Automatic Discovery of Optimal Discrete Character Models

James D. Boyko[1,2]

[1] Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan 48109, USA

[2] Michigan Institute of Data Science, University of Michigan, Ann Arbor, Michigan 48109, USA

Contact. *jboyko@umich.edu*

Running Head. AUTOMATIC DISCRETE CHARACTER MODEL DISCOVERY

Key Words. Machine learning; Discrete character evolution; Regularization; Ancestral state reconstruction; Phylogenetic comparative methods; hidden Markov model; Bias-variance trade-off

Abstract

Modeling discrete character evolution in a Markovian framework has become the standard in phylogenetic comparative methods. The increasing size and complexity of these models reflects a trend of analyses to include more taxa and more discrete characters. However, as complexity of the models increase, so do the number of potential model structures and number of estimable parameters, making it nearly impossible to consider all modeling options for a given dataset. To overcome this issue, I apply a combination of regularization and parameter sharing optimization to models of discrete character evolution. This allows for the automatic searching and optimization across different model structures without user specification. I test this framework under several simulation scenarios including hidden rates and multiple discrete characters. The results indicate that regularized models significantly outperform traditional approaches, yielding a far lower variance and a nearly tenfold reduction in the overall error of parameter estimates in the most extreme scenarios. I illustrate the power of automatic model selection by revisiting the ancestral state estimation of concealed ovulation and mating systems in Old-World monkeys. Using the dredge algorithm, I discover a previously unexamined model structure which has both better statistical performance and a differing ancestral state reconstruction when compared to default model sets. In general, these results highlight the dangers of an over-reliance on default model sets. The combination of automatic model selection and regularization help overcome problems of over-parameterization, and these results demonstrate that when inferences are drawn from a larger model space, they are both more statistically robust and biologically realistic.

Complex discrete character models are now widely applied in phylogenetic comparative methods (PCMs). These models, which were initially limited to relatively few characters and simple processes, have been expanded in an attempt to incorporate various biological processes and sources of variation. Complexity has been introduced to these models through correlated character evolution (Pagel 1994), hidden rate variation (Beaulieu et al. 2013; Boyko and Beaulieu 2021), state-dependent speciation and extinction (SSE) (Maddison et al. 2007; Beaulieu and O'Meara 2016), and the incorporation of continuous character information (May and Moore 2020; Boyko et al. 2023).

This increasing complexity has also led to higher generality. For example, hidden Markov models which were initially introduced to model two binary characters (Beaulieu et al. 2013), have been expanded to allow for any number of characters, observed states, and hidden states (Boyko and Beaulieu 2021). However, while this additional flexibility allows for customization of the phylogenetic comparative method to the system at hand, the large model space can make it challenging for biologists to find the most appropriate model for their particular dataset. The typical approach, multi-model inference framework (Burnham and Anderson 2002), has a biologist decide on a set of potentially realistic models. The relative support for each model is then evaluated based on their fit to the data using information criteria such as Akaike Information Criterion (AIC). This approach is powerful because it allows for model averaging, where inferences are made based on a weighted average of the models' included in the set (Burnham and Anderson 2002). However, the effectiveness of multi-model inference relies heavily on the appropriateness of the model set. Several PCMs have been criticized for high false positive rates due to the exclusion of the "correct" null hypothesis (Rabosky and Goldberg 2015). Although these criticisms have been addressed by introducing new model structures to serve as better null hypotheses (Beaulieu and O'Meara 2016; Boyko and Beaulieu 2022), the problem, when recast as a failure to include a complete model set (Boyko and Beaulieu 2022), suggests that there is still a vast space of unexplored model structures.

Often the model set chosen for discrete character evolution are nested, with the difference being which parameters are variable, and which are fixed. For example, for a two-character binary state dataset, the difference between a correlated model (k=8) and the independent model (k=4) is whether changes in the focal character depend on the state of the background character (Pagel and Meade 2006). In the independent model parameters representing this dependency process are fixed to be equal, while in the correlated model they are freely estimated (Pagel and Meade 2006; Boyko and Beaulieu 2022). In this way, one can think of discrete models adding complexity by adding parameters to increase "biological realism" and represent new processes or relationships between variables.

This increasing model complexity then leads to another, more technical, challenge. As the number of estimable parameters grows it reaches a point where the number of parameters (k) approaches the number of taxa (N). However, the rate at which the number of parameters increase often surpasses the rate at which information can be gained through data (Felsenstein 2012). In the case of correlated discrete character models, this is because each additional character requires considering its relationship to all other traits. For instance, the most complex discrete model for a single binary character has 2 parameters, while the most complex model with two binary characters has 12 parameters (k=8 if excluding dual transitions), and the most complex model with three binary characters has 56 parameters (k=24 if excluding dual transitions; Fig. 1). In each instance we have added a single character, but because we must consider the new character's relationship to all other existing characters, the model complexity (as measured by the number of parameters) outpaces the potential information gained from the new data. This is problematic because although likelihood-based methods are consistent estimators when N >> k, their performance deteriorates as models become more parameter-rich, potentially leading to unreliable and biased parameter estimates (Huelsenbeck et al. 2001). However, for complex models with a finite data, it is unlikely that all parameters will be essential or necessary to best model the data (Lemey et al. 2009; Gelman et al. 2013).

To address the challenges associated with increasing discrete model complexity, I introduce regularized discrete character models and a method to optimize parameter sharing structures. Regularization techniques, such as L1 (lasso) and L2 (ridge), constrain the magnitude of parameter estimates and encourage simpler, more generalizable models (Hoerl and Kennard 1970; Tibshirani 1996). By incorporating regularization, we can balance model complexity and goodness-of-fit, reducing the risk of overfitting and improving the stability of parameter estimates. Furthermore, the automatic evaluation of different parameter sharing structures can help mitigate model complexity and improve interpretability. I implement this framework within corHMM (Beaulieu et al. 2022) as corHMMDredge. To test the dredge framework, I conduct an extensive set of simulations to explore the bias-variance trade-off associated with regularization. Under regularization, it is expected that models will have increased generality and lower variance. I test these expectations by examining regularized model's accuracy of parameter estimates and predictions for ancestral states. Additionally, I perform a more detailed simulation test on a subset of historically important discrete models, such as the hidden rates model (Beaulieu et al. 2013) and correlated character evolution model(Pagel 1999; Pagel and Meade 2006; Boyko and Beaulieu 2021), to ensure that the dredge framework has acceptable false positive and negative rates. Finally, as a case study, I reanalyze the dataset from Pagel and Meade

(2006) testing whether female Old World monkey estrus advertisement is associated with multimale mating systems. I use this case study to demonstrate the dredge framework, guiding users through each step of the process including cross-validation and uncertainty estimation. I also compare profile likelihoods for regularized and non-regularized methods, as likelihood ridges were found for the best fitting unregularized model.

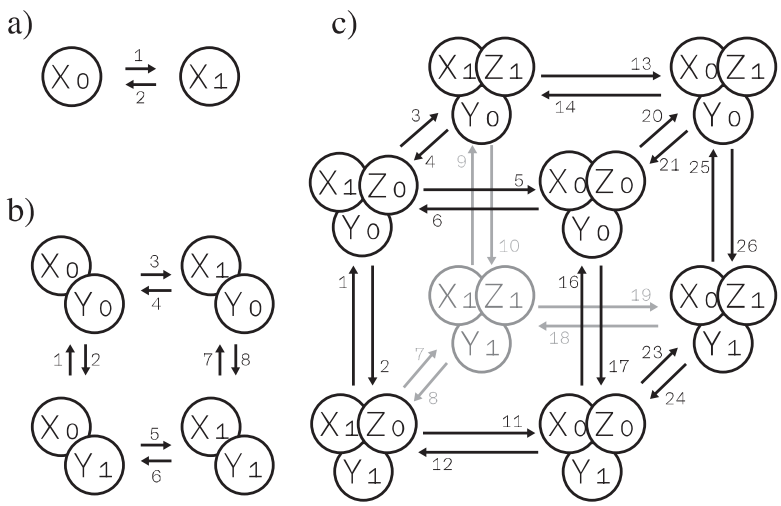

Figure 1– Visualization of the exponential increase in model complexity as additional binary characters are introduced. (a) A single binary character with two states (k=2). (b) Two binary characters, each with two states (k=8, excluding dual transitions). (c) Three binary characters, each with two states (k=26, excluding dual transitions). Adding characters exponentially increases the number of parameters due to the need to account for relationships between all characters. The disallowance of dual simultaneous transitions keeps the models from being even more complex.

Methods

Regularization and Parameter sharing

The likelihood of a discrete character model with its underlying framework as a continuous time Markov chain (CTMC) is calculated as $L = P(D|Q, \Phi, \pi)$, which is the probability of observing the data (D) given an instantaneous rate matrix (Q), root state frequency vector (π), and a phylogeny with a fixed topology and set of branch lengths (Φ). The data (D) consist of the observed character states (S) at the tips of the phylogeny, while the rate matrix (Q) contains the rates of character state transitions (q_{ij}). The likelihood function is then computed by integrating the product of transition probabilities along the branches of the phylogeny (more detailed descriptions can be found in (Pagel 1994; Lewis 2001; Felsenstein 2004). Though not mathematically necessary, it is also useful to consider a mapping matrix (M) which gives the structure of the discrete model by indicating which transition rates are

estimated and/or fixed to be equal. For example, if we consider a simple binary character with states 1 and 2, the instantaneous rate matrix is given by$Q = \begin{bmatrix} -q_{12} & q_{12} \\ q_{21} & -q_{21} \end{bmatrix}$. However, the mapping matrix can specify several alternative model structures. If $M_{er} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, then$q_{12} = q_{21}$ and we have what is commonly known as the "equal rates" model. An "all rates different" model can be indicated by$M_{ard} = \begin{bmatrix} 0 & 2 \\ 1 & 0 \end{bmatrix}$, where$q_{12} \neq q_{21}$and a "unidirectional" model can be constructed via $M_{dir} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, where $q_{21} = 0$. In this specific example, we have models of 1 ("equal rates" and "unidirectional") or 2 ("all rates different") estimable parameters. To determine which model structure is optimal (best balances goodness-of-fit and complexity) for the dataset, we can compute the maximum likelihood estimate (MLE) for each model structure and compare them using a information criterion such as AIC (Akaike 1974; Burnham and Anderson 2002).

This framework has been successfully applied for many years within PCMs (O'Meara 2012), but growing model complexity (e.g. Fig. 1) is making it untenable for users to define all possible relevant model structures or for method developers to construct a complete set of default models. As such, regularization techniques may be a necessary step in discovering optimal model structures for large and complex datasets. In corHMMDredge, I incorporate three regularization approaches, two of which are analogous to *l1* and *l2* regularization (Hastie et al. 2015), and the third is based on the null expectation that transition rates are all equal (*er*; Zhou et al. 2024). Specifically, the regularized likelihoods are defined as:

$$L_{l1}(D|Q,\Phi,\pi;\lambda) = L(D|Q,\Phi,\pi) - \lambda\frac{1}{n}\sum_{i=1}^{n} Q_{ii},$$

$$L_{l2}(D|Q,\Phi,\pi;\lambda) = L(D|Q,\Phi,\pi) - \lambda\frac{1}{n}\sum_{i=1}^{n} {Q_{ii}}^{2},$$

$$L_{er}(D|Q,\Phi,\pi;\lambda) = L(D|Q,\Phi,\pi) - \lambda\frac{1}{k}\sum_{i\neq j} \theta_i - \theta_j,$$

where $L(D|Q,\Phi,\pi)$denotes the standard likelihood, $\frac{1}{n}\sum_{i=1}^{n} Q_{ii}$ is the elementwise mean of the diagonal of the instantaneous rate matrix, $\frac{1}{n}\sum_{i=1}^{n} {Q_{ii}}^{2}$ is the squared elementwise mean of the diagonal of the instantaneous rate matrix, $\frac{1}{k}\sum_{i\neq j} \theta_i - \theta_j$ is the mean pairwise distances between all freely estimated parameters in their raw units, and λ is a hyper-parameter that adjusts the severity of the penalty, ranging

from 0 (no regularization) to 1 (full penalization). This penalization scheme uses the mean rather than the sum of the parameter values because the number of transition rates is a function of the number of possible discrete states and using the sum would cause the penalization term to be a function of the number of states rather than the overall complexity of the model (the number of parameters). Despite this adjustment, this regularization scheme is still particularly effective at removing unnecessary parameters such that the dredge algorithm can efficiently move from more complex to simpler models (Fig. 2). It is important to note that λ cannot be jointly estimated alongside the parameters via the maximized likelihood (Clavel et al. 2019). Instead, a form of cross-validation is necessary to tune λ (see *Phylogenetically informed k-fold cross validation*). This scheme is similar to Sanderson's penalized likelihood approach for estimating chronograms from phylogenetic trees (Sanderson 2002). Under penalized likelihood, the penalty is based on differences of rates and is essentially a regularization approach. Furthermore, the Sanderson method includes cross validation to estimate the equivalent of the λ hyper-parameter (which he also labeled λ, the smoothing parameter).

The second component of the dredge algorithm involves a method to explore and assess alternative model structures that utilize parameter sharing (Fig. 2). To implement parameter sharing, I adopt a strategy where the most complex possible model for C characters and S states is first fit to the data by maximizing L_{reg}. Next, any parameters which were estimated to be zero are omitted from the model and the most similar pair of parameters within regularized model are equated. If equating these parameters, thus making the model simpler, results in a AIC less than 2 units (or another specified threshold) worse than the current best model, the next most similar pair of parameters are equated, and the process continues. Note that this is a greedy approach, and an alternative approach could stochastically sample which rates to equate based on their distance or equate all pairs of most similar rates. Importantly, this algorithm also readily applies to hidden Markov models (Boyko and Beaulieu 2021). To evaluate whether additional rate variation is necessary through the presence of a hidden state, corHMMDredge will first find the optimal model for a single rate category model. Upon increasing the number of rate categories, the most complex possible two rate class HMM is evaluated by maximizing L_{reg}. At least one round of parameter sharing is tested before the HMM is compared to the best model within that particular rate class. To ensure the identifiability of HMMs, each rate class must have a uniquely valued parameter, and the transition rates between hidden states must be distinct from those between observed characters. Consequently, the minimum number of parameters for a two-rate-class HMM is three. If this results in a ΔAIC of less than 2 units from the previous best model, the search will continue and may increase rate categories until the AIC no longer improves.

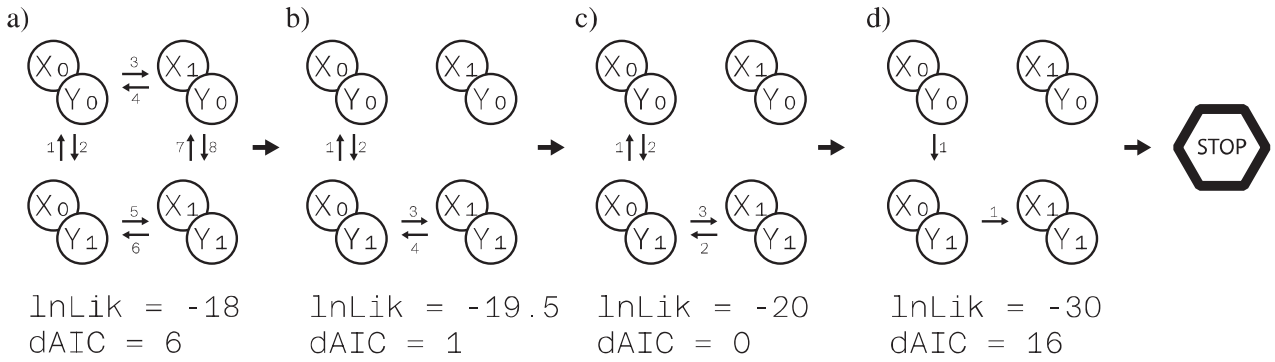

Figure 2 – A hypothetical example of the automatic model selection by dropping parameters via regularization and parameter sharing. a) The most complex model is evaluated. b) It is found that parameters 3,4,7,8 are all below a specified drop threshold and removed from the model. This results in a minor decrease in likelihood, but an improved AIC. c) Parameters from (b) are equated based on their similarity. Again, this comes at the cost of a decreased likelihood, but an improved AIC. (c) Based on model (c), parameters below the threshold are removed, but this time the AIC is not improved so the dredge algorithm halts. Note, that had no parameters of (c) been below the specified drop threshold, the dredge algorithm would have equated a pair of parameters and continued evaluating models.

When defining the initial model, it becomes important to distinguish between characters and states. Mathematically, it is unnecessary to distinguish between the two since the model is entirely defined by the S states since $S = \prod_{m=1}^{C} s_m$ where s_m is the number of states for each character m. Nonetheless, biologically it is often useful to think of characters as independently evolving units that can take on some number of states. This would mean that any character is comprised of a set of mutually exclusive states (e.g., the character "flower color" cannot have states both "red" and "blue"). Furthermore, this distinction is particularly relevant when specifying M to evaluate dependent relationships and avoid "dual transitions" (Pagel 1994). A dependent relationship is described by differences in the transition rates of a character in the presence or absence of the state of a different background character (e.g., if rates of transition in character 1, flower color, depend on the presence or absence of character 2, floral scent). Obviously, to evaluate dependent relationships, one must have a concept of separate characters and their mutually exclusive states. Regarding dual transitions, they are transitions in which the states of more than one character change in an instant in time (Pagel and Meade 2006). These are removed from the mapping matrix because it is assumed that two characters cannot change simultaneously in an instant of time. As such, the most complex model is not defined by $k = S \times (S - 1)$ parameters (which would be the case if all possible transitions were allowed). Rather, it is defined by $k = (\sum_{m=1}^{C}(s_m - 1)) \times S$ for a single rate class.

Phylogenetically informed k-fold cross validation

To optimize λ I choose to employ a phylogenetically informed k-fold cross-validation procedure across the tips of the phylogeny. K-fold cross-validation is a common way to optimize λ when using regularization (Hastie et al. 2009). It involves dividing the dataset into k separate subsets (or folds). The model is trained on k-1 of these folds with the remaining fold used as the test set to evaluate the models' performance. This process is repeated k times, with each of the k folds used exactly once as the test set. The results of these k tests are then averaged to produce a single estimate. In the context of a discrete character model the metric chosen for evaluation is reconstruction of tip state values. Each species within a k-fold is coded as having an unknown state and a reconstruction based on the fitted model is then used for evaluation. The score for a fold is determined by the average Jensen-Shannon divergence between the predicted and actual tip states. Specifically, $score_k = \frac{1}{N_k}\sum_{i=1}^{N_k} D\left(L_i, \hat{L_i}\right)$ where N_k is the number of tips in the k-th fold, L_i is the observed likelihoods at the tips of the k-th fold, $\hat{L_i}$ are the predicted states for the tips in the k-th fold and D is the Jensen-Shannon (JS) divergence (or any form of divergence more generally). The k scores are then averaged over all folds to obtain a single score for the model. The JS divergence is chosen because it a symmetric and bounded measure of difference between predicted and observed probability distributions while being robust to zero probabilities. An advantage of using tip values to measure the model fit is that the phylogenetic structure of the data is unchanged throughout the entire cross-validation procedure because no tips need to be explicitly dropped when fitting the model; they only need to be set to an unknown value. The samples within a given k-fold are chosen without replacement and with probabilities equal to $w = (I^T C^{-1})\vec{1}$ (Rohlf 2001) to ensure phylogenetically even sampling between the k-folds.

Simulation Study I: Bias and variance of parameter estimates

The expected behavior of a regularized model is that the introduction of a penalty term to the likelihood will reduce sensitivity to the specific dataset, resulting in lower variance and greater generalizability at the cost of a systematic error (bias). Though the behavior of regularization techniques is well documented in generalized least squares regression (Hastie et al. 2009), it is less well studied in the context of phylogenetic comparative methods (but see Khabbazian et al. 2016; Clavel et al. 2019) and particularly understudied when applied to non-regression frameworks such as the Markov models. In the context of discrete character models, the expected behavior is that rate estimates will tend towards zero, resulting in a downward bias of parameter estimates. Biologically, this means that a regularized model expects fewer changes in the character state through time and favors explanations of homology over homoplasy. The benefit of a regularized model is the increased generalizability should

lead to a model which more accurately reflects the dynamics of unsampled taxa and leads to better predictions of unknown character states (e.g., ancestral character estimation).

To test the bias-variance trade-off associated with the regularization framework introduced here, I conduct a simulation study to evaluate the bias, variance, and overall error of parameter estimates for regularized and unregularized models. I simulate data under three model structures which vary only in the number of characters included. Specifically, I simulate data for a single character with binary states (k=2), two characters with binary states (k=8), and three characters with binary states (k=24). It was not required that all state combinations had to be observed. All models are simulated under an all rates different (ARD) model with dual transitions disallowed. Parameters are sampled from a log-normal distribution, where the logarithms of the parameters are normally distributed with a mean of 0 and standard deviation of 0.25. Parameters are sampled 100 times for a given phylogenetic tree. Phylogenetic trees of size 100, 250, and 500 taxa are simulated 10 times under a pure birth model with a sampling fraction of 1 using the R-package TreeSim (Stadler 2019). Phylogenies are then rescaled to a height of 1. These same 30 phylogenetic trees (10 each for the different number of taxa) are used for the 3 different model structures. In total, each model structure is tested under 3000 simulated datasets for a total of 9000 simulations.

Discrete character models are fit using the R-package corHMM (Beaulieu et al. 2022) with the simulating model structure (i.e., ARD) specified. Models are fit with no penalty term (unregularized), an *er* penalty term, an *l1* penalty term, and an *l2* penalty term with $\lambda = 1$. The maximum likelihood estimates are then compared to the simulating values and their bias, variance, and Root Mean Squared Error (RMSE) evaluated. Note that hidden Markov models are not included in Simulation Study I because, for trees of this size, it would be difficult to simulate data which consistently had a strong signal of rate heterogeneity (Boyko and Beaulieu 2021). This means that regularized models would show an inflated negative bias as parameters in a second-rate class would be estimated as 0 when there is weak evidence of hidden rates.

Simulation Study II: Automatic model selection

Apart from reducing the variance of discrete character models, a primary aim of the dredge framework is to help users navigate large and complex model spaces. As such, in *Simulation Study II* I simulate data under three historically significant and widely applied model structures and evaluate whether the dredged model captures the most important features of the generating models. Specifically, I simulate data under: (1) a correlated model, which captures the influence of two or more traits on

each other (Pagel 1994); (2) an ordered and unidirectional model of trait change (e.g., Skinner et al. 2008); and (3) a hidden Markov model, where there is an unobserved factor causing rate variation in an observed character (Felsenstein and Churchill 1996; Beaulieu et al. 2013; Boyko and Beaulieu 2021). Each of these models has a unique set of criteria for the "most important features," depending on what is typically considered relevant in empirical hypotheses using the particular model structure. Below I go into detail about which features are considered relevant for this study in each case.

When the generating model is consistent with Pagel's (1994) correlated model, 100 datasets are simulated under mapping matrix M_{corr} transition rates of $q_{34} = 10$, $q_{21} = 5$, $q_{43} = q_{12} = 0.1$, and $q_{13} = q_{24} = q_{31} = q_{42} = 1$Note that the in-text subscript notation (e.g.q_{34}) intentionally differs from those presented in the mapping matrices (e.g., $q_{X_0 \to X_1|Y_1}$) with the former describing the rate's location in the matrix and the latter describing the type of transition. A unique pure birth phylogeny (birth rate 1 events/MY) with 250 tips is simulated for each dataset and rescaled to have a height of 1. To evaluate whether the dynamics of the correlated model are adequately captured, I quantify how often does the best fitting dredge model find that: (1) $q_{34} > q_{12}$, (2) $q_{21} > q_{43}$, and (3) $q_{21} > q_{34} > q_{ij}$. These comparisons are chosen because they are relevant to the interpretation of the correlated model. Evidence for the correlated model comes from tests of whether changes in the focal character, depend on the state of the background character (Pagel and Meade 2006). As a hypothetical example, consider the situation in which $X$is parity mode and$Y$is climatic preference in reptiles: $X_0$is oviparity, $X_1$is viviparity, $Y_0$is a warm climatic preference, and $Y_1$is a cold climatic preference. An important factor for the evolutionary transition from egg-laying (oviparity) to live-birth (viviparity) in reptiles is the climate a lineage occupies, with squamate viviparity evolving frequently in colder climates, such as those found at high altitudes and latitudes (Blackbum 1999). When a correlated model of discrete character evolution is applied to this hypothetical example, the simulation parameters outlined above indicate that the transition of oviparity to viviparity is faster when a lineage is in a cold environment (q_{34}) than when the lineage is in a warm environment (q_{12}) and that transitions from viviparity to oviparity are faster when lineages are in a warm environment (q_{21}) than when they are in a cold environment (q_{43}). If parity mode evolved independently of climatic environment, then the expectations is that the transition rates between oviparity and viviparity are the same regardless of which environment the lineages inhabit. If the corHMMDredge framework consistently finds $q_{34} > q_{12}$and $q_{21} > q_{43}$, then it would seem to adequately capture correlated trait dynamics. The final test ($q_{34} > q_{21} > q_{ij}$) determines whether the the fastest and slowest transition rates maintain the same order, though this is a less important comparison with respect to correlated character evolution.

$$M_{corr} = \begin{array}{c|cccc} & X_0Y_0 & X_1Y_0 & X_0Y_1 & X_1Y_1 \\ \hline X_0Y_0 & - & q_{X_0 \to X_1 | Y_0} & q_{Y_0 \to Y_1 | X_0} & - \\ X_1Y_0 & q_{X_1 \to X_0 | Y_0} & - & - & q_{Y_0 \to Y_1 | X_0} \\ X_0Y_1 & q_{Y_1 \to Y_0 | X_0} & - & - & q_{X_0 \to X_1 | Y_1} \\ X_1Y_1 & - & q_{Y_1 \to Y_0 | X_1} & q_{X_1 \to X_0 | X_1} & - \end{array}$$

To simulate under an ordered trait evolution model, I construct a mapping matrix M_{ord} which allows for transitions from state 1 to state 2 to state 3, where reversions are possible only in state 2. All transition rates are set to 1, such that $q_{12} = q_{21} = q_{23} = 1$ and a unique pure birth tree with 250 taxa is simulated with birth rate 1 for 100 datasets and trees are rescaled to a height of 1. To evaluate whether the dynamics of a ordered model are actually captured by the dredge framework, I conduct 4 tests: (1) , $q_{13} = 0$(2) $q_{21} > 0 \wedge q_{23} > 0$, (3) $q_{31} = q_{32} = 0$, and (4) $\mathrm{k} = 1$. The first 3 tests examine whether the correct structure of the ordered model is discovered by corHMMDredge, and the final test examines the parameter equating algorithm. A biological example of this type of model would be transitions between outcrossing and obligate selfing in angiosperms, through an intermediate facultative selfing state and the irreversibility of obligate selfing state. If this were the case, X_1 would represent out-crossing, X_2 would represent facultative selfing, and X_3 would represent obligate selfing. The first test asks whether corHMMDredge has accurately discovered that it is impossible to directly transition from outcrossing to obligate selfing. The second test naturally follows the first examining whether transitions from the intermediate facultative selfing state are allowed to transition to both obligate selfing and outcrossing. The third tests asks whether the obligate selfing is correctly inferred to be a sink state, incapable of reverting back to either facultative selfing or outcrossing. The final test assesses whether an equal rates model was found. This ordered model is somewhat unique among the historical models I am testing in that it is often clearly linked to apriori biological expectations related to developmental biology. And in those cases, it may not be appropriate to apply a dredge framework as the model set will be well defined. However, these sorts of ordered dynamics are possible in settings where developmental information is not readily available, and the goal of these simulations is to determine if the corHMMDredge framework can detect ordered models in general. This model structure is particularly important for testing the ability of the dredge framework to correctly drop parameters.

$$M_{ord} = \begin{array}{c|ccc} & X_1 & X_2 & X_3 \\ \hline X_1 & - & q_{1 \to 2} & - \\ X_2 & q_{2 \to 1} & - & q_{2 \to 3} \\ X_3 & - & - & - \end{array}$$

The final historically significant model being tested is a hidden Markov model where unobserved rate variation is introduced via a hidden factor. The mapping matrix, M_{HMM}, describes a simple HMM for a single character, X, with two states, 0 and 1. The structure of an HMM is not unlike a correlated model, except that in the correlated case, both characters are observed and for an HMM we are trying to detect the hidden character based on the rate differences in the observed character (Boyko and Beaulieu 2021). For simulations, transition rates are set to be $q_{12} = q_{21} = 10$, $q_{34} = q_{43} = 0.01$, and $q_{13} = q_{24} = q_{31} = q_{42} = 1$ and simulated on 100 unique pure birth tree with 250 taxa. To evaluate whether the dynamics of a hidden Markov model are captured by the corHMMDredge framework, I conduct 4 tests: (1) $k > 2$ (2) $q_{12} \wedge q_{21}/q_{34} \wedge q_{43} \geq 100$, (3) $q_{12} \vee q_{21} > q_{ij}$, and (4) . The first tests $q_{12} = q_{21} \wedge q_{34} = q_{43}$asks whether there is support for hidden rates at all. The second test, conditioned on the first, asks whether the difference between the fastest rate class 100 times is at least greater than the slower rate class. I note that the simulating difference is a 1000 times difference, but for this test I was less interested in the exact parameter estimation capabilities since biases in estimation were examined more thoroughly in *Simulation Study I.* The large difference (and arguably unrealistic) in rate classes is chosen to ensure that simulating data had a signal of rate heterogeneity. Test 3 examines whether the fastest rates belong to transition between observed states, rather than between hidden states. And finally, test 4 asks whether the transitions between observed states are correctly inferred to be the same within their respective rate class.

$$M_{HMM} = \begin{array}{c|cccc} & X_{0A} & X_{1A} & X_{0B} & X_{1B} \\ \hline X_{0A} & - & q_{X_0 \to X_1 | A} & q_{A \to B | X_0} & - \\ X_{1A} & q_{X_1 \to X_0 | A} & - & - & q_{A \to B | X_1} \\ X_{0B} & q_{B \to A | X_0} & - & - & q_{X_0 \to X_1 | B} \\ X_{1B} & - & q_{B \to A | X_1} & q_{X_1 \to X_0 | B} & - \end{array}$$

In addition to testing the ability to correctly infer model structures, I examined the influence of the different regularization schemes on ancestral state reconstruction. For each of the simulating datasets, I examined how far the probability distributions of the known ancestral states was from the best fitting dredge model's ancestral state estimate. To estimate the distance between ancestral state reconstructions, I use JS Divergence. To summarize these differences, I use median divergence and because accuracy of ancestral states can differ based on the age of the node, I examine ancestral state reconstruction divergence as a function of node age. I fit a simple linear regression to visualize potential differences between alternative regularization methods. However, parameter estimates from the linear regression fits should be taken with a grain of salt as the data points are certainly non-

independent because each phylogeny contributes several data points, and these data are related to one another due to common ancestry. Nonetheless, as a qualitative assessment it can still prove useful as I am more interested in the relative differences between regularization methods, and they should be equally influenced by this non-independence.

Empirical example: concealed ovulation and mating systems in Old-World monkeys

Concealed ovulation has been considered an adaptation in primates as either a means of promoting parental care by increasing confidence in paternity or improving male behavior towards their potential offspring by confusing paternity issues (Sillen-Tullberg and Moller 1993). Addressing whether concealed ovulation increases confidence or confusion in paternity issues is difficult to study at a phylogenetic scale, so previous examinations of this question have focused on the context in which the trait has evolved. An initial analysis by Sillen-Tullberg and Moller (1993) found that the common ancestor of all anthropoid primates showed signs of ovulation (or estrus advertisement) and had a multimale mating system. Concealed ovulation then evolved from this state 8 to 11 times in a nonmonogamous context and at most once in a monogamous context. In contrast, in their examination of this same question, Pagel and Meade (2006) found the common ancestor of all Old-World primates to have concealed ovulation and a monogamous mating system. This discrepancy was explained by the difference in methodology. Pagel and Meade (2006) had used rate estimates from Markov models and Sillen-Tullberg and Moller (1993) had used a parsimony reconstruction. Methodology notwithstanding, estimating the ancestral state of the Old-World primate common ancestor is crucial for the subsequent interpretation as it will imply whether estrus advertisement or concealed ovulation is the derived state in old world primates.

To examine the evolution of concealed ovulation and mating systems, I apply corHMMDredge to the dataset of Pagel and Meade (2006). The dredge models use *l1* regularization and I conduct phylogenetic cross-validation to determine the optimal λ value. For cross-validation I examine 5 folds for λ values of 0, 0.25, 0.5, 0.75, and Jensen-Shannon divergence is used as a metric of accuracy. The dredge results are compared to a standard k-state Markov model (Mk) with parameters allowed to completely differ (ARD). The Mk model allows for the dependent evolution of mating system and estrus advertisement. All model fits undergo 10 random restarts to increase the chances of finding the global optimum for the rate estimates. Parameters from the Mk model are dropped if they reach a lower bound so that profile likelihoods only examine the parameters which are most necessary to the model (this was done manually for the Mk model). Both the Mk model and dredge results are used to estimate ancestral states based on their respective maximum likelihood estimates. The root prior for both

frameworks is set to “maddfitz” in the corHMM R-package (FitzJohn et al. 2009). I examine the profile likelihoods of the rate estimates for both frameworks and consider whether, and how, the interpretations might differ within 95% confidence intervals of the MLE. To do this, I implement a new profile likelihood function within corHMM and examine 50 fixed values for each parameter spanning 100 orders of magnitude around the MLE. Profile likelihood operates by examining what the maximum likelihood would be given a fixed value and with the remaining parameters freely estimated.

Results

Simulation Study I: Regularized models substantially reduce variance

As anticipated, the introduction of *l1* and *l2* regularization led to a significant reduction in variance across all tested models. The variance averaged across all model structures (more detailed breakdowns can be found in Supplementary Tables) for *l1* regularized models was 2.4%, 10.8%, and 45.0% of the unregularized models’ variance for 100, 250, and 500 tips respectively. For *l2* regularized models, variance was 0.5%, 3.1%, and 18.4% of the unregularized models’ variance. For *er* regularized models, variance was 97.8%, 101.1%, and 102.4% of the unregularized models’ variance (raw variance values are reported in Table 1). These results indicate a substantial improvement in model stability for *l1* and *l2* regularization methods, but not for *er* regularization. However, these reductions in variance for *l1* and *l2* regularization were larger than anticipated and beg the question of what is driving the substantial variance of the unregularized models. The large variance of the unregularized models is primarily driven by parameter estimates approaching the upper bound of 100 transitions/MY. If estimates above 50 transitions/MY are excluded (641 model fits in total) variance is instead 13%, 31%, and 64% of the unregularized models’ variance for *l1* regularized models and 2.8%, 9.0%, and 26.5% for *l2* regularized models for 100, 250, and 500 tips respectively. The reduction of variance continues (57.5%, 71.2%, 80.7% for *l1* and 13.4%, 22.3%, 34.9% for *l2*) if one were to be stricter and exclude estimates above 10 transitions/MY. It is important to note that this study has the benefit of knowing with certainty that high transition rate estimates are incorrect and in an empirical setting ignoring transition rates above a specified threshold is difficult to justify.

The absolute average bias for phylogenies with 100 taxa and 250 taxa was found to be lower for *l1* and *l2* regularized models, contrary to the expected increase in negative bias (Table 1). This trend reversed for *l2* regularized models fit to datasets with trees of 500 taxa, but *l1* regularization consistently showed the lowest bias across all tested tree sizes. More surprisingly, the bias of *l1* regularized models was slightly positive on average, contrary to the expected downward bias, although

a negative bias is found if one uses the median difference instead of the arithmetic mean (Fig. 3). Furthermore, examining these parameters in log-space results in the expected behavior of a consistent negative bias, but since empirical studies rarely discuss parameter estimates outside of their raw units (transitions/MY), I focus my interpretations on the untransformed parameter units. *l2* regularized models did consistently show a moderate downward bias. When considering overall error, measured as the RMSE, *l1* and *l2* regularized models showed substantial improvements over unregularized and *er* regularized models (Table 1). This improvement in RMSE highlights the effectiveness of *l1* and *l2* regularization in achieving better generalizability and prediction accuracy, despite the potential trade-off with increased bias.

Table 1: Bias, variance, and overall error (RMSE) for regularized and unregularized models across phylogenies of varying sizes. Regularization methods l1 and l2 demonstrate significant reductions in variance and overall error compared to unregularized (Mk) and er regularized models. l1 regularization consistently achieves a low RMSE and bias across all tree sizes, with a slight positive bias for smaller phylogenies that becomes negative when using the median. l2 regularization similarly reduces RMSE and variance but introduces moderate downward bias.

Reg. Type	**N Taxa**	**Bias**	**Variance**	**RMSE**
None (Mk)	100	1.65	87.83	9.45
er	100	1.68	90.66	9.6
l1	100	-0.05	2.08	1.4
l2	100	-0.5	0.46	0.98
None (Mk)	250	0.48	15.91	3.92
er	250	0.47	15.47	3.87
l1	250	0.04	1.71	1.18
l2	250	-0.33	0.5	0.81
None (Mk)	500	0.16	2.69	1.45
er	500	0.16	2.71	1.46
l1	500	0.04	1.21	0.86
l2	500	-0.21	0.49	0.64

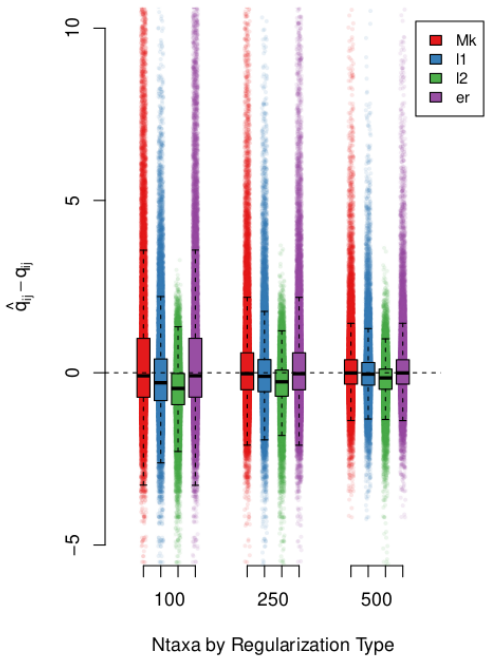

Figure 3 – Comparison of parameter estimates across different regularization methods and phylogeny sizes. The dashed line represents a perfectly unbiased estimate (0). Both l1 and l2 regularization show markedly reduced variance compared to unregularized and er regularized models. The median values of both l1 regularization and l2 regularization consistently demonstrate a moderate downward bias.

Simulation Study II: Similar performance for model selection

Both regularized and unregularized methods preformed equally well when it came to automatically inferring the generating model structure (Table 2). The most substantial differences are seen for hidden state models where *l1* and *l2* regularization models were better at detecting the magnitude of the differences of the rate classes ( ) and at detecting that the transitions between observed states were the same within their rate class ( ). Otherwise, it is worth considering the overall performance of automatic model selection and determining which tests posed the greatest difficulties. For the dependent model, only ~50% of the dredges were able to detect that (the relative magnitude of the rates). Although this is not a particularly important test to pass to identify correlated evolution, there is biological significance to rate differences. For the ordered model, only one test had greater than 50% pass rate (Table 2). This test ( ) is also one in which the model only needs to avoid dropping parameters and should be expected to pass by default. It is more concerning that the pass rate was only around 50% for and because these are the defining features of the ordered model. However, just

because the generating model was an ordered model does not mean that the simulated data will most support that model. The randomness and finiteness of the generated data means that one cannot expect the generating model to always be the most supported. Viewing these results relative to the standard Mk model, the regularization approaches matched traditional unregularized methods.

Table 2: Results from structure tests comparing regularized and unregularized methods for model inference. All methods, including Mk, l1, l2, and er regularization, performed similarly in identifying the generating model structure. Regularized approaches, particularly l1 and l2, demonstrated advantages in hidden state models by better detecting the magnitude of rate class differences and the equality of transitions within rate classes. However, challenges remained for tests defining specific models, such as the ordered model.

Generating model	**Focal Test**	***Mk***	***l1***	***l2***	***er***
dependent	$q_{34} > q_{12}$	0.88	0.88	0.88	0.86
dependent	$q_{21} > q_{43}$	0.8	0.88	0.88	0.83
dependent	$q_{21} > q_{34} > q_{ij}$	0.45	0.53	0.52	0.45
ordered	$q_{13} = 0$	0.45	0.44	0.4	0.47
ordered	$q_{21} > 0 \wedge q_{23} > 0$	0.94	0.91	0.93	0.94
ordered	$q_{31} = q_{32} = 0$	0.47	0.48	0.48	0.5
ordered	$q_{ij} = 1$	0.32	0.32	0.29	0.34
hidden states	$k > 2$	0.78	0.77	0.77	0.79
hidden states	$q_{12} \wedge q_{21}/q_{34} \wedge q_{43} \geq 100$	0.48	0.62	0.58	0.48
hidden states	$q_{12} \vee q_{21} > q_{ij}$	0.75	0.7	0.72	0.75
hidden states	$q_{12} = q_{21} \wedge q_{34} = q_{43}$	0.36	0.64	0.67	0.4

There was little effect of regularization on the accuracy of an ancestral state estimate (Table 3). This is surprising because regularization is expected to increase the generality of the model by reducing the variance of the estimates. The reduction of variance was confirmed by S*imulation Study 1*, but this did not translate into an improved ancestral reconstruction which itself can be viewed as a prediction based on the model and current data. It is possible this is due to phylogenetic trees having more nodes towards the present and the increased generality playing less of a role as the prediction is not as deep in time for a difference to manifest. This is given some credibility as the accuracy of ancestral state predictions as a function of time was better for *l1* and *l2* regularized models than for either unregularized (*Mk*) or *er* regularized models (Table 3). However, I hesitate to fully embrace any formal statistical significance of this result because the nodes being reconstructed are non-independent. If the set of simulations is somehow biased towards the regularized models, the pseudo-replication of having many nodes per single phylogenetic tree could inflate the sample size and give more power to statistical tests than is warranted. Nonetheless, qualitatively, regularization did have the expected effect in which

its model predictions were better than an unregularized model as ancestral reconstructions got nearer to the root of the phylogeny (Table 3; Fig. S1).

Table 3: Ancestral state estimation results across generating models and regularization methods. Lower values indicate better performance. Regularization had minimal impact on overall accuracy for ancestral state reconstruction, with similar results across Mk, l1, l2, and er methods. Notably, l1 and l2 regularization outperformed other methods when considering predictions closer to the root of the phylogeny, as indicated by the lower slope values in deeper ancestral reconstructions.

Generating Model	**Estimator**	**Mk**	**l1**	**l2**	**er**
dependent	slope	0.21	0.2	0.16	0.22
dependent	JS_dist	0.06	0.06	0.06	0.06
ordered	slope	0.11	0.1	0.08	0.11
ordered	JS_dist	0.03	0.03	0.03	0.03
hidden states	slope	0.16	0.15	0.1	0.17
hidden states	JS_dist	0.22	0.22	0.26	0.21

A multimale mating system and concealed ovulation ancestor

The evolutionary history of mating system and estrus advertisement differed depending on which modeling framework was used. These differences applied to both the best supported model structure and the ancestral state estimation. The dredge approach with regularization was able to find a better supported model structure than default settings ($\Delta AIC = -10.9$). This AIC comparison came from a comparison of two unregularized models, one of which used the model structure automatically detected by the dredge algorithm. The estimated the ancestral state at the root under the dredge model was found to be a multimale mating system with concealed ovulation (marginal probability 0.818), whereas the standard Mk model under default settings inferred the root state to have estrus advertisement with multimale mating (marginal probability 0.61). Given this different root state, it is unsurprising that much of the evolutionary history of these two characters differed throughout the internal nodes (Fig. 4). Most of these differences can be attributed to whether estrus advertisement or concealed ovulation is the derived state. Transition rates were generally similar between the two approaches (Fig. 5), with the exception that the dredge model estimated the transition from monogamous to multimale (given estrus advertisement was present in the lineage) to be 0, whereas that particular transition was the fastest under the standard corHMM approach (0.185 transitions/MY). Additionally, the dredge model suggested that the transition from concealed ovulation to estrus advertisement given a multimale mating system was 0.014 transitions/MY, whereas the standard corHMM model inferred a transition rate of 0. Importantly, the rate estimates were generally of the same magnitude despite the regularization penalty downwardly biasing values (Fig. 5). Finally, the

profile likelihoods for both parameters of the dredged results were better behaved with clean likelihood peaks and reasonable confidence intervals than the default Mk model (Fig. 5).

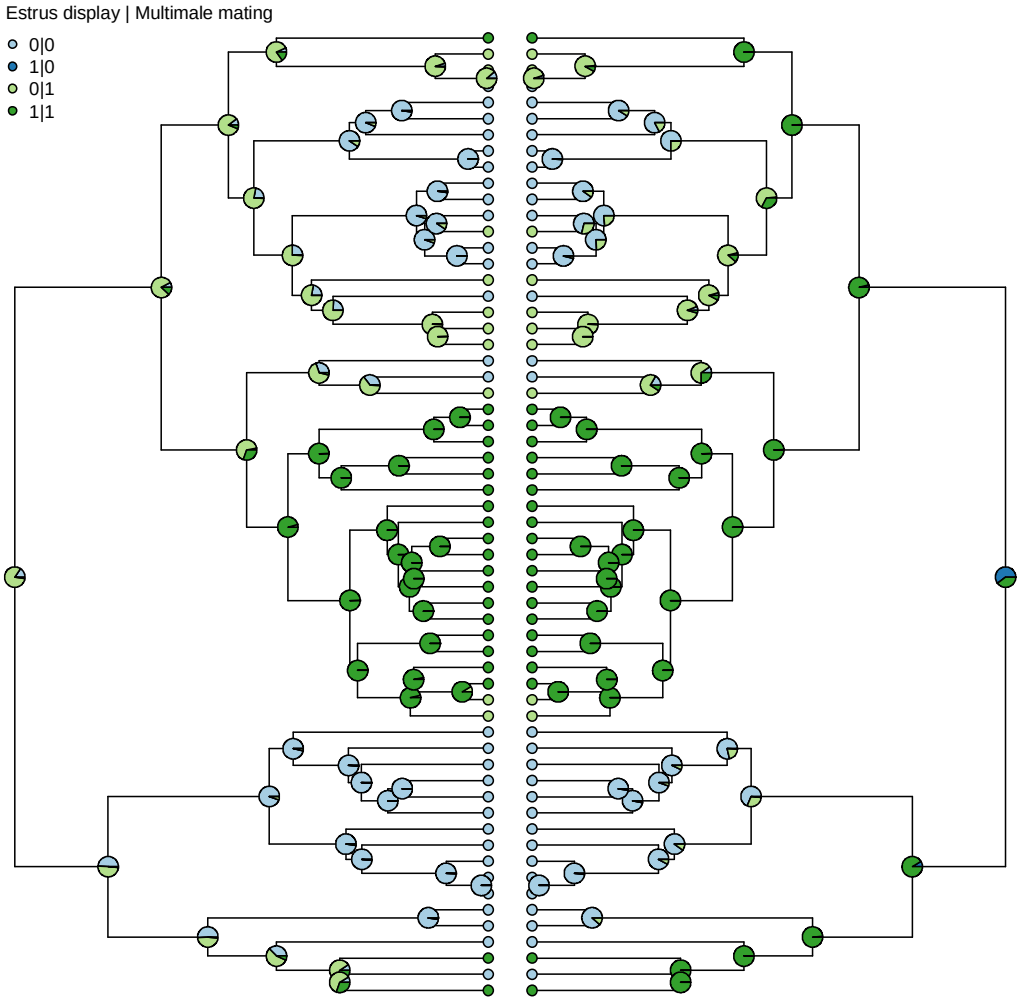

Figure 4 – Empirical reconstructions of ancestral states under two modeling frameworks: dredge with regularization (left) and standard corHMM (right). Light colors represent the absence of estrus display, while dark colors indicate its presence. Blue signifies the absence of multimale mating, and green indicates its presence. The dredge model reconstructs the root state as no estrus display with multimale mating (light green), whereas the default discrete models suggest near-equal probabilities of estrus display with a slight preference for monogamy (dark blue and dark green). These differences reflect distinct interpretations of ancestral reconstruction as influenced by the modeling approach.

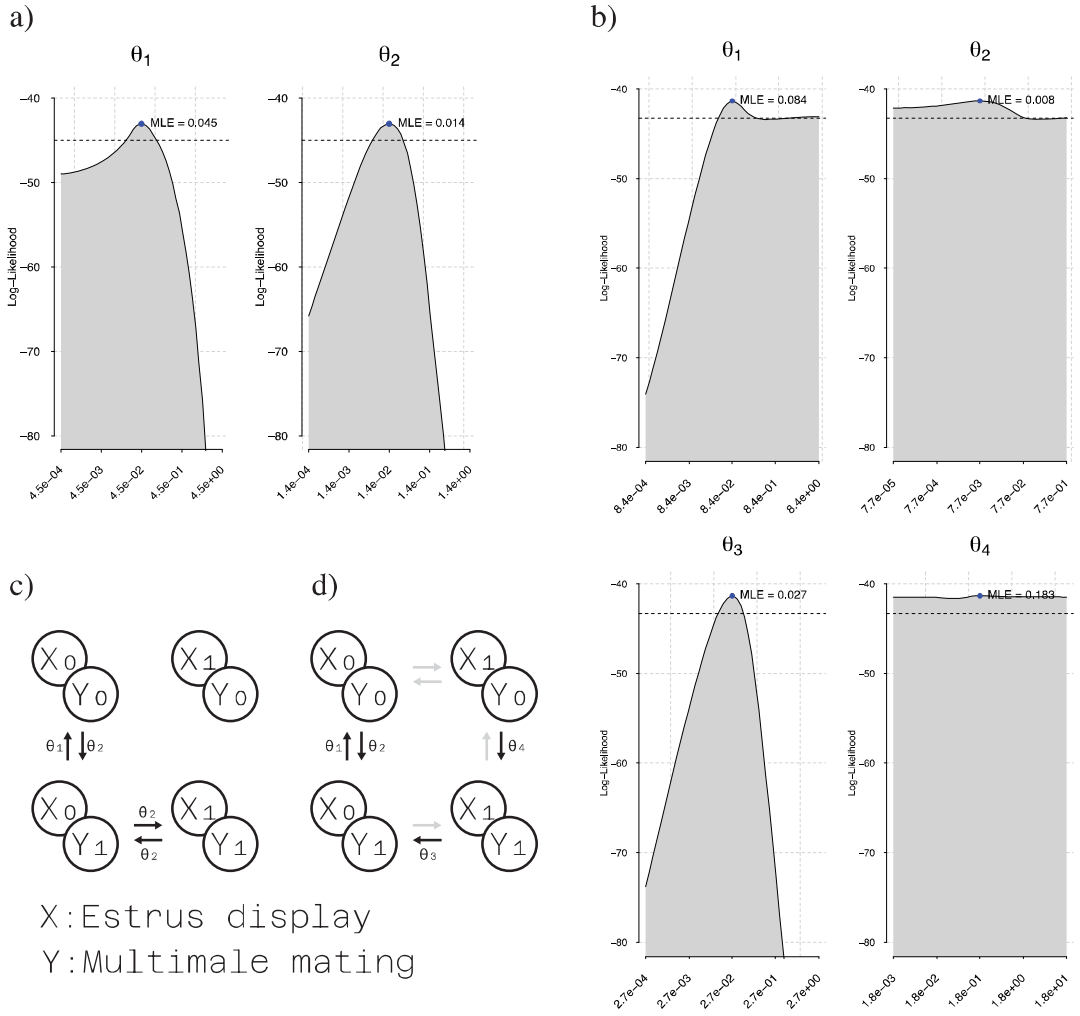

Figure 5 - Comparison of likelihood profiles and model structures for transition rates under the dredge model (a, c) and the standard corHMM model (b, d). Panels (a) and (b) show the profile likelihoods, with the dredge model producing cleaner peaks and narrower confidence intervals. Panels (c) and (d) depict the optimal model structures inferred by each approach, with transitions estimated near zero represented in grey. The dredge model assigned a transition rate of 0 for the shift from monogamous to multimale mating when estrus advertisement was present, in contrast to the standard corHMM model's highest transition rate of 0.185 transitions/MY for the same scenario. Similarly, the dredge model estimated a low transition rate (0.014 transitions/MY) for concealed ovulation to estrus advertisement under multimale mating, while the corHMM model inferred no transitions. Despite regularization downwardly biasing rate estimates, overall rate magnitudes were consistent across models, with improved parameter behavior under the dredge framework.

Discussion

A good model set is crucial for ancestral state estimation

The dredge algorithm was able to find a unique, non-standard, model structure which differed in its inferred history of Old-World monkey evolution from the default settings (Fig. 4ab). The dredge model estimated the ancestral state at the root to be a multimale mating system with concealed ovulation (marginal probability of 0.818), whereas the standard Mk model under default settings inferred the root state to have estrus advertisement (marginal probability of 0.99), but the estimate was less certain for the mating system with multimale mating being slightly preferred (marginal probability of 0.61). These differences highlight how large of an effect examining a more complete model set can have. If one were to only examine the default model settings, not only would the rate estimates differ but the entire evolutionary history of these characters changes. Furthermore, 3 of the 4 parameters estimated under default settings showed evidence of likelihood ridges spanning several orders of magnitude around the MLE (Fig. 5). This large amount of uncertainty is indicative of a lack of information needed to freely estimate these parameters and demonstrates the value of automatically searching model space to find optimal model structures.

Biologically, whether one used the default settings or the dredged model, both results would have differed from what has been previously presented as the ancestral state of Old-World monkeys. Sillen-Tullberg and Moller (1991) found that the root state was likely to be estrus advertisement with multimale mating system and therefore discussed the evolution of concealed ovulation in this context. In contrast, Pagel and Meade (2006) had found that the root state was concealed ovulation with monogamous mating. In the absence of direct observation, it is nearly impossible to say which of these outcomes is more correct and since the modeling results presented here are an intermediate outcome between the two, no consensus can be claimed. However, revisiting the hypotheses of Sillen-Tullberg and Moller (1991) under the dredged model results do provide some interesting alternative interpretations for the evolution of estrus advertisement and mating system.

Sillen-Tullberg and Moller (1991) had considered concealed ovulation to be the derived state and found that it had almost always evolved in non monagomous lineages. This provided evidence that concealed ovulation was more likely to be a prerequisite for the evolution of monogamy, rather than a consequence of it (as had been hypothesized by Burley 1979). Interestingly, despite a completely different ancestral root state, this interpretation still holds under the new modeling results. In fact, the root state being inferred as concealed ovulation with multimale mating expedites the evolutionary

pathway proposed by Sillen-Tullberg and Moller (1991) with monogamy only ever evolving in lineages with concealed ovulation. The primary difference is that rather than first evolving concealed ovulation from estrus advertisement, the ancestral state was already concealed ovulation. These modeling results also suggest that estrus advertisement evolved primarily from lineages which already had multimale mating systems. This finding somewhat contradicts the hypothesis that signs of ovulation will disappear in non-monogamous mating systems because paternity confusion will be beneficial (Hrdy 1979). That is one possible outcome and is inferred to occur at a rate of 0.0139 transitions/MY (1|1 → 0|1), but from the concealed ovulation and multimale state, a lineage is just as likely to regain overt ovulation signs as it was to have lost them in the first place (0.0139 transitions/ MY; 0|1 → 1|1). This is because under the dredged model structure, concealed ovulation with multimale mating, is an intermediate state. Nonetheless, a lineage with a multimale mating system was more than 3 times more likely to become monogamous as it was to develop estrus advertisement.

Are ancestral states more robust than rate estimates?

One of the more surprising results of this study was that regularization did not noticeably improve ancestral state reconstruction. This was despite regularization reducing the variance of model estimates as one would hope. In fact, often rate estimates which were several orders of magnitude apart could have similar marginal probabilities for ancestral states despite the estimates suggesting very different tempos of phenotypic evolution. This is to be expected when both rate estimates are relatively low, but following along the likelihood ridge (Fig. 5) and increasing a single parameter by 1000 orders of magnitude led to no noticeable differences in the ancestral history (Fig. S2). This could be indicative of rate estimates being more sensitive to overly complex models than ancestral states. However, this is not to say that ancestral states should be taken at face value. For instance, within the Old-World monkey case study, many of the ancestral state reconstructs at internal nodes completely changed depending on which model structure was used. So, although the ancestral estimation may have been more robust to rate variation along a likelihood ridge, it is far from robust to alternative models. Furthermore, this could reflect the relative importance of model structure versus accurate parameter estimation. Inferences from a wide range of rate estimates were roughly equivalent, but slight differences in model structure required a substantially different biological interpretation. Additionally, and a far deeper reason that ancestral state reconstruction should not be taken at face value, is that it is frequently reliant on the phenotypic variation observed in extant taxa. Fossil evidence can have an unprecedented impact on the estimation of ancestral states (Finarelli and Flynn 2006), but only if it is available and included. It should be noted that the inclusion of fossil taxa is itself not a silver bullet as

they will disproportionately influence rate and ancestral state estimates. This can be invaluable, especially if the phenotypic data comes from outside the extant distribution. But it means that the correct placement of the fossil taxon within the tree is of utmost importance.

The cost of regularization and eliminating very high transition rates

The dredge algorithm, and regularization more generally, will result in a downward bias proportional to the magnitude of the transition rate. It is worth considering what, if any, is the cost of this downward bias. In empirical settings, it is not uncommon to have transition rates approaching and being estimated at an upper bound. When interpreted as 100s of state transitions occurring over short time intervals, this may be cause for concern because it often seems biologically implausible. However, high transition rates do not necessarily predict that many transitions of that type have occurred. Often, high transition rates may correspond to only a few inferred transitions on a tree, (though they will have occurred very rapidly). The two inferences, extremely high transition rates and relatively few transitions, may seem to be contradictory, but they are perfectly compatible since a lineage must first be in the initial state to transition and the initial state need not be common. Instead, a more serious cause for concern is the amount of information available to infer the transition rate if that transition only occurs rarely throughout the clade's history. This was the case for the Old world monkey dataset above, as several parameter estimates seemed to have very little information and fell along a likelihood ridge. The combination of regularization and model structure searching removes this ridge, but also has the benefit of estimating rates which better match intuition that a transition which occurs rarely throughout the clade has a corresponding transition rate that is also low. Ultimately, a regularization scheme moves towards a more parsimonious view of character evolution, with rate estimates tending to be biased towards explanations of homoplasy over homology. Though this bias is minute.

The value of multiple characters

Part of the value of a dredge approach lies in its ability to identify important relationships between characters by eliminating or equating parameters that lack significant support. This is particularly useful for phenotypic complexes with multiple interacting characters (e.g., pollinator syndromes). However, a dataset containing many characters also introduces a large state space with numerous testable model structures. One solution to address this issue is to independently model each character and then compile the results into a cohesive view. This approach is computationally tractable and can yield well-behaved parameter estimates. But modeling each character independently ignores correlated character evolution and dependent relationships can influence both biological inferences and

ancestral state estimation (Boyko and Beaulieu 2021). When the evolution of one character affects the rate of change in another, independent modeling may lead to uncertainty in certain parts of the phylogeny. By accounting for correlated character evolution, shared information between characters can improve inference (Boyko and Beaulieu 2021). Of course, correlation between discrete characters is a hypothesis that should first be tested. These tests should account for character-independent rate variation (Boyko and Beaulieu 2023), which will introduce even more parameters than standard independent models (Pagel 1994). Searching a state space that includes multiple characters and hidden rate classes can be a daunting task even for experienced comparative biologists and it can be tempting to rely on the most used default model sets. However, there is greater flexibility in potential dependent relationships beyond typical fixed model sets. Correlation tests have traditionally compared independent and correlated models, but not all characters need to show dependent relationships for there to be evidence of correlation (see also Pagel and Meade 2006). A dredge framework is valuable here, as it searches potential model structures and identifies only the necessary dependent relationships.

Choosing the "best" model

Modeling approaches within PCMs often use model averaging, which derives inferences from multiple models (e.g. Caetano et al. 2018). This technique is beneficial because it improves predictive power by accounting for uncertainty across models. Regularization is also well-known for increasing predictive power, as it helps prevent overfitting by penalizing model complexity. So why not simply summarize all models that have been fit during the dredge process? One technical challenge arises from the label switching problem. For example, one model might assign high transition rates to rate class 1, while another assigns them to rate class 2. Averaging across these models without accounting for this arbitrary class assignment could lead to incorrect conclusions. This issue can be mitigated in some cases, such as in diversification models, by focusing on tip rates rather than internal parameters (Title and Rabosky 2019; Vasconcelos et al. 2022). However, a further complication with model averaging is the potential for pseudo-replication of model structures. The model space for more complex models (e.g., those with two rate classes or correlated character evolution) is inherently larger than that for simpler models (e.g., with one rate class or independent evolution). As a result, the total number of models available for averaging may disproportionately represent more complex models, even when there is roughly equal support for simpler alternatives. This can distort model averaging, as the AIC weights may reflect the sheer number of models rather than the actual support for distinct, biologically meaningful models. The most extreme example of this would be if identical models are refit, effectively double counting them. However, this problem can also occur if models are nearly identical,

such as when two parameters are equated (e.g., k=7 vs. k=8). These models might still receive significant support despite being only minor variations of the most complex model. Handling this model structure pseudo-replication becomes particularly important for automated model selection.

When Burnham and Anderson proposed multi-model inference (MMI), they emphasized fitting a set of biologically plausible models (Burnham and Anderson 2002). Each model was justified based on specific hypotheses and each model tested the contribution of different factors. For exploratory analyses on large model spaces, the rationale behind model averaging becomes less clear. In a dredge run, one is not testing specific hypotheses but rather comparing parameter estimates across many models. The final model of a dredge run does not necessarily reflect the combination of several well-supported biological models but is the byproduct of optimizing the bias-variance trade-off through AIC and encouraging sparse parametrizations through regularization. This is valuable from an inference perspective since the more complete a model set is, the less likely it is for unanticipated biases to creep into the results (Maddison and FitzJohn 2015; Rabosky and Goldberg 2015). Nonetheless, the dredge approach and MMI need not be mutually exclusive. In cases where a set of biologically plausible hypotheses are known beforehand, a MMI approach remains an excellent option. And so long as procedures are clearly documented, there are advantages to combining the approaches by considering which model structures are biologically reasonable and determining what the optimal model structure could be for a given dataset. Because these models are ultimately phenomenological, the biological insights we draw from the best-supported model, and how these insights shape our understanding of evolutionary processes, are ultimately derived from subsequent interpretations and considerations in concert with other pieces of evidence (Gardner and Organ 2021). This perspective leaves room for quantifying the patterns in a general way via comparative modeling results, without necessarily forcing too much biological meaning on any particular parameter estimate.

Conclusion

Determining which models to test in an empirical setting is an incredibly valuable process in which biologists take hypotheses and formally structure them so that they can be compared. This process can help clarify thinking around biologically relevant interactions between variables and will lead to more robust inferences. However, if only default models are considered, there is a risk of overlooking plausible and important model structures. This issue is exacerbated by the fact that the knowledge necessary to manipulate PCMs is often difficult to acquire and may be hidden in highly

technical texts, making it challenging for biologists to explore and customize model sets effectively (Cooper et al. 2016). It is evident that a comprehensive model set is important for trustworthy inferences in comparative biology (FitzJohn et al. 2009; Rabosky and Goldberg 2015; Beaulieu and O'Meara 2016; Boyko and Beaulieu 2022), but the growing complexity of discrete character models makes it challenging for users to determine which models are potentially realistic and important to consider. Here I have shown how the corHMMDredge framework may help alleviate this burden. This framework will enable biologists to focus more on model interpretations rather than model construction and may even lead to the discovery of model structures which imply unique hypotheses that would have not been considered otherwise.

Acknowledgements

The author thanks Thais Vasconcelos and the Smith lab for their feedback on the overall clarity of the manuscript. The author would also like to thank Brian O'Meara for his thoughtful comments on an earlier version of this manuscript and pointing out the similarity of this approach and Sanderson's (2002) penalized likelihood. Finally, the author thanks Rudolf Schill and Yutong Wang for general discussions about regularization approaches.

Disclosure statement

The author declares no conflicts of interest.

Data availability

Code and simulation results are available on Dryad:

https://doi.org/10.5061/dryad.kprr4xhfc

Code for corHMMDredge function is available in corHMM:
https://github.com/thej022214/corHMM/blob/master/R/corHMMDredge.R

Code for the profile likelihood is available in corHMM:
https://github.com/thej022214/corHMM/blob/master/R/contour_plots.R

References

Akaike H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 19:716–723.

Beaulieu J., O'Meara B., Oliver J., Boyko J. 2022. corHMM: Hidden Markov Models of Character Evolution. .

Beaulieu J.M., O'Meara B.C. 2016. Detecting Hidden Diversification Shifts in Models of Trait-Dependent Speciation and Extinction. Systematic Biology. 65:583–601.

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. 62:725–737.

Blackbum D.G. 1999. Viviparity and Oviparity: Evolution and Reproductive Strategies. .

Boyko J.D., Beaulieu J.M. 2021. Generalized hidden Markov models for phylogenetic comparative datasets. Methods in Ecology and Evolution. 12:468–478.

Boyko J.D., Beaulieu J.M. 2022. Reducing the Biases in False Correlations Between Discrete Characters. Systematic Biology.:syac066.

Boyko J.D., O'Meara B.C., Beaulieu J.M. 2023. A novel method for jointly modeling the evolution of discrete and continuous traits. Evolution. 77:836–851.

Burley N. 1979. The Evolution of Concealed Ovulation. The American Naturalist. 114:835–858.

Burnham K.P., Anderson D.R. 2002. Model Selection and Multimodel Inference. New York, NY: Springer New York.

Caetano D.S., O'Meara B.C., Beaulieu J.M. 2018. Hidden state models improve state-dependent diversification approaches, including biogeographical models. Evolution. 72:2308–2324.

Clavel J., Aristide L., Morlon H. 2019. A Penalized Likelihood Framework for High-Dimensional Phylogenetic Comparative Methods and an Application to New-World Monkeys Brain Evolution. Systematic Biology. 68:93–116.

Cooper N., Thomas G.H., FitzJohn R.G. 2016. Shedding light on the 'dark side' of phylogenetic comparative methods. Methods in Ecology and Evolution. 7:693–699.

Felsenstein J. 2004. Inferring phylogenies. Inferring phylogenies. p. 664–664.

Felsenstein J. 2012. A Comparative Method for Both Discrete and Continuous Characters Using the Threshold Model. The American Naturalist. 179:145–156.

Felsenstein J., Churchill G.A. 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. Molecular Biology and Evolution. 13:93–104.

Finarelli J.A., Flynn J.J. 2006. Ancestral State Reconstruction of Body Size in the Caniformia (Carnivora, Mammalia): The Effects of Incorporating Data from the Fossil Record. Systematic Biology. 55:301–313.

FitzJohn R.G., Maddison W.P., Otto S.P. 2009. Estimating Trait-Dependent Speciation and Extinction Rates from Incompletely Resolved Phylogenies. Systematic Biology. 58:595–611.

Gardner J.D., Organ C.L. 2021. Evolutionary sample size and consilience in phylogenetic comparative analysis. Systematic Biology. 70:1061–1075.

Gelman A., Carlin J.B., Stern H.S., Dunson D.B., Vehtari A., Rubin D.B. 2013. Bayesian Data Analysis. Boca Raton London New York: Chapman and Hall/CRC.

Hastie T., Tibshirani R., Friedman J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. New York, NY: Springer.

Hastie T.H., Tibshirani R., Wainwright M. 2015. Statistical Learning with Sparsity The Lasso and Generalizations. .

Hoerl A.E., Kennard R.W. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics. 12:55–67.

Hrdy S.B. 1979. Infanticide among animals: A review, classification, and examination of the implications for the reproductive strategies of females. Ethology and Sociobiology. 1:13–40.

Huelsenbeck J.P., Ronquist F., Nielsen R., Bollback J.P. 2001. Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology. Science. 294:2310–2314.

Khabbazian M., Kriebel R., Rohe K., Ané C. 2016. Fast and accurate detection of evolutionary shifts in Ornstein–Uhlenbeck models. Methods in Ecology and Evolution. 7:811–824.

Lemey P., Salemi M., Vandamme A.-M. 2009. The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. Cambridge: Cambridge University Press.

Lewis P.O. 2001. A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data. Systematic Biology. 50:913–925.

Maddison W.P., FitzJohn R.G. 2015. The unsolved challenge to phylogenetic correlation tests for categorical characters. Systematic biology. 64:127–136.

Maddison W.P., Midford P.E., Otto S.P. 2007. Estimating a Binary Character's Effect on Speciation and Extinction. Systematic Biology. 56:701–710.

May M.R., Moore B.R. 2020. A Bayesian Approach for Inferring the Impact of a Discrete Character on Rates of Continuous-Character Evolution in the Presence of Background-Rate Variation. Systematic Biology. 69:530–544.

O'Meara B.C. 2012. Evolutionary Inferences from Phylogenies: A Review of Methods. Annual Review of Ecology, Evolution, and Systematics. 43:267–285.

Pagel M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proceedings of the Royal Society of London. Series B: Biological Sciences. 255:37–45.

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

Pagel M., Meade A. 2006. Bayesian Analysis of Correlated Evolution of Discrete Characters by Reversible Jump Markov Chain Monte Carlo. The American Naturalist. 167:808–825.

Rabosky D.L., Goldberg E.E. 2015. Model Inadequacy and Mistaken Inferences of Trait-Dependent Speciation. Systematic Biology. 64:340–355.

Rohlf F.J. 2001. Comparative Methods for the Analysis of Continuous Variables: Geometric Interpretations. Evolution. 55:2143–2160.

Sanderson M.J. 2002. Estimating Absolute Rates of Molecular Evolution and Divergence Times: A Penalized Likelihood Approach. Molecular Biology and Evolution. 19:101–109.

Sillen-Tullberg B., Moller A.P. 1993. The Relationship between Concealed Ovulation and Mating Systems in Anthropoid Primates: A Phylogenetic Analysis. The American Naturalist. 141:1–25.

Skinner A., Lee M.S., Hutchinson M.N. 2008. Rapid and repeated limb loss in a clade of scincid lizards. BMC Evolutionary Biology. 8:310.

Stadler T. 2019. TreeSim: Simulating Phylogenetic Trees. .

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology. 58:267–288.

Title P.O., Rabosky D.L. 2019. Tip rates, phylogenies and diversification: What are we estimating, and how good are the estimates? Methods in Ecology and Evolution. 10:821–834.

Vasconcelos T., O'Meara B.C., Beaulieu J.M. 2022. A flexible method for estimating tip diversification rates across a range of speciation and extinction scenarios. Evolution. 76:1420–1433.

Zhou Y., Gao M., Chen Y., Shi X. 2024. Adaptive Penalized Likelihood method for Markov Chains. .

Supp Figures

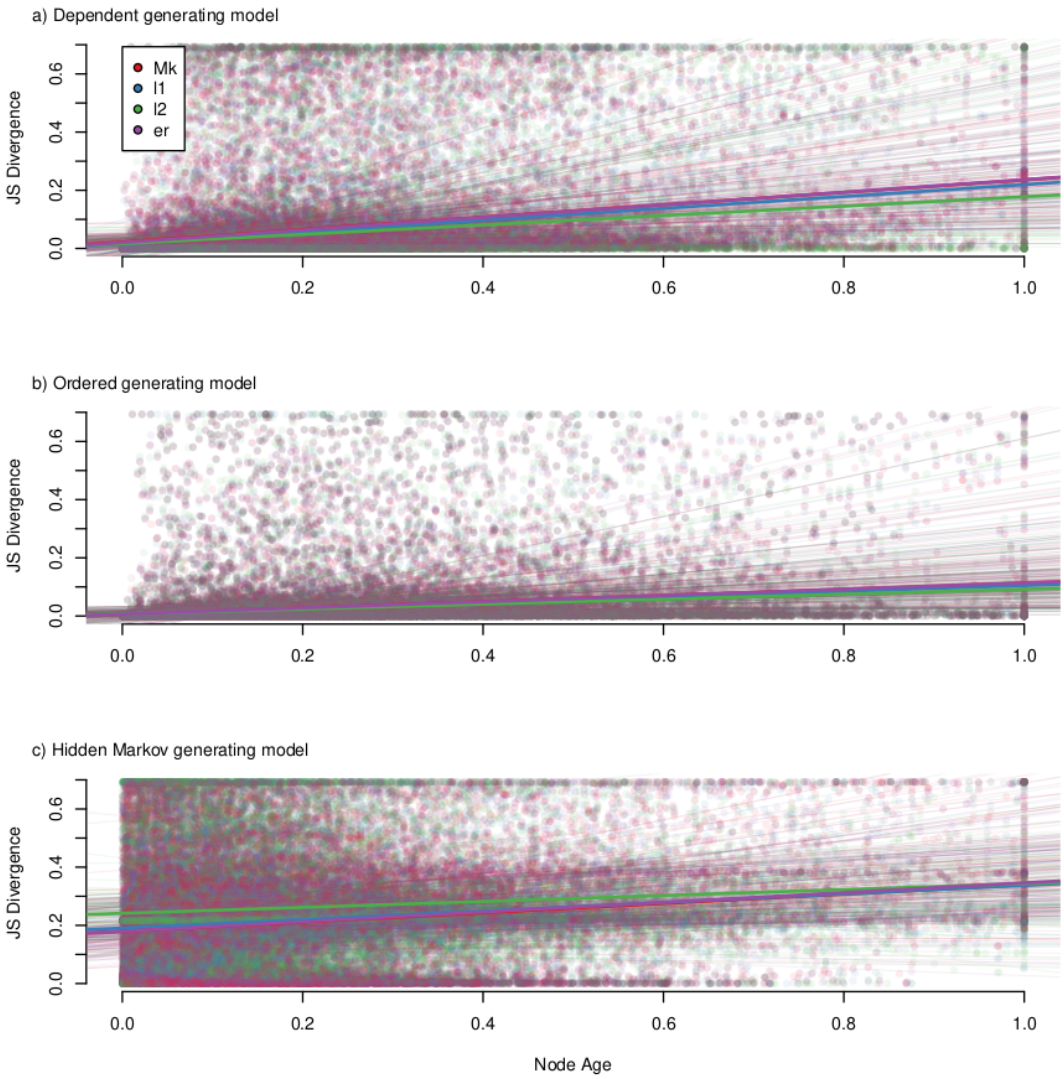

Figure S 1 – divergence of ancestral state as a function of node age. Solid lines are the average estimate for all simulations and each dot represents one node.

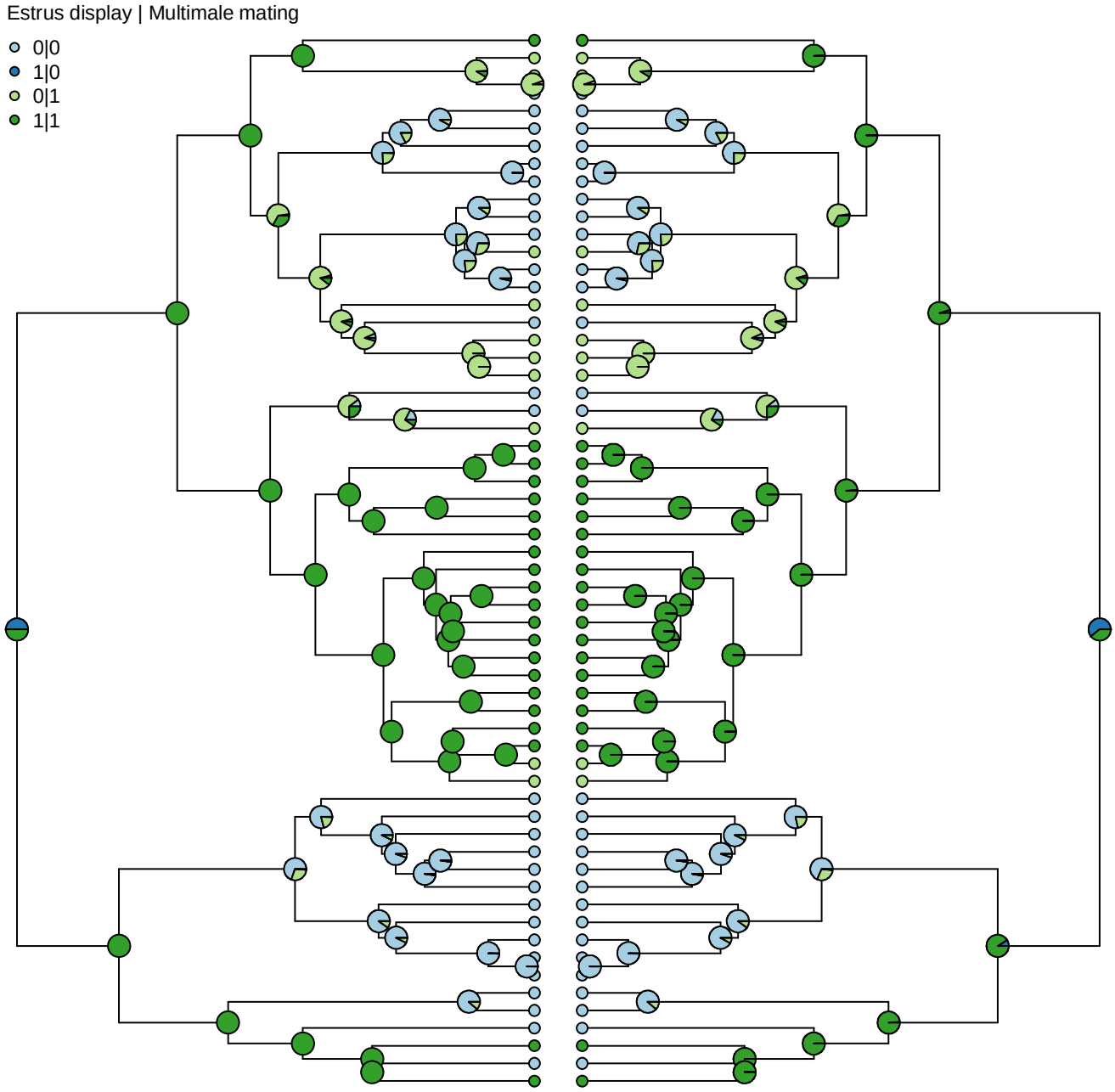

Figure S 2 - left is 95% confidence interval rate choice (high rates, but from within 95% confidence interval) and right is the standard corhmm fit.