OXFORD UNIVERSITY PRESS | Systematic Biology

Morphological Datasets Fit a Common Mechanism Much More Poorly than DNA Sequences and Call Into Question the Mkv Model

Journal:	*Systematic Biology*
Manuscript ID	USYB-2018-192.R1
Manuscript Type:	Regular Manuscript
Date Submitted by the Author:	23-Oct-2018
Complete List of Authors:	Goloboff, Pablo; INSUE-CONICET, Pittman, Michael; The University of Hong Kong, Department of Earth Sciences Pol, Diego; CONICET - Museo Paleontológico Egidio Feruglio, Xu, Xing; Institute of Vertebrate Paleontology and Paleoanthropology Chinese Academy of Sciences, Key Laboratory of Vertebrate Evolution and Human Origins
Keywords:	phylogenetics; mkv model; common mechanism; parsimony

RUNNING-TITLE: Morphological datasets reject Mkv model

Morphological Datasets Fit a Common Mechanism Much More Poorly than DNA Sequences and Call Into Question the Mkv Model

Pablo A. Goloboff[1]*, Michael Pittman[2], Diego Pol[3], and Xing Xu[4]

[1] Unidad Ejecutora Lillo (UEL), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), S.M. Tucumán, Argentina. E-mail: pablogolo@yahoo.com.ar.

[2] Vertebrate Palaeontology Laboratory, Department of Earth Sciences, University of Hong Kong, Pokfulam, Hong Kong

[3] Museo Egidio Feruglio, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Trelew, Argentina

[4] Key Laboratory of Vertebrate Evolution and Human Origins, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing, China

* Corresponding Author

Abstract

The Mkv evolutionary model, based on minor modifications to models of molecular evolution, is being increasingly used to infer phylogenies from discrete morphological data, often producing different results from parsimony. The critical difference between Mkv and parsimony is the assumption of a "common mechanism" in the Mkv model, with branch lengths determining that probability of change for all characters increases or decreases at the same tree branches by the same exponential factor. We evaluate whether the assumption of a common mechanism applies to morphology, by testing the implicit prediction that branch lengths calculated from different subsets of characters will be significantly correlated. Our analysis shows that DNA (38 datasets tested) is often compatible with a common mechanism, but morphology (86 datasets tested) generally is not, showing very disparate branch lengths for different character partitions. The low levels of branch length correlation demonstrated for morphology (fitting models without a common mechanism) suggest that the Mkv model is too unrealistic and inadequate for the analysis of morphological datasets.

KEYWORDS: phylogenetics, Bayesian analysis, morphological data, Mkv model

Discrete morphological characters, despite the predominance of molecular datasets, continue playing an important role in inferring phylogenetic trees (e.g. as the sole source of evidence for most fossil taxa). Parsimony (implemented in PAUP*, Swofford 2002, or TNT, Goloboff and Catalano 2016) is widely used for morphological data. The Mkv model (Lewis 2001), based on minor modifications to models of molecular evolution, is being increasingly used for phylogenetic inference (Wright and Hillis 2014, O'Reilly et al. 2016, Puttick et al. 2017), even when it is often acknowledged that morphology and molecules may evolve in very different ways (e.g. Lee 2016, Zhang 2018). The Mkv model is implemented in several major phylogeny programs such as PAUP*, MrBayes (Ronquist et al. 2012), or RAxML (Stamatakis 2014). The Mkv model critically differs from parsimony in assuming a "common mechanism" (CM, Tuffley and Steel 1997), in which the probability of change in different tree branches varies simultaneously for all characters, exponentially depending on the "length" of the branch (expected number of changes per character, the product of time and instantaneous rate, both affecting all characters equally; for details, see Swofford et al. 1996, Felsenstein 2004). This assumption of a CM is in fact what Lewis (2001: 915-916) considered that systematists would likely find most unrealistic. Eliminating this commonality assumption causes parsimony and likelihood to select the same tree (e.g. with the "no-common-mechanism" model of Tuffley and Steel 1997, Steel 2013; NCM); if the data have indeed not evolved with common branch lengths (e.g. with heterotachy), parsimony may produce better results than model-based methods that assume homogeneity (Kolaczkowski and Thornton 2004, Goloboff et al. 2017).

Although there have been no empirical comparisons of molecules and morphology in terms of their fit to a CM, patterns of change in discrete morphological characters seem not to follow this assumption of commonality. Sets of characters highly variable in a group are often almost invariable in another, where different characters become highly variable instead (e.g. Farris 1983:15, Sereno 2009), suggesting that the Mkv model may be inappropriate for morphological data (Goloboff and Pol 2005, Nyakatura and

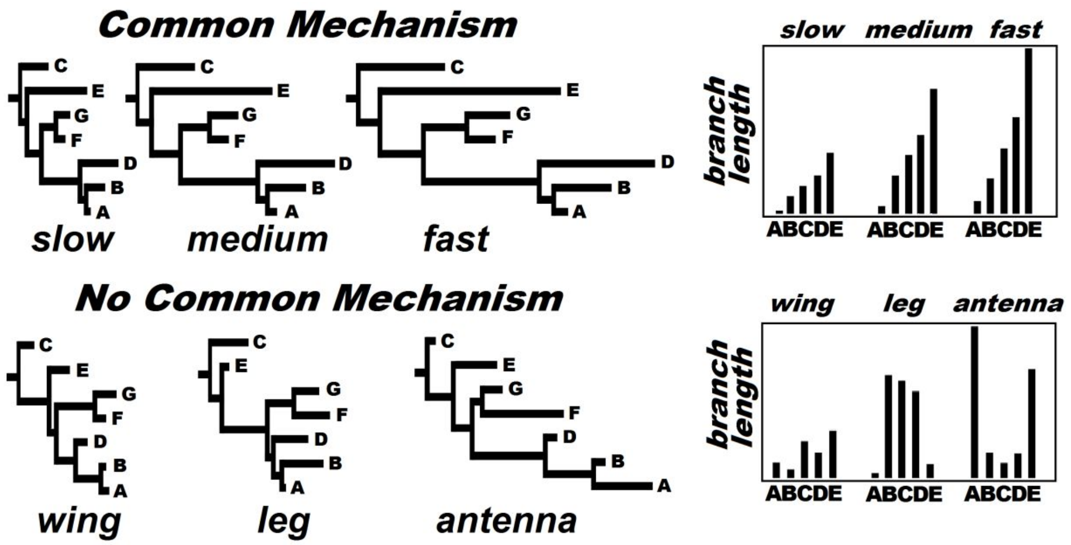


Figure 1. Differences between common and no-common mechanisms. Under a common mechanism, there can be slower and faster characters (e.g. with a gamma distribution), but branch lengths (expected changes per character, product of time and instantaneous rate of change for the branch) increase or decrease for all the characters together. This is shown in the barplot diagram, with five branches of the tree, A–D, ordered in increasing length. Without a common mechanism, there can be characters with different overall rates (e.g. wing, leg, and antenna), but the expected changes show no correlation between the different characters. The branch leading to taxon A is intermediate in the first character, shortest in the second, and longest in the third, while the branch leading to taxon B is shortest in the first character, longest in the second, and intermediate in the third, and the shapes of length distributions vary for the three characters.

Bininda-Emonds 2012). With methods like the discretized gamma distribution (see details in Felsenstein 2004), the Mkv model allows for rate heterogeneity among characters, but this still assumes that the expected changes per character increase or decrease, together, for faster and slower evolving characters, along the same branches of the tree, as illustrated in Figure 1. The patterns of change in morphological characters would seem instead to depart strongly from that CM, both at the level of character partitions, and individual characters. Multiple (unlinked) partitions (Duchene et al. 2014, Lanfear et al. 2017) allow expected changes per character at a branch to change separately in each partition, but are rarely used in morphological datasets and continue requiring both the commonality assumption within each partition and a prior identification of the correct partitions.

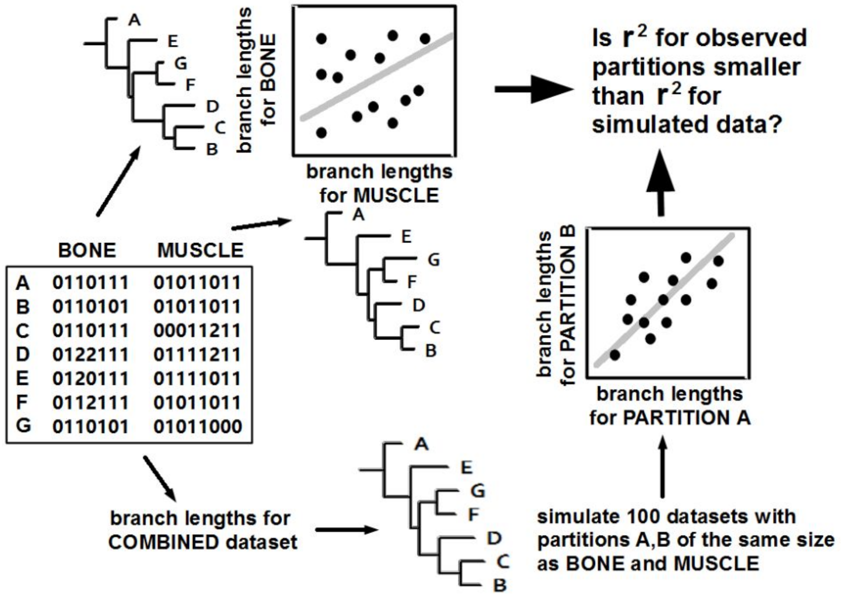
The present study evaluates, for the first time, the assumption of a CM for morphological datasets. Bayesian model selection has been applied in some studies to evaluate differences between morphological partitions, but only to assess among-character rate variation (e.g Harrison and Larsson 2015), or the fit of different partitions to alternative rate parameters (Lanfear et al. 2017, Clarke and Middleton 2008), instead of critically evaluating the adequacy of a CM. **Model selection may be problematic when both models compared are incorrect (Yang and Zhu 2018), and can only be used to compare two alternative models (instead of testing whether a single model has an acceptable fit). The latter becomes particularly difficult when the alternatives to CM are to be sought among phylogenetic methods approaching parsimony: Tuffley and Steel's (1997) NCM is equivalent to parsimony, but (as noted by Holder et al. 2010: 478; see also Sober 2004) NCM is too highly parameterized to be ever selected, and probably not the only way to characterize parsimony –yet no currently available implementation emulates parsimony methods with fewer parameters, to enable a more meaningful comparison of likelihoods.**

Given those difficulties, we use here an approach based on statistical hypothesis testing, to assess the adequacy of the Mkv model for morphological datasets. The CM of the Mkv model predicts that branch lengths for different subsets of data will be correlated, and our test is based on evaluating whether that prediction is met in empirical datasets. The paper begins by outlining the test and its justification, then applies it to morphology and DNA sequences, first to partitions predefined on the basis of contiguity (DNA) or anatomy (morphology), then to randomly defined subpartitions (within predefined partitions, and whole datasets). As these tests show that the vast majority of morphological datasets do not conform to a CM, we then apply a similar test to evaluate the alternative: whether the degrees of correlation between branch lengths in morphological datasets could have been produced by models without a CM. These tests reject a pure NCM, but an alternative model for generating datasets without a CM (which we call the *episodic* model, with character changes restricted to certain parts of the tree; see below) produces a correlation between numbers of character changes in each partition that is well within the values observed in real morphological datasets. Finally, we examine the relative performance of phylogenetic methods on datasets simulated under the episodic model, and show that parsimony tends to perform on par or better than Bayesian analysis.

METHODS AND MATERIALS

Datasets.– Source and details in the Supplementary Material. A total of 86 morphological datasets was used, with 26–188 taxa, and 80–4541 characters. For 8 of the morphological datasets, it was possible to define partitions on the basis of anatomy (with 2–8 partitions per dataset, 40–1451 characters per partition). For sequences, a total of 38 datasets for 35 different genes, with 60–500 taxa and 305–2218 characters, was examined. These molecular datasets were prepartitioned in 2–5 partitions of 100–500 contiguous positions, depending on the length of the sequences (only one case, with very short sequences, used two partitions of 50 positions). These molecular partitions were created leaving out the initial 100 positions (which, due to alignment, often contain large proportions of missing entries), except in the shortest sequences (where partitions started at position #50).

Branch length tests.– Branches shorter for one partition and longer for the other are evidence of at least heterotachy (or, at most, the complete absence of a CM). The strength of the observed correlation can be measured with the r^2 statistic; the observed r^2 was then compared with that for partitions of the same size generated on a model tree with the same branch lengths as the combined dataset; if observed r^2 is matched with a low probability, then the CM of the Mkv model can be confidently rejected. Figure 2 displays the


A
E
G
F
D
C
B
branch lengths for BONE
branch lengths for MUSCLE
Is r^2 for observed partitions smaller than r^2 for simulated data?
A
E
G
F
D
C
B
branch lengths for PARTITION B
branch lengths for PARTITION A
branch lengths for COMBINED dataset
A
E
G
F
D
C
B
simulate 100 datasets with partitions A,B of the same size as BONE and MUSCLE

	BONE	MUSCLE
A	0110111	01011011
B	0110101	01011011
C	0110111	00011211
D	0122111	01111211
E	0120111	01111011
F	0112111	01011011
G	0110101	01011000

Figure 2. General scheme of the test to evaluate significance of heterogeneity between branch lengths for different partitions.

procedure for testing branch length homogeneity in two partitions. The only similar evaluation of which we are aware is that of Clarke and Middleton (2008), who compared branch lengths for different morphological partitions; however, they did not evaluate the significance of the differences in branch lengths by reference to a specific model of evolution.

Even when the data have been generated by a model with a CM (e.g. Mkv or JC69) the expected homogeneity in branch lengths for the simulated partitions will depend on both the branch lengths of the model tree, as well as the numbers of characters in the two partitions being compared. To the extent that the branch lengths of the model tree are more dissimilar, the correlation between branch lengths for two sets of characters generated on the same model tree will be stronger; when all branch lengths of the model tree are identical, character changes can be located equiprobably on any tree branch, resulting in very low correlation. On the other hand, to the extent that there are more characters in the partitions, branch lengths will more accurately converge to the values in the model tree, thus increasing the correlation between the branch lengths for both partitions. Therefore, a proper test cannot be based solely on the observed value of r^2 for the correlation between branch lengths for two partitions: the values of r^2 must be compared against the values expected under the specific situation being tested, i.e. using the same numbers of characters of the observed partitions, and a model tree with the same branch lengths as the combined dataset.

For completeness, most of the tests were repeated calculating branch lengths with most parsimonious reconstructions (MPR). In this case, the scripts calculated branch lengths simply as the number of characters in the partition unambiguously changing along the branch, divided by the total number of characters in the partition.

Calculation of Branch Lengths.– Branch lengths for the results reported were calculated using maximum likelihood, unless noted otherwise. Taxa with missing entries for all characters in one (or both) partition(s) were pruned from the tree, and the branch lengths were calculated on the resulting reduced tree. This was necessary only in few comparisons. TNT scripts (Goloboff et al. 2008) automatically created Nexus files and called PAUP* with commands to calculate and save branch lengths in Newick format, then reading back the branch lengths into TNT, for further processing. For morphological datasets, invariant characters were excluded (for different pairwise comparisons between partitions, some of the variable characters in a partition could become invariant if some taxa with only missing entries in the other partition are deactivated). For morphological data, branch lengths were calculated with default PAUP* options (in the absence of invariant characters, PAUP* defaults to the Mkv model, estimating the proportion of invariant characters automatically). For sequence data, the simplest model (JC69, Jukes and Cantor 1969) was invoked, with *lset nst=1 rates=equal basefr=equal*, which is the closest equivalent to the Mkv model (except for the estimation of invariant characters, which has a minimum effect on branch length proportionality). Invoking more complex DNA models and adding more parameters to be estimated seemed unnecessary, given that the goal of the analysis is only evaluating the heterogeneity in branch lengths for different partitions.

Model tree.– The datasets for calculating the statistical distribution of the correlation between partitions with the same numbers of characters as the observed partitions were simulated using the observed tree as model. The “observed” tree is the published tree, when available, or a most parsimonious tree for the combined dataset otherwise (in the case of phylogenomic datasets, this is a tree for the dataset combining all the genes). We did experiments to confirm that the test does not strongly depend on the topology of the tree used to calculate branch lengths, so even if the observed tree is slightly different from the correct phylogeny, the results of the test continue being valid (See Supplementary Material). This makes the test radically different from "empirical" comparisons where real datasets are analyzed with different methods of phylogenetic inference and the resulting groupings are evaluated on whether they agree with groupings presumed to be correct prior to the analysis (e.g. Puttick et al.'s 2017 discussion of results for 4 empirical datasets). No presumption of prior knowledge is needed for the present correlation test, which considers only the fit of the model to the dataset, not the accuracy of the trees produced by assuming the model.

Common Model Tested Between Pre-Defined Partitions

We first tested 8 large published matrices, containing partitions corresponding to anatomical regions or organ systems with numerous characters (40 characters per partition was considered as the minimum for appropriate testing). Given the different numbers of partitions per dataset, a total of 79 pairwise comparisons were possible. The vast majority of these partitions (Figs. 3, 4) have much more pronounced differences in numbers of character changes along branches than expected under the Mkv model (only 3.8% of cases fail to reject the Mkv model as null model with α=0.01; Fig. 3a). The results of a similar test performed on DNA sequences (38 datasets for 35 different genes, with partitions defined by contiguity, 66 possible comparisons) are very different, with a common mechanism accepted for 42.4% of comparisons (Fig. 3a), over ten times more frequently than for morphology. The average probabilities of observed r^2 values under a CM are also much higher for DNA than for morphological datasets (Fig. 3c, 4). Therefore, branch lengths for partitions of DNA sequences are clearly much less heterogeneous than for morphological data.

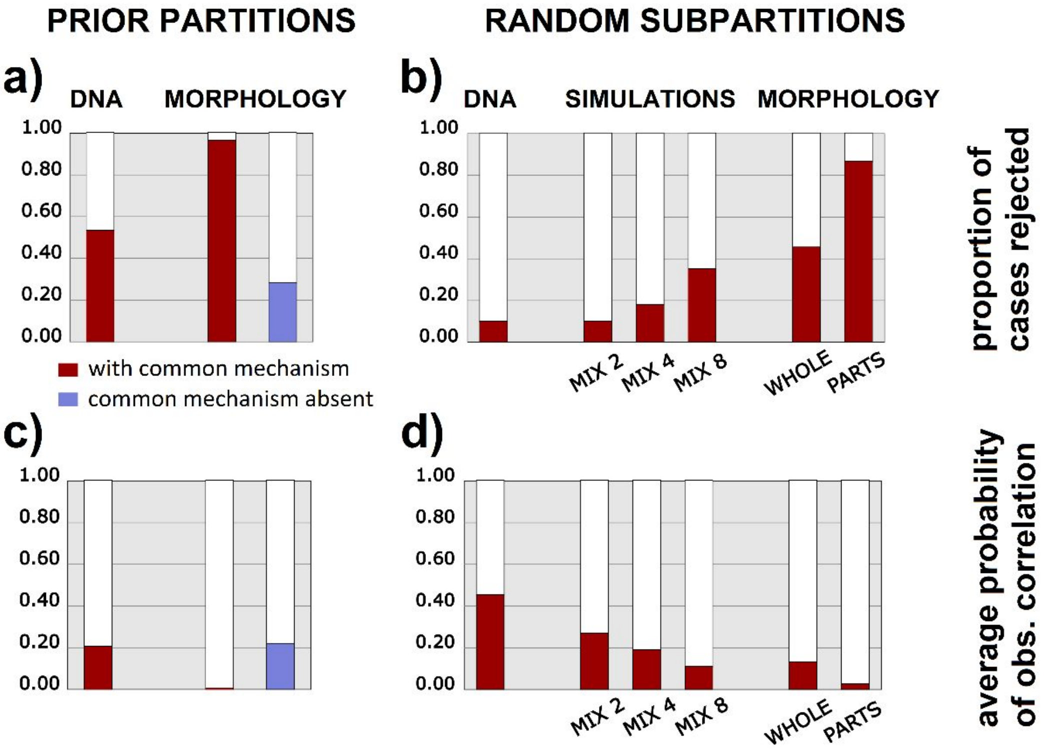


Figure 3. Proportion of cases where different models are rejected with α=0.01 by branch length tests (a, b), and average probabilities of observed correlation (c, d). Prior partitions (a, c) correspond to characters grouped on the basis of anatomy in the case of morphology, and on the basis of contiguity in the case of DNA. For morphology, the common mechanism model is Mkv; for DNA, its closest equivalent, JC69. The model without a common mechanism is the Episodic model described in the text. Random subpartitionings (b, d) for DNA were tested on a mid-sequence group of positions, on whole datasets simulated with mixtures of 2–8 independent sets of branch lengths (MIX), on the partitions predefined on the basis of anatomy (PARTS), and on whole datasets when no anatomical partitions could be predefined (WHOLE).

Our tests evaluate multiple instances, and not all comparisons are fully independent, because some of those imply combinations of partitions. Appropriate corrections for confidence levels on individual cases would have required making prohibitively time consuming simulations (i.e. with many more replications per test). Our interest, however, is not in the significance of individual comparisons, but rather in the collective results, and the differences between morphological and molecular datasets. While corrections for multiple tests might have lowered somewhat the rejection rate of homogeneity, a correction would equally affect the comparisons for morphology and sequences, so that the differences between the degree to which branch length homogeneity is, or is not rejected, by each type of dataset, would have remained equally strong.

An important caveat of the test performed here is that the observed branch lengths were calculated using a single rate category (i.e. no gamma parameter). There are indications (e.g. Marshall et al. 2006; Nguyen et al. 2017) that taking into account among-site rate heterogeneity improves estimations of branch lengths. A more accurate appraisal would perhaps have analyzed both the observed and simulated datasets allowing rate heterogeneity, simulating data under the same gamma values estimated for the combined dataset (instead of the single rate now used); this would have made evaluations significantly slower, would have required modifications to the functions of TNT that simulate data under a CM, and would have added another layer of complexity (and thus, potential errors) to the estimations. It seems doubtful, however, that using a gamma correction would have changed much the evaluations. The test focuses on correlations between branch lengths for two partitions, and the main numerical effect of applying a gamma correction is to alter the absolute values of all branch lengths by roughly the same factor, with only minor modifications to their proportionality. This is indeed a problem when the interest is in calculating the correct values of branch lengths for each partition (e.g. as in the study of Nguyen et al. 2017), but does not have a strong effect on the values of correlation (changing only the regression slopes). The best indication that the use of a single rate category did not bias the comparisons in the case of morphology is in the results for DNA sequence data: those analyses did not use, either, a gamma parameter for among-site rate variation, yet they produced a high proportion of cases where correlation between estimated branch lengths was within the range expected under the single-rate model. This suggests that the effect of a test considering among-site rate variation would have been minor, and that the same differences between DNA and morphological datasets would have been obtained.

The results obtained when comparing branch lengths for the partitions calculated with MPR are, overall, similar to those obtained with likelihood, with the same difference between DNA and morphological datasets. The probability of obtaining the observed correlation for the morphological datasets (p-values under the episodic and Mkv models), and for DNA sequences (p-values under JC69) is shown in Figure 4, for each individual comparison. The similarity in results obtained using two methods as different as MPR and maximum likelihood also suggests that the rejection of a CM in morphology does not strongly depend on method used for calculating branch lengths (including the use of a gamma parameter).

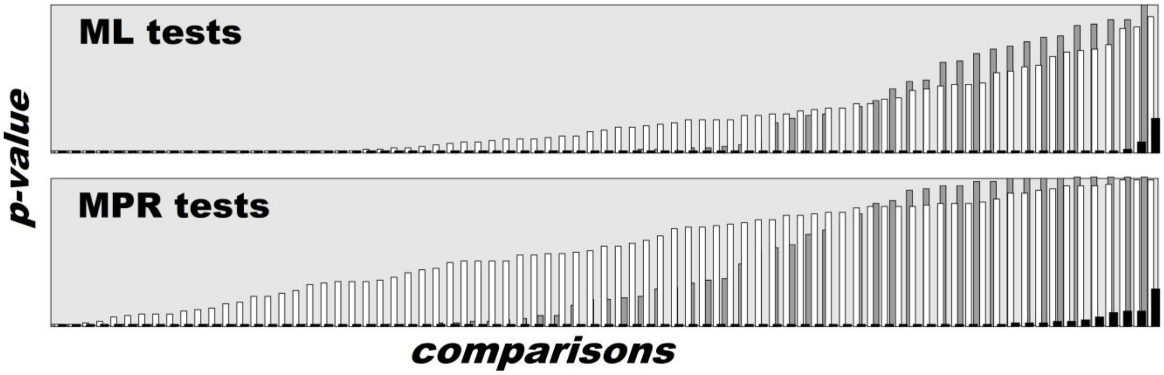


Figure 4. Plot showing the probability (P-values) of obtaining a correlation between branch lengths for two partitions as strong as the observed one, under different models for generating data, for morphological partitions defined by anatomy, and DNA partitions defined by contiguity, arranged in increasing order of P-value. Probabilities calculated both with likelihood (ML) and most parsimonious reconstructions (MPR). Black, morphological partitions tested against Mkv model; gray, DNA, tested against JC69; white, morphological partitions tested against episodic model. Gray and black bars are models with a common mechanism, white bars are for a model lacking a common mechanism.

Common Model Tested Within Partitions and Entire Datasets

Some studies have already demonstrated (with different methods; Clarke and Middleton 2008, Tarasov and Genier 2015, Lee 2016) heterogeneity in branch lengths for predefined partitions, so a meaningful evaluation must test whether a CM is in effect *within* individual partitions. Two subpartitions containing similar proportions of characters evolving under two completely different sets of branch lengths will have similar mixtures of rates, combining to provide a common average “rate” for each branch (Kolaczkowski and Thornton 2004), similar for both subpartitions. The internal heterogeneity of such mixtures cannot be detected by the present test (or any test we know), unless the correct partitions are known in advance –seldom the case for morphological data. Some partitioning schemes will produce the opposite effect, of making datasets generated from a single set of branch lengths to appear heterogeneous (e.g. by separating the characters in two groups, depending on which half of the tree they have more changes), but those partitionings are unlikely to be obtained at random. Thus, a conservative test can randomly subpartition characters, comparing the degree of branch length correlation between the random subpartitions with that expected under a CM; mixtures with similar proportions of two (or a few) sets of distinct branch lengths will often appear relatively homogeneous under such a test, for the mixtures will be sampled in roughly similar proportions. Therefore, rejection of branch length homogeneity in a majority of randomly chosen subpartitions is especially meaningful: more than just a few alternative rates, such a result suggests the absence of a CM altogether.

For testing random subpartitions, only the partitions with 80 or more characters were considered, dividing in two evenly-sized subpartitions. The results for random subpartitions are summarized in Figures 3b, 3d. Figure 5 shows the results of testing each subpartition individually; Figure 6 shows the average results for all the subpartitions of each partition (or dataset). For 86.4% of cases, random subpartitionings of the anatomically defined partitions produced a heterogeneity beyond (α=0.01) expected under the CM of the Mkv model (white boxes in Fig. 5a). Given that the test based on random subpartitioning requires no prior definition of partitions, an additional set of 78 morphological datasets (for which partitions could not be easily defined on the basis of anatomy) were tested as a whole. A CM was rejected (α=0.01) in 43.8% of all subpartitions (white boxes in Fig. 5b). For molecular datasets, instead, random subpartitionings (for 200 mid-sequence positions) reject a CM in only 9.8% of cases (Fig. 5c). To give these results further context, we simulated datasets (200 characters) under a Mkv model but with independent sets of branch lengths; as expected, the proportion of

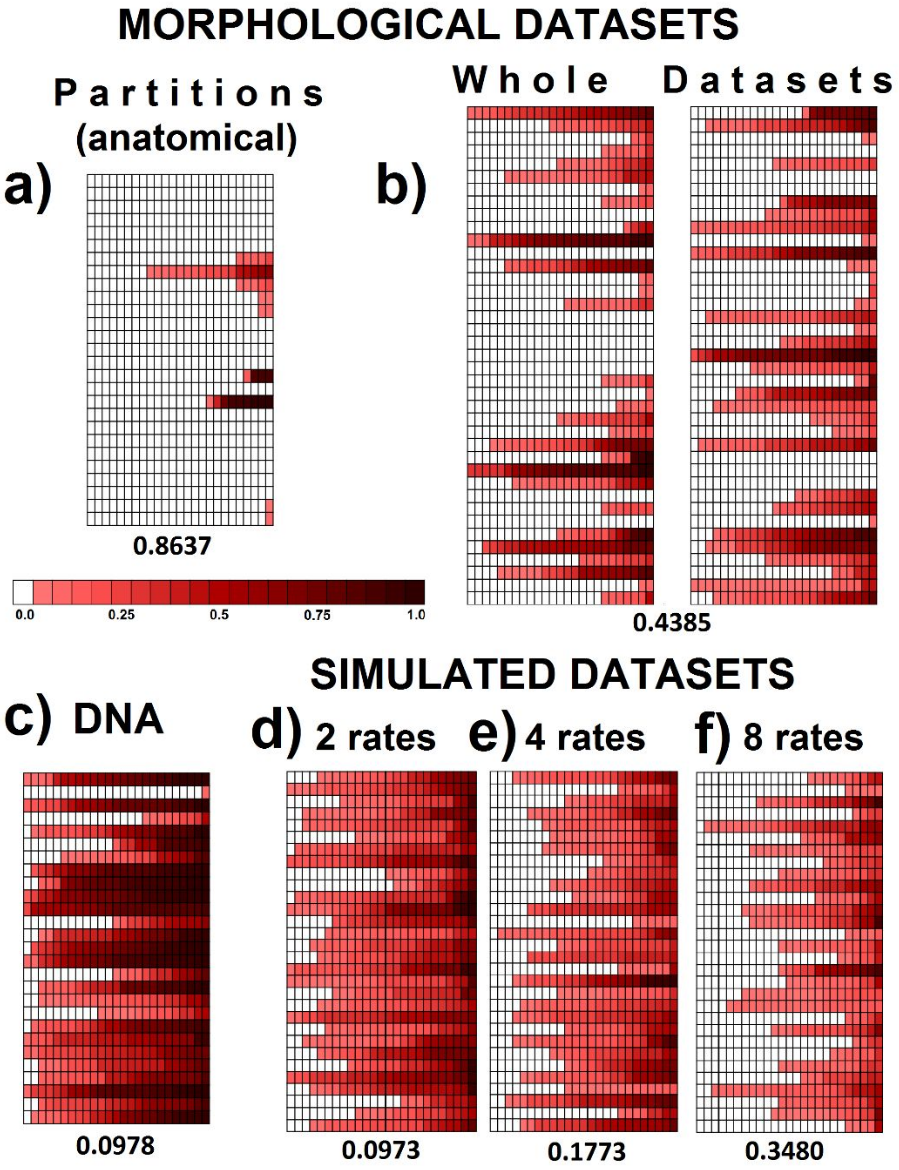


Figure 5. Plots of subpartition tests (25 per partition/dataset). Every row corresponds to a dataset or partition, every **individual** box corresponds to a subpartition. **The color of each box indicates the probability of obtaining the branch length correlation in the subpartition under a common mechanism (white color is $p < 0.05$, with a darker color as p increases). The numbers below frames correspond to proportion of subpartitions where common mechanism is rejected with α=0.01 (i.e. lower proportions correspond to cases where the common mechanism is less likely to have generated the data). (a) Subpartitions of partitions predefined on the basis of anatomy; (b) Whole datasets; (c) Molecular datasets; (e-f) Datasets simulated with 2, 4 and 8 independent sets of branch lengths.**

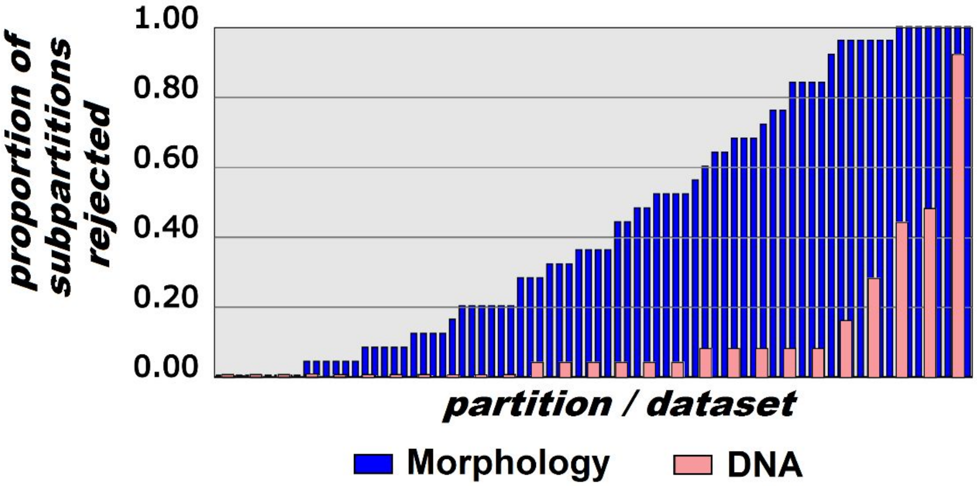


Figure 6. Comparison of proportion of subpartitions with a common mechanism rejected (at α=0.01) per partition (or dataset), for morphology and DNA.

cases where a single CM could be rejected on random subpartitions increased with the number of independent sets of branch lengths (Figs. 5d–f), reaching up to 34.8% for mixtures of 8 independent sets (Fig. 5f). This is still well below the rejection rate for morphological datasets, suggesting that (on average) morphological characters evolved with even larger deviations from a single CM.

Testing Models Without a Common Mechanism

The homogeneity of branch lengths for DNA sequences can be expected from theoretical considerations and previous empirical work evaluating the CM in sequences (Huelsenbeck et al. 2008). The results for morphological datasets, in contrast, strongly refute the CM (and hence the Mkv model), both in datasets taken as a whole, and most importantly, within partitions defined on the basis of anatomy. Whether the data evolve under a CM is indeed relevant for phylogenetic inference based on morphology:

simulations show that Bayesian inference works best when the data evolve homogeneously (Wright and Hillis 2014, O'Reilly et al. 2016, Puttick et al. 2017), but parsimony may work best (Kolaczkowski and Thornton 2004, Goloboff et al. 2017) when they do not.

Note that the probability distribution of character patterns of both NCM and the "Ultra-Conserved-Mechanism" (UCM, with a CM and all tree-branches having exactly the same length for all characters) are exactly identical, as shown by Huelsenbeck et al. (2008) and Steel (2011). Either of those models will have any change equiprobably located (CEL) on any tree branch, which is how Goloboff et al. (2017) generated their data. Given that parsimony is an appropriate method if the data do evolve under NCM (as shown by Tuffley and Steel 1997, Steel 2011), it follows that so it is under the equivalent (but less strongly parameterized) UCM or CEL, which generate the same probability distributions for character patterns. Note that CEL is a statement of the product of evolution (i.e. on how character changes will be located on tree branches), more than a statement of process; this product may be achieved by different processes (NCM, UCM, and possibly others).

The equiprobability of location of changes on any tree branch in the simulations of Goloboff et al. (2017) is a uniform distribution which (given the difficulties in modelling morphology) can be defended as an initial reference assumption, and produces no branch length correlation between partitions. Parsimony is then a well-justified method, but the model is also rejected by morphological datasets: branch length correlation between partitions is higher than expected (with 41 of the 79 comparisons between predefined partitions rejecting the model with α=0.05). The generating model can be made more realistic with characters equiprobably changing in every branch but only within a certain region of the tree, thus following the mosaicism proposed by Farris (1983: 15) and Goloboff et al. (2018). This model (Fig. 7) assumes that, during evolution, the possibility

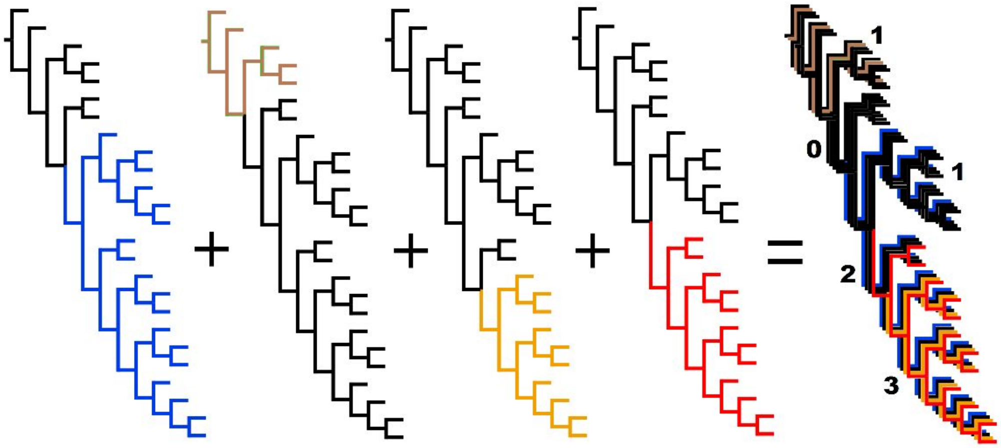


Figure 7. Episodic model. Colors indicate regions of the tree where a character (or group of characters) can change; black branches indicate regions where characters cannot change. Within the colored region, a change has the same probability of being located in any of the branches. The example shows 4 pivots (i.e. points where change becomes possible or impossible); because of the interaction between pivots, different branches of the tree have different numbers of characters (indicated on the rightmost tree) that could possibly change.

that formerly invariable characters become variable (or viceversa) can be triggered in *episodic* events; the point at which the character becomes variable (or invariable) is a node in the tree acting as a *pivot*. Morphological characters, by its very hierarchical nature (Maddison 1993, De Laet 2005, Brazeau et al. 2017) and by being subject to selection shifting from stabilizing to directional along time or changes in the developmental constraints, may well be liable to such episodic evolution. This *episodic* model is reminiscent of the covarion model (Fitch and Markowitz 1970, and successive modifications), differing in that character changes within regions of variability can be equiprobably located at any possible branch, thus lacking a CM and a formal branch-length parameter, consequently being more suited for morphological data. In the presence of multiple pivots affecting groups of characters, some branches will have larger numbers of synapomorphies for each group, generating a correlation between branch lengths for different partitions, and the correlation observed for empirical partitions is mostly within that expected from the episodic model (with only 27.8% of comparisons rejecting the model when half the characters are affected by pivots and half are not; see Figs. 3a, 4). This does not prove that an episodic model is the best general explanation for morphological patterns of character change, but at least the model is not as widely rejected as the Mkv or NCM models. More interestingly theoretically, the model shows that trees with some correlation between changes per branch for different partitions can result from models that do not assume a CM. The episodic model is used here solely to generate data, not to infer trees; likelihood inference assuming that model has not been implemented, and (by analogy to the covarion model) may suffer from identifiability problems (as noted by Gruenheit et al. 2008 for standard covarion models) unless significant restrictions are imposed.

Implications for Choice of Phylogenetic Methods

The episodic model resembles the NCM, UCM, or any other process leading to CEL, except that change is restricted to some parts of the tree. Given this similarity, we conjecture that only multiple pivots per character could produce inconsistency for parsimony if the model truly generated the data. In other words, if only one pivot per character occurs in the tree, parsimony can be justified just like the model with changes in each character occurring equiprobably over all the tree. With a single pivot per character, several tree branches may have more changes by virtue of being intermediate between pivots, but each of those long branches would have changes in different groups of characters (just like the synapomorphies for the long branches leading to e.g. Cetacea and Chiroptera correspond to different characters; Goloboff et al. 2018), so that they would be unlikely to attract.

The fact that models with a CM are strongly rejected by morphological data, and some models without a CM are not, is relevant for the choice of phylogenetic method. Previous studies where Bayesian analysis outperformed parsimony (e.g. Wright and Hillis 2014, O'Reilly et al. 2016, Puttick et al. 2017) had generated their data with a CM. In addition, for implied weighting (Goloboff 1993), O'Reilly et al. (2016) and Puttick et al. (2017) chose the worst concavity value (k=2, close to a clique, contrary to recommendations of Goloboff 1995: 99) and did not eliminate poorly supported groups (Fig. 8a). With a milder concavity and poorly supported groups eliminated, Bayesian analysis with the Mkv model outperforms implied weighting by a much smaller difference (Fig. 8b), but by a difference nonetheless, when the data are generated with a CM. When the data are generated instead with the half-episodic model, which does not assume a CM, parsimony tends to produce (as in the unrestricted model of Goloboff et al. 2017, and in agreement with expectations) slightly better results than Bayesian analysis (see Fig. 8c). As the number of characters increases (Fig. 8d), **both methods improve their results, but** Bayesian analysis has a slightly poorer performance for every statistic, perhaps as a result of the departure from the CM assumed by the Mkv model becoming more evident (given the large amounts of data).

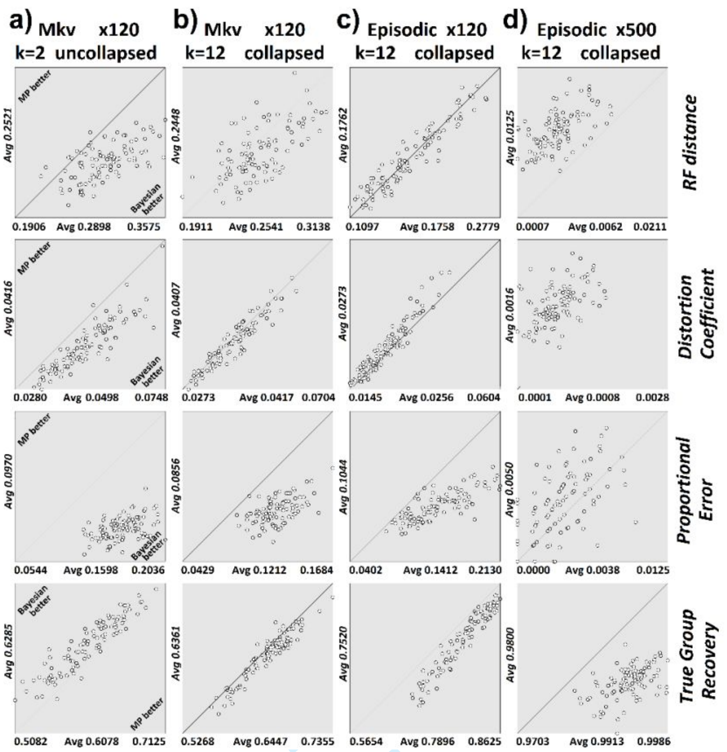


Figure 8. Comparison between implied weights and Bayesian analysis, using different methods for simulating and analyzing data (columns), and four different statistics to evaluate performance (rows). Proportional error is the number of incorrect groups found, divided by the number of groups in the inferred tree. The values of different statistics for Bayesian analysis are plotted against implied weights parsimony; by plotting the values for implied weighting on the x-axis, and those for BI on the y-axis, the deviation from the diagonal allows the difference in performance between the two methods to be easily detected. Datasets generated with both the Mk model of Lewis (2001) (columns A, B), and with the half-episodic model (C, D). Each of 100 points represents the average of 10 simulations with the same numbers of taxa and characters (to reduce dispersion, for a total of 1,000 simulated datasets). As the datasets are generated with the half-episodic model (lacking a common mechanism), the number of characters increases, and a concavity value of k=12 is used for implied weighting (instead of k=2, the worst performing value, chosen by O'Reilly et al. 2016 and Puttick et al. 2017 for their comparisons), parsimony outperforms Bayesian analysis **by a smaller margin, but more consistently**. The average values for each statistic are indicated in the x-axis for implied weighting, and on the y-axis for Bayesian analysis.

CONCLUSIONS

Our findings provide the first empirical demonstration, in a phylogenetic framework, of the differences in modes of evolution of molecules and morphology. While models that lack a CM (such as the episodic model) can produce degrees of branch length correlation between partitions that are in line with those observed in real datasets, the CM assumed by the Mkv model is strongly rejected by the morphological datasets. Of course, as generally acknowledged, a model need not reflect reality perfectly to be a useful aid in estimation, but a model still needs to have *some* basis in reality. If it is accepted that “all models are wrong, but some are useful”, then one must also accept that some models are *not* useful. The extent to which the CM assumed by the Mkv model deviates from reality seems strong enough to suspect the model may well do more harm than good. It is possible that violations of its assumptions rarely mislead Bayesian inference of trees in practice**; our simulations show that MrBayes seems rather robust to such violations. Such robustness may well be a result of the mechanics of the Markov chain and subsequent tree summarization, more than the result of assuming the Mkv model. If this is correct, MrBayes with the “parsimony” model might well produce (for datasets generated without a CM) trees of about the same quality as those produced with the Mkv model (a possibility that has not hitherto been examined in detail).** But one of the advantages claimed for model-based methods is that (by incorporating biological knowledge about evolutionary processes; Huelsenbeck et al. 2011) they allow estimating more than just tree topologies. Unrealistic assumptions built into phylogenetic models, therefore, can also affect studies of character mapping, dating of nodes on given trees, calculation of probabilities of specific evolutionary events, and even how taxonomists think of characters or diagnose groups. Thus, in light of the evidence against the common mechanism assumption, we strongly advise against the uncritical use of the Mkv model.

Supplementary Material

Material and methods, datasets, results, and scripts are available at the Dryad repository, doi:10.5061/dryad.3680n0c.

Acknowledgments

The study benefited from discussions at the International Pennaraptoran Dinosaur Symposium (held at the University of Hong Kong and supported by Kenneth HC Fung and First Initiative Foundation) and at the International Symposium of Archosaurian Phylogeny (held at the Institute of Vertebrate Paleontology and Paleoanthropology, Beijing). We also thank discussion, advice, or comments on the ms. by Salvador Arias, Santiago Catalano and Claudia Szumik (UEL); Lone Aagesen (IBODA); Julián Faivovich and Martín Ramírez (MACN); Leandro Jones (UNP); Adriana Pérez (FCEN-UBA); Mark Simmons (Colorado State University); Maureen O'Leary (Stony Brook University); Gonzalo Giribet (Harvard University); David Swofford (Duke University). The comments of AE Simon Ho and three reviewers (Joseph Brown, and two Anonymous) were also very helpful. This study was supported by a Proyecto de Unidad Ejecutora (PUE0070 to PAG) from Consejo Nacional de Investigaciones Científicas y Técnicas; National Science Foundation, National Aeronautics and Space Agency, and Fundação de Amparo à Pesquisa do Estado de São Paulo (via a grant to J. Cracraft and L. Lohman, on "Assembly and evolution of the Amazonian biota and its environment: an integrated approach", with participation from PAG); Research Grant Council of Hong Kong's General Research Fund (17103315) and University of Hong Kong's University Research Committee Seed Fund for Basic Research (both to MP), and the National Science Foundation of China (41688103, 41120124002 and 91514302 to XX).

REFERENCES

Brazeau, M., Guillerme, T., Smith, M. 2017. Morphological phylogenetic analysis with inapplicable data. BioRxiv https://doi.org/10.1101/209775 [Note: this is under review in Syst.Bio, so it will likely be citable as a paper by the time our paper comes out]

Clarke, J., Middleton, K. 2008. Mosaicism, Modules, and the evolution of birds: results from a Bayesian approach to the study of morphological evolution using discrete character data. *Syst. Biol.* **57**, 185–201.

De Laet, J. 2005. Parsimony and the problem of inapplicables in sequence data. In: *Parsimony, Phylogeny and Genomics (*V. Albert, ed.), Oxford University Press, pp 81–116.

Duchene, S., Molak, M., Ho, S. 2014. ClockstaR: choosing the number of relaxed-clock models in molecular phylogenetic analysis. *Bioinformatics* **30**, 1017–1019.

Farris, J. 1983. The logical basis of phylogenetic analysis. In: Platnick, N., Funk, V. (Eds.), *Advances in Cladistics II*. Columbia University Press, New York, NY, pp. 7–36.

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.

Fitch, W., Markowitz, E. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem. Genet.* **4**, 579–593.

Goloboff, P. 1993. Estimating character weights during tree search. *Cladistics* **9**, 83–91.

Goloboff, P. 1995. Parsimony and weighting: a reply to Turner and Zandee. *Cladistics* **11**, 91–104.

Goloboff, P., Pol, D. 2005. Parsimony and Bayesian phylogenetics. In: *Parsimony, phylogeny, and genomics* (Victor Albert, ed.). Oxford University Press, pp. 148-159.

Goloboff, P., Farris, J. 2008. Nixon, K. TNT, a free program for phylogenetic analysis. *Cladistics* **24**, 774–786.

Goloboff, P., Catalano, S. 2016. TNT version 1.5, including a full implementation of geometric morphometrics. *Cladistics* **32**, 221–238.

Goloboff, P., Torres, A., Arias, S. 2017. Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology. *Cladistics*, https://doi.org/10.1111/cla.12205.

Goloboff, P., Torres, A., Arias, S. 2018. Parsimony and model-based phylogenetic methods for morphological data: comments on O'Reilly et al. (2017). *Palaeontology* doi.org/10.1111/pala.12353.

Gruenheit, N., Lockhart, P., Steel, M., Martin, W. 2008. Difficulties in testing for covarion-like properties of sequences under the confounding influence of changing proportions of variable sites. *Mol. Biol. Evol.* **25**, 1512–1520.

Harrison, L., Larsson, H. 2015. Among-character rate variation distributions in phylogenetic analysis of discrete morphological characters. *Syst. Biol.* **64**, 307–324.

Holder, M., Lewis, P., Swofford, D. 2010. The Akaike Information Criterion will not choose the No Common Mechanism model. *Syst. Biol.* 59, 477–485.

Huelsenbeck, J., Annè, C., Larget, B., Ronquist, F. 2008. A Bayesian perspective on a non-parsimonious parsimony model. *Syst. Biol.* **57**, 406–419.

Huelsenbeck, J., Alfaro, M., Suchard, M. 2011. Biologically inspired phylogenetic models strongly outperform the no common mechanism model. *Syst. Biol.* **60**, 225–232.

Jukes, T., Cantor, C. 1969. Evolution of protein molecules. In Munro, N. (editor), *Mammalian protein metabolism*. Vol. 3, New York, Academic Press, pp. 21–132.

Kolaczkowski, B., Thornton, J. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* **431**, 980–984.

Lanfear, R., Frandsen, P., Wright, A., Senfeld, T., Calcott, B. 2017. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol. Biol. Evol.* **34**, 772–773.

Lee, M. 2016. Multiple morphological clocks and total evidence tip-dating in mammals. *Biol. Lett.* **12**, 20160033. http://dx.doi.org/10.1098/rsbl.2016.0033.

Lewis, P. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* **50**, 913–925.

Maddison, W. 1993. Missing data versus missing characters in phylogenetic analysis. *Syst. Biol.* 42, 576–581.

Marshall, D., Simon, C., Buckley, T. 2006. Accurate branch length estimation in partitioned Bayesian analyses requires accommodation of among-partition rate variation and attention to branch length priors. *Syst. Biol.* **55**, 993–1003.

Nguyen, L.-T., von Haeseler, A., Minh, B. 2017. Complex models of sequence evolution require accurate estimators as exemplified with the invariable site plus gamma model. *Syst. Biol.* **67**, 552-558.

Nyakatura, K., Bininda-Emonds, O. 2012. Updating the evolutionary history of Carnivora (Mammalia): a new species-level supertree complete with divergence time estimates. *BMC Biol.* **10**, 1–31.

O'Reilly, J., et al. 2016. Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. *Biol. Lett.* **12**, 20160081, 1–5. https://doi.org/10.1098/rsbl.2016.0081.

Puttick, M., et al. 2017. Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. *Proc. R. Soc. B* **284**, 20162290.

Sereno, P. 2009. Comparative cladistics. *Cladistics* **26**, 624–659.

Swofford, D. 2002. *PAUP*: Phylogenetic analysis using parsimony (* and other methods)*. Version 4. Sunderland (MA): Sinauer Associates.

Swofford, D., Olsen, G., Waddell, P., Hillis, D. 1996. Phylogenetic inference. In: Hillis, D., Moritz, C., Mable, B. (Eds.), *Molecular Systematics*, second ed. Sinauer, Sunderland, MA, pp. 407–514.

Ronquist, F., et al. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542.

Sober, E. 2004. The contest between parsimony and likelihood. *Syst. Biol.* 53, 644–653.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313.

Steel, M. 2011. Can we avoid "sin" in the house of "no common mechanism"? *Syst. Biol.* **60**, 96–109.

Tarasov, S., Génier, F. 2015. Innovative Bayesian and parsimony phylogeny of dung beetles (Coleoptera, Scarabaeidae, Scarabaeinae) enhanced by ontology-based partitioning of morphological characters. *PLOS One* **10**, e0116671.

Tuffley, C., Steel, M. 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bull. Math. Biol.* **59**, 581–607.

Wright, A., Hillis, D. 2014. Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. *PLoS ONE* **9**, e109210. https://doi.org/10.1371/journal.pone.0109210.

Yang, Z., Zhu, T. 2018. Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees. *Proc. Nat. Acad. Sc.* **115**, 1854–1859.

Zhang, J. 2018. Neutral Theory and Phenotypic Evolution. *Mol. Biol. Evol.* **35**, 1327–1331.

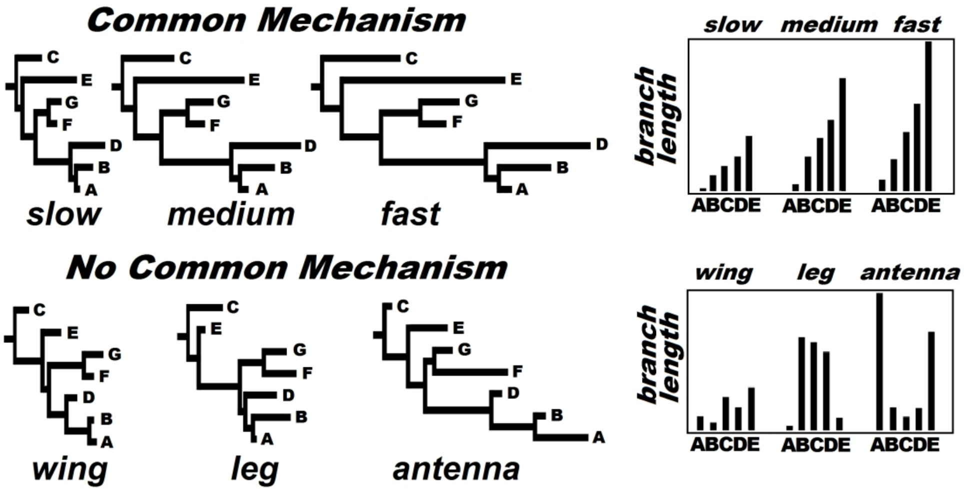
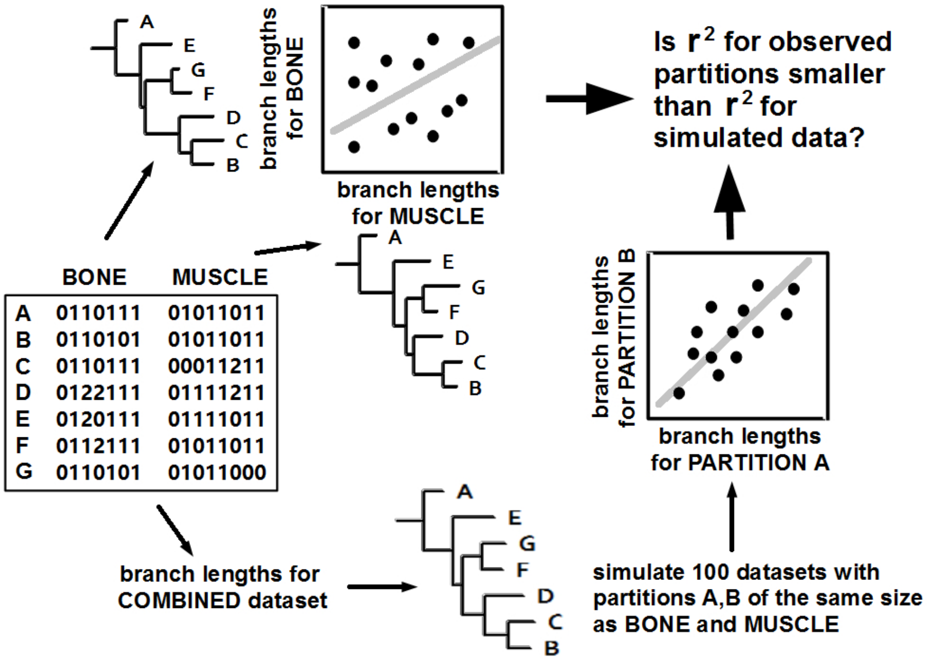


Figure 1

757x382mm (96 x 96 DPI)


A
E
G
F
D
C
B
branch lengths for BONE
branch lengths for MUSCLE
Is r^2 for observed partitions smaller than r^2 for simulated data?
A
E
G
F
D
C
B
branch lengths for PARTITION B
branch lengths for PARTITION A
branch lengths for COMBINED dataset
A
E
G
F
D
C
B
simulate 100 datasets with partitions A,B of the same size as BONE and MUSCLE

	BONE	MUSCLE
A	0110111	01011011
B	0110101	01011011
C	0110111	00011211
D	0122111	01111211
E	0120111	01111011
F	0112111	01011011
G	0110101	01011000

Figure 2

278x210mm (96 x 96 DPI)

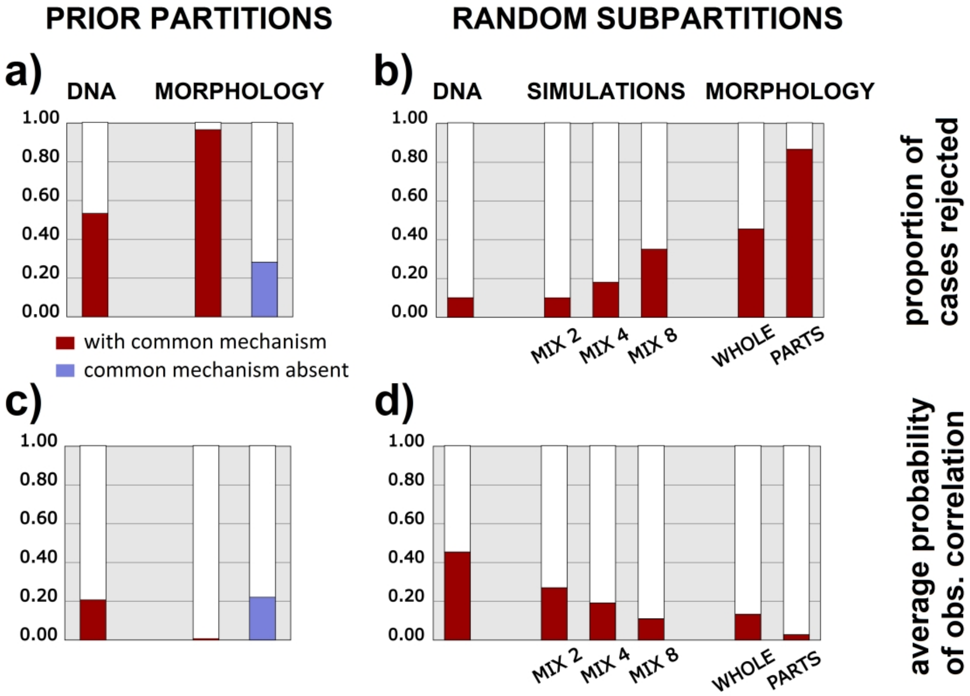


Figure 3

1280x927mm (96 x 96 DPI)

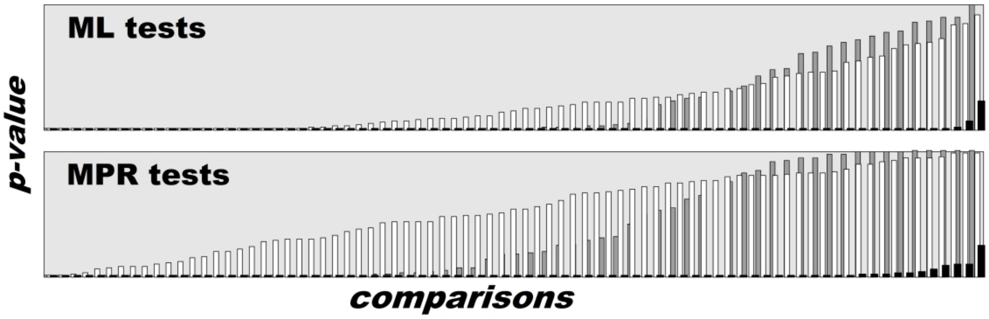


Figure 4

1320x438mm (96 x 96 DPI)

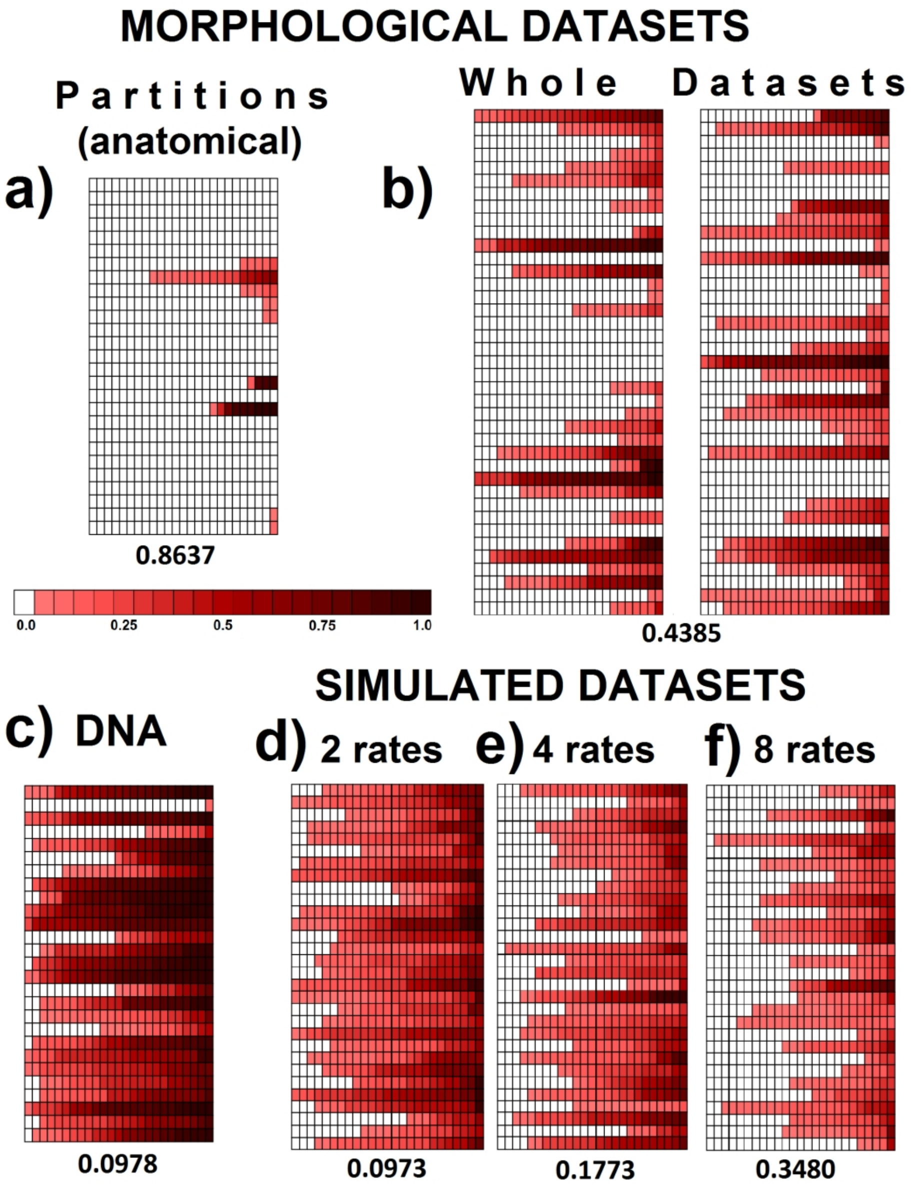


Figure 5

546x702mm (96 x 96 DPI)

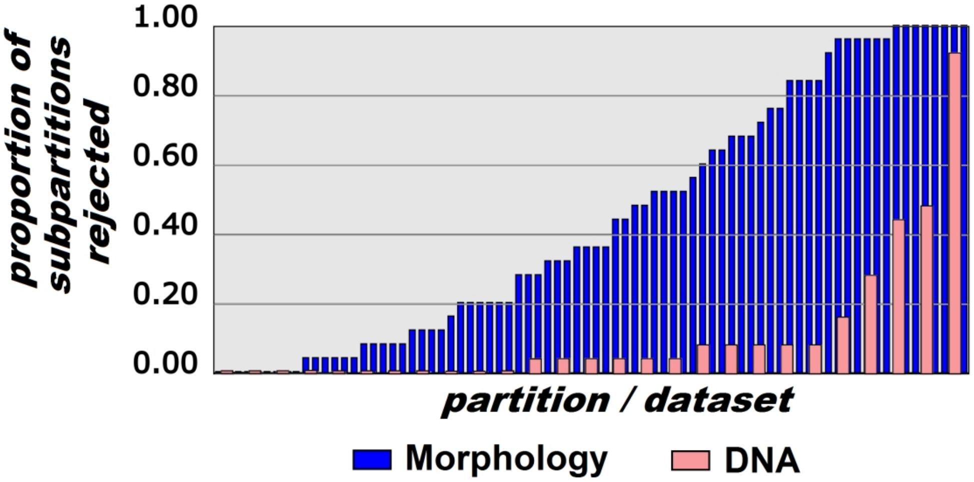


Figure 6

712x358mm (96 x 96 DPI)

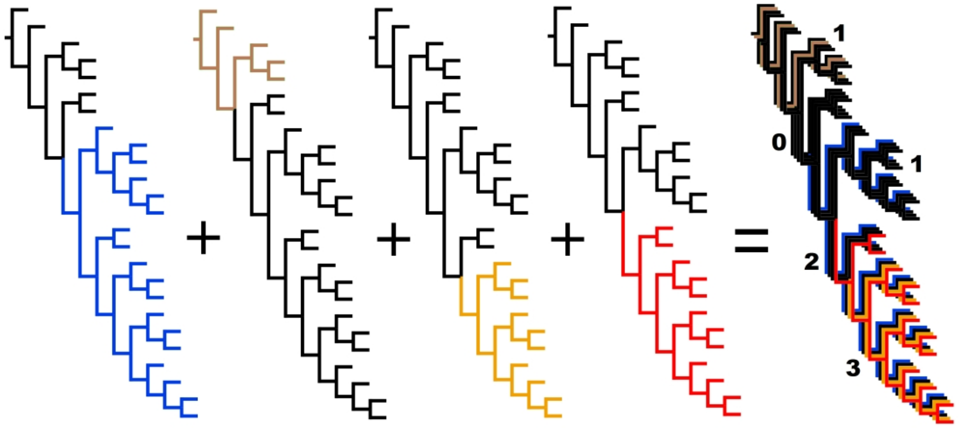


Figure 7

1301x618mm (96 x 96 DPI)

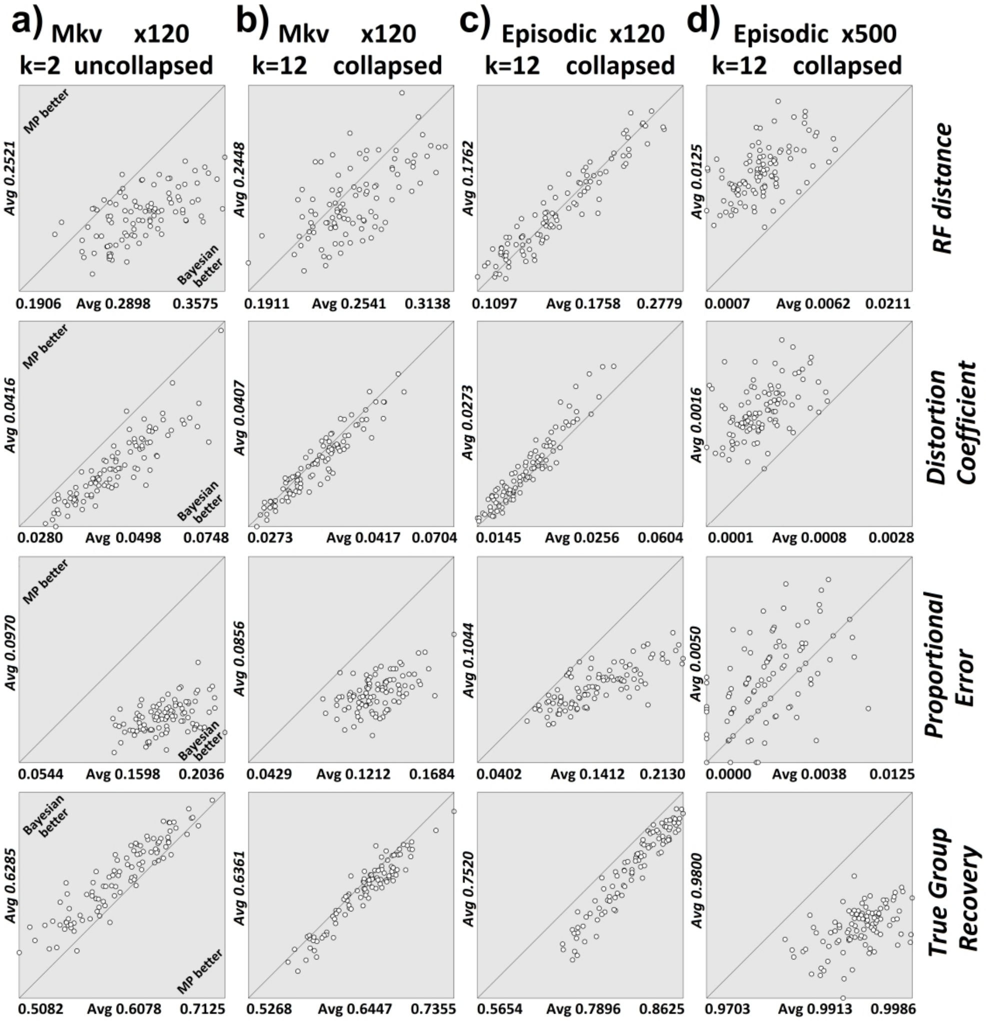


Figure 8

1508x1547mm (96 x 96 DPI)