Bayesian Analyses In Phylogenetic Paleontology: Interpreting the Posterior Sample

Graeme T. Lloyd
April M. Wright

Introduction:

Inferring the tree of life, and using phylogenetic trees to understand organismal form and function, is a primary challenge of biology. Phylogenetic trees provide a historical perspective on relationships between taxa, and performing comparative biological studies without a phylogenetic context is misleading (Felsenstein 1985). While many researchers in comparative biology have moved to using molecular data to estimate phylogenetic trees, fossils remain the sole source of character data for many extinct groups. Fossils provide the only direct evidence for historical trends in many groups (Cobbett and Wills 2007; Koch and Parry 2019). Excluding fossils from trees has been demonstrated to bias the inferences we draw from phylogenetic comparative analyses (Slater, Harmon, and Alfaro 2012). Therefore, how to best include fossils in phylogenetic analyses remains a vital and thriving research topic.

Phylogenetic analyses of discrete morphological data have traditionally been conducted under the maximum parsimony criterion, but Bayesian approaches are increasingly being employed instead. In part this is because Bayesian implementations allow not just estimation of topology, but more complex analyses, such as divergence time estimation. Bayesian methods assume an explicit model of evolution, and allow the researcher to set priors on each parameter in their model. In effect, this allows researchers to incorporate prior information about the value of parameters into their analysis. The most basic and restrictive model of morphological evolution is the Mk model (Lewis 2001). This model assumes equal rates of change between each state at a character. In practice, these assumptions are not much different than the basic assumptions of an equal-weights parsimony step matrix. However, these methods calculate a rate-based branch length in expected substitutions per site, allowing for the accommodation of multiple changes along a branch (e.g., a `0` state to a `1` state and back again), something parsimony cannot do. More elaborate models that allow for relaxation of the core assumptions of the Mk model have been published and evaluated (Nylander et al. 2004; Wright, Lloyd, and Hillis 2016).

Comparing the results of a Bayesian analysis and a parsimony analysis can be difficult. Bayesian methods estimate what is called the *posterior sample*. The posterior sample is a set of trees and associated model parameters that is plausible given the data, the model, and any priors. Phylogenetic trees are often estimated using Markov-Chain Monte Carlo (MCMC) sampling (cite Metropolis). Under this algorithm, a tree and model parameters are proposed and evaluated given the model and priors. In general, if the tree and model parameters have a better score than the previous tree and model parameters, the new solution is kept. MCMC is considered a memoryless process; that is, the next proposed tree and model parameters is not chosen based on previously-sampled trees. A good tree may therefore be visited many times in a phylogenetic estimation. In fact, how often a particular tree or set of relationships appears in the posterior sample is considered to be a measure of support for that tree. Usually the researcher will compute a summary tree for publication, though tools now exist to visualize sets of phylogenetic trees.

This is in stark contrast to parsimony. The aim of parsimony is to estimate the most parsimonious tree. The most parsimonious tree is the one that minimizes the number of character changes in the dataset implied by the tree. Under the criterion of maximum parsimony, a tree is proposed, and the changes implied in the morphological character matrix by that tree are counted. A tree is considered better than another if it has a lower parsimony score. Many datasets will have one tree that minimizes the parsimony score. In this case, a point estimate of the phylogeny is typically published by the author. However, in some datasets several most parsimonious trees may be returned, due to either lack of signal or conflicting signal. In this case, a summary tree is usually created. The most common of these is some form of consensus tree, which displays

all phylogenetic relationships that are present in some proportion of the most parsimonious trees. Common variations include a strict consensus tree, in which all clades on the summary tree are represented in all parsimonious trees, and majority rule consensus trees, in which clades on the summary tree are represented in more than 50% (or some larger fraction) of the set of most parsimonious trees.

How to compare a Bayesian posterior sample and a set of most parsimonious trees is a fraught topic. There are many aspects of parsimony analyses that are not strictly comparable to Bayesian analyses. For example, synapomorphies are interpreted differently in a Bayesian analysis, as multiple changes can occur along a branch. Bayesian analyses consider branch lengths (in substitutions per site for an undated tree) to be parameters of the model. Therefore, tree summarizations take these values into account (Bouckaert et al. 2014) Puttick????), not solely the topology. Support values for bipartitions in the tree are calculated as part of a Bayesian estimation, being the number of times that a set of relationships is contained in the posterior sample. The assumption is that the sample, not simply the "best" tree, contains vital information.

Because of these differences, formulating a fair comparison between parsimony and Bayesian trees is difficult. Most studies to date have focused on intrinsic comparisons, those comparisons about the tree itself. For example, most simulation studies to date have simulated data along a given tree or set of trees. Then, from the simulation data, a tree has been estimated under both parsimony and Bayesian methods. Finally, a summary tree for each method is computed and compared to the tree under which it was simulated. Often, this focuses on the behavior of the researcher, comparing a Bayesian consensus tree to a parsimony consensus tree (self-flagellate). Most phylogenetic estimates in published articles are presented as point estimates. Because parsimony trees are most commonly published as majority-rule or strict consensus trees, computing this type of summary for both treatments (parsimony and Bayesian) and comparing them is fairly straightforward. Although it is worth noting that a topology can appear multiple times in a Bayesian sample but never more than once in a parsimony sample. Comparisons have typically involved tree-based metrics, such as the Robinson-Foulds (Robinson and Foulds 1979, 1981), which supplies the number of bifurcations that differ between two trees. While this approach makes a degree of sense, it also does not include or account for most of the results associated with Bayesian estimation (the posterior sample).

An underutilised comparative tool between trees is their stratigraphic congruence - how well they match the order of appearance of taxa in the fossil record. Sansom et al. (2018) used stratigraphic congruence to compare parsimony and Bayesian summary trees to assess which analytical method produces trees that best fit the fossil record. The most attractive feature of this approach is that it employs extrinsic criteria, evaluating how well the tree fits data external to that used to infer the tree. Stratigraphic congruence methods (Table 1) use various measures to determine how well an estimated tree fits first and last appearance data for the fossil tips on this tree. Sansom et al. estimated trees for empirical datasets under both Maximum Parsimony and Bayesian estimation, computed consensus trees, and calculated stratigraphic fit for a sample of 500 trees from the posterior, then averaged those stratigraphic congruences. This is a novel way to independently compare the fit of trees produced under different optimality criteria, although previously it has been used to compare competing hypotheses of relationships (Brochu and Norell 2000). In this manuscript, we extend this approach to evaluate the whole posterior sample for stratigraphic fit and, critically, plot stratigraphic fit of individual trees in the posterior or the set of most parsimonious trees in "treespace" (Hillis, Heath, and John 2005). We conclude that averaging a posterior sample to a summary statistic is an oversimplification, and encourage the use of more sophisticated visual tools, such as RWTY (Warren et al. 2016) to incorporate the variation in any one metric.

Methods

Dataset Filtering

Empirical datasets were downloaded from graemetlloyd.com/matr.html, the same starting repository used by Sansom et al. (2018). We initially excluded any data sets that were molecular, phenetic, ontogenetic or meta-analytical as these do not represent data sets intended for morphological phylogenetic inference (our focus here). We additionally removed data sets with polymorphic or inapplicable characters as these are

currently not well dealt with by Bayesian implementations. In the repository, and more broadly amongst morphological phylogenetics, many datasets are derived from older datasets with little or no modification. For example, a matrix in the repository may have derivative matrices in which taxon or characters sampling was expanded from (see Figure S1 of (Hartman et al. 2019) for an excellent illustration of how complicated this issue can become). For this study, if there were multiple dataset derivatives, we chose the largest or most expansive dataset using the same criteria applied by Wright et al. (2016) and the appropriate metadata available in the XML files associated with each data set (Lloyd et al. 2016). The same XML files contain taxonomic reconciliations - links between OTU names used in the trees with taxa entered into the Paleobiology Database, the source used for our (and (Sansom et al. 2018)) age data. Using the API (Peters and McClennen 2016) we further filtered the data down to just those where all taxa are both reconciled and have age data available in the database (i.e., at least one fossil occurrence is attributable to each OTU). The final pool comprised 128 different data sets.

Phylogenetic Analysis

Each data set was analysed under both Maximum Parsimony and Bayesian criteria.

For the Bayesian approach we analysed each dataset under the Mk model in the software RevBayes (Höhna et al. 2014, @Hohna2016). We partitioned each dataset according to the number of states per character in order to specify an appropriate Q-matrix. Among-character rate variation was parameterized according to a Gamma distribution with four discrete categories. Branch lengths were drawn from an exponential distribution. Datasets were run for one million generations, and then assessed visually for convergence using Tracer.

Maximum Parsimony trees were re-estimated for each data matrix using TNT (Goloboff and Catalano 2016). Implicit enumeration was used where there were 24 or fewer tips. For larger tip counts 20 replicates of new technology searches followed by a round of tree bisection-reconnection were applied with maxtreees capped at 100,000.

Stratigraphic Congruence

We calculated stratigraphic congruence values for each tree in the Bayesian posterior sample, and each most parsimonious tree in the dataset. For each dataset, we used the `PaleobiologyDBOccurrenceQuerier` function in the `metatree` R package (github.com/graemetlloyd/metatree) to query the Paleobiology Database for minimum and maximum ages of each tip in the analysis. Using the R package `strap` (Bell and Lloyd 2015), we calculated several stratigraphic congruence metrics: Minimum Implied Gap (MIG), Stratigraphic Consistency Index (SCI), Relative Completeness Index (RCI), Gap Excess Ratio (GER), Manhattan Stratigraphic Measure (MSM*), and Wills' modifications of GER (GERt and GER*). An explanation of these metrics can be seen in Table 1.

Metric	Reference(s)	Meaning	Range
Stratigraphic Consistency Index (SCI)	(Huelsenbeck 1994)	Proportion of nodes for which the oldest descendent of that node is younger than the oldest descendent of that node's ancestor	0 to 1, with one being perfectly consistent
Minimum Implied Gap (MIG)	[Norell and Novacek (1992); Norell, Novacek, and Wheeler (1992)	The sum of the branch lengths excluding tip durations	Positive numbers in millions of years

Metric	Reference(s)	Meaning	Range
Relative Completeness Index (RCI)	(Benton and Storrs 1994)	MIG score proportional to the summed length of tip durations	All real numbers
Manhattan Stratigraphic Measure (MSM*)	[Siddall (1998); Pol and Norell (2001)	MIG for the maximally stratigraphically consistent possible tree divided by the actual MIG	0 to 1, with one being the most consistent
Gap Excess Ratio (GER)	(Wills 1999; Wills, Barrett, and Heathcote 2008)	MIG minus the best possible stratigraphic fit, scaled by the contrast between the best and worst fit values	0 to 1, with one being the most consistent

Table 1: - Summary of the common stratigraphic congruence metrics used. Here we prefer the MIG as our primary cmparative metric as it both captures absolute fit to the fossil record and is the primary driver of three other metrics (RCI, MSM, GER).

For each set of topologies and stratigraphic congruence results, we plotted a treespace plot in the R package `RWTY`. `RWTY` calculates the topological differences between the trees in the posterior sample using the Robinson-Foulds metric. This metric calculates the number of splits that are on one tree, but not the other. Then, `RWTY` uses a multi-dimensional scaling algorithm (Gower 1966) to plot these differences to 2-D space. We modified `RWTY` to color points according to their stratigraphic consistency score. We visualized the set of most parsimonious trees and the Bayesian posterior sample in the same treespace plots.

The full set of code and data used are available at: github.com/graemetlloyd/ProjectWhalehead.

Results

For direct comparison with Sansom et al. (2018), we made boxplots of the distributions of several stratigraphic congruence measures (Supplementary Figures SX-SZ?). We made these figures comparing the posterior sample and the set of most parsimonious trees for each dataset, and each stratigraphic congruence metric shown on Table 1. Note that a major difference between a Bayesian sample and a parsimony sample is that the former can include duplicate trees (those visited multiple times by the MCMC, whereas a parsimony analysis only returns unique topolog(ies)). One such boxplot can be seen in Fig. 1A. The spread of stratigraphic congruence metrics is much higher in the Bayesian analysis and this is typical of the majority of data sets. For most metrics, the median of the parsimony sample is lower than the median of the Bayesian sample. However, in most instances the stratigraphic congruence of the most parsimonious tree or set of trees is contained within the quartiles of the boxplot.

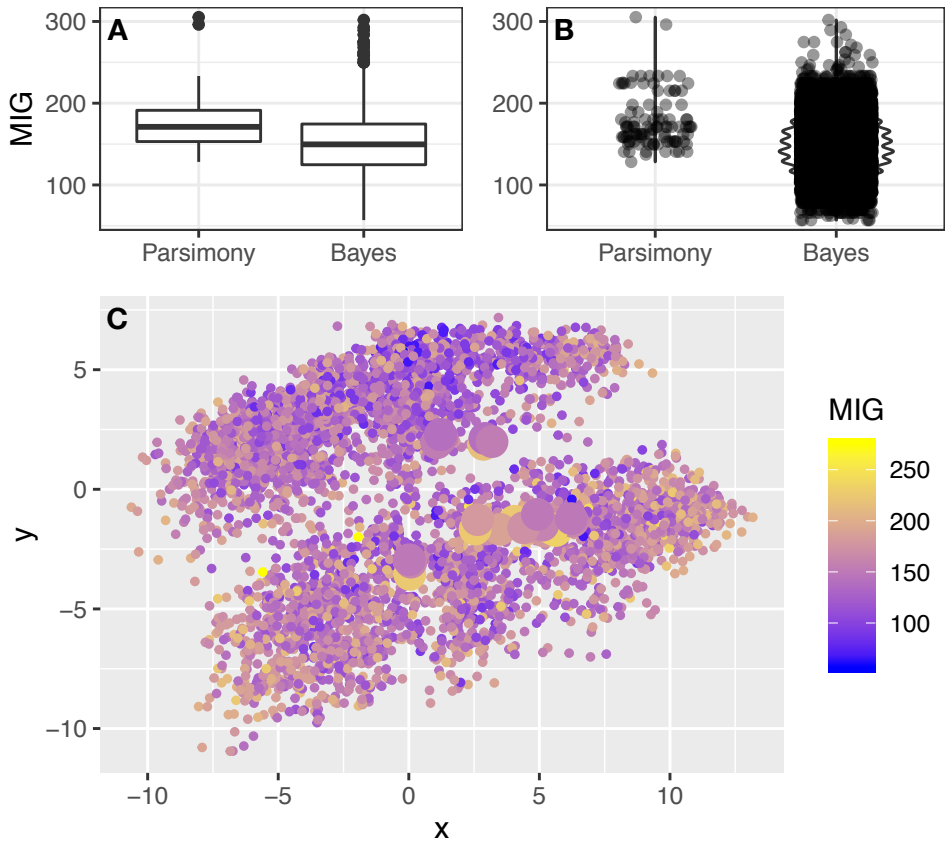

Fig. 1: Plots showing Minimum Implied Gap score (MIG) for the posterior sample and set of most parsimnonious trees for the Yates (2003) dataset. Panel A shows boxplots of the distributios of MIG scores, as in Sansom et al. (2018). Panel B shows the same data, but as a violin plot with points overlain. Panel C shows a treespace visualization, with points colored by MIG score. Large points indicate parsimony trees.

We also visualized these data as treespace plots. An example treespace plot can be seen in Fig. 1C, and the full set of treespace plots is available in the supplemental material. These multi-dimensional scaling graphs demonstrate that the Bayesian posterior samples contain many more trees, including many more trees that are substantially different from one another, than the parsimony estimations do. In most estimations, the posterior sample contains the parsimony trees, in addition to other solutions plausible under the model. As shown on Fig. 1C, it is very possible (and even common) in both Bayesian and maximum parsimony estimation to have topologies with good stratigraphic fit plotted near trees with poor stratigraphic fit (and for dissimilar trees to have similar or even identical stratigraphic fits). This indicates that in some treespaces, there may be little topological difference between a tree that is quite good with respect to stratigraphic fit and a tree with poor stratigraphic fit. In other words, the landscapes of stratigraphic congruence is generally highly volatile - something boxplots fail to capture. As in the boxplots, the treespace plots largely indicate that the parsimony trees fall within the range of solutions explored in a Bayesian search.

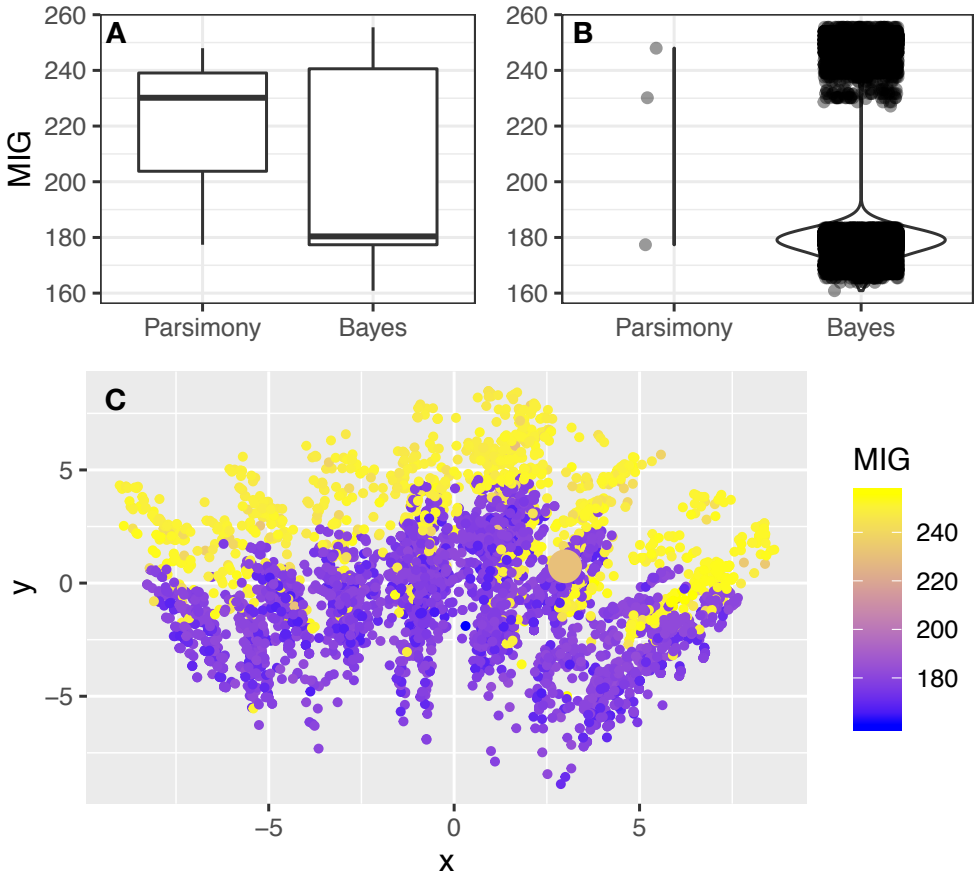

Fig. 2: Plots showing Minimum Implied Gap score (MIG) for the posterior sample and set of most parsimonious trees for the Demar (2013) dataset. Plot types and labels are the same as in Fig. 1. In contrast to Fig. 1, there are defined regions of treespace in which trees with better MIG scores are found.

Discussion

Summarizing a posterior sample

The aim of a Bayesian analysis is not a single point estimate of a solution. Rather, it is to examine solutions and outcomes that are plausible given the model and the data. This is particularly important in phylogenetics, where we are estimating lineage divergences that occurred tens or hundreds of millions of years ago. We do this from scarce, and likely biased data, using models that may or may not adequately capture reality. To responsibly present a solution under these conditions must mean incorporating uncertainty in that solution. It is prudent to avoid using a single point estimate summary of a posterior sample, whether a summary tree or an average value computed across several trees.

The authors of this work are not immune to using a point estimate of topology from a posterior sample to compare to a point estimate of a parsimony topology (see Wright and Hillis 2014).
In some cases, this is a pragmatic choice, given that consensus summaries are common lin the literature. But the properties of a Bayesian analysis are different than a parsimony analysis. A Bayesian analysis represents solutions that are plausible under a model. A parsimony analysis, on the other hand, pre-filters trees by whether or not they are the best under the maximum parsimony criterion. This means that the Bayesian

sample will almost certainly contain more sub-optimal trees than the parsimony criterion. This has the effect, when stratigraphic congruence measures are averaged for the posterior sample, of pulling down the mean of the distribution of stratigraphic consistency values.

There is an additional complication to interpreting a Bayesian posterior distribution: that the distribution itself is meaningful. Because a Markov Chain Monte Carlo analysis is memoryless, and solutions are sampled in proportion to their likelihood and the prior, a good solution will be visited many times. In fact, popular metrics of phylogenetic support, such as the posterior probability are calculated directly from the posterior sample for this reason. That the distribution of solutions itself contains information means that approaches, such as randomly sampling trees from the posterior, may give some sense of the variation. But they can also fail to approximate the true distribution of the posterior, particularly given that the variance in a posterior sample is unlikely to be normally distributed.

However, it is also not reasonable for researchers to examine every single tree in a posterior sample. How can a researcher draw reasonable conclusions from their posterior sample? Over the past five years, modern and open-source treespace visualization software has become available. As demonstrated in Fig. 1C (a treespace plot made from the posterior sample estimated from the Yates (2003) dataset), treespace visualizations show the distance between trees in the sample, and enable researchers to color points by other factors (formally generating so-called tree "landscapes"; (St. John 2016)). As can be seen in Fig. 1C, both the Bayesian and parsimony analyses predominantly sampled two islands of tree space. The points in this graphic show that topologically similar trees can have wildly different Minimum Implied Gap scores. For this dataset, it does not appear that certain regions of treespace produce better stratigraphic congruence. This is in contrast to Fig. 2C (a treespace plot made from the posterior sample estimated from the (Demar Jr 2013) dataset), in which some areas of treespace clearly contain trees that have better fit to the stratigraphic record.

Stratigraphic congruence and optimality criterion

Which analytical method is most congruent with stratigraphy is an all together murkier question when the fullness of the posterior sample is considered. A table of averages for Yates (2003) and Demar (2013) is provided in Table 2. Here we limit our consideration to just the MIG metric, as this is both a simple way to capture the key difference (total implied missing history in millions of years) and is the primary variable driving three of the four other metrics (GER, MSM*, and RCI). The other (SCI) is not favoured here as it treats all implied gaps in the record the same regardless of their duration. As can be seen from Table 2, sometimes the MIG favours the Bayesian posterior sample, sometimes the set of most parsimonious trees. Overall, 47 of 128 datasets (36.7%) had lower average MIG using Bayesian methods (Table S1). However, for 120 of 128 datasets (93.8%) the lowest MIG tree belonged to the Bayesian sample and for 127 of 128 (99.2%) datasets the Bayesian sample included the highest MIG tree.

Tree	Mean MIG	Max MIG	Min MIG
Demar - Bayesian	219	248	177
Demar - Parsimony	197	255	161
Yates - Bayesian	149	302	57
Yates - Parsimony	178	305	128

Table 2: Summary of the MIG values for the two figured example data sets. Note that this illustrates that neither Bayesian nor parsimony inferred trees consistently have the best fit (lowest MIG), contra Sansom et al. (2018).

An average may not be providing a good accounting of the variation in the results for each dataset. In Fig. 2A and 2B, we show a dataset from Demar et al (2013). The treespace for this dataset can be seen in Fig. 2C. This is a highly structured treespace: because Bayesian MCMC samples in proportion to the posterior probability, we can infer from the shape of this treespace visualization that there are two peaks to this distribution that contained fairly good trees. There are three most-parisimonious trees; two are sampled from one peak with poor stratigraphic fit, one is sampled from the peak with fairly good stratigraphic fit. In

this case, averaging is unlikely to produce a value that represents either peak adequately, and the average in this case is a value that doesn't belong to a tree that was sampled in the analysis at all.

In the case of the Yates (2003) dataset, treespace is sampled much more evenly. In this case, taking an average is likely to represent the sample a bit better. Even so, there are a number of problems with comparing means. As shown in the boxplots of Fig. 1, the mean and variance for the parsimony estimates are almost wholly subsumed in the Bayesian posterior sample. This is largely expected: the Bayesian posterior sample encompasses all trees sampled in the analysis (typically thinned by some proportion as they are exported from the tree estimation software). The set of most parsimonious trees contains a sample of the best trees according to an optimality criterion (fewest implied evolutionary steps). Assuming both estimation criteria are using the same data to sample the same treespace, we would expect that averaging among the best trees should produce better stratigraphic congruence. However, looking across all the datasets, only a slight improvement is seen from this biased averaging (Table S1, Sansom et al. 2018, their Fig. 1).

Which analytical method produces better stratigraphic congruence is the wrong question. A better question to ask may be "How can researchers explore and quantify variation in their sets of solutions?". Here we argue a key but underappreciated visualization tool (treespace) may be especially useful. Implementations of such tools have been in place for well over a decade (Hillis et al. 2005), but we are unaware of them being used in any published analysis of samples of trees produced from morphological data. We also note that we see many different "kinds" of treespace across the data sets examined here. For example, single or multiple-tree islands, smooth or volatile gradients of straigraphic congruence, parsimony samples enveloped by Bayesian samples, or parsimony and Bayesian samples occupying different parts of the space.

In this manuscript we have shown that "parsimony versus Bayesian" comparisons do not have a simple outcome that is consistent across all empirical data sets. Instead, we present a modified version of an underutilised set of tools already available in the R programming language that can better help workers visualise and understand large posterior samples, either on their own or in the context of maximum parsimony trees. In particular, we urge the exploration of treespace visualisations to better understand the output from phylogenetic inference. We show that these tools reveal a huge variety of outcomes in the occupation, shape and landscape of such spaces. It is our expectation that increased application of these approaches will reveal new and interesting

Conclusion

In writing this manuscript, we used and modified a set of tools already available in the R programming language. These tools, such as the package `RWTY`, enabled us to read in many large posterior samples, and to calculate treespace plots across 128 empirical datasets. The code and data available for this are freely available on GitHub. We would like to close this manuscript by noting that the tools to perform complex formatting and subsetting datasets, including large ones such as a Bayesian posterior sample, are more accessible than ever. We look forward to many years of interesting analyses about how different methods explore and sample phylogenetic tree space.

References

Bell, Mark A, and Graeme T Lloyd. 2015. "Strap: An R Package for Plotting Phylogenies Against Stratigraphy and Assessing Their Stratigraphic Congruence." *Palaeontology* 58: 379–89.

Benton, Michael J., and G. W. Storrs. 1994. "Testing the Quality of the Fossil Record: Palaeontological Knowledge Is Improving." *Geology* 22: 111–14.

Bouckaert, Remco, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A. Suchard, Andrew Rambaut, and Alexei J. Drummond. 2014. "BEAST 2: A Software Platform for Bayesian Evolutionary Analysis." *PLOS Computational Biology* 10 (4): 1–6. https://doi.org/10.1371/journal.pcbi.1003537.

Brochu, Christopher A, and Mark A Norell. 2000. "Temporal Congruence and the Origin of Birds." *Journal of Vertebrate Paleontology* 20 (1): 197–200.

Cobbett, Wilkinson, A., and M. A. Wills. 2007. "Fossils Impact as Hard as Living Taxa in Parsimony Analyses of Morphology" 56: 753–7666.

Demar Jr, David G. 2013. "A New Fossil Salamander (Caudata, Proteidae) from the Upper Cretaceous (Maastrichtian) Hell Creek Formation, Montana, Usa." *Journal of Vertebrate Paleontology* 33 (3): 588–98.

Felsenstein, Joseph. 1985. "Phylogenies and the Comparative Method." *An*, 1–15.

Goloboff, Pablo A, and Santiago A Catalano. 2016. "TNT Version 1.5, Including a Full Implementation of Phylogenetic Morphometrics." *Cladistics* 32 (3): 221–38.

Gower, John C. 1966. "Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis." *Biometrika* 53 (3-4): 325–38.

Hartman, Scott, Mickey Mortimer, William R Wahl, Dean R Lomax, Jessica Lippincott, and David M Lovelace. 2019. "A New Paravian Dinosaur from the Late Jurassic of North America Supports a Late Acquisition of Avian Flight." *PeerJ* 7: e7247.

Hillis, D. M., T. A. Heath, and K. S. John. 2005. "Analysis and Visualization of Tree Space." *Sysbio* 54 (3): 471–82.

Höhna, Sebastian, Tracy A Heath, Bastien Boussau, Michael J Landis, Fredrik Ronquist, and John P Huelsenbeck. 2014. "Probabilistic Graphical Model Representation in Phylogenetics." *Systematic Biology* 63 (5): 753–71.

Höhna, Sebastian, Michael J Landis, Tracy A Heath, Bastien Boussau, Nicolas Lartillot, Brian R Moore, John P Huelsenbeck, and Fredrik Ronquist. 2016. "RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language." *Systematic Biology* 65 (4): 726–36.

Huelsenbeck, John P. 1994. "Comparing the Stratigraphic Record to Estimates of Phylogeny." *Paleobiology* 20 (4): 470–83.

Koch, N. M., and L. A. Parry. 2019. "Death Is on Our Side: Paleontological Data Drastically Modify Phylogenetic Hypotheses." *In Review*.

Lewis, Paul O. 2001. "A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data." *Systematic Biology* 50 (6): 913–25.

Lloyd, GT, DW Bapst, M Friedman, and KE Davis. 2016. "Probabilistic Divergence Time Estimation Without Branch Lengths: Dating the Origins of Dinosaurs, Avian Flight and Crown Birds." *Biology Letters* 12 (11): 20160609.

Norell, Mark A, and Michael J Novacek. 1992. "Congruence Between Superpositional and Phylogenetic Patterns: Comparing Cladistic Patterns with Fossil Records." *Cladistics* 8 (4): 319–37.

Norell, Mark A, MJ Novacek, and QD Wheeler. 1992. "Taxic Origin and Temporal Diversity: The Effect of Phylogeny." *Extinction and Phylogeny. Columbia University Press, New York*, 89–118.

Nylander, Johan AA, Fredrik Ronquist, John P Huelsenbeck, and Joséluis Nieves-Aldrey. 2004. "Bayesian Phylogenetic Analysis of Combined Data." *Systematic Biology* 53 (1): 47–67.

Peters, Shanan E, and Michael McClennen. 2016. "The Paleobiology Database Application Programming Interface." *Paleobiology* 42 (1): 1–7.

Pol, Diego, and Mark A Norell. 2001. "Comments on the Manhattan Stratigraphic Measure." *Cladistics* 17 (3): 285–89.

Robinson, DF, and LR Foulds. 1979. "Comparison of weighted labelled trees." *Combinatorial Mathematics, VI (Proc. Sixth Austral. Conf., Univ. New England, Armidale, 1978), Lecture Notes in Mathematics* 748: 119–26.

———. 1981. "Comparison of phylogenetic trees." *Math. Biosci* 53 (1-2): 131–47.

Sansom, Robert S, Peter G Choate, Joseph N Keating, and Emma Randle. 2018. "Parsimony, Not Bayesian Analysis, Recovers More Stratigraphically Congruent Phylogenetic Trees." *Biology Letters* 14 (6): 20180263.

Siddall, Mark E. 1998. "Stratigraphic Fit to Phylogenies: A Proposed Solution." *Cladistics* 14 (2): 201–8.

Slater, Graham J, Luke J Harmon, and Michael E Alfaro. 2012. "Integrating Fossils with Molecular Phylogenies Improves Inference of Trait Evolution." *Evolution: International Journal of Organic Evolution* 66 (12): 3931–44.

St. John, Katherine. 2016. "The Shape of Phylogenetic Treespace." *Systematic Biology* 66 (1): e83–e94.

Warren, DL, A Geneva, DL Swofford, and R Lanfear. 2016. "Rwty: R We There yet." *A Package for Visualizing MCMC Convergence in Phylogenetics*.

Wills, Matthew A. 1999. "Congruence Between Phylogeny and Stratigraphy: Randomization Tests and the Gap Excess Ratio." *Systematic Biology* 48 (3): 559–80.

Wills, Matthew A, Paul M Barrett, and Julia F Heathcote. 2008. "The Modified Gap Excess Ratio (Ger*) and the Stratigraphic Congruence of Dinosaur Phylogenies." *Systematic Biology* 57 (6): 891–904.

Wright, April M., Graeme T. Lloyd, and David M. Hillis. 2016. "Modeling Character Change Heterogeneity in Phylogenetic Analyses of Morphology Through the Use of Priors." *Sysbio* 65 (4): 602–11.