Syst. Biol. 72(2):476–488, 2023

https://doi.org/10.1093/sysbio/syac066
Advance Access Publication September 29, 2022

Reducing the Biases in False Correlations Between Discrete Characters

James D. Boyko[1,2,*] and Jeremy M. Beaulieu[1]

[1]*Department of Biological Sciences, University of Arkansas, Fayetteville, AR 72701, USA and*
[2]*Michigan Institute of Data Science, University of Michigan, Ann Arbor, MI 48109, USA*
[*]*Correspondence to be sent to: Department of Biological Sciences, University of Arkansas, Fayetteville, AR 72701, USA; E-mail: jboyko@uark.edu*

Received 4 April 2022; reviews returned 14 September 2022; accepted 27 September 2022
Associate Editor: Stacey Smith

Abstract.—The correlation between two characters is often interpreted as evidence that there exists a significant and biologically important relationship between them. However, Maddison and FitzJohn (in The unsolved challenge to phylogenetic correlation tests for categorical characters. Syst. Biol. 2015;64:127–136) recently pointed out that evidence of correlated evolution between two categorical characters is often spurious, particularly, when the dependent relationship stems from a single replicate deep in time. Here we will show that there may, in fact, be a statistical solution to the problem posed by Maddison and FitzJohn naturally embedded within the expanded model space afforded by the hidden Markov model (HMM) framework. We demonstrate that the problem of single unreplicated evolutionary events manifests itself as rate heterogeneity within our models and that this is the source of the false correlation. Therefore, we argue that this problem is better understood as model misspecification rather than a failure of comparative methods to account for phylogenetic pseudoreplication. We utilize HMMs to develop a multirate independent model which, when implemented, drastically reduces support for correlation. The problem itself extends beyond categorical character evolution, but we believe that the practical solution presented here may lend itself to future extensions in other areas of comparative biology. [Macroevolution; model adequacy; phylogenetic comparative methods; rate heterogeneity].

Correlated or dependent evolution on a macroevolutionary scale is defined as a change in a character state (e.g., plumage color) that is linked to the presence of a particular state in a separate character (e.g., beak color). In other words, the evolution of character X can be said to be dependent on character Y if, in the presence of a particular state of Y (e.g., Y_0; black plumage), shifts within character X (e.g., X_0 to X_1; orange beak to red beak) occur at a different rate from when the lineage is in an alternative state of Y (e.g., Y_1; white plumage). For example, a shift from X_0 to X_1 may occur more quickly when paired with Y_1 than with Y_0 resulting in a distribution with many character pairs X_1Y_1. It is often the case that these sorts of dependent relationships between characters seem obvious, especially if the observations of many species are consistent.

However, what happens when all observations of the pair come from, for example, a single clade? In other words, there may have been many species in which X_1Y_1 is observed, but they all came from one peculiar clade of waterfowl. Since the strength of the relationship is related to the number of individual observations, their phylogenetic nonindependence raises concerns about the validity of the proposed correlation. This fact was well understood as early as Darwin (1859), and the tools for dealing with the resulting statistical nonindependence have been available to comparative biologists since the foundational work of Felsenstein (1985). Nevertheless, this issue of "phylogenetic pseudoreplication," where species are nonindependent due to their shared ancestry, served as the basis for the concerns raised by Maddison and FitzJohn (2015) regarding tests of dependent character evolution.

Maddison and FitzJohn (2015) demonstrated that the most widely used phylogenetic method for detecting correlated evolution between categorical characters (Pagel 1994), almost always indicates strong evidence of correlation when singular events deep in time can account for the codistribution of two characters. To demonstrate their point, they fit correlated models to data sets generated under their so-called, "Darwin's" and the "Unreplicated Burst" scenarios (Fig. 1a,b). Darwin's scenario results in the perfect codistribution of two characters, which in practice, might occur when testing for correlations between two synapomorphies (e.g., presence/absence of middle ear bones and fur). Under the Unreplicated Burst scenario, only one of the two characters has phylogenetically replicated change. This scenario occurs when one of the characters is a synapomorphy for the clade, with the other character undergoing several changes within the focal clade. Both Darwin's and the Unreplicated Burst scenarios can be contrasted to data sets simulated with a model of correlated evolution, where there are several repeated and independent instances of the correlation arising (Fig. 1d). However, the issue is that, when applied to either Darwin's or the Unreplicated Burst scenario, commonly used comparative methods (Pagel 1994) will almost always indicate strong evidence of correlation despite the dependent relationship arising from little more than a single event deep in time.

There is considerable interest in understanding and, ultimately, finding a resolution to the problem posed by Maddison and FitzJohn (2015). Recently, Uyeda et al. (2018) suggested that for Darwin's scenario, the relatively long periods of stasis between the two characters (i.e., minimal trait change) are the primary cause for their significant dependent relationship. In fact, they showed

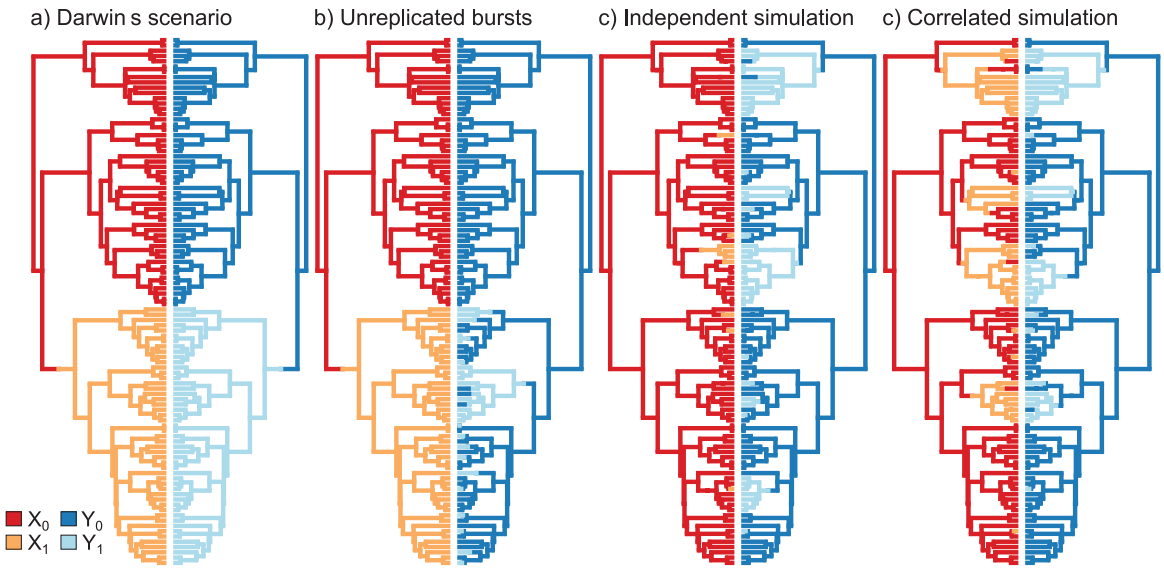

FIGURE 1. The two problematic scenarios from Maddison and FitzJohn (2015) for the evolution of characters X and Y. Character X is painted on the left phylogeny using red and orange for state X_0 and X_1, whereas character Y is painted on the right phylogeny using dark blue and light blue for state Y_0 and Y_1. (a) Darwin's scenario is depicted as a single event deep in time that has led to the codistribution of X_0Y_0 outside of the focal clade and X_1Y_1 within the focal clade. (b) Unreplicated Bursts scenario is where a single event deep in time has led to the codistribution of X_0Y_0 outside of the focal clade and X_1Y_0 and X_1Y_1 within the focal clade. (c) One realization of a data set simulated under an independent model. (d) One realization of a data set simulated under a correlated model. Notice how in this case there is a consistent pattern of correlation and multiple independent origins of the potential correlated characters (X_0Y_0 and X_1Y_1) throughout the phylogeny.

that the probability of selecting a character-dependent model (i.e., a model of correlated evolution between the two characters) over a character-independent model (i.e., a model where the two characters are explicitly not correlated) was proportional to the ratio between the length of the branch where the shift occurred and the total length of the tree. The nature of this ratio ensured that a correlated model would always be supported in cases where singular evolutionary events led to a codistribution of characters. Another study by Gardner and Organ (2021) tested a variety of correlated models beyond Markov models and examined the structure of data sets which are susceptible to the problem of false dependence. They found that all the tested comparative methods produced erroneous correlations when data sets were phylogenetically pseudoreplicated and suggested that this was due to an inability to estimate parameters associated with transitions between unobserved states.

In both of these studies, the authors have addressed the problem by encouraging scientists to think critically about their models and data sets before conducting a comparative analysis. While this recommendation is certainly admirable and correct, it is not a direct and satisfying solution to the statistical problems presented so far, as no amount of methodological vigilance will ever prevent analyses from being marred by phylogenetic pseudoreplication. However, prior analyses have limited model comparisons to only a few models and have overlooked the very large set of alternative Markov models which can also be consistent with correlation or independence depending on the model's structure. These alternative models have been discussed previously (Pagel 1994; Pagel and Meade 2006) and, as we will show, the inclusion of a few examples within the model set can play a crucial role in ensuring a fair test of correlation. These underrepresented models, in addition to the enormous model space provided by hidden Markov models (HMMs) for addressing rate heterogeneity across the tree (Beaulieu et al. 2013; Boyko and Beaulieu 2021), can severely reduce the bias toward correlation noted by Maddison and FitzJohn (2015). We acknowledge that the problem itself extends beyond categorical character evolution, but we believe that the practical framework presented here may lend itself to future extensions in other areas.

We draw on two important insights as they relate to models of categorical character evolution. The first is that model space is severely underexplored and that the inclusion of more complex, character-independent models within our modeling set helps reduce evidence of false correlation. We reiterate previous findings (Uyeda et al. 2018; Gardner and Organ 2021) that estimates of transition rates to and from unobserved character states are not statistically identifiable, revealing that the canonical correlated model is overparameterized in phylogenetically pseudoreplicated data sets like Darwin's scenario (Fig. 1a). When only two or three of the four possible character state combinations are observed, we produce models nested within the correlated and independent models that are overwhelmingly favored over both. Second, the issue of false dependent relationships is not one of stasis *per se*, but rather, a failure to account for rate heterogeneity. We demonstrate that an explicit character-independent HMIM provides significant

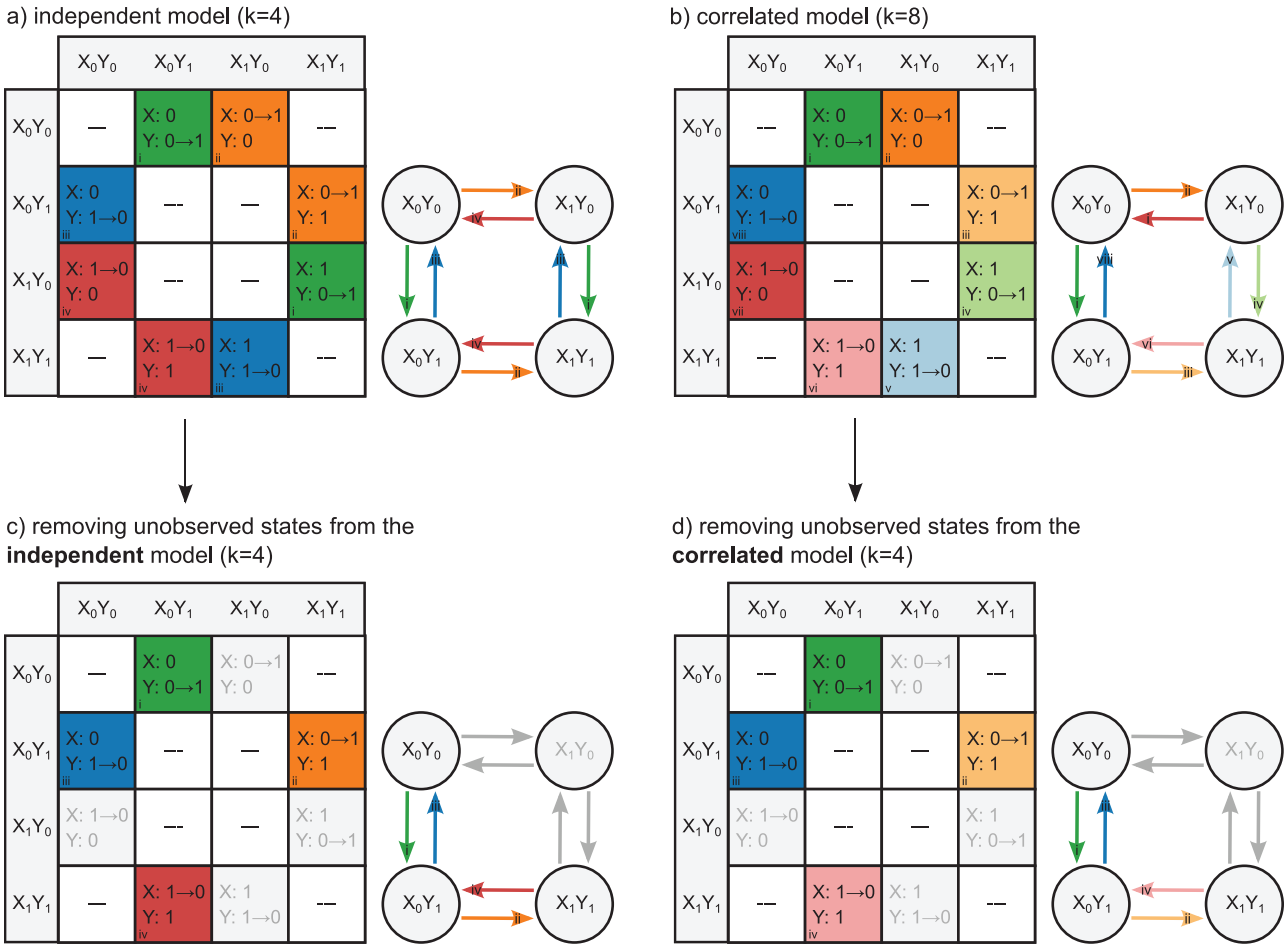

FIGURE 2. Representations of the different transition rate matrices, Q, with k number of parameters associated with each. Where transitions are fixed to occur at the same rate, the squares are colored to be the same. Unique parameters are also indicated with a roman numeral in the bottom left corner of the square. To the right of each matrix, a ball-and-stick representation of the model is presented with colors and parameter numbers matching the transitions indicated in the matrix, Q. The ball-and-stick representation is organized such that internal arrows represent transitions from 1 to 0, and external arrows represent transitions from 0 to 1. Additionally, arrows that cross the vertical midpoint indicate transitions in character X, whereas transitions across the horizontal midpoint indicate transitions in character Y. (a) An independent model with four unique parameters, which fixes transitions within a character such that changes in X or Y do not depend on the state of the other character. (b) A correlated model with eight unique parameters. This model allows transitions within a character to depend on the state of the other character. (c) A model which removes transitions to and from an unobserved state from the independent model (a). (d) A model that removes transitions to and from an unobserved state from the correlated model (b). In (c) and (d), the unobserved state is based on the Unreplicated Burst scenario where X_0Y_1 is not observed.

evidence for models of independent evolution in cases where a correlated model would have previously been supported. This is because under the classic Pagel (1994) framework, support for correlation comes from both a dependent relationship between characters and a strong signal of rate heterogeneity. By amending the Pagel (1994) framework with a model which allows for rate heterogeneity independent of a focal character, we correct the bias toward correlation.

Correlated Models Depend on Observations of Intermediate States

While much has been written about the specifics of Pagel's model, we briefly review aspects of it in order to better illustrate our point—namely, that certain transition rates are not estimable and that their inclusion may be an additional cause of false correlations uncovered by Maddison and FitzJohn (2015). The correlated or dependent model of discrete character evolution, introduced by Pagel (1994), uses a continuous-time Markov process to estimate the rate of transitions between character states (Fig. 2a,b). With a single binary character, X, the transition rate matrix, denoted as $\mathbf{Q}$, is a simple 2 × 2 matrix that contains all the information necessary to estimate the probability of a transition occurring between two states of character X over a given period of time. At its most complex, $\mathbf{Q}$ would contain two transition rates: from state X_0 to state X_1, and from state X_1 to state X_0. If we introduce a second binary character, Y, the number of possible observed state combinations is expanded—that is, the possible

observed state combinations become X_0Y_0, X_0Y_1, X_1Y_0, and X_1Y_1. Consequently, this requires an expansion of **Q** to a 4 × 4 matrix, to account for all the possible transitions between state combinations. This model is considerably more complex than the previous one, as the number of transitions goes from a minimum of 2 to a maximum of 12. However, the model introduced by Pagel (1994) is constrained specifically for the purpose of detecting correlations between characters by examining whether the state of one variable affects the probability of change in the other. To do this, dual transitions (i.e., changes in both X and Y occurring in a single time step) are removed. As noted by Pagel (1994), setting dual transition rates to zero does not rule out dual transitions over long periods of time. Rather, a dual transition from X_0Y_0 must first pass through state X_0Y_1 or X_1Y_0, before finally transitioning to X_1Y_1. Equating the rates of transitions between particular pathways allows for the construction and testing of an independent model (Pagel and Meade 2006). A model of independent evolution is nested within the correlated model but assumes that the transition rates between states of a character are equal to one another regardless of the state of the other character (e.g., $[X_0$ to $X_1 \mid Y_0] = [X_0$ to $X_1 \mid Y_1]$; Fig. 2a,b). In other words, if these two characters, X and Y, are independent, the presence of one character will have no influence on the change of the other and thus model selection criteria should choose the simpler model..

Using this specific nested framework, we were able to replicate the results of Maddison and Fitzjohn (2015). Specifically, we generated 100 data sets for Darwin's scenario and the Unreplicated Bursts scenario. Phylogenies were simulated under a $\lambda = 1$ and $\mu = 0.5$ until 100 extant taxa were reached, and each resulting tree was then evaluated for a focal monophyletic group between 40 and 60 taxa. For Darwin's scenario, extant species within the focal clade were assigned X_1Y_1, and species outside the clade were assigned X_0Y_0. We simulated Unreplicated Bursts by assigning all species outside the focal clade X_0, and all species within the clade X_1. Next, character Y was simulated at a rate of 100 transitions per million years. Outside of the focal clade, species were assigned Y_0 whereas, within the focal clade, the simulated data resulted in both Y_0 and Y_1. We used *corHMM* (Beaulieu et al. 2013; Boyko and Beaulieu 2021) to fit and compare the four-state independent model (Fig. 2a) against the four-state correlated model (Fig. 2b) using Akaike Information Criterion (AIC). In all cases, we found overwhelming support for the correlated model for both Unreplicated Bursts and Darwin's scenario data sets (see Supplemental Materials). The mean AIC weight for the correlated model under Darwin's scenario was 92.52% and under Unreplicated Bursts, it was 99.96%. As expected, an independent model was never favored over a correlated model in either scenario.

For Darwin's scenario, setting aside the critical analytical issues regarding phylogenetic pseudoreplication, we had additional concerns with the structure of the data and how this might impact estimates of transition rates. Under any continuous-time Markov process, the estimates of the transition rates among all possible character combinations are reflective of the observed state frequencies and distribution at the tips. But, what if two of the four character combinations are not observed at all? Here we are referring to the two combinations, X_0Y_1 and X_1Y_0, not observed in any of the tips under Darwin's scenario. There may be biological reasons for not observing intermediate state combinations. For example, these combinations may be at some selective disadvantage, resulting in rapid transitions to another, more viable character combination (e.g., X_0Y_0 or X_1Y_1). Alternatively, it could be that one or both combinations are never possible due to some underlying genetic or developmental reasons (e.g., certain fruit character combinations, see Beaulieu and Donoghue 2013). However, whatever biological meaning is attributed to the lack of intermediate character state observations, in this case, is beside the point. There are identifiability issues with including transitions to and from these unobserved state combinations in the model, calling into question fitting the correlated model to these types of data (also see Gardner and Organ, 2021). That is to say, if we never see intermediate state combinations at the tips, how can the model ever favor one pathway over the other?

To illustrate this point, we examined the likelihood surface of one of the data sets simulated under Darwin's scenario and fit under Pagel's correlated model (Fig. 3). Whether starting from X_0Y_0 or X_1Y_1, transition rate estimates to either of the unobserved character combinations fall along a ridge of equal likelihood, where changing the rate of transition to one unobserved state determines the rate for the transitions to the other unobserved state. When a lineage transitions into one of the states, the likelihood surface for transitions out of these states to either state X_0Y_0 or X_1Y_1 is completely flat, with all rates ranging from 0.1 to 100 transitions per unit time all having nearly identical likelihoods. In other words, at least in the case of Darwin's scenario, they are not *identifiable*. The preferred model estimates for various transition rates arise simply by chance of the optimization procedure.

One immediate solution is to simply remove the unobserved character combinations from the model completely. From a modeling perspective, removing unobserved states removes the parameters that fall along the likelihood ridge and should lead to a model that ends up being well estimated. This is accomplished by removing transition rate estimates to and from the unobserved character states (in the case of Darwin's scenario, a dual transition between X_0Y_0 and X_1Y_1 must be included). Consequently, the question of whether independent or correlated models better explain the data becomes irrelevant as the two models collapse into one another when unobserved states are removed (Fig. 2c,d). This is clearly seen when the collapsed model is applied to an Unreplicated Burst scenario. Whether one starts with an independent model (Fig. 2a) or a correlated model (Fig. 2b), once unobserved states are removed, comparing alternative transition pathways

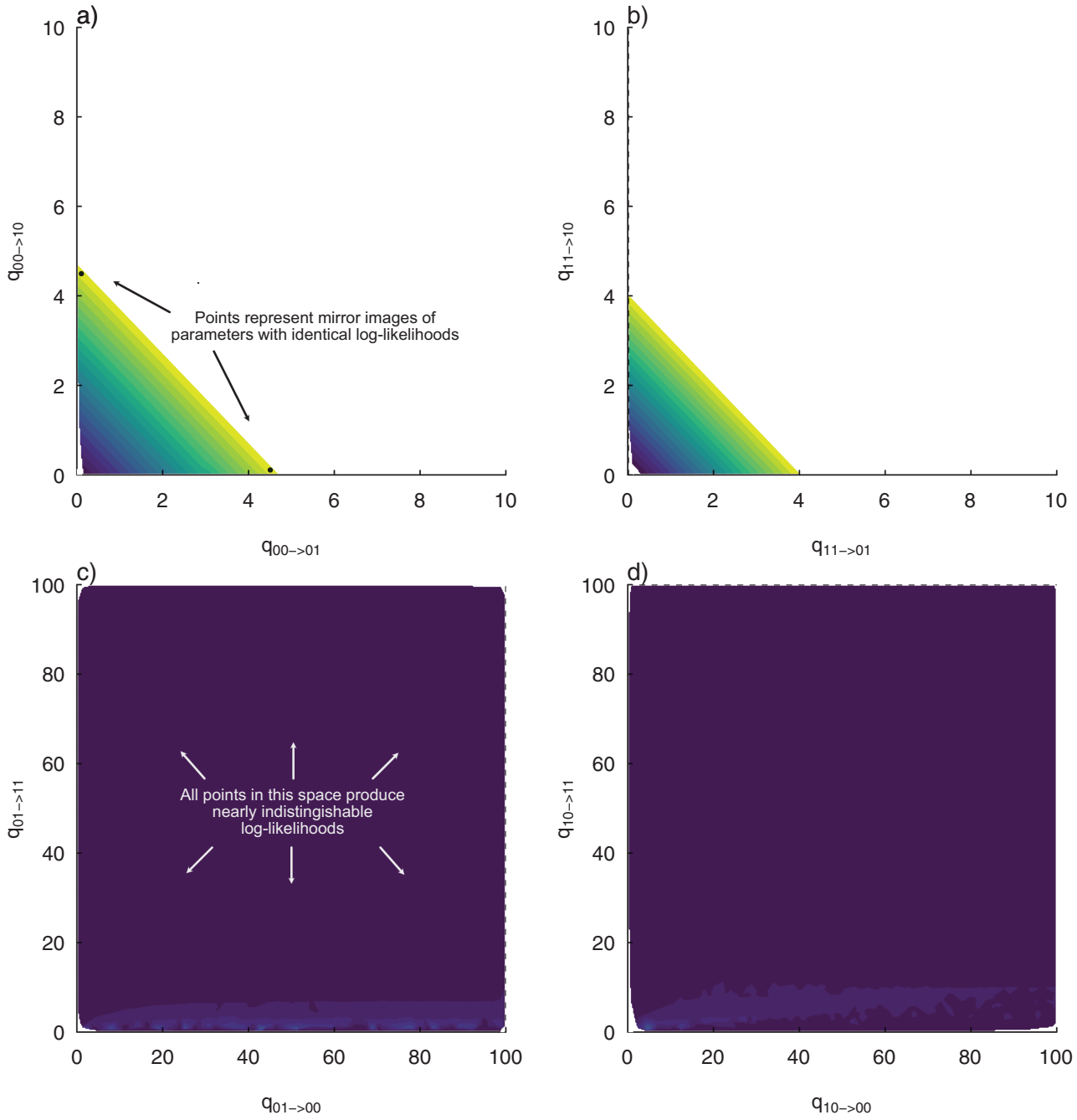

FIGURE 3. An example contour plot of a correlated model when applied to 1 of the 100 Darwin's scenario data sets. The colors indicate the log-likelihood surface for points that are two log-likelihood units away from the maximum likelihood (darker colors indicating support near the maximum likelihood). Each surface is constrained such that a particular pair of parameters is fixed but all remaining transition rates are free to find their MLE. We sampled 5000 pairs of points for a particular parameter pair from a Latin hypercube sampling design. (a) Transitions from X_0Y_0 to either intermediate state result in several likelihood ridges, as indicated by the linear bands of support. We highlight this "ridge" effect by showing two points that represent mirror images of transitions from X_0Y_0 to X_0Y_1 and X_0Y_0 to X_1Y_0. Both points produce identical log-likelihoods, as does any pair of points that occur when sliding along that particular band. (b) Transitions from X_1Y_1 to an intermediate state result in several likelihood ridges. (c) Transitions from X_0Y_1 to either X_0Y_0 or X_1Y_1 result in a completely flat likelihood surface, as indicated by the entire search space producing strong support that is nearly indistinguishable from the maximum likelihood estimate. (d) Transitions from X_1Y_0 to either X_0Y_0 or X_1Y_1 also result in a completely flat likelihood surface.

between X_0Y_0 and X_1Y_1 are no longer possible. For example, take transitions between states of character X. Both the correlated and independent models estimate transitions from X_0 to X_1 as necessarily linked to Y_0, since X_0Y_1 is not observed in the data set. The collapsed model is neither an independent nor correlated model because it no longer describes the coevolution of multiple characters, but instead transitions among states of

a single composite character. Put another way, rather than having two characters with two mutually exclusive states (X_0 or X_1 and Y_0 or Y_1), we are left with three mutually exclusive states of a single character (X_0Y_0 or X_0Y_1 or X_1Y_1). There is no way to make meaningful comparisons between possible intermediate pathways in which one character influences the other.

Including a collapsed model as part of our model set drastically changes the results. We found complete support for a collapsed state model for both Darwin's scenario and Unreplicated Bursts (see Supplemental Materials). The average AIC weight for the collapsed model is 99.7% under Darwin's scenario and 100.0% under an Unreplicated Burst scenario. This suggests that the support for the correlated models over simpler independent models is the result of parameter constraint. Specifically, in an independent model, transitions between observed states are constrained to be identical to transitions between unobserved states (e.g., X_0Y_0 to X_0Y_1 must be identical to X_0Y_1 to X_1Y_1, even if X_0Y_1 is never observed). In contrast, the correlated model is not subject to these constraints. This is, of course, the important distinction between the two models and what allows us to test for correlated evolution. However, when exclusively modeling observed state combinations the independent and correlated models become equivalent descriptions of the evolutionary process and are, therefore, indistinguishable from the given data.

Although we do so here for illustrative purposes, the collapsed model is typically not necessary to include within our modeling set even if we are interested in testing for evidence of correlated evolution without observations of intermediate states. We expect that the data sets in which empiricists are generally interested in testing for correlation do not lack observations of intermediate states and that the reason for Maddison and FitzJohn (2015) using such scenarios was to illustrate a consistent bias in our methods.

Rate Heterogeneity Is Necessary When Testing for Correlation Between Categorical Variables

Beyond not being able to directly test for character correlation, a major issue for the collapsed model described above is that in Darwin's scenario, a single observation of X_0Y_1 and X_1Y_0 removes the possibility of collapsing the model structure. As we will show, with only a single observation of intermediate character combinations, support for the correlated model over an independent model remains substantial. Even so, the results above highlight information limitations and that the strong evidence for correlated models may be due to a lack of viable alternative independent models rather than being irrefutable evidence of correlation (see also Gardner and Organ 2021).

It is worth considering again the possible explanations of the data under Darwin's scenario. One possibility is that the characters X and Y evolve slowly and that their codistribution is the result of two independent events deep in time. The probability of this scenario has been explored in-depth and its implausibility is a major contributor to the recurrent issues of false correlation when comparing correlated and independent models (Uyeda et al. 2018). We propose a complementary explanation for the correlated model's support: the independent model structure fixes the transition X_0 to X_1 to always be the same rate regardless of the state of Y (Fig. 2a), whereas a correlated model structure allows transitions from X_0 to X_1 to vary depending on the state of Y (Fig. 2b). Part of the support for the correlated model, therefore, comes from the fact that these data sets contain a signal of multiple transition rates for each character. The most likely description of the process is one in which the transition rates from X_0 to X_1 and Y_0 to Y_1 are high within the focal clade and occur low outside of the focal clade. The relative stasis of X_0 outside the focal clade and the rapid transition to X_1 within the clade suggests that changes in X are not consistent throughout the tree.

HMMs are a natural way to deal with this kind of rate heterogeneity across the tree. The underlying mathematical framework of an HMM is no different than a typical Markov model. They utilize a rate matrix, $\mathbf{Q}$, to estimate the probabilities of transitioning between discrete states and arrive at the likelihood of the model given the observed data set (Felsenstein and Churchill 1996). However, HMMs introduce a so-called "hidden state," which can represent any number of unobserved factors, biological or otherwise. Based on the presence or absence of this hidden state, changes between observed states are allowed to vary. In the most extreme cases, the absence of the hidden state may halt the evolutionary process and result in periods of stasis. For example, Marazzi et al. (2012) conceptualized the hidden state as a "precursor" trait and only in its presence could extrafloral nectaries (EFNs) emerge. The precursor state was never directly observed and the information for its presence or absence of the hidden state came from the rate heterogeneity of EFNs transitions. In some parts of the tree, the model EFNs emerged rapidly and in others, there were periods of stasis. Of course, HMMs are more general than either halting or actuating the evolutionary process and are used to quantify rate heterogeneity without the necessity of stasis (e.g., comparing fast, slow, or intermediate rates as in Beaulieu et al. 2013). The key point here is that they allow for rate heterogeneity that is unlinked to another observed character.

We developed and tested a hidden Markov independent model (HMIM) which accounts for rate heterogeneity while maintaining the independence of the observed focal characters X and Y (Fig. 4). In our view, the inclusion of our model within the evaluated set better levels the playing field between correlated and independent models. For example, if we focus on character X, our proposed model utilizes hidden states to vary transition rates between X_0 and X_1 based on an unobserved character. This is similar to the way that the

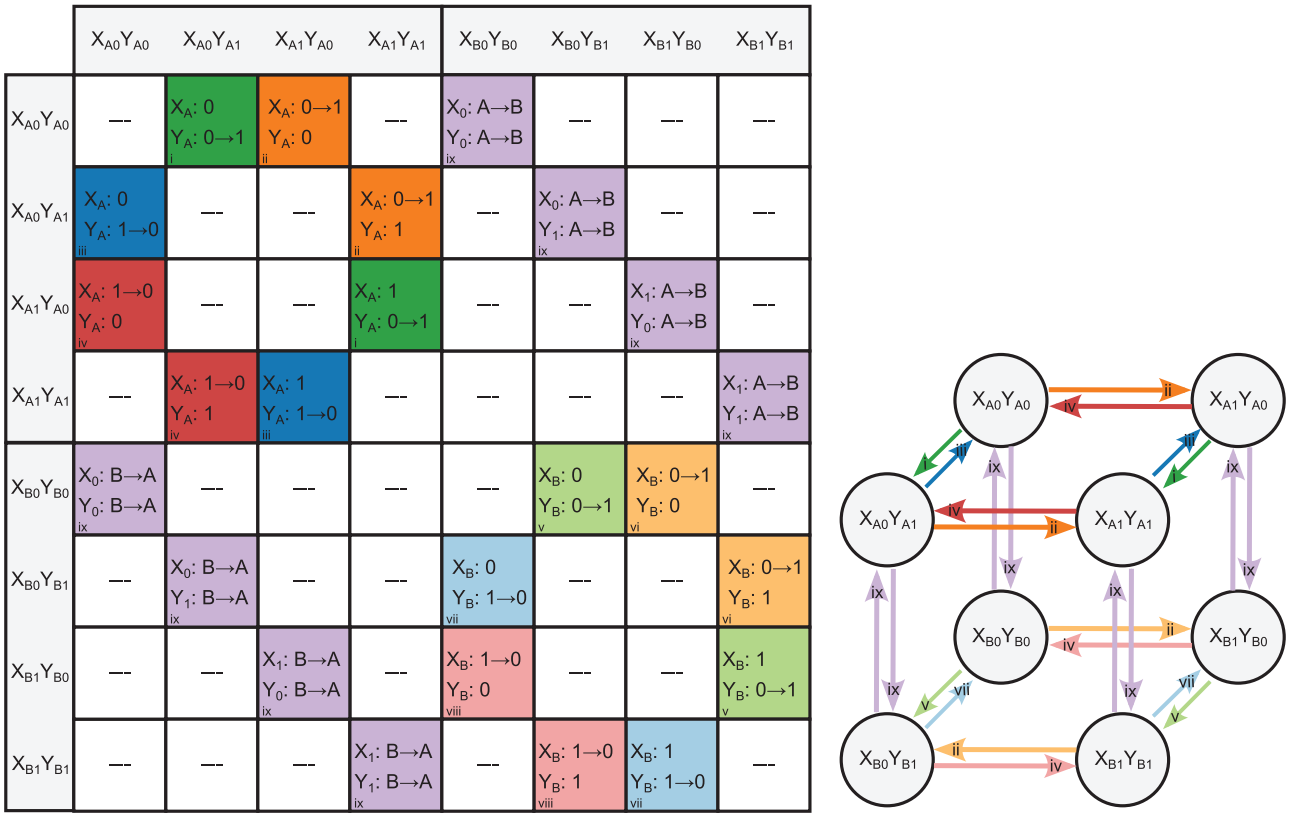

FIGURE 4. The HMIM that allows transitions within a character to have rate heterogeneity without it necessarily being linked to an observed character. This matrix can be read as a block matrix, with 4 × 4 blocks representing transitions between observed characters following an independent model (top left and bottom right) and transitions between hidden rate classes A and B (top right and bottom left). The independent model is essentially duplicated in the top left (blue and green) and bottom right (red and orange) of the block matrix with transitions occurring between these different types of independent models (purple). Here, transition rates between the hidden states are fixed to be the same (parameter ix), but it is straightforward to allow the transition between rate class A and B to differ.

correlated model allows transition rates between X_0 and X_1 to differ based on the observed state of Y. If the cause of false correlation was, as we suspect, not accounting for rate heterogeneity, then both the hidden-state independent and correlated model should be preferable to the simple independent model and evidence of correlation between X and Y should be greatly reduced.

We first removed the possibility of collapsing the Markov model by modifying Darwin's scenario. We defined the focal clade as being the monophyletic group where all observations of X_1Y_1 occur and randomly add the intermediate state observations of X_0Y_1 and X_1Y_0 within the focal clade (which refer to as "inside" hereafter), outside of the focal clade (which we refer to as "outside" hereafter), and both within and outside the focal clade (which refer to as "both" hereafter) (Fig. 5). Next, we verified that this modified Darwin's scenario still suffers from the problems of the original Darwin's scenario by comparing the independent and correlated models *sensu* Pagel (1994). We then added the HMIM to the model set and evaluated two questions: (1) when comparing independent models to one another, is there evidence of rate heterogeneity? and (2) is support for the correlated model reduced when compared to an independent model with rate heterogeneity? In addition to AIC weight, we utilized evidence ratios (ERs) to explore the relative likelihood of our models. ERs are a simple extension of AIC weights, but as a means of evaluation, are important here since they allow us to focus on evaluating the relative evidence of pairs of models irrespective of other models in the set (Burnham and Anderson 2002). The evidence for model i over model j is the ratio between their AIC weights: ER = w_i/w_j and it can help quantify whether the best model in our comparison is convincingly best. With alternative samples, a convincingly best model is likely to be chosen again from sample to sample. However, if evidence for a model is low, we expect model selection uncertainty to be high. Following Burnham and Anderson (2002), an ER of greater than 2.7 is used as a guide to justify judging support for one model being better than another. This also neatly corresponds to a ΔAIC = 2. We emphasize that this value should *not* be misconstrued as a significant test in a frequentist sense since we are not evaluating the probability of rejecting a null hypothesis.

For all modified Darwin's scenarios, we found substantial evidence (ER > 2.7) for a correlated model over a single-rate class independent model (Fig. 5). The

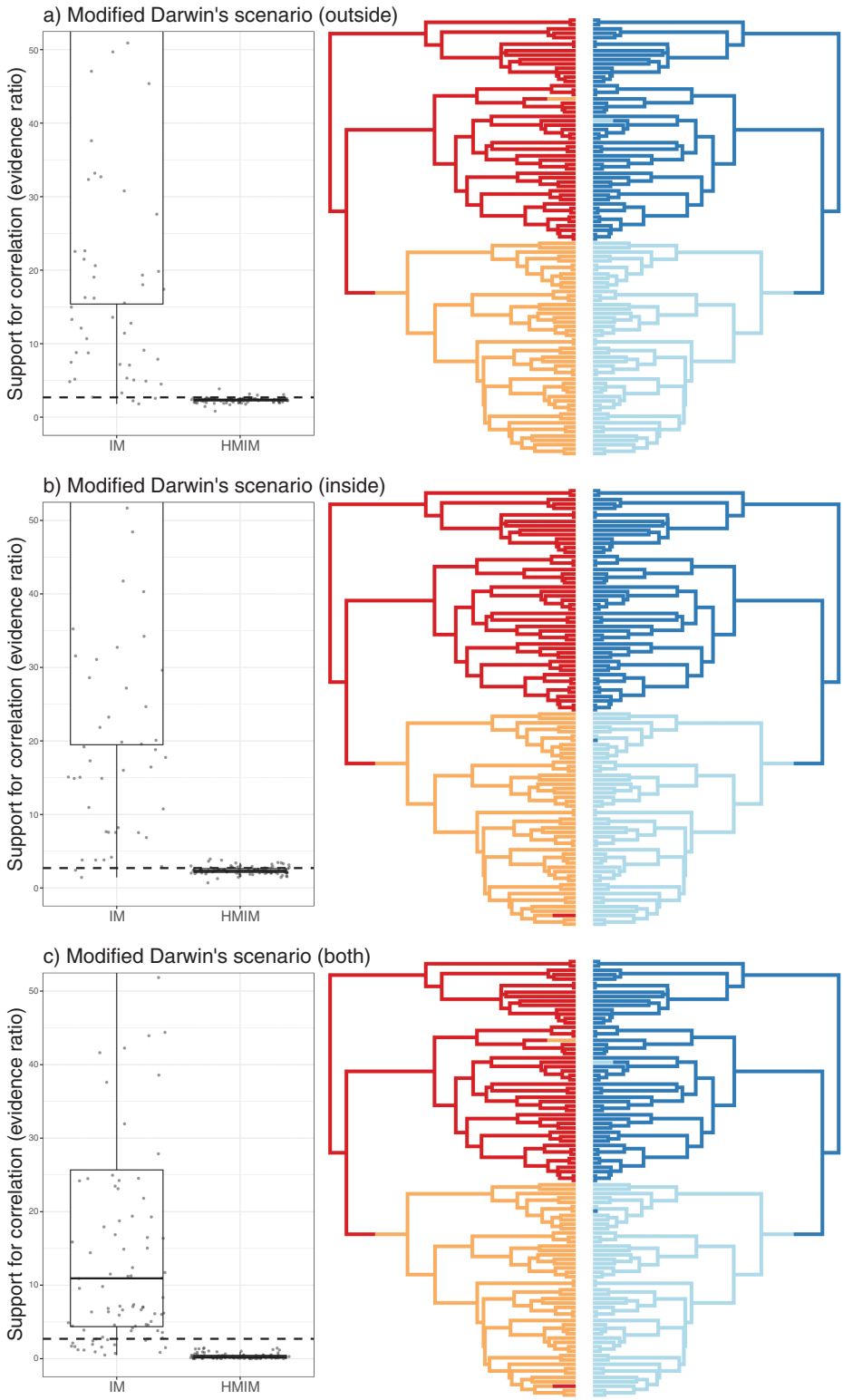

FIGURE 5. The amount of evidence for correlation when comparing a correlated model to ether an independent model (IM) or HMIM. The models are fit to data of the modified version of Darwin's scenario where a single observation of X_0Y_1 and X_1Y_0 is added outside of the focal clade (a), inside of the focal clade (b), and both within and outside of the focal clade (c). Evidence ratios for each model comparison are plotted as box plots to the left of the simulation scenario. In all cases, the evidence ratio of the correlated model over the independent model is substantially greater than 2.7 (left box plot) but, the correlated model receives much less support over the hidden Markov independent model (right box plot).

geometric mean ER for the correlated model over the single-rate independent model was $ER_{outside}$ = 59.51, ER_{inside} = 78.16, ER_{both} = 11.44 (Fig. 5), thus we again successfully recreated the conditions of Maddison and FitzJohn (2015) under a modified Darwin's scenario. Next, we examined the evidence for rate heterogeneity by comparing a single-rate independent model to the HMIM. We found substantial evidence for rate heterogeneity across all scenarios, with all mean ERs of the HMIM over the standard independent model well over 20, indicating substantial support for rate heterogeneity ($ER_{outside}$ = 24.45, ER_{inside} = 24.33, ER_{both} = 50.45). Finally, we tested whether there is still conclusive evidence of correlation between characters if we include the hidden-state independent model within our modeling set. We found that the evidence for a correlated model over the HMIM was greatly reduced when compared to the single-rate class independent model (Fig. 5; $ER_{outside}$ = 2.43, ER_{inside} = 3.21, ER_{both} = 0.22; Fig. 5). In fact, with only two observations of each intermediate state combination (X_0Y_1 and X_1Y_0), support for the HMIM over the correlated model was substantial (evidence for HMIM over a correlated model: ER_{both} = 4.41).

However, for both the inside and outside scenarios (as well as the original Darwin's and Unreplicated Bursts scenarios), support for a correlated model was still greater than the HMIM (Table 1). Specifically, we found that the difference in support for a correlated model over an HMIM is between 1 and 2 AIC units for all but Darwin's scenario, where ΔAIC = 8.8. We will address Darwin's scenario in detail below, but for the other scenarios, it is interesting to note the near identical likelihoods of the HMIM and correlated model. For reference, a ΔAIC of 2 corresponds to exactly the penalty of adding one additional parameter, and this is the number of parameters that differs between the HMIM and correlated model. Like our findings when examining the collapsed model, this suggests that the HMIM and correlated model explain the data equally well and there may not be enough information to determine whether there is a dependent relationship between the focal characters.

In summary, our findings thus far highlight three important insights: 1) There is indeed substantial evidence of rate heterogeneity, and that this is leading to a biased signal of false correlation, 2) including an HMIM will, at least, muddle evidence for correlation, and 3) for data sets similar to Unreplicated Bursts and Darwin's scenario, there may be a lack of information in to distinguish between the signals of correlated and independent evolution.

Exploring Model Space Using the Flexible Hidden Markov Model Framework

It still concerns us that for the original and two of the modified Darwin's scenarios (specifically the "outside" and "inside" sets; see Fig. 5), support for the correlated model was often greater than the hidden-state independent model. Specifically, the difference in support between the correlated model and the HMIM was ΔAIC = 8.8 under a strict Darwin's scenario corresponding to substantial evidence of correlation. To further examine this issue, we rely on the inherit flexibility of Markov and HMMs to structure a model specifically to address Darwin's scenario. We apply what we have learned thus far, with regards to the over-parameterization and the necessity of rate heterogeneity, and add a new set of simplified models. Although these simplified models can have some utility in empirical settings, they are primarily used here to demonstrate the flexibility and necessity of the HMM approach in creating a robust model set.

Model space has been underexplored and there are many nested model structures that are consistent with either independence or correlation depending on their constraints (see also Pagel and Meade 2006). Here we

Table 1. Average ΔAIC values for 100 data sets with standard deviations shown in brackets.

Scenario	Darwin's	Unreplicated bursts	Modified Darwin's (outside)	Modified Darwin's (inside)	Modified Darwin's (both)
Collapsed	**0.0 (±0.0)**	**0.0 (±0.0)**	NA	NA	NA
Independent	17.9 (±12.3)	36.8 (±9.0)	14.3 (±3.6)	14.8 (±4.0)	15.6 (±4.7)
Simplified independent	13.9 (±2.3)	67.3 (±15.8)	10.3 (±3.6)	10.8 (±4.0)	11.6 (±4.7)
Correlated	12.0 (±0.2)	8.0 (±0.1)	6.1 (±0.5)	6.1 (±0.7)	10.8 (±2.6)
Simplified correlated	13.9 (±2.3)	30.0 (±8.2)	9.8 (±3.6)	10.4 (±4.1)	11.6 (±4.7)
Hidden Markov independent	20.8 (±6.8)	9.2 (±0.4)	7.9 (±1.2)	8.4 (±3.3)	7.8 (±2.2)
Simplified hidden Markov independent	5.5 (±0.1)	36.3 (±9.1)	**0.0 (±0.0)**	**0.0 (±0.0)**	**0.0 (±0.0)**
Correlated hidden Markov	29.7 (±0.3)	24.9 (±0.8)	22.9 (±0.7)	23.5 (±0.8)	23.2 (±1.4)
Simplified correlated hidden Markov	18.8 (±2.1)	34.3 (±7.7)	14.2 (±3.3)	14.3 (±2.8)	15.7 (±3.5)

Notes: Each column represents a scenario described in the main text and each row represents a different Markov model structure which may be consistent with independence or correlation. For each scenario, eight or nine models were fit to the data sets. The collapsed model is fit only when not all potential state combinations are directly observed and therefore are not fit in modified scenarios. A ΔAIC of 0 indicates the best model and models within two AIC units of each other are generally considered good fits to the data (Burnham and Anderson 2002).

Bold indicates the model with the lowest AIC.

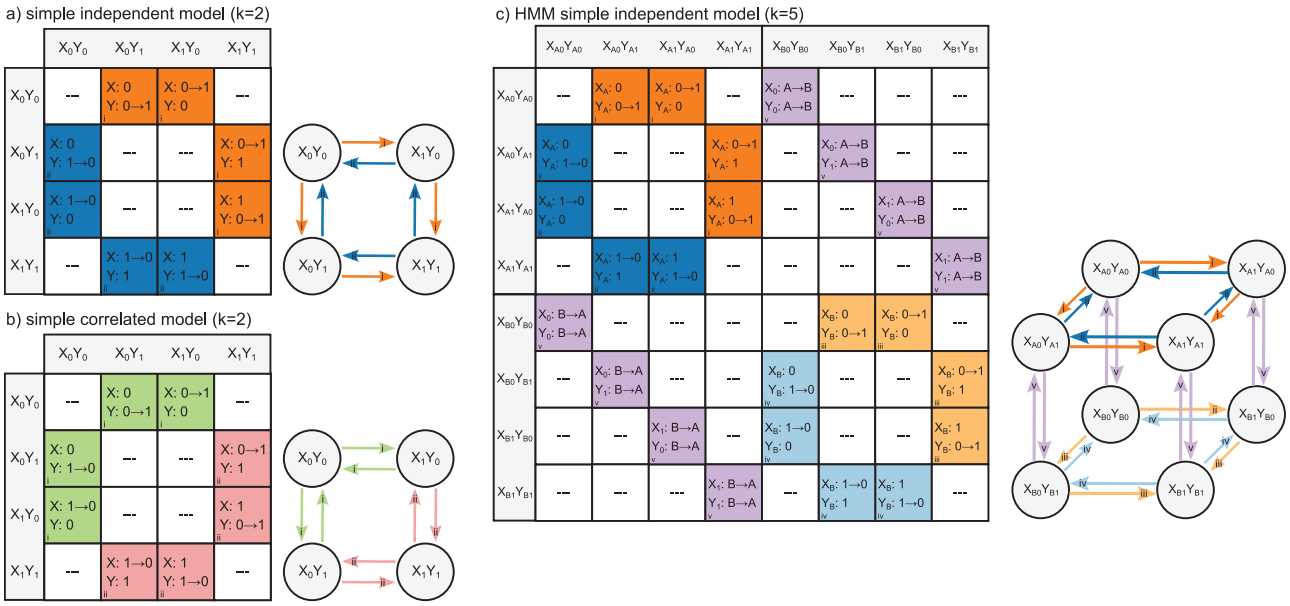

FIGURE 6. (a) A simplified independent model. In this model, transitions from 0 to 1 all occur at the same rate and transitions from 1 to 0 all occur at the same rate. (b) A simplified correlated model. Under this model, transitions between states of character X and Y depend on the background state of the other character. (c) A simplified hidden Markov independent model, where the simple independent model of (a) is used in the hidden Markov framework which allows for rate heterogeneity independent of focal characters. The same can be done for the simple correlated model (not shown).

describe two constrained versions of the independent and correlated models that achieve an efficient description of the data. One simplified version of the correlated model suggests that when either character X or Y is in state 0, rates of change are slower or faster than when either character is in state 1 (Fig. 6b). We refer to this as the "simplified correlated" model and it represents the simplest way to model a dependent relationship between two binary characters. Next, we created a "simplified independent" model of equal parameterization to the simplified correlated model, which equates all changes from 0 to 1 regardless of the character and the same is done for changes from 1 to 0 (Pagel and Meade 2006; Fig. 6a).

The structures of these simplified models have certain qualities that may make them apt descriptions of data like Darwin's scenario. Primarily, these models suggest that changes between states 0 and 1 do not necessarily depend on the specific identity of character X or Y since they are constrained to be equal. Considering the redundancy of a data set composed of two synapomorphies, it is obvious that there is little to no information that distinguishes the two characters—that is, it makes no difference whether one analyzes character X or character Y since their distributions are identical. The simplified models make that assumption explicit. It is also important to note that the simplified independent model and simplified correlated model maintain independence and dependence *sensu* Pagel (1994). The background state of the unchanging character does not influence changes in the case of the simplified independent model, whereas the background state of the unchanging character will influence rates of change in the case of the simplified correlated model (Pagel and Meade 2006). Finally, we can introduce rate heterogeneity by modeling the simplified independent and correlated models as two rate class HMMs (Fig. 6c).

Returning to the modified Darwin's scenario data sets, consistent and overwhelming support for the simplified HMIM was found across all scenarios (Table 1). The average AIC weight of the simplified HMIM when fit to modified Darwin's scenarios are $w_{\text{outisde}} = 89.6\%$, $w_{\text{inside}} = 90.2\%$, and $w_{\text{both}} = 93.5\%$. The set of models applied to this data included all models discussed thus far as well as more complicated versions of those previously described (such as a standard correlated model with multiple rate classes). Additionally, to ensure that these models are not biased towards being favored across all data sets, we simulated data under a simplified correlated, simplified independent, and simplified HMIM s. We then fit each model to these data sets and found that the generating model is consistently chosen as the best fitting model (see Supplemental Materials).

Support for a simplified HMIM over alternative correlated models was also found when we consider the unmodified Darwin's scenario (Table 1). This is because there is no information to detect whether transitioning between X_0Y_0 and X_1Y_1 happens more rapidly through an intermediate state of X_1Y_0 or X_0Y_1. This distinction is necessary for correlated models, but these two pathways are equivalent in the simplified HMIM. The hidden state in Darwin's scenario, then, represents the presence of some character that changes the dynamics of how the state X_1Y_1 accumulates – that is, the hidden state may represent the "trait" of "being a mammal." The subsequent simultaneous evolution of fur

and inner ear bones are then represented by higher transition rates from X_0Y_0 to X_1Y_1 in the mammalian rate class. However, there is no information present in Darwin's scenario to tell us whether fur or inner ear bones evolved as a consequence of one another. They simply shared the evolutionary dynamics and consequences of being a mammal.

With regards to the parameter estimates, there is no way to directly assess how exactly accurate they are, because the generating parameters are unknown given the contrived nature of Darwin's scenario. Nevertheless, we can make a back-of-the-envelope calculation to determine whether our estimates are at least reasonable and consistent with our expectations. The estimate for a transition to "being a mammal" (i.e., a transition in hidden rate class) was, on average, 0.0113 transitions per million years, which corresponds to an expected transition occurring, on average, every 88.5 Myr. With the average total branch length being 133.9 Myr across our 100 taxon trees that corresponds to an expectation of roughly one transition. We also found a relatively low standard deviation (±0.002), indicating this did not vary dramatically among data sets.

Finally, it is important to emphasize that our model set is also incomplete. We have examined only a handful of structures from a very large model space, and it is possible that there are equally good explanations of the data within the expanded model space. Nonetheless, these findings suggest that a model set without character-independent rate heterogeneity will consistently produce a statistical bias toward correlation noted by Maddison and Fitzjohn (2015). However, this bias can be greatly reduced by empiricists by accounting for, and expecting the presence of, rate heterogeneity through the use of HMMs.

A Broadly Applicable Framework

The issue discussed herein is recognized as being broadly applicable to several comparative methods that test for associations between variables (FitzJohn 2010; Rabosky and Goldberg 2015; Uyeda et al. 2018; Nakov et al. 2019; Gardner and Organ 2021). It is concerning that such a significant issue has seemingly gone unresolved for so long given comparative methods are of critical importance for understanding macroevolutionary patterns. However, in our view, the prevalence of the problems identified over the past few years is due to a singular overarching cause, namely, model misspecification, which occurs when a model, or set of models, is incomplete. Within the context of their model sets, authors of previous studies have correctly portrayed and analyzed the correlation bias of modeling dependence between discrete characters (Maddison and FitzJohn 2015; Uyeda et al. 2018; Gardner and Organ 2021). However, the danger of model misspecification is that the inferences drawn from an incomplete set are highly susceptible to unforeseen biases—a fact that will hold true in both theoretical and empirical contexts.

Here, we are arguing that the model set is incomplete without the inclusion of models that allow for rate heterogeneity that is independent of the focal characters. The canonical character-independent model of Pagel (1994) has no way to account for multiple rates of evolution, whereas support for a correlated model can come from both evidence of correlation and evidence of rate heterogeneity. Additional support for correlation as a consequence of hidden rate heterogeneity is not exclusive to Pagel's model and has been seen in other phylogenetic comparative methods. This stems from the fact that PCMs often test for correlation by comparing rates in the presence and absence of focal characters. Not being able to account for character-independent rate heterogeneity has led to consistently biased evidence towards correlation within state-dependent speciation extinction models (Beaulieu and O'Meara 2016). In that case, the biased association was between diversification rates and phenotype (Rabosky and Goldberg 2015), but the cause is the same. Models in which there are no differences in diversification rates are compared to models which tested for the presence of a correlation between character and diversification rate (which necessarily allows for multiple rates of diversification). Whether it be speciation or phenotypic evolution or both, if rates vary as a rule of macroevolution (Simpson 1944), then the inclusion of models which allow rate heterogeneity independent of focal characters is necessary within any model set.

One difference between the problem of false correlation in SSE models and the problems within simpler Markov models is the narrative surrounding them. In the case of SSE models, the problem was viewed as a high false positive rate (Rabosky and Goldberg 2015), whereas in the case of discrete character evolution, we are led toward viewing rate heterogeneity through the lens of single unreplicated evolutionary events (Maddison and FitzJohn 2015). However, both points contribute to the same problem and if we view single evolutionary events as examples of where evolution has changed in tempo or mode, then the inclusion of HMMs as a way forward arises naturally from the problem.

Since we as comparative biologists are involved in historical science, we will inevitably encounter single evolutionary events of large importance. However, it must be recognized that data sets that are susceptible to biases from singular events are not amenable to most phylogenetic comparative tests. Although here we have reduced the statistical biases associated with false correlations, there is no amount of methodological massaging that will allow for a satisfying test of macroevolutionary correlation between two synapomorphies. This is because comparative methods rely on several independent replicates of correlation such that the associations found between the variables may be considered robust even when extended beyond the data set used for the analysis. If there is only one example of the correlation arising in the entire data set, we should not have confidence in extending our inferences beyond the

clade and should be wary of the correlation even within the focal clade. However, that is not to say there is no mechanistic reason for an association between synapomorphies. It is entirely possible that two characters that share identical evolutionary histories have an underlying biological link. Nonetheless, conclusions about the potential links between these characters cannot come from studies conducted on a macroevolutionary scale, and they should instead be investigated on a smaller scale (Beaulieu and O'Meara 2018, 2019; Donoghue and Edwards 2019). Additional lines of evidence and a more mechanistic explanation will be necessary in order for a conclusion of correlation to be satisfying (Gardner and Organ 2021). In a sense, the hidden rate classes of our proposed framework may represent lineage-specific factors that, once present, readily allow for a shift in the tempo and mode of a lineage's evolution (Maddison and FitzJohn 2015; Ogburn and Edwards 2015).

Concluding Remarks

Sparked by an appreciation of the limitations of PCMs, several commonly used phylogenetic comparative methods have seen critical challenges recently, which have led to advancements useful for both developers and users (Boettiger et al. 2012; Maddison and FitzJohn 2015; Rabosky and Goldberg 2015; Louca and Pennell 2020). Here, too, the critiques of classic tests of correlation (Pagel 1994) are not wrong, and the recommendations of past studies remain useful (Maddison and FitzJohn 2015; Uyeda et al. 2018; Gardner and Organ 2021). There will be data sets where distinguishing between correlation and independence is simply not possible without lines of evidence outside of comparative biology (Uyeda et al. 2018; Gardner and Organ 2021). What we have demonstrated here is that the statistical bias toward correlation is primarily due to a misspecification of the model set and a failure to account for character-independent rate heterogeneity. We have highlighted that the inclusion of less frequently used Markov model structures in the model set can be critical for the quality of the inferences being made. We acknowledge that choosing a diverse set of models *a priori* is not always straightforward, but both likelihood and Bayesian methods will only be as effective as the plausibility of the models set being analyzed (Burnham and Anderson 2002). We know that a homogeneous process over millions of years and across thousands of lineages is incorrect (Eldredge and Gould 1972) and that the individual parts of an organism do not evolve independently (Levins and Lewontin 1985). While we may not be able to always specify each of these individual processes, we must try to incorporate them in our modeling. Accounting for rate heterogeneity through HMMs is a simplified way that we can bring realism to our modeling while also making statistically consistent and unbiased estimates of evolutionary parameters. From there, undoubtedly more work will be necessary (e.g., Goldberg and Foo 2020). But comparative analyses must at the very least attempt to account for what we know about macroevolution while making us aware of the wonderful idiosyncrasies of evolutionary history.

Funding

This work was funded by the National Science Foundation grants DEB–1916558.

Acknowledgements

We thank Jacob Gardner and an anonymous reviewer for their comments on an earlier draft of this work. We would also like to thank Brian O'Meara for his comments and discussion of the ideas presented here. Finally, we thank Thais Vasconcelos and Eric Hagen for their helpful comments and edits of this work at various stages.

Data Availability

Data and a user guide are available from the following github repository: https://github.com/jboyko/2022_unsolved-challenge.

References

Beaulieu J.M., Donoghue M.J. 2013. Fruit evolution and diversification in campanulid angiosperms. Evolution 67:3132–3144.

Beaulieu J.M., O'Meara B.C. 2018. Can we build it? Yes we can, but should we use it? Assessing the quality and value of a very large phylogeny of campanulid angiosperms. Am. J. Bot. 105:417–432.

Beaulieu J.M., O'Meara B.C. 2019. Diversity and skepticism are vital for comparative biology: a response to Donoghue and Edwards (2019). Am. J. Bot. 106:613–617.

Beaulieu J.M., O'Meara B.C., Donoghue M.J. 2013. Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in Campanulid angiosperms. Syst. Biol. 62:725–737.

Boettiger C., Coop G., Ralph P. 2012. Is your phylogeny informative? Measuring the power of comparative methods. Evolution 66:2240–2251.

Boyko J.D., Beaulieu J.M. 2021. Generalized hidden Markov models for phylogenetic comparative datasets. Methods Ecol. Evol. 12:468–478.

Burnham K.P., Anderson D.R. 2002. Model selection and multimodel inference: a practical information-theoretic approach. New York: Springer.

Darwin C. 1859. On the origin of species, 1859. London, England, UK: Routledge.

Donoghue M.J., Edwards E.J. 2019. Model clades are vital for comparative biology, and ascertainment bias is not a problem in practice: a response to Beaulieu and O'Meara (2018). Am. J. Bot. 106:327–330.

Eldredge N., Gould S.J. 1972. Punctuated equilibria: an alternative to phyletic gradualism. Models Paleobiol. 1972:82–115.

Felsenstein J. 1985. Phylogenies and the comparative method. Am. Nat. 125:1–15.

Felsenstein J., Churchill G.A. 1996. A hidden Markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol. 13:93–104.

FitzJohn R.G. 2010. Quantitative traits and diversification. Syst. Biol. 59:619–633.
Gardner J.D., Organ C.L. 2021. Evolutionary sample size and consilience in phylogenetic comparative analysis. Syst. Biol. 70:1061–1075.
Goldberg E.E., Foo J. 2020. Memory in trait macroevolution. Am. Nat. 195:300–314.
Levins R., Lewontin R. 1985. The dialectical biologist. Cambridge, MA, USA: Harvard University Press.
Louca S., Pennell M.W. 2020. Extant timetrees are consistent with a myriad of diversification histories. Nature 580:502–505.
Maddison W.P., FitzJohn R.G. 2015. The unsolved challenge to phylogenetic correlation tests for categorical characters. Syst. Biol. 64:127–136.
Marazzi B., Ané C., Simon M.F., Delgado-Salinas A., Luckow M., Sanderson M.J. 2012. Locating evolutionary precursors on a phylogenetic tree. Evolution 66:3918–3930.
Nakov T., Beaulieu J.M., Alverson A.J. 2019. Diatoms diversify and turn over faster in freshwater than marine environments. Evolution 73:2497–2511.
Ogburn M.R., Edwards E.J. 2015. Life history lability underlies rapid climate niche evolution in the angiosperm clade Montiaceae. Mol. Phylogenet. Evol. 92:181–192.
Pagel M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proc. R. Soc. B: Biol. Sci. 255:37–45.
Pagel M., Meade A. 2006. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov Chain Monte Carlo. Am. Nat. 167:808–825.
Rabosky D.L., Goldberg E.E. 2015. Model inadequacy and mistaken inferences of trait-dependent speciation. Syst. Biol. 64:340–355.
Simpson G.G. 1944. Tempo and mode in evolution. New York, NY, USA: Columbia University Press.
Uyeda J.C., Zenil-Ferguson R., Pennell M.W. 2018. Rethinking phylogenetic comparative methods. Syst. Biol. 67:1091–1109.