gene partitions. The projection of trees from the "shuffled" data set was then compared to the projection of trees obtained from the original alignment (Fig 1). This test was performed on an alignment of 42 salamander mtDNA sequences from another study.

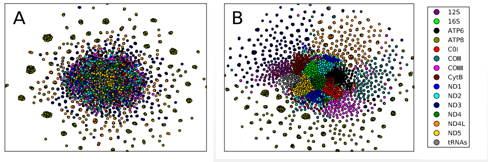

FIGURE 1. Projections of the RF-distances among bootstrap trees from separate analyses of 15 mtDNA data partitions using A) a test data set generated by shuffling columns in an alignment of Salamander sequences and B) in the original salamander alignment [18]. Colors correspond to the bootstrap trees favored by each separate data partition.

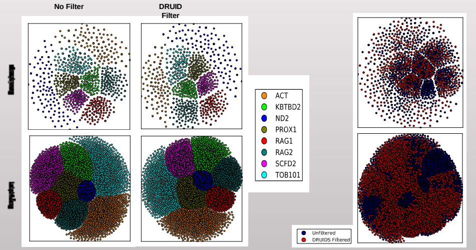

FIGURE 2. Projections of bootstrap and Bayesian trees obtained from the analysis of unfiltered and DRUIDS filtered alignments. Each locus was analyzed independently. RF-distances were calculated on concatenated sets of trees obtained from each analysis and RF-distances were projected using CCA and Stochastic 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.

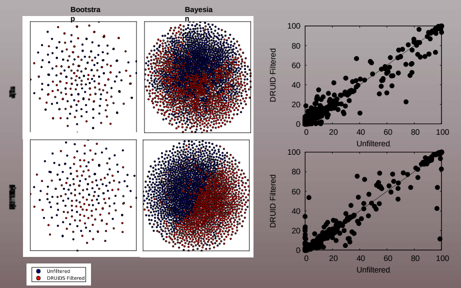

FIGURE 3. Projections of bootstrap and Bayesian trees obtained from the KBTBD2 and ACT loci using the unfiltered and DRUIDS filtered alignments. The bivariate scatter plots on the right represents the posterior probabilities of tree bipartitions with and without DRUIDS filtered characters. The top biplot represents ACT bipartitions and the bottom represents KBTBD2 bipartitions.

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 3. 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 [19] and Random Index Methods suggest that filtering the data does not lessen the distinction, which is congruent with our visualizations.

ML Bootstrap versus Bayesian MCMC

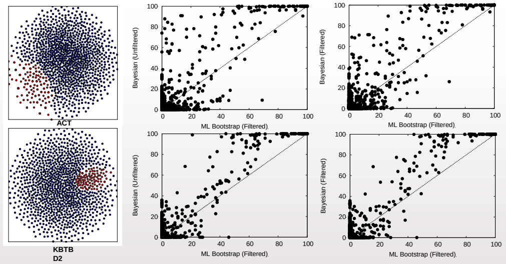

FIGURE 4. Plots on the left show the relationship among bootstrap trees (Red) and Bayesian MCMC trees (Blue). Bivariate plots show the relationship among bootstrap support values and Bayesian posterior probabilities for each bipartition. The DRUIDS filter did not obviously reconcile the difference between bootstrap support values and Bayesian posterior probabilities.

Observations

- Removing nonstationary characters from the Elasmobranch multiple sequence alignment did not significantly alter the relationship among bootstrap and Bayesian phylogenetic trees favored by different loci.
- Bootstrap trees from the DRUIDS filtered and the unfiltered alignments were indistinguishable.
- Bayesian MCMC trees from the DRUIDS filtered and the unfiltered alignments were noticeably different for each locus, suggesting that Bayesian MCMC analyses are more susceptible to model misspecification than are bootstrap analyses.
- Using DRUIDS to filter the Elasmobranch multiple sequence alignment did not help to reconcile differences between Bayesian MCMC posterior probabilities and bootstrap support values.

TreeScaper Software

FIGURE 5. A view of 2- 3D output generated by the software package TreeScaper. TreeScaper is freely available at:

http://bpd.sc.fsu.edu/index.php/diagnostic-software

References

1. 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.
2. Bull, J.J. *et al.* (1993) Partitioning and Combining Data in Phylogenetic Analysis. *Syst. Biol.*, 42, 384–397.
3. Nylander *et al.* (2004) Bayesian phylogenetic analysis of combined data. *Syst. Biol.*, 53, 47-67.
4. 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.
5. Maddison, W.P. (1997) Gene Trees in Species Trees. *Syst. Biol.*, 46, 523–536.
6. Sanderson, M.J. and Kim, J. (2000) Parametric phylogenetics?, *Syst. Biol.* 49, 817–829.
7. Hillis, D. et al. (2005)Analysis and visualization of tree space. *SYSTEMATIC BIOLOGY* **54**, 471-482.
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. Swofford, D. L. (2002). *PAUP*: phylogenetic analysis using parsimony (* and other methods).* version 4.0 b10.
10. Ronquist, F., et al. (2012). MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Sys Bio*. doi:10.1093/sysbio/sys029
11. Robinson, D.F. & Foulds, L.R. (1981) Comparison of phylogenetic trees. *Math. Biosci* **53**, 131–147.
12. Huang, W., et al. (2010). TreeScaper: software to visualize tree landscapes. *http://bpd.sc.fsu.edu/index.php/diagnostic-software*
13. Pettis, K.W., et al. (1979).An intrinsic dimensionality estimator from near-neighbor information. *IEEE Trans Pattern An and Machine Intel* **1**, 25-37.
14. Verveer, P.J. & Duin, R.P.W. (1995). An evaluation of intrinsic dimensionality estimators. *IEEE Trans on Pattern An and Machine Intel* **17**, 81–86.
15. Grassberger, P. & Procaccia, I. (1983) Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena* **9**, 189-208.
16. Camastra, F. & Vinciarelli, A. (2002) Estimating the intrinsic dimension of data with a fractal-based method. *IEEE Trans on Pattern An and Machine Intel* **24**, 1404–1407.
17. Levina, E. & Bickel, P.J. (2004) Maximum likelihood estimation of intrinsic dimension. *Ann Arbor MI* **48109**, 1092.
18. Zhang, P., et al. (2008).Phylogeny and biogeography of the family Salamandridae (Amphibia: Caudata) inferred from complete mitochondrial genomes. *MOLE PHYLO & EVO* **49**, 586-597.
19. Van Der Maaten, L.J.P. (2007) Dimensionality reduction: A comparative review. *Preprint.*

Acknowledgements

This work is supported in part by grants EF-0849861 and DEB-1132229 from the National Science Foundation.