Generalized Hidden Markov Models for Phylogenetic Comparative Datasets

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

[1]*Department of Biological Sciences, University of Arkansas, Fayetteville, AR, 72701, USA*

**To whom correspondence should be addressed (jboyko@uark.edu).*

Running head: corHMM: Generalized hidden Markov models

1 Hidden Markov Models for phylogenetic comparative datasets

Abstract

1. Hidden Markov models (HMM) have emerged as an important tool for understanding the evolution of characters that take on discrete states. Their flexibility and biological sensibility make them appealing for many phylogenetic comparative applications.
2. Previously available packages placed unnecessary limits on the number of observed and hidden states that can be considered when estimating transition rates and inferring ancestral states on a phylogeny.
3. To address these issues, we expanded the capabilities of the R package `corHMM` to handle *n*-state and *n*-character problems and provide users with a streamlined set of functions to create custom HMMs for any biological question of arbitrary complexity.
4. We show that increasing the number of observed states increases the accuracy of ancestral state reconstruction. We also explore the conditions for when an HMM is most effective, finding that an HMM outperforms a Markov model when the degree of rate heterogeneity is moderate to high.
5. Finally, we demonstrate the importance of these generalizations by reconstructing the morphology of the ancestral angiosperm flower. Exactly opposite to previous results, we find the most likely state to be a spiral perianth, spiral androecium, whorled gynoecium. The difference between our analysis and previous studies was that our modeling allowed for the correlated evolution of several flower characters.

Keywords: ancestral state reconstruction, discrete character evolution, hidden Markov models, hidden states, stochastic map.

1. Introduction

Hidden Markov models (HMMs) are important centerpieces in many biological applications (Eddy, 2004; Yang Lou, 2017). They provide a natural framework for comparative biologists, particularly for relaxing assumptions about homogeneous evolution through time and across taxa without vastly increasing the number of parameters (e.g., Felsenstein & Churchill, 1996; Galtier, 2001; Penny, McComish, Charleston, & Hendy, 2001; Beaulieu, O'Meara, & Donoghue, 2013; Beaulieu & O'Meara, 2016). For instance, simple models of binary character evolution make sense for small, young clades, because a single set of transition rates seems like a reasonable assumption. However, homogeneous rates are unlikely to explain the evolution of the same character across a much larger and older clade in which transition rates may differ dramatically among subclades, perhaps due to correlations with traits that were not included in the model. This observation was the motivation for the development of the hidden rate model (HRM) of Beaulieu *et al.* (2013), which uses a hidden Markov approach to objectively locate regions of a phylogeny where hidden factors have either promoted or constrained the evolutionary process for a binary character.

Within comparative biology, HMMs have been applied as both standalone models (Beaulieu et al., 2013) and in combination with other phylogenetic models (e.g., hidden state-dependent speciation and extinction models, Beaulieu & O'Meara, 2016). Hidden Markov models can be used to address many problems in comparative biology (Siepel & Haussler, 2005) and their flexibility allows biologists to create models tailored to their specific hypotheses. However, previous implementations of HMMs for comparative methods have placed limitations on the number of observed and hidden states. For instance, the implementation of the HRM model of Beaulieu et al. (2013) is restricted only to the analysis of binary characters. There is no

mathematical basis for limiting the number of observed states or hidden states in an HMM, and such constraints necessitate a simplification of datasets and candidate models.

We describe a new version of `corHMM` that implements *n*-state HMMs. This does not require new algorithms or a different likelihood function. Instead, we optimized and generalized existing code so users can create custom HMMs for any biological question of arbitrary complexity. We have also added a number of "quality of life" improvements that make `corHMM` much easier to use and interpret, including an implementation of stochastic character mapping (simmap; Bollback, 2006). Additionally, we demonstrate the effectiveness of HMMs to identify rate heterogeneity when it is present, and we outline the informational advantages of increasing the number observed and hidden states in discrete character data sets. Finally, to demonstrate the importance of this generalization we apply `corHMM` to reconstruct the morphology of the ancestral angiosperm flower.

2. Materials and Methods

2.1 Generalizing HMMs

From a technical standpoint, hidden Markov models have a hierarchical structure that can be broken down into two components: a "state-dependent process" (Fig. 1a,b) and an unobserved "parameter process" (Fig. 1c)(Zucchini, MacDonald, & Langrock, 2017). In comparative biology, for characters that take on discrete states the standard "state-dependent process" is a continuous-time Markov chain with finite state-space (CTMC-FS). The benefit of a Markov model is its simplicity — to calculate the probabilities of observed discrete states at the tips of a phylogeny all that is required is a tree, a transition model describing transitions among a set of observed states, and frequencies at the root (O'Meara, 2012; Fig. 1a,b). The observed states

could be any discretized trait such as presence or absence of extrafloral nectaries (Marazzi et al., 2012), woody or herbaceous growth habit (Beaulieu et al., 2013), or diet state across all animals (Román-Palacios, Scholl, & Wiens, 2019). However, a simple Markov process that assumes homogeneity through time and across taxa is often not adequate to capture the variation of real datasets (e.g. Beaulieu *et al.*, 2013). Under an HMM, observations are generated by a given state-dependent process, which in turn depends on the state of the parameter process. In other words, the observed data are the product of several processes occurring in different parts of a phylogeny and the parameter process is way of linking them. It is initially unknown what the parameter process corresponds to biologically, hence the moniker "hidden" state. Nevertheless, the information for detecting hidden states comes from the differences in how the observed states change. As long as the transitions between observed states of different lineages are more adequately described by several Markov processes rather than a single process, there will be information to detect hidden states (see *3.1 Performance in Simulations*).

The likelihood of any HMM is obtained by maximizing the standard likelihood formula, $L = P(D|\mathbf{Q},T)$, for observing character states , D, across a set of extant taxa, given the continuous-time Markov model **Q**, and a fixed topology with a set of branch lengths (denoted by T). For a binary character, **Q** is a 2×2 transition matrix representing the transition rates, whose entries define transitions between the character states, 0 and 1. To form an HMM, we expand **Q** to accommodate both observed and hidden states. Formally, the HMM can be generalized to include any number of observed states (e.g., 0, 1, 2), and hidden states (e.g., A, B, C). Following Beaulieu and O'Meara (2016), the state space is defined as o being the index of the observed state, $o \in 0,1,\ldots,\alpha$, and h as the index of the hidden state, $h \in A,B,\ldots,\beta$. Thus, a given model will have, in general, $|o| \times |h|$ states. In `corHMM`, the model **Q** is defined by amalgamating each

of the state-dependent processes and the parameter process specified in the model. For example, if we have state-dependent matrices, $\boldsymbol{R}$,

$$\boldsymbol{R}_1 = \begin{pmatrix} - & 1 & 2 \\ 1 & - & 3 \\ 2 & 3 & - \end{pmatrix}, \ \boldsymbol{R}_2 = \begin{pmatrix} - & 5 & - \\ 4 & - & 7 \\ - & 6 & - \end{pmatrix},$$

that are related by a parameter-process $\boldsymbol{P}$,

$$\boldsymbol{P} = \begin{pmatrix} - & 9 \\ 8 & - \end{pmatrix} = \begin{pmatrix} - & r_{R1\to R2} \\ r_{R2\to R1} & - \end{pmatrix},$$

where entries $r_{R1\to R2}$ and $r_{R2\to R1}$ define transition rates between the state-dependent processes, we can extend Eq. 2 of Tarasov (2019) to amalgamate these processes,

$$\mathbf{Q} = A(\boldsymbol{R}_1, \boldsymbol{R}_2, \boldsymbol{P}) = \begin{pmatrix} - & 1 & 2 & 9 & - & - \\ 1 & - & 3 & - & 9 & - \\ 2 & 3 & - & - & - & 9 \\ 8 & - & - & - & 5 & - \\ - & 8 & - & 4 & - & 7 \\ - & - & 8 & - & 6 & - \end{pmatrix}.$$

This matrix can be understood as a block matrix where the diagonal blocks are the state-dependent processes $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$, and the off-diagonal blocks are the parameter process, $\boldsymbol{P}$, that describe transitions between $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$. The resulting $\mathbf{Q}$ can simply be evaluated using the same likelihood formula as before. It should be noted that the observed character states are set to 1 when the trait is consistent with the observed data, and 0 otherwise. For example, we would set the probability to 1 for both 1A and 1B for all species exhibiting state 1.

The general formulation of an HMM can easily be extended to examine the correlated evolution of multiple characters (Pagel, 1994). For example, consider a case of two binary characters where trait 1 defines the presence or absence of fleshiness of fruits, and trait 2 defines whether or not the fruits are animal-dispersed. At most there are four binary combinations of these characters (i.e., 00, 01, 10, and 11). But, it can also be coded as a single multistate character, where 1=dry fruits not dispersed by animals, 2=dry fruits dispersed by animals,

3=fleshy fruits not dispersed by animals, and 4=fleshy fruits dispersed by animals. Therefore, transforming binary combinations to multistate characters also applies for two characters with a different number of observed states. In other words, one character could be binary (e.g., dry vs. fleshy fruit) and the other could be multistate (e.g., fruits dispersed mechanically, by wind, or by animal).

2.2 Simulation Study

We conducted a set of simulations to address two broad goals. The first was to demonstrate, from an informational standpoint, the advantage of increasing the number of observed states by comparing two-state, three-state, and four-state datasets. The second goal was to demonstrate the ability of hidden Markov models to detect varying degrees of rate heterogeneity. We then link these goals together by demonstrating that while HMMs naturally address rate heterogeneity in discrete characters, they can also recover some of the informational content of unobserved characters through the use of hidden states. These simulations are in no way exhaustive, but represent a set of reasonable questions that many beginning users might have about the behavior of HMMs.

2.2.1 Increasing the number of observed characters or states

To test the behavior of two-state, three-state, and four-state datasets we relied on ancestral state reconstruction (ASR) at nodes. ASR is a widely-utilized feature of `corHMM`, and it is important to know the accuracy of multistate ancestral reconstructions. Additionally, using ancestral states gives us a direct means to compare models with different datasets. A 500-tip phylogeny was simulated (birth rate set to 1 event Myr^{-1}, and death rate of 0.5 events Myr^{-1}) to

be used as a fixed tree. Datasets were simulated using transition rates sampled from a truncated normal distribution ($\mu = 1$, $\sigma = 0.5$), which were then scaled to have mean rates of 0.1, 1.0, or 10 transitions Myr^{-1} by dividing the rate matrix by the sum of the diagonal and then multiplying by the desired scalar. This resulted in a range of evolutionary rates. For each transition model, 100 datasets were simulated. The transition rates of each dataset were then estimated and their maximum likelihood estimates were used to infer marginal probabilities of each character state across the tree. This procedure was repeated 10 times.

An underappreciated concern with evaluating models that differ in the number of observed states is that the probability of guessing the correct state without any additional information is simply $1/k$ states. This could, in theory, inflate the accuracy of datasets with fewer states even though the tip states themselves provide no information about the ancestral states when the rates are high (Schultz, Cocroft, & Churchill, 1996; Sober & Steel, 2011, 2014). To deal with this issue, we also calculated the mutual information, measured in bits, about ancestral states from each dataset and model (Cover & Thomas, 1991; Sober & Steel, 2011). Specifically, mutual information is a measure of how much ancestral state uncertainty is reduced by knowing the tip states. The initial uncertainty, or unconditional entropy, is set by the model – given a model of evolution and no knowledge of the extant tips, how uncertain is the best guess of the ancestral states? The remaining uncertainty after ASR, or conditional entropy, is given by the combination of the model and the tip states – given the model of evolution and knowledge of the extant tips, how uncertain is the best guess of the ancestral states? It is important to note that information, just like ancestral state reconstruction, is highly correlated with the model of evolution, and thus any results related to information will take on the assumptions of the model.

We define information as the difference between the unconditional entropy of the node states, $H(X_v)$, and the entropy of the node states conditioned on the data, $H(X_v|X_h = D)$ (Cover & Thomas, 1991). The unconditional entropy of node v is defined as:

$$H(X_v) = -\sum_{i=1}^{k} \pi[X_v = i] log_2(\pi[X_v = i]),$$

where $\pi[X_v = i]$ is the prior probability of a node taking a particular state. For the root, the prior depends on user choice, as there are several options (Yang, Kumar, & Nei, 1995; Pagel, 1999; FitzJohn, Maddison, & Otto, 2009). Here we assume the prior probability on the root node is the expected equilibrium frequency, π, which is calculated directly from the transition model by solving $\pi \boldsymbol{Q} = 0$. This aligns our expectation of the root node with all other internal nodes such that, in the absence of information from the tips, the probability of a particular state is assumed to be drawn from the equilibrium frequencies. In other words, the information of the tip states decreases as rates increase and, ultimately, the probability of a node state becomes completely determined by the model. We define the conditional entropy as:

$$H(X_v|X_h = D) = -\sum_{i=1}^{k} P[X_v = i|X_h = D] log_2(P[X_v = i\ |X_h = D]),$$

where $P[X_v = i|X_h = D]$ is the conditional probability that a node is fixed as being in state i given the probability of observing the tip data (which is just the marginal probability of state i). In particular, we are interested in the average entropy of a node for all states $i \ldots k$, given we observe a particular dataset, $X_h = D$. Thus, the conditional entropy will vary by node, but the unconditional entropy is set by the model. To produce a measure of mutual information between the observations at the tips and estimates at internal nodes, we take the difference between the conditional entropy and the unconditional entropy and average across all nodes. However, the

unconditional entropies will be greater for datasets that include more states because unconditional entropy sets the upper limit of what is possible to learn. This alone could contribute to large informational differences between models with different numbers of observed states. Therefore, we also measure the proportion of maximum information gained $(\frac{mutual\ information}{unconditional\ entropy} \times 100\%)$.

2.2.2 Evaluating hidden Markov models

We evaluated the ability to detect rate heterogeneity by simulating data under an HMM. As outlined above (see *2.1 Generalizing HMMs*), there are two major axes along which an HMM differs from standard Markov models. First, we varied the magnitude of the difference in the state-dependent process by simulating data under a model where there was: (1) no difference between the state-dependent processes ($\boldsymbol{R}_1$=$\boldsymbol{R}_2$), (2) a 2-fold difference in rates between the state-dependent processes (e.g. if $\boldsymbol{R}_1$'s mean rate was 1 Myr^{-1}, $\boldsymbol{R}_2$ mean rate would be 2 Myr^{-1}), (3) a 10-fold difference between the state-dependent processes, and (4) a covarion-like trait model in which transitions among nucleotide states occur freely, all transition rates are zero, and evolution is essentially "turned off" (Penny et al., 2001). For all simulation scenarios, we set the parameter-process to have equal transition rates between state-dependent processes. In addition to examining ancestral state reconstruction at nodes, we also used the new `makeSimmap` to assess how well the model captures the expected number of character changes within and among all branches in the tree. For each of the 150 datasets simulated above, we evaluated 100 simmaps per model by counting the number of transitions for a given simmap.

Next, we tested the impact of the magnitude of the asymmetry in the underlying parameter-process. We simulated data where the state-dependent process always differed by 2-

fold, but for the underlying parameter-process there was: (1) no difference in transition rate ($r_{R1\rightarrow R2}$=$r_{R2\rightarrow R1}$), (2) 1.5× faster transition rate to the slower rate class ($r_{R1\rightarrow R2}$>$r_{R2\rightarrow R1}$), (3) 2× faster transition rate to the slower rate class, (4) 10× faster transition rate to the slower rate class. For each of the models described, we used the same 500-tip phylogeny as before to simulate 150 datasets.

Finally, we examined how much information is available when we allow for hidden states to be observed at the tips. We used the same data generated from simulations examining state-dependent differences, but this time we did not remove the hidden state. We then fit a Markov model to this full dataset and compared it to models in which the "second character" remained unobserved.

2.3 Case study: reconstructing the ancestral angiosperm flower

2.3.1 Background

Understanding the origin of flowering plants is widely considered to be one of the most important goals of systematic botany. In a recent study, Sauquet et al. (2017) compiled an extensive database of floral characteristics and attempted to reconstruct the morphology of the ancestral angiosperm flower. The paper's conclusions proved controversial. Sokoloff, Remizowa, Bateman, & Rudall (2018) disputed the original study's claim that the ancestral flower had a whorled perianth, whorled androecium, and spiral gynoecium, instead preferring the hypothesis that the ancestral flower was either entirely whorled or entirely spiraled. We do not dispute the biological claims put forth by either side. However, a major limitation of the original study, by the authors' own admission, was the fact that "no comparative method exists yet to account for the potential correlation of more than two discrete characters" (although see Beaulieu

& Donoghue, 2013). This represents a major problem for the successful reconstruction of the ancestral state because flowers are highly integrated structures with potentially several developmental constraints (Sauquet et al., 2017, 2018; Sokoloff et al., 2018). Treating the phyllotaxy of the perianth, androecium, and gynoecium as independent represents a major assumption with potentially large consequences on the ancestral state reconstruction.

2.3.2 Worked example and methods

We limited ourselves to including only the characters related to the phyllotaxy of the perianth, androecium, and gynoecium. Although it is possible to include other characters, given the corresponding increase of parameter space, we suspect that we would not have the power to accurately infer the model and ancestral states. An additional simplification was the exclusion of polymorphic species. Once again, `corHMM` is capable of analyzing such a dataset, but the rarity of these species would make maximum likelihood estimation far more uncertain. Thus, we treat the phyllotaxy of the perianth, androecium, and gynoecium as either “whorled” or “spiral”. We use the C series phylogeny of Sauquet et al. (2017) in which *Amborella* is constrained as the sister of angiosperms and *Monocotyledoneae*, *Ceratophyllaceae*, and *Eudicotyledoneae* are constrained to form a monophyletic group. In total, 291 taxa have a complete description of our focal characters.

In our case, we have three data columns each with two observed states. Because this dataset contains two or more columns of trait information, each column is automatically interpreted as an evolving character. In these cases, `corHMM` will also automatically remove dual transitions from the model since that would constitute two or more evolutionary events (Pagel, 1994; Maddison, Midford, Otto, & Oakley, 2007). For example, a lineage with a whorled

perianth, whorled androecium, and whorled gynoecium cannot evolve directly to have a spiral perianth, spiral androecium, and spiral gynoecium. It is forced to gain the spiral state as three separate evolutionary events. In our analysis without hidden states we include three different model structure: `model="ER"` (equal rates), `model="SYM"` (symmetric rates), and `model="ARD"` (all rates differ). The other options used (`rate.cat=1` and `nstarts=10`) specify that no hidden states are to be used and that the maximum likelihood search will be performed 10 additional times with different initial parameters. For example, the `corHMM` call specifying an all rates differ model without hidden states would be:

```
CorRes_1RC.ARD <- corHMM(phy=phy, data=data, rate.cat=1,
nstarts=10, model = "ARD")
```

We also include a set of analyses in which hidden states are present because it is likely that there are unobserved characters which influence the evolution of the angiosperm flower. We include four hidden state models: *ER/ER*, *SYM/SYM*, *ARD/ARD*, and *ER/ARD*. Each of these models allows for the possibility of rate heterogeneity through the inclusion of a hidden state, however the state-dependent processes differ. The *ER/ER* model can be thought of as a drift model where differences between $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ represent different rates of change, but ultimately state change is random. The *SYM/SYM* model specifies that within character changes are equally probable, but some characters may change faster than others. The *ARD*/ARD model specifies that there could be an optimal phenotype, but the optimal state may differ depending on whether the lineage is in $\boldsymbol{R}_1$ or $\boldsymbol{R}_2$. Finally, *ER/ARD* is a hybrid model which includes aspects of random drift and selection towards an optimal phenotype. To create these models we use tools available

in `corHMM` and specify our model through the `rate.mat` option. To create the *ER/ARD* model we first obtain a generic `rate.mat` object using the new `getStateMat4Dat` function:

```
LegendAndRateMat <- getStateMat4Dat(data)
```

The `getStateMat4Dat` function produces a list that contains `$legend`, which indicate the unique character combinations recognized by `corHMM`, and `$rate.mat`, which is a rate index matrix describing a single rate class. We can modify `$rate.mat` to customize transitions for our desired model. Any of the state-dependent processes can be based on the initial rate index matrix (from `getStateMat4Dat`) and then subsequently modified to specify the differences between the rate categories. In $\boldsymbol{R}_1$ we assume that all transition rates are equal, and in $\boldsymbol{R}_2$ we assume that all rates differ. This model describes a mixture of a drift-like process ($\boldsymbol{R}_1$) and a process in which there is an optimal phenotype ($\boldsymbol{R}_2$).

```
R1 <- getStateMat4Dat(data, model = "ER")$rate.mat
R2 <- getStateMat4Dat(data, model = "ARD")$rate.mat
```

We can group all of our rate classes together in a list. The first element of the list corresponds to $\boldsymbol{R}_1$, the second to $\boldsymbol{R}_2$, and so on:

```
StateMats <- list(RateCat1, RateCat2)
```

To obtain the parameter process matrix, we have implemented a separate function `getRateCatMat`. The only input is the number of hidden states to include in the model. By default, this function will assume that all transitions among the specified number of rate classes occur independently. In our example, we will generate a matrix that specifies how transitions between $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ occur. Note that $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ could represent a biologically-relevant, but unmeasured factor, such as temperate or tropical environments, island or mainland, presence or absence of a trait (e.g., woody vs. herbaceous life forms). For illustrative purposes, we will specify that the transition rate from $\boldsymbol{R}_1$ to $\boldsymbol{R}_2$ is the same as the rate from $\boldsymbol{R}_2$ to $\boldsymbol{R}_1$:

```
RateClassMat <- getRateCatMat(2)
RateClassMat <- equateStateMatPars(RateClassMat, c(1,2))
```

We now have all the components necessary to create the full model using the `getFullMat` function. This function requires that the first input be a list of the state-dependent processes and the second argument be the parameter process:

```
ER.ARD_rate.mat <- getFullMat(StateMats, RateClassMat)
```

Three additional features warrant brief discussion. First, it is important to note that users are not limited to the default models (*ER*, *SYM*, *ARD*). Any state-dependent matrix can be modified to create specific models that suit the user's hypothesis. For example, a model in which transitions between states is irreversible is not included within the default models but is nonetheless straightforward to create. Second, to help ensure that the user-specified rate matrix is

consistent with their verbal model, we have implemented a new function for plotting a decomposed version of the model: `plotMKModel`. The user can input either the rate matrix they intend to use for modeling, or the resulting `corHMM` object. In both cases, the plotting output will be in two parts: (1) ball-and-stick diagrams of the state-dependent processes (Fig. 1a,b) and the parameter process (Fig. 1c), and (2) a set of rate matrices that describe the model in matrix form (Fig. 1d-f). Finally, we have implemented a new function, `makeSimmap`, which is based on Bollback (2006). Although simmaps are closely related to ancestral state reconstruction, a character history not only generates a hypothesis about the ancestral states but is an effective way to understand the tempo of evolution. This is particularly important for HMMs because the rates of evolution can vary drastically across the tree (see *Performance in simulation* below). We choose not to implement any new plotting functions, instead `makeSimmap` produces a `simmap` object which is formatted such that it can be used with other R packages such as phytools (Revell, 2012). For additional capabilities, options, and biological examples we refer readers to the detailed vignette now provided as part of the R package.

3. Results

3.1 – Performance in Simulation

3.1.1 – Increasing the number of observed characters or states

Overall, the accuracy of an ancestral state reconstruction is a function of the transition rates as well as the number of states allowed in the model (Fig. 2a). For example, all datasets generally inferred the correct ancestral state at low rates. However, when viewed in terms of information, datasets that contained just two states showed detectable informational loss when compared to the three- and four-state datasets. In fact, across all scenarios — low, intermediate,

and especially at the highest rates — datasets with more states consistently showed more informational gain relative to the maximum information content for a given number of states (Fig. 2b). We suspect this largely reflects the impacts of homoplasy when the number of character states are restricted in the model (Sanderson & Donoghue, 1989; Steel & Penny, 2005). This is not to say that more character states are always necessary for accurate ASR, rather we demonstrate that there are cases when additional characters or character states enhance the accuracy of an ancestral state reconstruction and those datasets have a signal of increased information.

3.1.2 – Evaluating hidden Markov models

The accuracy of ancestral state estimation, based solely on reconstructing character states at nodes, appears largely unaffected by the inclusion of hidden states regardless of differences in the state-dependent processes (Fig. 3a). However, the amount of information gained depends on both the use of an HMM and the presence of strong differences between the state-dependent processes (Fig. 3b). These seemingly contradictory results are best explained by examining the average number of transitions based on the simmap character histories. Datasets fit under an HMM more accurately estimated the tempo of evolution compared to the standard Markov model. We also found that when the generating model does not have state-dependent differences, the HMM does not pickup significant rate variation (Fig. 4a-c) and resembles the character history implied by the standard Markov model. These findings are reassuring and suggest that HMMs are supported in datasets where rate heterogeneity is present. Although a standard Markov model performs well when there are not state-dependent differences (Fig. 3c & Fig. 4b),

it clearly underperforms compared to the HMM when there are state-dependent differences (Fig. 4d-f).

Surprisingly, we found little effect of altering the transition rate bias of the parameter process on either ancestral state reconstruction (Fig. 5a) or information content (Fig. 5b). The apparent association between ancestral state accuracy and bias is likely a consequence of asymmetric direction of evolution towards the slower rate class (as the bias increased, the time spent in a slower rate class increased). This is evident from fewer transitions occurring as the asymmetry was increased (Fig. 5c).

Unsurprisingly, observing the "second character" states increased the amount of information (Fig. 6). However, as the state-dependent processes became more distinguishable, the informational gap between an HMM and including observing the second character decreased. In other words, when the evolution of an observed character changes across the phylogeny, an HMM is able to extract additional information from a dataset.

3.2 – Case study: reconstructing the ancestral angiosperm flower

The best fitting model was one without hidden states and all rates differing (Table 1). The most likely state at the root is a spiral perianth, spiral androecium, whorled gynoecium (Fig. 7a). This result differs Sauquet et al.'s (2017) which suggested a whorled perianth, whorled androeciuim, and spiral gynoecium as the most likely state. The difference between our result and those of Sauquet et al. (2017) is likely a direct consequence of allowing for correlated evolution among several characters. However, neither result matched the expectations of Sokoloff et al. (2018), who predicted that the ancestral state of the angiosperm flower was likely to be either entirely whorled or entirely spiral. Although our best fitting model did not match

403 those expectations, if we include hidden states we found that the root state to be either whorled
404 perianth, spiral androecium, and spiral gynoecium or entirely spiral (Fig. 7b). It is likely that the
405 hidden rates model is more biologically appropriate because there are several characters which
406 influence flower evolution not included in our analyses. But, in order to find support for a hidden
407 rate model it is often necessary to have several examples of each state-dependent process within
408 the dataset. This is more likely as a dataset grows larger and the appropriateness of a hidden
409 state, as well as the power to detect rate heterogeneity, increases.
410
411 **4. Discussion**
412 Hidden Markov models are an essential tool for inferring character states across
413 phylogenies. The new version of `corHMM`, expands the array of potential uses of HMMs by
414 increasing the number of possible character states and allowing users to construct custom
415 models. In addition, we demonstrated the informational advantages of using hidden Markov
416 models versus simple Markov models. Users interested in hypothesis-driven model construction
417 are encouraged to read through the vignette associated with the `corHMM` package. This vignette
418 fully describes how to use the package and includes several examples of how to take a biological
419 hypothesis and codify it into an explicit HMM.
420 Information theory has mainly been discussed in a theoretical context and rarely used in
421 practice to understand empirical trait evolution (Mossel, 2003; Mossel & Peres, 2003; Townsend
422 & Naylor, 2007; Sober & Steel, 2011, 2014; Gascuel & Steel, 2014). In this paper we have
423 introduced a measure for the amount of information that the tips provide the nodes during
424 ancestral state reconstruction. Two important caveats of this measure of information is that the
425 data and model are taken as fixed. These are not uncommon assumptions in phylogenetic

comparative methods. For example, if one is to interpret an ancestral state reconstruction it comes with the implicit assumption that the model accurately describes the evolution of the traits (Beaulieu & O'Meara, 2019). Mutual information, as we have defined it, only provides information relative to the specified model and specified tips. A model which is more uncertain about any ancestral state, such as an equal rates model, is liable to have a more informative ancestral state reconstruction because any deviation from an uninformed ancestral state is due to the particular values of the tip states. This does not make the equal rates model better than alternatives nor do we advocate for the use of information to assist in model selection. Instead, mutual information provides insight into the interaction between the model and tip states. Higher information of particular nodes could be indicative of an area of a phylogeny where the model is poorly suited and thus the tips provide the major explanation of the ancestral state. Mutual information is also highly correlated with the rates of evolution and has the intuitive property that as rates of evolution (or time) increase the information that the tips provide to the nodes decreases (Sober & Steel, 2011).

It is important to have a biologically realistic model of trait evolution when conducting an ancestral state reconstruction. With the generalizations made to `corHMM` we have provided two distinct ways to increase the realism of phylogenetic comparative modeling. First, we have allowed for the correlated evolution of several characters and states. Whether traits are correlated because of underlying pleiotropy leading to genetic correlation (Conner et al., 2011) or similar ecological contexts leading to convergent morphological evolution (Mahler, Revell, Glor, & Losos, 2010), at the macroevolutionary scale they are better understood in a holistic context rather than independently evolving subunits. Second, the inclusion of hidden states allows for more detailed descriptions of the evolutionary process. State-dependent processes can differ in

449 both rate and structure and thus provide a description of heterogeneity in the tempo and mode of
450 evolution.

451 We have demonstrated the informational and accuracy advantages of including additional
452 states and characters in a simulation setting (*3.1.1 Increasing the number of observed characters*
453 *or states)*. However, it remained to be seen whether additional characters would impact the
454 ancestral state in an empirical example and whether those results match biological expectations.
455 The controversy surrounding the phyllotaxy of the ancestral angiosperm flower is a particularly
456 appropriate case study for the generalized version of `corHMM`, as it not only allows for the
457 dependent evolution of several discrete characters but also includes hidden states as a fitting
458 addition to help describe the heterogeneous evolution of angiosperms. The effects of including
459 correlated character evolution or hidden states are evident from our ancestral state reconstruction
460 of the angiosperm flower. The results of our best model were exactly opposite the original
461 study's finding (Sauquet et al., 2017). Additionally, although models which included hidden
462 states were not best supported, the ancestral state reconstruction produced by these models
463 matched the expectations put forward based on developmental data (Sokoloff et al., 2018). This
464 demonstrates that reconstructions of ancestral states change when correlated trait evolution are
465 incorporated and might be more biologically realistic when hidden states are included.

466

467 **5. Conclusion**

468 Although there is a growing consensus that phylogenies and their associated methods are
469 being used in ways that exceed what they can infer (Losos 2011; Maddison and FitzJohn 2015;
470 Rabosky and Goldberg 2015; Cooper et al. 2016), we have shown that there is still under-utilized
471 information in phylogenetic comparative datasets. First, HMMs extract signals of rate

heterogeneity when it is present and, equally important, do not falsely locate signals where they are absent. Second, increased trait depth adds new information and consistently improves ancestral state reconstruction estimates. Indeed, as datasets continue to grow, so will the analytical power that biologists have for testing complex models of evolution. Finally, the inclusion of correlated trait evolution and hidden states is relevant beyond theoretical considerations and we have shown that these generalizations can completely change the results of an ancestral state reconstruction in empirical datasets. Although hidden Markov models are not a perfect substitute for real observation of a hidden character, they make for a tractable and a biologically reasonable description of heterogeneity in the evolutionary process over long time scales.

Availability

This open source software is written entirely in the R language and is freely available through the Comprehensive R Archive Network (CRAN) at http://cran.r-project.org/.

Acknowledgements

We thank members of the Beaulieu lab and colleagues at the University of Arkansas for their comments and for general discussions of the ideas presented here. We would also like to specifically thank Andrew Alverson, Brian O'Meara, and Adam Siepielski for their insightful critiques and helpful edits on an earlier version of this manuscript.

Conflict of Interest

None declared.

Author Contributions

J.D.B. and J.M.B. designed research; J.D.B performed research and analyzed data; and J.D.B. and J.M.B. wrote the paper.

References

Beaulieu, J. M., & Donoghue, M. J. (2013). Fruit Evolution and Diversification in Campanulid Angiosperms. *Evolution*, *67*(11), 3132–3144. doi:10.1111/evo.12180

Beaulieu, J. M., & O'Meara, B. C. (2016). Detecting Hidden Diversification Shifts in Models of Trait-Dependent Speciation and Extinction. *Systematic Biology*, *65*(4), 583–601. doi:10.1093/sysbio/syw022

Beaulieu, J. M., & O'Meara, B. C. (2019). Diversity and skepticism are vital for comparative biology: a response to Donoghue and Edwards (2019). *American Journal of Botany*, *106*(5), 613–617. doi:10.1002/ajb2.1278

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. *Systematic Biology*, *62*(5), 725–737. doi:10.1093/sysbio/syt034

Bollback, J. P. (2006). SIMMAP: stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics*, *7*(1), 88.

518 Conner, J. K., Karoly, K., Stewart, C., Koelling, V. A., Sahli, H. F., & Shaw, F. H. (2011). Rapid
519 Independent Trait Evolution despite a Strong Pleiotropic Genetic Correlation. *The*
520 *American Naturalist*, *178*(4), 429–441. doi:10.1086/661907
521 Cooper, N., Thomas, G. H., & FitzJohn, R. G. (2016). Shedding light on the 'dark side' of
522 phylogenetic comparative methods. *Methods in Ecology and Evolution*, *7*(6), 693–699.
523 doi:10.1111/2041-210X.12533
524 Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley.
525 Eddy, S. R. (2004). What is a hidden Markov model? *Nature Biotechnology*, *22*(10), 1315–1316.
526 doi:10.1038/nbt1004-1315
527 Felsenstein, J., & Churchill, G. A. (1996). A Hidden Markov Model approach to variation among
528 sites in rate of evolution. *Molecular Biology and Evolution*, *13*(1), 93–104.
529 doi:10.1093/oxfordjournals.molbev.a025575
530 FitzJohn, R. G., Maddison, W. P., & Otto, S. P. (2009). Estimating Trait-Dependent Speciation
531 and Extinction Rates from Incompletely Resolved Phylogenies. *Systematic Biology*,
532 *58*(6), 595–611. doi:10.1093/sysbio/syp067
533 Galtier, N. (2001). Maximum-Likelihood Phylogenetic Analysis Under a Covarion-like Model.
534 *Molecular Biology and Evolution*, *18*(5), 866–873.
535 doi:10.1093/oxfordjournals.molbev.a003868
536 Gascuel, O., & Steel, M. (2014). Predicting the Ancestral Character Changes in a Tree is
537 Typically Easier than Predicting the Root State. *Systematic Biology*, *63*(3), 421–435.
538 doi:10.1093/sysbio/syu010
539 Losos, J. B. (2011). Seeing the Forest for the Trees: The Limitations of Phylogenies in
540 Comparative Biology: (American Society of Naturalists Address). *The American*
541 *Naturalist*, *177*(6), 709–727. doi:10.1086/660020
542 Maddison, W. P., & FitzJohn, R. G. (2015). The Unsolved Challenge to Phylogenetic
543 Correlation Tests for Categorical Characters. *Systematic Biology*, *64*(1), 127–136.
544 doi:10.1093/sysbio/syu070

Maddison, W. P., Midford, P. E., Otto, S. P., & Oakley, T. (2007). Estimating a Binary Character's Effect on Speciation and Extinction. *Systematic Biology*, *56*(5), 701–710. doi:10.1080/10635150701607033

Mahler, D. L., Revell, L. J., Glor, R. E., & Losos, J. B. (2010). Ecological opportunity and the rate of morphological evolution in the diversification of Greater Antillean anoles. *Evolution: International Journal of Organic Evolution*, *64*(9), 2731–2745.

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*(12), 3918–3930. doi:10.1111/j.1558-5646.2012.01720.x

Mossel, E. (2003). On the Impossibility of Reconstructing Ancestral Data and Phylogenies. *Journal of Computational Biology*, *10*(5), 669–676. doi:10.1089/106652703322539015

Mossel, E., & Peres, Y. (2003). Information flow on trees. *The Annals of Applied Probability*, *13*(3), 817–844. doi:10.1214/aoap/1060202828

O'Meara, B. C. (2012). Evolutionary Inferences from Phylogenies: A Review of Methods. *Annual Review of Ecology, Evolution, and Systematics*, *43*(1), 267–285. doi:10.1146/annurev-ecolsys-110411-160331

Pagel, M. (1994). Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond. B*, *255*(1342), 37–45. doi:10.1098/rspb.1994.0006

Pagel, M. (1999). The Maximum Likelihood Approach to Reconstructing Ancestral Character States of Discrete Characters on Phylogenies. *Systematic Biology*, *48*(3), 612–622.

Penny, D., McComish, B. J., Charleston, M. A., & Hendy, M. D. (2001). Mathematical Elegance with Biochemical Realism: The Covarion Model of Molecular Evolution. *Journal of Molecular Evolution*, *53*(6), 711–723. doi:10.1007/s002390010258

Rabosky, D. L., & Goldberg, E. E. (2015). Model Inadequacy and Mistaken Inferences of Trait-Dependent Speciation. *Systematic Biology*, *64*(2), 340–355. doi:10.1093/sysbio/syu131

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, *3*(2), 217–223. doi:10.1111/j.2041-210X.2011.00169.x

Román-Palacios, C., Scholl, J. P., & Wiens, J. J. (2019). Evolution of diet across the animal tree of life. *Evolution Letters*, *3*(4), 339–347. doi:10.1002/evl3.127

Sanderson, M. J., & Donoghue, M. J. (1989). Patterns of Variation in Levels of Homoplasy. *Evolution*, *43*(8), 1781–1795. doi:10.1111/j.1558-5646.1989.tb02626.x

Sauquet, H., von Balthazar, M., Doyle, J. A., Endress, P. K., Magallón, S., Staedler, Y., & Schönenberger, J. (2018). Challenges and questions in reconstructing the ancestral flower of angiosperms: A reply to Sokoloff et al. *American Journal of Botany*, *105*(2), 127–135. doi:10.1002/ajb2.1023

Sauquet, H., von Balthazar, M., Magallón, S., Doyle, J. A., Endress, P. K., Bailes, E. J., … Schönenberger, J. (2017). The ancestral flower of angiosperms and its early diversification. *Nature Communications*, *8*(1), 16047. doi:10.1038/ncomms16047

Schultz, T. R., Cocroft, R. B., & Churchill, G. A. (1996). The Reconstruction of Ancestral Character States. *Evolution*, *50*(2), 504–511. doi:10.1111/j.1558-5646.1996.tb03863.x

Siepel, A., & Haussler, D. (2005). Phylogenetic Hidden Markov Models. In *Statistical Methods in Molecular Evolution* (pp. 325–351). New York: Springer-Verlag. doi:10.1007/0-387-27733-1_12

Sober, E., & Steel, M. (2011). Entropy increase and information loss in Markov models of evolution. *Biology & Philosophy*, *26*(2), 223–250. doi:10.1007/s10539-010-9239-x

Sober, E., & Steel, M. (2014). Time and Knowability in Evolutionary Processes. *Philosophy of Science*, *81*(4), 558–579. doi:10.1086/677954

Sokoloff, D. D., Remizowa, M. V., Bateman, R. M., & Rudall, P. J. (2018). Was the ancestral angiosperm flower whorled throughout? *American Journal of Botany*, *105*(1), 5–15. doi:10.1002/ajb2.1003

Steel, M., & Penny, D. (2005). *Maximum parsimony and the phylogenetic information in multistate characters*. New York: Oxford University Press.

Tarasov, S. (2019). Integration of Anatomy Ontologies and Evo-Devo Using Structured Markov Models Suggests a New Framework for Modeling Discrete Phenotypic Traits. *Systematic Biology*. doi:10.1093/sysbio/syz005

Townsend, J. P., & Naylor, G. (2007). Profiling Phylogenetic Informativeness. *Systematic Biology*, *56*(2), 222–231. doi:10.1080/10635150701311362

Yang Lou, X. (2017). Hidden Markov Model Approaches for Biological Studies. *Biometrics & Biostatistics International Journal*, *5*(4). doi:10.15406/bbij.2017.05.00139

Yang, Z., Kumar, S., & Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, *141*(4), 1641–1650.

Zucchini, W., MacDonald, I. L., & Langrock, R. (2017). *Hidden Markov models for time series: an introduction using R*. Chapman and Hall/CRC.

Table 1 – Model rankings from the maximum-likelihood analysis of the ancestral angiosperm flower. Models separated by a "/" indicate a hidden rates model and the split distinguishes between the two state-dependent process (for example, *ER/ARD*, represents a hidden rate model where R1 is an equal rates model and R2 is an all rates differ model).

Model	AICc	AICc Weight	Mean Rate (events/Myr)	Mutual Information (bits)	Proportion of Max Information (%)
ER	247.18	0	0	2.54	98.26
SYM	167.45	0	28.57	2.53	97.87
ARD	149.17	0.93	7.07	0.52	82.02
ER/ER	164.63	0	0.05	3.49	97.35
SYM/SYM	173.85	0	10.02	3.38	94.28
ARD/ARD	182.47	0	5	1.73	81.47
ER/ARD	154.45	0.07	2.51	2.27	89.5

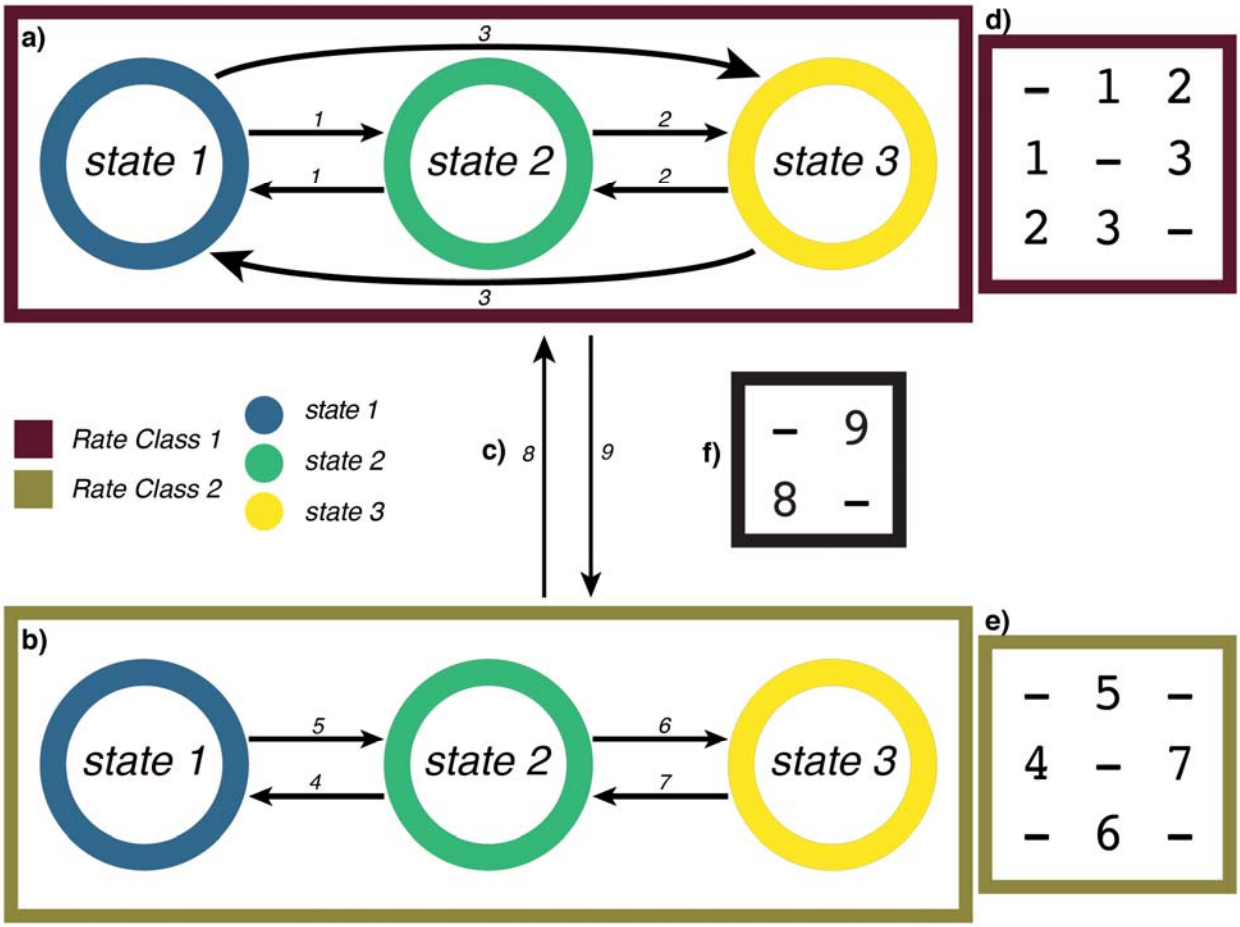


Figure 1. A decomposed HMM containing 3 observed states and 2 hidden rate classes. R1 is one state-dependent process that describes transitions to and from observed states as being equal (a, d), whereas R2 is a state-dependent process that describes state 2 as a necessary intermediate (b, e). The parameter process that relates R1 and R2 and describes the transitions between R1 and R2 (c, f).

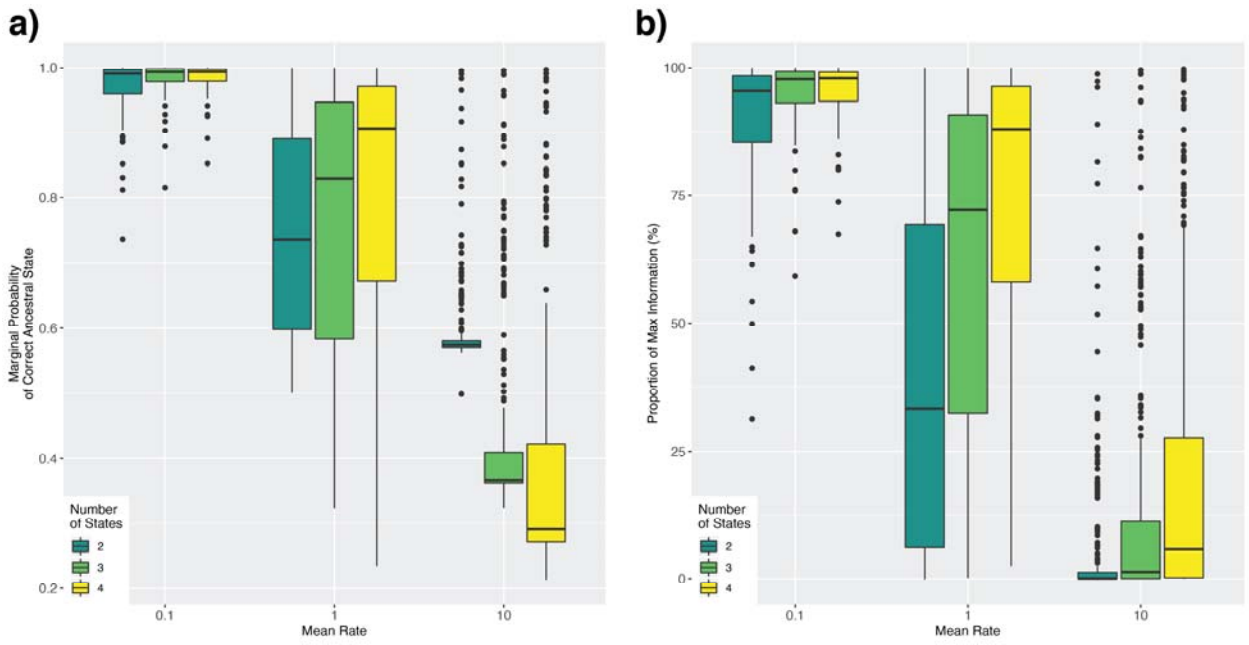


Figure 2. Performance of standard and hidden Markov models depending on the number of states in the dataset and mean rate. Each dataset was simulated under a mean rate of 0.1, 1, or 10 transitions Myr^{-1}, with 2, 3, or 4 observed states and no hidden states. a) The marginal probability of estimating the correct ancestral state. b) The proportion of information gained about ancestral states from each dataset and model.

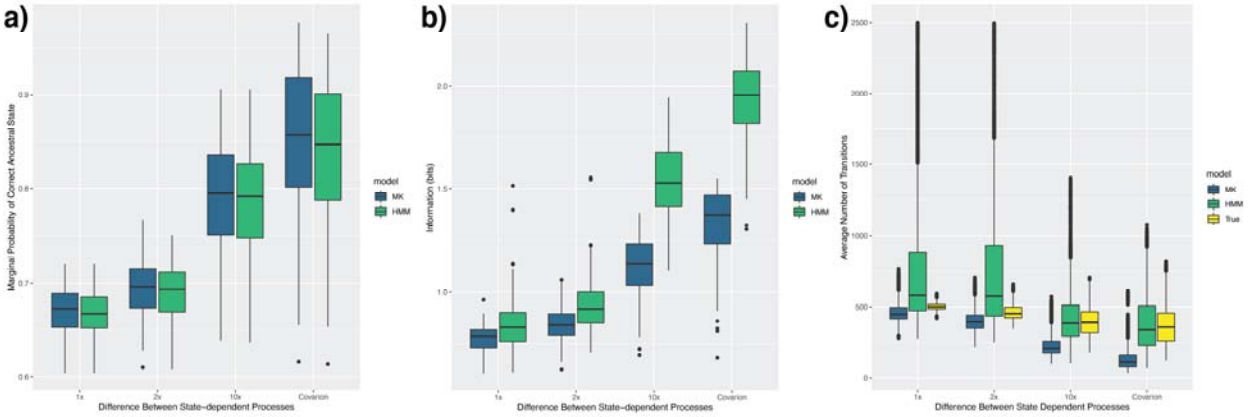


Figure 3. Comparison of fits from a Markov and HMM model when an HMM is the generating model. We vary the difference between state-dependent processes from no difference (1x) to complete asymmetry where state transitions occur in one state-dependent process only (i.e., "covarion" model; see *Performance in simulation*). a) The marginal probability of the correct ancestral state. b) The average amount of information (bits) for ancestral states from each dataset and model. c) The number of transitions averaged over 150 simmaps.

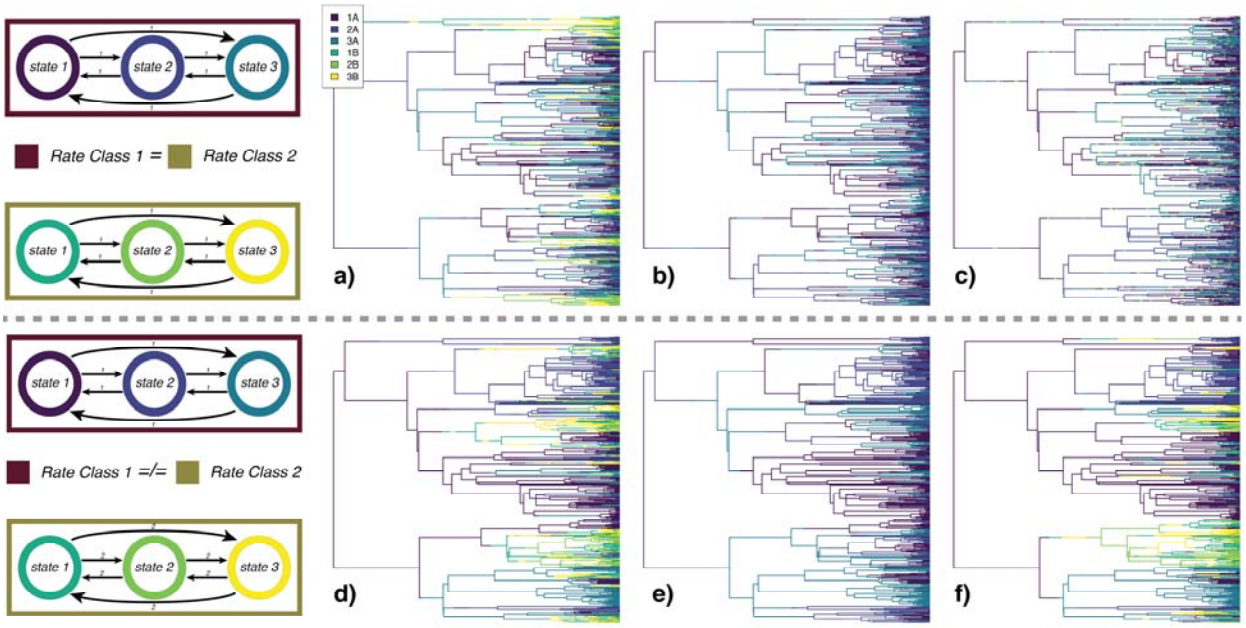


Figure 4. Stochastic maps demonstrating the effect of differences in the magnitude of two state-dependent processes. The first row shows data is simulated where there were no differences in the state-dependent processes, where (a) is the true generating model, (b) is one example of the character history simulated under the MLE from standard Markov model, and (c) is one example of the character history simulated under the MLE of an HMM. The second row is the same, but with data simulated with a 10-fold difference between the state-dependent processes. A Markov model does not contain a distinction between the hidden classes A and B, thus it is displayed only in terms of the states 1A, 2A, and 3A. Comparing the HMM in (c) and (f) demonstrates that an HMM will only detect a hidden state when it influences the observed, state-dependent, process.

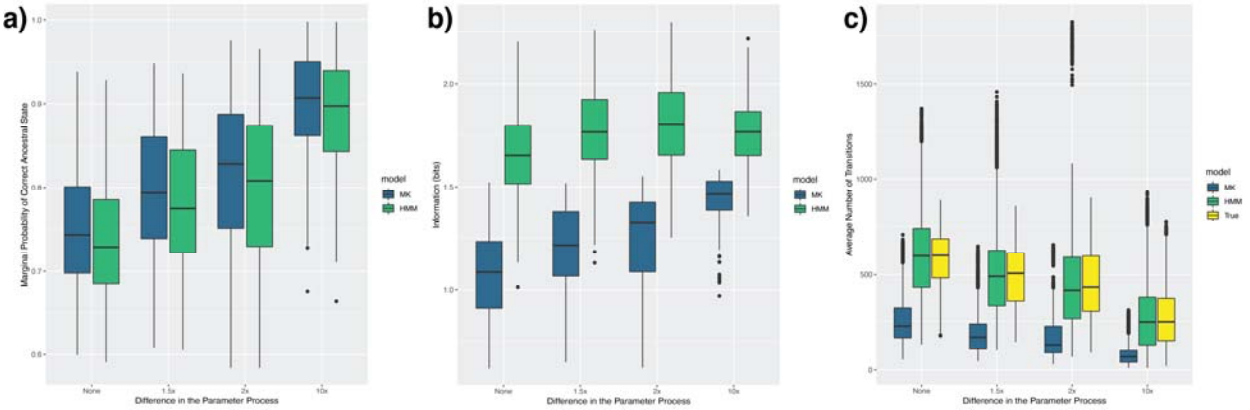
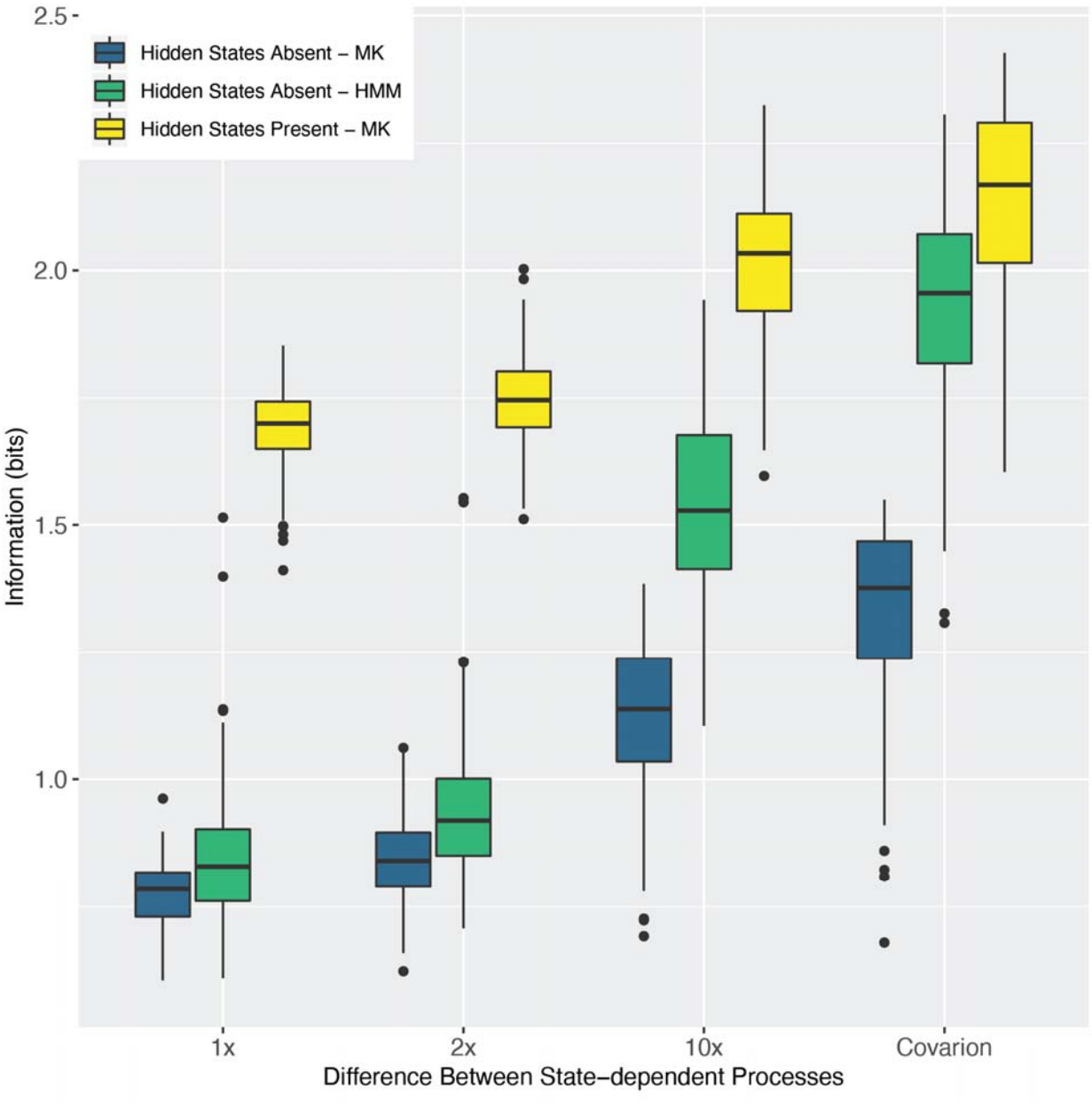


Figure 5. Comparison of a standard Markov and HMM model when an HMM is the generating model. We vary the bias from rate class 1 to rate class 2 from no difference (none) to a 10-fold difference (10x). a) The marginal probability of the correct ancestral state. b) The average amount of information the tips provide the nodes. c) The number of transitions averaged over 150 simmaps.



684

685 **Figure 6**. Average information when a hidden state is either directly observable or unobserved.
686 If the hidden state is unobserved, we compare the information gained when fitting a Markov
687 model (MK) or a hidden Markov model (HMM). When the hidden state is directly observable we
688 fit a standard MK. When the hidden state is directly observed, the datasets are comprised of 6
689 discrete states.

690

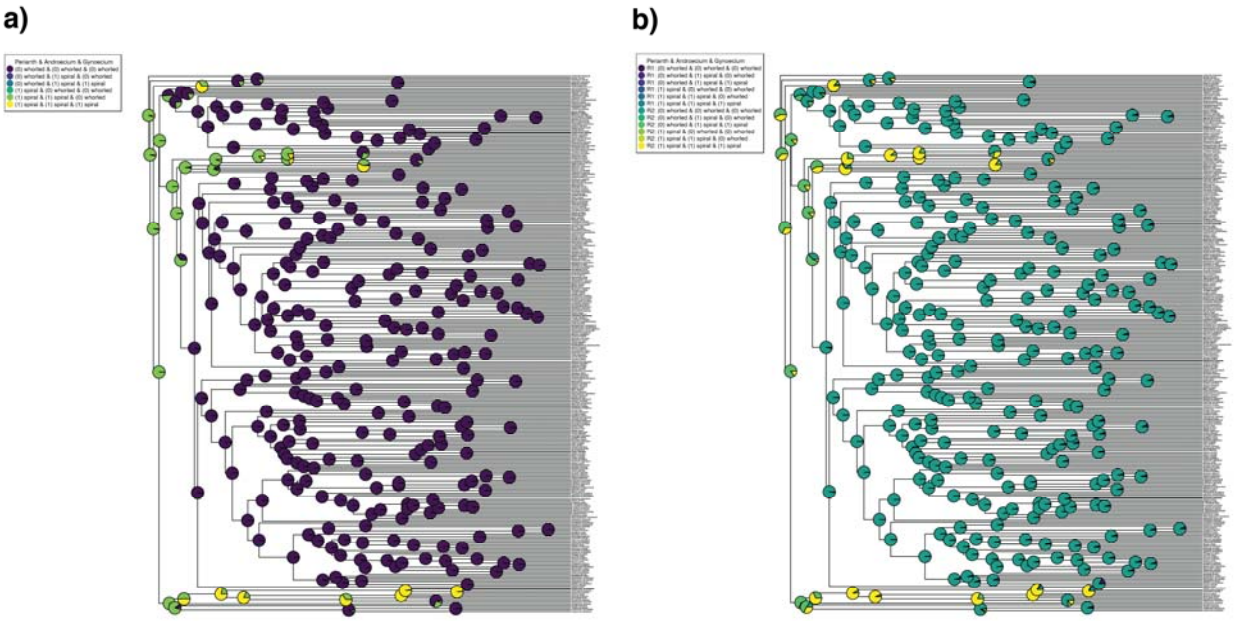


Figure 7. Ancestral state reconstructions based on two models. a) an all rates differ model without any hidden states (relative probability based AICc, w = 0.93). The most likely state at the root is a spiral perianth, spiral androecium, whorled gynoecium (marginal probability = 94%). b) A hidden rate model where the state-dependent processes are an equal rates model and an all rates differ model (w = 0.07). The ancestral reconstructions which received the majority of the support were whorled perianth, spiral androecium, and spiral gynoecium (marginal probability = 57%) or spiral perianth, spiral androecium, and spiral gynoecium (marginal probability = 43%).