05-Apr-2025  
  
Dear Dr Boyko:  
  
Decision on USYB-2024-270, Automatic Discovery of Optimal Discrete Character Models:  
  
Major revisions.  
  
Thank you for your Systematic Biology submission. It has been reviewed by Associate Editor Professor Josef Uyeda and two expert reviewers. Their comments are listed at the end of this letter. The reviewers and the AE provide some excellent constructive suggestions that I am sure you will appreciate. I really want to emphasize that I found the suggestions from both reviewers and the AE extremely constructive.  
**Thank you for the opportunity to revise our manuscript. I am grateful to you, Associate Editor Josef Uyeda, and the two expert reviewers for the thoughtful and highly constructive feedback. I agree that the suggestions have substantially improved the quality, clarity, and impact of this work. I have addressed all comments in the revised manuscript and provide a point-by-point response below. Original comments are numbered and will appear in plain text. My responses will be below in bolded text.**

E.1. The main concerns are (1) the lack of discussion to the RJ-MCMC approach of Pagel, and (2) a better description of the models and algorithms that were newly developed. I absolutely agree with this assessment.

**A: I have taken steps to correct both problems. First, I have revised the introduction to include a discussion of the reversible-jump MCMC (RJMCMC) approach of Pagel and Meade (2006), contextualizing this work and highlighting the key similarities and differences in the methods section. Second, in response to this and other comments (see R1.1, R1.4), I have overhauled the core search algorithm. The manuscript now details a new simulated annealing (SA) approach for exploring model structures, which offers theoretical guarantees of finding a global optimum. The methods section now includes a detailed description of this new algorithm and a direct comparison to the mechanisms used in the RJMCMC framework.**

E.2. I also struggled a bit to fully understand each specific step in the dredge procedure. Maybe some more visual examples, also including the hidden states, could help.

**A: Agreed, I have addressed this in 3 ways. First, I have included a simple conceptual figure (Fig. 2) which displays a simple version of how corHMMDredge will propose moves between model structures. Second, I’ve included detailed pseudocode in the supplemental materials for readers interested in the exact details of the process. Finally, my empirical example, now includes a visualization of the exact dredge process that proceeded for the model evaluation (Fig. 4c). This includes an examination of a two-rate class model (hidden states).**   
  
Some minor comments from my own reading:  
E.4: Please check in the entire manuscript that in-line equations and math symbols have spaces around them.  
**A: Apologies for the formatting errors. This was an artifact of converting from an open-text format to a .docx file. I have carefully gone through the manuscript and corrected the spacing around all in-line equations and mathematical symbols**

E.5: line 33-39: This could have more broader background and citations. Also, please define what you mean by complex characters.

**A: I meant multistate, rather than complex. I have changed to wording.**

E.6: line 73: Could you please consider adding the equation k=(2^s)\*(2^s-1), where s is the number of binary characters.

**A: For clarity, I have removed this section of the text (replaced by the simulated annealing description), so I have not included this equation.**

E.7: line 77: I would not say "likelihood-based" but instead Maximum Likelihood, as Bayesian inference is also likelihood-based and might show different behaviour.

**A: Corrected.**

E.8: line 111: Please cite the original pruning algorithm by Felsenstein

**A: Corrected.**

E.9: line 117-119: Please add citations.

**A: Corrected.**

E.10: line 166: Could you explain your approach to ensure the identifiability better? More specifics would help.

**A: Apologies, I have added some additional description and the relevant citation which describes the issue in detail.**

E.11: line 190-209: Perhaps you could briefly emphasize how your cross-validation approach connects and is different from this one <https://doi.org/10.1093/sysbio/syad004>

**A: Thank you for pointing out this interesting paper. I had not seen it. I have added a mention to the methods section. The cited paper is a far more in-depth discussion of cross-validation approaches to phylogenetics and demonstrates that cross-validation are more adequate than alternatives for selecting the best approximation of the data-generating process. This paper ends up focusing on leave-one-out-CV as it is asymptotically equivalent to wAIC. The major differences between this method and the one presented here are both my choice to focus on k-fold CV and the goals of this CV. One of the main goals of my implementation esd to preform cross-validation without having to randomly shuffle the phylogenetic data or remove any tips, hence the choice to essentially mask different sets of tips across the k-folds. It seems like there could be some interesting theoretical ties between wAIC and JS divergence (since JS divergence is based on Kullback–Leibler (KL) divergence and wAIC is also based on KL divergence), but that is well beyond the scope of this paper.**

E.12: Results: The Mk model represents the unconstrained model, i.e., the ARD model? Why not call it ARD? Also, in my understanding the Mk is the same as the ER model.

**A: You are correct, I have corrected this throughout the manuscript. I now refer to the unconstrained model as the "all-rates-different" (ARD) model and the single-rate model as the "equal-rates" (ER) model.**

E.13: Results: Since for rates the fold changes is more important than the actual change, please consider re-evaluating the simulations by computing the bias and variance on the the log-rates.

**A: This is an excellent point. While my original choice was made to parallel empirical settings where rates are often reported directly, log-rates are indeed more appropriate for evaluating bias and variance. I have now re-analyzed and presented the simulation results using both original units and log-rates.**

E.14: Simulations: Would it be possible to plot, e.g., a histogram, the number of transition across the phylogeny? I'm curios if the shown power has to do with the number of transitions.

**A: I agree that the number of transitions is a key factor influencing statistical power, however the simulation protocol that I used does not return the number of transitions or their location. This would need to be examined in a more detailed follow up study.**

E.15: Empirical Example: Could you provide confidence intervals for the parameter estimates?

**A: Added.**

Associate Editor Recommendation: Recommendation #1: Major Revision  
  
Associate Editor's comments to author (if any):  
Associate Editor: Uyeda, Josef  
Comments to the Author:  
Dear Dr. Boyko,  
Thank you for your submission "Automatic Discovery of Optimal Discrete Character Models". Your manuscript has been reviewed by two expert reviewers, and both find that the manuscript is very timely and likely of substantial interest to the readers of Systematic Biology. I concur with these assessments. Both reviewers also have provided extensive comments to improve the manuscript.

**A: Thank you.**

AE.1: Reviewer 1 points out the similarity of the approach to the Reversible-Jump approach developed by Pagel and Meade. No discussion is made of this approach or a Bayesian framework, despite the similarity in objectives.

**A: I thank you and Reviewer 1 for highlighting this omission. I have now thoroughly revised the manuscript to address this. The introduction now contextualizes this work by discussing the RJMCMC framework. Furthermore, the methods section now contains a direct comparison between the proposal mechanisms in the new simulated annealing algorithm and those used in the RJMCMC approach (e.g., merging/splitting rates). This provides a much clearer picture of how this framework relates to the existing Bayesian alternative (see also our responses E.1 and R1.2).**

AE.2: Furthermore, both reviewers ask for how the dredge approach works for a "negative control" of two independently evolving characters.

**A: This is a good point. The original simulation design for the correlated evolution test did actually include a negative control, but I had not explicitly reported the success rate for correctly identifying these independent relationships. Specifically, only half of the non-zero rate parameters represented correlated changes, while the other half represented independent transitions. I have now revised the results section to report the success rates for identifying both dependent and independent rate structures, demonstrating that the method performs well in both cases.**

AE.3: Line 156 - should be "within the regularized model are"

**A: Corrected.**  
  
AE.4: Line 265, 419, 422 - The order here for condition 3 doesn't make sense, and is in conflict with line 282. Should be q34 (=10) > q21 (=5) > qij (=1?); although it's not clear what qij refers to.

**A: The final test (q34 (=10) > q21 (=5) > qij (=1?)) was meant to determine if the model correctly inferred the relative rates with qij being all other rates not q34 or q21. This test has been removed in favor of focusing on the correlated trait dynamics (finding dependent and independent relationships).**

AE.5: Line 467 - "better behaved" and "clean likelihood peaks and reasonable confidence intervals". I see what is meant regarding information but it gives the impression we can evaluate models by the cleanliness of their peaks and the reasonableness of their CI's, but it's unclear to me exactly what model selection philosophy supports that, or how those can be quantitatively evaluated. I'm not sure that's the implication that's intended, but it's possibly what could be received by the reader. True models can presumably sometimes have wide confidence intervals and poorly behaved likelihood peaks. The answer is provided later in Line 484, but again no citations to indicate that this necessarily "demonstrates the value" of the approach. *How is this different from, for example, choosing the model in statistical phylogenetics that maximizes your bootstrap scores or posterior probabilities in your topology ?* This would be a very poor way of evaluating models, as we want our models to reflect our true uncertainty on our target of inference, not arbitrarily choose models that necessarily give us "cleaner" results.

**A: This is a great point. I did not intend to imply that model selection should be based on the aesthetics of the likelihood surface or the narrowness of confidence intervals. I agree that this is not a valid model selection philosophy. I have removed this section and rewritten later sections to be clear that one should not choose a model to arbitrarily obtain “cleaner” results. My revised discussion now focuses on the fact that by expanding the set of candidate models, our method can identify simpler, well-supported models whose parameters are more readily informed by the data (i.e., have higher Fisher information), without forgoing standard model selection criteria like AIC.**

AE.6: The use of "standard corHMM model" throughout is a bit confusing/misleading. The model used is a standard Pagel model for correlated traits implemented in corHMM, and does not contain hidden states. Nothing about a Pagel model is the standard approach for how corHMM models discrete traits is it? It's one of many models that can be implemented in the package.

**A: I have removed these phrases. I now refer to specific models by their conventional names, such as “all-rates-different” (ARD) or “equal-rates” (ER), to avoid ambiguity.**

AE.7: Figure 5 - It's kind of ambiguous if the lower right panel of profile likelihoods falls under the b) or d). I'd also say that it's somewhat hard to compare them because the x-axes are in strange and not common units. e.g. it takes a fair amount of work to compare Theta\_1 in a) and b) and Theta\_2 to Theta\_3 in a) and b).

**A: Good point. However, this figure has been removed in response to your previous comments (see AE.5).**

AE.8: Line 479 - Adding to Reviewer 1's flagging of this issue, this is an additional location in the text where an apparent conflict between what is stated and what is in Figure 4. It says there is a slight preference for multimale mating, but in the figure it says it's monogamy that is slightly preferred. Figure 4 shows monogamy being preferred.

**A: You are correct. This was an error in the text. However, this figure has now been removed because I no longer make a comparison to an arbitrary default model.**

AE.9: Line 497 - In light of the simulation study, the great uncertainty among models, and the very variable outcomes to ancestral state estimation, I think this discussion overinterprets the results. It is striking to say on the one hand that the dredged model changes the ancestral state reconstructions to a previously undescribed one but that is supporting a previous hypothesis, discuss its implications at length, and then in the next section, point out that regularization doesn't improve ancestral state estimation. Are readers to cite this study as providing evidence for one hypothesis of primate evolution over another, or should we simply view any ancestral state reconstruction with this dataset with great skepticism?

**A: This is a fair critique. I agree that I have overinterpreted the ancestral state reconstruction (ASR) results from a single empirical case. The primary message should be about the importance of model selection, not about a definitive answer to the primate mating system question. This point is further underscored by my new results, which include hidden states and identify a different, strongly supported model. I have substantially toned down the discussion, reframing it to emphasize how sensitive ASRs can be to the choice of model and that readers should indeed be skeptical of any single reconstruction from this single dataset.**  
  
AE.10: Line 525 - "state reconstructions" not "state reconstructs"

**A: Corrected.**

AE.11: Line 524-530 - This discussion is not clear to me. That the marginal reconstructions can change across slight differences in models is one thing, but it does not speak to accuracy like your simulation results do, because we don't know the truth there. However, you're not comparing the same thing in this discussion. On the one hand you're talking about the consistency of the reconstructions across alternative models in the empirical study, while in the other you're discussing accuracy of reconstructions in the simulations. It's not clear to me what you're saying in Line 527. For example, suppose that in your simulation study your Mk model compared to your L2 model both recover ancestral states at the same rate under a particular set of conditions. It's unclear to me what this number would mean for how often they differed in their reconstructions at a set of key nodes that change biological interpretations. It could easily be quite common, I would think. The fact is the empirical study is one realization. Overall, I'm just confused on the relationship between accuracy and "robustness" of ancestral estimations that is being implied. I may be being dense here.

**A: You are right to point out the flawed logic. I was indeed conflating the consistency of reconstructions (across models in the empirical case) with the accuracy of reconstructions (tested in simulation). This study was not designed to test the relationship between these two concepts. As such, I have removed this section from the discussion. I have retained a more modest and speculative version of the point as a potential avenue for future research, rather than presenting it as a finding of this study.**

AE.12: Line 556 - Shouldn't this be "homology over homoplasy"?

**A: Yes, thank you for pointing that out.**

AE.13: Line 584-586 - Is this referring just to hidden state models? Otherwise I don't know what rate classes mean in this context.

**A: Yes, this referred to hidden state models.**    
  
AE.14: Figure S2 - Related to the discussion above about lines 524-530, I don't think I understand the procedure for Figure S2. From which models 95% CI are these high rates sampled? Jointly or independently for each transition rate? Or is it for the corHMMDredge model? I think the discussion of increasing rates is worth discussing within the context of phylogenetic signal. If all transition rates are arbitrarily large, the assumption is that the process has no phylogenetic signal and all nodes should be a their stationary distribution. This is usually what I assume leads to likelihood ridges (or as stated, low frequency states). However, that's clearly not the case for Figure S2, or the empirical example generally. Is it the case for the simulation results? Should this be factored in to how they are summarized? Presumably, across complicated model structures, many equivalent stationary distributions could be obtained with a variety of high transition rates configurations.

**A: Apologies for the lack of clarity. As you and other reviewers have pointed out, I over-interpreted this result, and the simulation study was not designed to properly test this question. Given these concerns, and to streamline the manuscript, I have removed this figure and the associated discussion from the paper.**

Reviewer(s)' comments to author (if any):  
Reviewer: 1  
  
Comments to the Author  
Automatic discovery of optimal discrete character models  
Systematic Biology  
Review, January 2025  
  
Reviewer expertise: phylogenetic comparative methods, discrete character models, vertebrate evolution  
  
Evolutionary models for discrete (or categorical) characters are ubiquitous in evolutionary biology, anthropology, linguistics, and beyond; however, as these models increase in complexity, so do the number of potential models to explain a given dataset. The author presents a dredge framework for automatically detecting the optimal model structure that balances model fit (likelihood) and complexity (number of parameters). The framework uses a regularization scheme to penalize more complex models and a parameter sharing method for equating and reducing the number of rate parameters. The study conducts simulations to compare the performance of the regularized models over traditional ones, and how often the automatic model selection recovers the models used to simulate the data. Then, the author applies this framework to the study of primate mating systems and compares their results with those of previous studies.  
  
R1.1: Overall, the dredge framework is a welcome and complementary contribution to modeling the evolution of discrete characters. I particularly appreciate the integration of this framework into the corHMM R package to include hidden Markov models and >2 binary characters. I agree with the author's goal of automating the search for a wider set of potential models while balancing model fit and complexity. I think this is a promising approach. However, I'm concerned about the dredge procedure's ability to explore a wide enough set of potential models. The current manuscript also lacks comparisons with the reversible-jump Markov chain Monte Carlo procedure described by Pagel and Meade (2006), which attempts to accomplish the same goal. And, in some ways, I would argue the RJMCMC procedure is better equipped to do this than the current implementation of the dredge procedure.

**A: I thank the reviewer for this insightful comment and agree that a robust model exploration strategy is critical. In response, I have fundamentally revised the dredge procedure, replacing the original stepwise search with a simulated annealing (SA) algorithm. This new approach is directly inspired by the logic of Pagel and Meade (2006) and is much better equipped to explore a wide range of model structures. For a discrete problem like model selection, SA is theoretically guaranteed to find the globally optimal model given sufficient time. This provides a robust search of the model space while retaining the speed of a maximum likelihood framework and generalizability of the original implementation. The manuscript has been updated to detail this.**

**Major concerns**  
R1.2 One of my main concerns is that, while referencing Pagel and Meade (2006) many times in the manuscript, and even using its data, the author refrains from mentioning that Pagel and Meade (2006) already devised an approach for automatically exploring different model structures while accounting for model complexity. There are a few details that distinguish Pagel and Meade’s model from the proposed dredge framework: 1) it is implemented in a Bayesian framework using reversible-jump Markov chain Monte Carlo (RJMCMC); 2) it doesn’t settle on one optimal model but quantifies the frequency of accepted models proposed throughout the MCMC; and 3) that model currently doesn’t allow for >2 binary characters or HMM. Despite these key differences, I found the lack of comparisons or discussion of the RJMCMC procedure surprising given that it seems to accomplish what the author has set out to do. This manuscript would be much improved by adding that context and some comparisons with Pagel and Meade’s RJMCMC procedure. For example, how does the author's parameter sharing procedure compare with the merging/splitting and augmenting/reducing operations of Pagel and Meade’s RJMCMC model?

**A: The reviewer is correct, and I thank them for pointing out this oversight. The original manuscript was focused on the problem of using fixed, standard model sets, and I failed to adequately connect this proposed solution to the work of Pagel and Meade (2006). I have now revised the introduction and methods sections to address this (see also response to A.1). The text now explicitly discusses their RJMCMC framework. When I introduce the SA-based dredge, I point out that it performs a similar heuristic search for optimal model structures but within a maximum likelihood context. I also include a direct comparison of the specific model proposal mechanisms, aligning the moves of corHMMDredge with the moves of the RJMCMC.**   
  
R1.3 The two approaches (the author’s dredge framework, and Pagel and Meade’s RJMCMC) also seem to differ in modeling philosophy. The author states the dredge approach is more parsimonious (lines 554-556 of PDF document), whereas Pagel and Meade’s RJMCMC uses a Bayesian framework to assess the frequency and probability of competing models. In my opinion, a parsimony-inspired framework seems antithetical to a model-based approach to study evolution. However, regardless of opinion on which framework is best to use, I think it would make this manuscript more holistic if the author discussed and compared these two approaches and philosophies.

**A: I agree that there is a philosophical difference between this approach and the Bayesian framework. To clarify, my preference for parsimony is for statistical parsimony, as embodied by information criteria like AIC, which balances model fit against complexity.**

**I acknowledge that the RJMCMC approach excels at model averaging by directly integrating over uncertainty to quantify the posterior probability of various models. The SA-based dredge, in contrast, is designed for model selection and comparison. Its goal is to efficiently identify a set of high-quality candidate models from a vast search space. Furthermore, I absolutely agree with the reviewer that examining a set of well-supported models is crucial for robust inference. However, applying formal model averaging (whether Bayesian or frequentist via AIC weights) to hidden Markov models is complicated by the “label switching” problem. Because the likelihood remains identical if the labels of two hidden rate classes are swapped, it becomes difficult to average parameters meaningfully across models. Considering this, the dredge framework identifies and returns all unique, well-supported models, allowing researchers to assess the robustness of their conclusions by comparing interpretations across this set.**

**I have now added a discussion to the manuscript that better compares the model selection and model averaging philosophies, clarifies the intended role of our dredge framework, and addresses the practical challenge of label switching in the Choosing the “best” model section.**  
  
R1.4: I’m also concerned that the author’s parameter sharing procedure doesn’t adequately explore the suite of potential models. Namely, the procedure only moves in one direction toward reducing the model and, therefore, does not evaluate other combinations of rate parameters that do not directly reduce from the initial model. For example, other simple models may fit the Old World primate data just as well as the one chosen through the dredge procedure (Figure 5c), but the rate parameters of those models were removed from consideration because they were estimated as close to 0 in the initial model. In contrast, Pagel and Meade’s RJMCMC allows rate parameters to either split after previously being equated or to be allowed back into the model (augment) after being reduced or removed from the model (i.e., set to 0). Indeed, Pagel and Meade (2006) find that there are many simple models with two or fewer non-zero rate classes that equally explain the primate data. Many of these models overlap with the model chosen by the author’s dredge procedure (if you compare Figure 5c with Table 2 in Pagel and Meade 2006). Of course, Bayesian RJMCMC is more computationally intensive than the author’s dredge framework; however, if the author’s intended goal is to explore a larger model space while overcoming problems of over-parameterization, I’m afraid the dredge framework in its current implementation may not accomplish this as thoroughly as RJMCMC.

**A: The reviewer’s point is well-taken. The original implementation did indeed lack the ability to “climb back up” in model complexity, which was a significant limitation. This concern was a primary motivation for re-implementing the entire search using simulated annealing (SA). The SA procedure explicitly incorporates three types of moves: 'drop' and 'merge' to simplify the model, and a 'free' move to increase model complexity. This 'free' move serves the same purpose as the 'augmenting' and 'splitting' proposals in Pagel and Meade (2006), ensuring the search is not unidirectional. The probabilistic nature of SA, which allows it to accept temporary increases in model complexity, combined with these bidirectional moves, ensures a much more exhaustive exploration of the model landscape. This directly addresses the reviewer’s concern and makes the dredge framework a far more robust tool for model discovery.**

R1.5: Lastly, I appreciate the author’s use of simulations to demonstrate the effectiveness of the dredge framework. However, for the second simulation study, I recommend the author include a negative control group, where the author simulates two randomly evolving characters independently and tests how often the automatic model selection procedure selects an independent model.

**A: Thank you for this recommendation. I agree that including a negative control is essential for demonstrating the method’s specificity. In fact, the original simulation design for the correlated evolution test implicitly included a negative control: only half of the non-zero rate parameters represented correlated changes, while the other half represented independent transitions. However, in the original submission, I had not reported on the success rate of correctly identifying these independent relationships. I have now revised the manuscript to explicitly report these results. The analysis shows that this method is indeed successful at recovering both the true dependencies and the true independencies, which I believe addresses the reviewer's concern without the need for an entirely new simulation.**

R1.6: I found the change in the use of the mapping matrix and Q matrix to be a bit confusing. For example, when the author described the correlated model (lines 258-283), they predominately used the notation for the rate’s location in the Q matrix (e.g., q1,2) without providing the same notation in the matrix on line 284. It might help the reader to give the location notation in the same matrix or state very clearly that q1,2 is on the first row and second column. Ideally, they could have both notations or matrices side-by-side so the reader doesn’t have to count or remember where the rates are located.

**A: This is a helpful suggestion for improving clarity. I have removed the rate notation (e.g., q1,2). I now present the Q-matrix with the transition type notation next to the simulating rate values, so the reader does not have to count rows and columns.**

R1.7: Figure 4: The standard model in the figure does not seem to match the standard model described in the text (lines 455-458). In the figure, the standard corHMM slightly prefers a root state with monogamy and estrus display (dark blue). However, the results describe a model with a 61% marginal probability of the root state having multimale mating and estrus display (lines 455-458). Is this the wrong figure?

**A: Thank you for catching this error. The text was indeed incorrect and did not match the figure. However, this figure has been removed in line with other revisions.**

R1.8: Lines 27, 341, 351, 355, 475, 488, and 524: Inconsistent hyphenation of Old World monkeys (or primates) throughout the text. I don’t think it needs to be hyphenated.

**A: Thank you, corrected.**   
  
R1.9: Line 50: Please remove the apostrophe after ‘models’

**A: Corrected.**  
  
R1.10: Line 63: Replace ‘by’ with ‘as’

**A: Corrected.**  
  
R1.11: Line 94: Add a space between ‘model’ and before the start of the parentheses

**A: Corrected.**  
  
R1.12: Line 154: Spaces are needed around ‘C’ and ‘S’

**A: Corrected.**  
  
R1.13: Line 155: Please define Lreg. Is this the likelihood for any of the regularized models?

A: **A definition has been added. I also point the readers to the equations where Lreg is actually different for the different regularization schemes.**  
  
R1.14: Line 169: Space is needed after ‘AIC’

**A: Corrected.**  
R1.15: Figure 2: Please define the shorthand ‘lnLik’ and ‘dAIC’ in the figure caption.  
**A: This figure has been replaced in response to other comments which desired a clearer description of the dredge algorithm. I have kept this comment in mind for all abbreviations in figure captions.**

R1.16: Figure 2: “(c) Based on model (c)" -- I think the first ‘(c)’ should be ‘(d)’.

**A: See response to R1.15**  
  
R1.17: Line 203: Add ‘is’ after “because it”

**A: Corrected.**  
R1.18: Line 260: Add a period and space after 1  
**A: Corrected.**

R1.19: Line 260: Add a comma and space after ‘e.g.’

**A: Corrected.**  
  
R1.20: Line 262: Add an ‘=’ sign after ‘birth rate’

**A: Corrected.**  
  
R1.21: Line 267: Remove the comma after ‘character’

**A: Corrected.**  
  
R1.22: Lines 269-270: Spaces may be needed before or after each of the Xs and Ys. There may be some formatting issues with the equations in Word.

**A: Yes, there were. I have corrected all the formatting errors.**   
  
R1.23: Line 289: Remove the comma after ‘(1)’

**A: Corrected.**  
  
R1.24: Line 290: Add a comma and space before ‘(2)’

**A: Corrected.**  
  
R1.25: Lines 317-318: Test #4 is misplaced in the next sentence on line 318.

**A: Corrected.**  
  
R1.26: Line 324: Make ‘transition’ plural

**A: Corrected.**  
  
R1.27: Line 332: Is JS Divergence a function in corHMM?

**A: Yes, I have added it for this manuscript. However, it is currently not exported. It can be accessed through corHMM:::** **js\_divergence and it uses corHMM:::kl\_divergence to calculate the divergence. If the reviewer believes it would be of general interest, I’d be happy to export the function and add documentation.**   
  
R1.28: Line 451: Are these the “default settings” in corHMM? Please remind the reader that this is an all rates different model. I recommend that the author continues referring to this as the 'all rates different model' because other programs or packages may have different default settings.

**A: Yes, this was pointed out in other comments as well. I have removed all mentions of “default corHMM settings” or “corHMM models” and have replaced them with the common model names (all rates different, equal rates, etc.)**  
  
R1.29: Lines 451-452: I’m confused by the AIC comparison here. Why is this AIC used to support regularization if the AIC score was from a comparison between two unregularized models?

**A: Good point, I have removed the statement. I had wanted to demonstrate that there are “better” models beyond the typical set tested (ER, SYM, ARD) and hence running dredge without regularization (for valid AIC comparisons). But you are right to point out that this is confusing, and I don’t believe it is necessary. This section as has also changed substantially with new empirical results where I tested an additional rate class.**   
  
R1.30: Line 453: Remove ‘the’ after ‘estimated’

**A: Corrected.**   
  
R1.31: Line 477: Again, I recommend that the author refers to this model as the 'all rates different model' even though it is the default setting for this particular R package.

**A: Agreed, corrected throughout the manuscript.**   
  
R1.32: Figure S1: I agree that there are issues of non-independence here, and I’m glad the author acknowledges this. I also recognize that the author is trying not to make a big deal out of these results (because of the non-independence issue). However, it is very difficult to see the points in this plot, and the author doesn’t provide any slopes, p-values, or R2 values, which makes it difficult to believe even their qualitative claims of an increase in predictive performance toward the root.

**A: Given both of our concerns of the validity of these statistical tests and the fact that the simulation design was not intended to test performance of ancestral state estimation, I have decided to remove this from the manuscript all together.**   
  
R1.33: Figure S2: I assume the left is the upper 95% CI, but is this for the standard all rates different model or from a regularized model?

**A: This is from the standard all rates different model. The AE pointed out the lack of clarity in this figure and it was removed from this submission due to concerns about readers conflating “good” model behavior as a criteria of model selection.**

Reviewer: 2  
  
Comments to the Author  
This manuscript is an interesting and timely contribution. When analyzing discrete traits, the number of potential evolutionary models (Q matrices) is vast, which makes navigating the model space challenging. Comparative phylogenetics has largely lacked robust methods for selecting the optimal model from this pool. To my knowledge, this is the first study that attempts to address this problem using maximum likelihood framework.  
  
The author  proposes using regularization techniques (lasso and ridge), which have been widely employed in other statistical fields, to penalize parameters in the Q matrices in order to identify the optimal model. Additionally, this appears to be the first application of these techniques to Q matrices and CTMCs. Similar approaches exist in Bayesian frameworks using a horseshoe prior, which shares similarities with the lasso method. Consequently, the study is novel and will be of high interest to the readership of Systematic Biology, particularly those working with discrete traits.  
  
Overall, the manuscript is well written; however, I believe it would benefit from additional details and discussion in some areas. I recommend its publication in Systematic Biology once these points are addressed, placing my recommendation between minor and major revision.  
**A: Thank you.**  
  
R2.1: Clarification on the Dredge Algorithm (L152). I am a bit lost of how the dredge algorithm functions, particularly how the regularization component is linked to model exploration. From my understanding, the algorithm calculates the penalized likelihood for a broader model, then merges parameters and estimates the merged model. Could the authors clarify whether the algorithm compares the wider model with the merged model using AIC or only the penalized log-likelihood? Additionally, how is AIC calculated for a penalized model—does it include the lambda parameter? It would be helpful to know the criteria for merging and dropping parameters; and an estimate of how many models the dredge algorithm explores on average for a single dataset.

**A: I apologize that the original description was unclear. As detailed in our responses to other reviewers (e.g., R1.1), I have completely revised the search algorithm to use simulated annealing. The new methods section, along with a new conceptual figure (Fig. 2) and supplementary pseudocode, provides a much more detailed explanation. To answer your specific questions in the context of the new algorithm: the SA search uses the penalized log-likelihood to navigate the model space. After the search is complete, the best-fitting model structures are refit with penalization to obtain regularized likelihoods and AIC scores for final comparison. The AIC is calculated in the standard way (2k - 2lnL) on the final, penalized model.**  
     
R2.2: I would appreciate more practical tips for using cross-validation. For example, does the user need to repeat the entire dredge algorithm with varying values of the lambda parameter and then select the set that minimizes the chosen statistic (e.g., JS)? A separate section briefly describing the best practices for the dredge algorithm  would be very useful.

**A: This is an excellent suggestion. I have added details on the cross-validation procedure and recommendations for empiricists. I have also included an example of cross-validation for the empirical rest cases (Table 3).**   
     
R3.3: Dredge and Hypothesis Testing. Correct me if I am mistaken, but it appears that dredge explores the model space by sequentially penalizing models and eventually selects the one with the best AIC. While this approach is reasonable, there is no guarantee—either theoretically or practically—that the dredge algorithm will converge on the true model, as there are no simulations presented to show the convergence rate. It seems that the decision is made based on parameter estimates that reduce variance, essentially dredge is a mix of approaches: AIC + parameter estimations. I would welcome a more detailed discussion of these issues. For instance, if dredge selects model M1 as the best model, but a user identifies an alternative model M2 (which was not explored by dredge) that has a better AIC, which model should be chosen? What practical considerations should guide this decision? Can dredge mislead hypothesis testing?

**A: The reviewer is correct that the original stepwise procedure offered no theoretical guarantee of finding the global optimum. This was a significant concern that, in part, motivated the decision to overhaul the search algorithm. In the revised manuscript, I now employ a simulated annealing (SA) search. For a discrete problem like model selection, SA has theoretical guarantees of converging on the globally optimal solution, given a sufficiently slow cooling schedule. I now discuss this property in the methods section. To the reviewer’s point about hypothesis testing, I have added the discussion a warning that corHMMDredge is a model exploration and selection tool, intended to complement, not replace, hypothesis-driven science. If a user has a priori biological hypotheses, there is no need for that set to be compared to the dredge-selected models.**

R2.4: Furthermore, in L. 605, the author states "In a dredge run, one is not testing specific hypotheses but rather comparing parameter estimates across" . Does it mean that the main values of dredge is parameter estimation rather than hypothesis testing? To me, it seems like it is more for parameter estimation. I think it is important for users to know it before hand. Consider adding an explanation to Introduction that clarifies the usage of dredge.

**A: Thank you for asking for clarification on this point. The original wording was ambiguous. I have revised the introduction and discussion to be clearer: corHMMDredge is primarily a tool for model selection and model exploration. While accurate parameter estimation is a downstream benefit of finding a better-fitting model, its main purpose is to navigate the vast space of potential models to identify the best candidates for formal hypothesis testing and comparison.**

R2.5:  It would be helpful to include a comparison of AIC scores between dredge-selected models and non-dredge models in Tables 1–2.

**A: There was no model selection done for Table 1. However, I have included the model selection table for the empirical case study in the supplemental material.**   
     
R2.6: In the simulations, the author demonstrated the performance of dredge for correlated models. It would be interesting to see how the dredge algorithm performs when the generative model is independent but contains several  parameters. Will dredge converge in identifying the independent model?

**A: This is a good point, and similar to one raised by R1.5. As I noted in that response, the main simulation study already included independent evolutionary processes. I have now revised the results to explicitly report the high success rate of our method in correctly identifying these true independent relationships, demonstrating that it does not falsely infer correlations.**  
     
R2.7: In Figures 4 and 5, dredge selects a model where the state X1Y0 is inaccessible (resulting essentially in a three-state model as shown in Fig. 5c), which is logical because this state is absent among the tips of the tree in Fig. 4. Generally, as a rule of thumb, in CTMCs, a model with an unobserved state always has a worse  likelihood than a model where that state is removed. Thus, to a certain extent, the performance of the models in Figs. 5c and 5d is deterministic. While it is encouraging that dredge can identify such cases—given that many users may not be familiar with this property—it would be interesting to see a discussion on this topic. How should such "deterministic" correlations, indicated by zero transitions, be treated in practice as such cases might be common?

**A: This is an interesting point. The reviewer is correct that if a state is entirely unobserved at the tips, models that make this state inaccessible will almost certainly have a better likelihood. I find it encouraging that the dredge algorithm can automatically identify this. However, the biological interpretation of such a finding is a complex issue. Whether it reflects a true evolutionary constraint versus a sampling artifact is tough to say. I am pleased that corHMMDredge helps uncover these patterns, but it does not in itself resolve their interpretation.**

R2.8. Unfortunately, I was unable to locate any code associated with the analyses presented in the paper despite following the links provided. I encourage the author to publish an R tutorial that explains how to implement the dredge algorithm.

**A: I apologize for this. The code and a detailed R vignette are included in the revised submission as supplementary materials and will be made available on a public repository upon acceptance. The vignette has been polished and provides a step-by-step guide to implementing the methods described in the paper.**

R2.9: (L127–128, L145) It would be helpful for readers to have more details about the mechanism of the regularization technique. In particular, please explain how the method penalizes models and provide an intuitive interpretation of the lambda parameter.

**A: Thank you for the suggestion. I now clarify that regularization works by adding a penalty term to the likelihood function. The lambda parameter controls the strength of this penalty: a lambda of zero means no penalty (standard ML), while a very large lambda will force parameters towards zero.**

R.10: (L89)  Should this be “corHMMDredge function”?

**A: Yes, thank you. Corrected.**

R.11: (L134)  What does “k” represent in the formula? Please clarify the notation.

**A: The number of freely estimated parameters. Added to main text.**

R.12: (L155) Please explain what is meant by Lreg​.

**A: Corrected.**

R.13: (L200) Does the term “observed likelihood” refer to the character states at the tips (e.g., vectors such as (1,0) or (0,1) for a binary character? Clarification here would be appreciated.

**A: Yes, that’s correct. Though for an HMM the vector could be (0.5,0,0.5,0)/(0,0.5,0,0.5) and for unknown states it would be (0.5,0.5).**

R.14: (L203)   it a symmetric” should be  “it is a symmetric.”

**A: Corrected.**

R.15: (L209)  Please explain the terms in the equation for "w=…"

**A: It is inversely proportional to the amount of phylogenetic similarity, so for the text I have described it as “phylogenetic uniqueness”.**

R.16: (L276)  Consider providing a Q matrix that defines terms such as q34​, as these terms are not introduced  elsewhere.

**A: Reviewer 1 made a similar point, so I have provided a point of comparison as suggested in R1.6.**

R.17: (L282)   “whether the the” remove “the.”

**A: Corrected.**

R.18. (L62)  The Mk model is typically understood to refer to a morphological model with a single rate parameter (similar to the Jukes-Cantor model) used for tree inference. Using “Mk” in this context may be confusing. I recommend renaming  to e.g.,  Pagel’s model or a correlated model.

**A: Yes, thanks for point this out. Corrected.**

R.19: (L378)  When referring to a “reduction in variance,”  clarify which variance is being discussed (e.g., total variance or another specific type).

**A: Excellent point, corrected.**

R.20: (L395)  Define the “absolute average bias.”

**A: Corrected.**

R.21: (L428)  Explain what the values in the right-hand columns of Table 2 represent.

**A: Thanks for pointing this out! Corrected.**