(2021), **0**, 0, *pp.* 1–11

Markov models provide many solutions for appropriate modeling of inapplicable morphological characters in phylogenetics

Sergei Tarasov

Finnish Museum of Natural History, Pohjoinen Rautatiekatu 13, FI-00014 Helsinki, Finland

Email: sergei.tarasov@helsinki.fi

Abstract

This paper describes new models for coding inapplicable characters for phylogenetic analysis with morphological data. I show that these new solutions fall into two categories that require use of different Markov models. Their implementation in the Bayesian framework using popular software RevBayes (Höhna et al., 2016) is provided. A combination of the new models and previous solutions (Tarasov, 2019) is capable to model all main type of dependencies between morphological characters.

Key words: Markov models, RevBayes, likelihood, Bayesian, inapplicable characters, parsimony.

Introduction

Recent publications have revitalized interest to the long-standing problem of modeling inapplicable characters for phylogenetic analysis with morphological data (Maddison, 1993; Hawkins et al., 1997). They offer new approaches for parsimony (Brazeau et al., 2019; Hopkins and St John, 2021; Goloboff et al., 2021) and maximum

likelihood (ML) methods (Tarasov, 2019). The inapplicability refers to the hierarchical dependency of traits due to anatomical dependency of the respective body parts. For example, tail color is absent (i.e., inapplicable) if a tail is absent but can be blue or red if the tail is present. Evolutionary this means that color trait evolves only within the certain limits of the controlling character – when the tail is present (hereafter called "embedded dependency" [ED]). The study of Goloboff et al. (2021) has criticized the ML solution using structured and hidden Markov models (SMM+HMM) in Tarasov (2019) for inappropriately treating ED. Additionally, Goloboff et al. (2021) claimed that ED would require complex computations under ML inference because it cannot be expressed with infinitesimal rate matrix (Q) that is used for specifying Markov model of character evolution. I agree with Goloboff et al. (2021) that SMM+HMM do not explicitly imply ED but disagree that they treat ED inappropriately; SMM+HMM can model various dependencies and behave appropriately with respect to ED as was shown in the simulations and analytical derivations (Tarasov, 2019). The explicit ED solution was not reviewed in that paper due to its focus on HMM.

Given the recent interest, it is the best time as ever to provide the explicit ED model as another solution for the inapplicable characters. Below, I demonstrate that it can be straightforwardly constructed using the techniques of SMMs for character amalgamation and lumpability given in Tarasov (2019). These techniques are central to modeling morphological characters since they equip them with a mathematically consistent closure property (Tarasov, 2020) – characters can be added together (i.e., amalgamated) to yield another character and a character can be divided by the state aggregation to yield one or few characters. The provided ED model comes in two flavours depending on the type of character hierarchy. It can be represented using single rate matrix Q and does not need expensive computations as mentioned in Goloboff et al. (2021). I give practical example of using it for tree inference in RevBayes (Höhna et al., 2016) and assess its performance with simulations.

Deriving Embedded Dependency (ED) Model using the Techniques of Structured Markov Models

In the derivations below, I begin with the general ED model and then consider its parametrization for tree inference. There are two types of ED model: (i) *qualitative (ql)* type when all states of a qualitative character (as e.g., shape or color) depend on an absence/presence character; and (ii) *absence-presence (ap)* type when absence/presence of one character depends on the absence/presence of another. These two types require different Markov models that can be derived using the SMM techniques of lumpability, state aggregation and amalgamation (for details see Tarasov (2019)). I evaluate these models in the context of the tail color and tail armor cases that were investigated by me earlier (Tarasov, 2019). In the mathematical notations below, bold font is used for vectors, while capitals are for matrices.

General ED model

Necessary conditions.— Following the tail color example (Maddison, 1993), there are two characters: T — tail presence $\{absent(a), present(p)\}$, and C — tail color $\{red(r), blue(b)\}$. In the embedded dependency (Fig. 1A), the dependent character C, is allowed to evolve only within the time interval when the controlling character T is in the state p; if the controlling character is in the state a then C is absent too. This implies the following conditions for ED model to hold: 1) T is independent of C, 2) C is dependent on T, 3) the probability of observing r or b state depends only on the last time interval of having the tail present but not the previous tail occurrences (Fig. 1A).

Qualitative Type: Tail Color Case.— Suppose that C and T evolve under the embedded dependency and according to the initial probability vectors $\boldsymbol{\pi_T} = (\pi_{T_1}, \pi_{T_2})$ and $\boldsymbol{\pi_C} = (\pi_{C_1}, \pi_{C_2})$ and these rate matrices:

$$T = \begin{array}{cc} a & p \\ \end{array} \begin{pmatrix} - & \alpha \\ \beta & - \end{pmatrix} \begin{array}{c} a \\ p \end{array}, C = \begin{array}{cc} r & b \\ \end{array} \begin{pmatrix} - & \gamma_1 \\ \gamma_2 & - \end{pmatrix} \begin{array}{c} r \\ b \end{array}.$$

The following rate matrix with three states $\{a, r, b\}$ fully characterizes the general ED model for the tail color case:

$$A_{ql}(T, C) = Q_{ql} = \begin{array}{ccc} a & r & b \\ \end{array} \begin{pmatrix} - & \lambda_1 & \lambda_2 \\ \beta & - & \gamma_1 \\ \beta & \gamma_2 & - \end{pmatrix} \begin{array}{c} a \\ r \\ b \end{array}.$$

This matrix has the initial vector $\boldsymbol{\pi_{eq}} = (1 - \pi_{T_2}, \pi_{T_2}\pi_{C_1}, \pi_{T_2}\pi_{C_2})$. Its rates λ_1 and λ_2 include the transition rates between the states a and p in T and the initial probabilities of C, which are $\lambda_1 = \pi_{C_1}\alpha$, $\lambda_2 = \pi_{C_2}\alpha$. It was obtained from T and C matrices by amalgamating them as a qualitative type $A_{ql}()$ of ED.

Note, Q_{ql} fulfils all three necessary conditions given above. It is lumpable with respect to the aggregation of states $\{a, \{r, b\}\}$ which yields the matrix T and indicates the independence of T from C (condition 1). In turn, C has no lumpable aggregations indicating its dependency on T (condition 2). The fulfilment to the condition 3 is given in the Appendix.

Absence-Presence Type: Tail Armor Case.— Consider the tail armor case, in which there are four species that possess a tail with blue or red armor, a tail without armor, and no tail. It can be characterized by the following three rate matrices that are similar to the tail color case but where A indicates absence/presence of the armor:

$$T = \begin{array}{cc} a & p \\ \end{array} \begin{pmatrix} - & \alpha_1 \\ \beta_1 & - \end{pmatrix} \begin{array}{c} a \\ p \end{array}, A = \begin{array}{cc} a & p \\ \end{array} \begin{pmatrix} - & \alpha_2 \\ \beta_2 & - \end{pmatrix} \begin{array}{c} a \\ p \end{array}, C = \begin{array}{cc} r & b \\ \end{array} \begin{pmatrix} - & \gamma_1 \\ \gamma_2 & - \end{pmatrix} \begin{array}{c} r \\ b \end{array}.$$

It is logical to assume that in the amalgamated matrix for T and A, a change from "no tail" a_1a_2 to "armor and tail present" p_1p_2 should go though a stepping-stone transition

"tail present, no armor" p_1a_2, to allow gradual grow of the hierarchical complexity and permit only one change during an infinitesimal time interval. This absence-presence type of the amalgamation $A_{ap}()$ differs from the qualitative one and yields the following rate matrix for T and A:

$$A_{ap}(T,A) = Q_{ap} = \begin{array}{c} \begin{array}{ccc} a_1a_2 & p_1a_2 & p_1p_2 \end{array} \\ \begin{pmatrix} - & \alpha_1 & 0 \\ \beta_1 & - & \alpha_2 \\ \beta_1 & \beta_2 & - \end{pmatrix} \end{array} \begin{array}{c} \\ a_1a_2 \\ p_1a_2 \\ p_1p_2 \end{array} . \tag{0.1}$$

Note, it also fulfils the three necessary conditions given above. This matrix relates to the qualitative type through $\lambda_1 = \alpha$, $\lambda_2 = 0$; it penalizes the direct change from a_1a_2 to p_1p_2. The dimensionality of $A_{ap}()$ grows for binary characters as $N(characters) + 1$.

To construct the rate matrix for their entire tail armor case, the qualitative and absence-presence amalgamations should be repeated successfully:

$$Q_{armor} = A_{ap}(T, A_{ql}(A,C)) = \begin{array}{c} \begin{array}{cccc} a_1a_2 & p_1a_2 & p_1r & p_1b \end{array} \\ \begin{pmatrix} - & \alpha_1 & 0 & 0 \\ \beta_1 & - & \alpha_2 & 0 \\ \beta_1 & \beta_2 & - & \gamma_1 \\ \beta_1 & \beta_2 & \gamma_2 & - \end{pmatrix} \end{array} \begin{array}{c} \\ a_1a_2 \\ p_1a_2 \\ p_1r \\ p_1b \end{array} .$$

Custom hierarchy.— The general equations for constructing the qualitative and absence-presence amalgamations are:

$$A_{ql}(Q_c, Q_d) = \begin{array}{c} \begin{array}{cc} a & p \end{array} \\ \begin{pmatrix} -\alpha & \boldsymbol{\pi_d}\alpha \\ \boldsymbol{\beta^T} & Q_d - I\beta \end{pmatrix} \end{array}, \quad A_{ap}(Q_c, Q_d) = \begin{array}{c} \begin{array}{cc} a & p \end{array} \\ \begin{pmatrix} \beta_c - \alpha_c & \boldsymbol{\hat{\alpha}_c} \\ \boldsymbol{\beta_c^T} & Q_d \end{pmatrix} \end{array} - I\beta_c,$$

where Q_c and Q_d are the controlling and dependent characters, I is identity matrix, $\boldsymbol{\beta}$ is a vector of β rates, superscript $\boldsymbol{T}$ is a matrix transpose and $\boldsymbol{\hat{\alpha}_c} = (\alpha_c, 0, \cdots, 0)$.

Note, the qualitative and absence-presence amalgamations are different from those given in Tarasov (2019) for modeling general cases of independent and correlated character evolution. However, the combinations of both should be used to construct complex

character hierarchies.

Tree inference with ED models

The general types of the ED models given in the previous section can be used in inferring ancestral character states on a known topology. However, if the topology itself is the focus of an analysis, the general ED can be straightforwardly adopted for phylogenetic inference by converting it into an Mk-type model (Lewis, 2001). This conversion is done by constraining its rate matrix to include one free parameter (that is usually interpreted as a branch length), setting its initial vector to contain the equilibrium distribution at the root, and conditioning its likelihood on observing only variable characters in the data.

For the tail color case this results in the conventional Mk-model with three states. For the tail armor case the conversion qives the follwoing model:

$$Mk_{armor} = \begin{array}{c} \begin{array}{cccc} a_1a_2 & p_1a_2 & p_1r & p_1b \end{array} \\ \begin{pmatrix} - & 1 & 0 & 0 \\ 1 & - & 1 & 0 \\ 1 & 1 & - & 1 \\ 1 & 1 & 1 & - \end{pmatrix} \end{array} \begin{array}{c} a_1a_2 \\ p_1a_2 \\ p_1r \\ p_1b \end{array} .$$

The initial vector with the equilibrium distribution for Mk_{armor} is $\boldsymbol{\pi}$=(12/24, 8/24, 3/24, 1/24). Note, Mk_{armor} is not a time-reversible model, which means that its likelihood should be calculated on a rooted tree (Felsenstein, 1981; Yang et al., 2006).

Materials and Methods

Inference

I implemented Mk_{armor} for the tail armor case in RevBayes (the scripts are available at: `https://github.com/sergeitarasov/MorphoModels`). Since this model is non-reversible and needs a rooted tree for likelihood calculation, I used the global molecular clock model of Zuckerkandl and Pauling (1962) with Yule process as a prior

distribution for the tree topology, with the age of the tree root set to 1.0. The inference was performed using a dataset with fifty identical characters referring to the tail armor case as described earlier in Tarasov (2019).

Results and Discussion

The expected topology for the tail armor case is shown in Fig. 1B. This topology was inferred with high posterior probabilities for all clades using Mk_{armor} model. Note, Mk_{armor} model was capable to correctly recover the expected root of the tree (i.e., the species with no tail).

The new ED models – absence-presence and qualitative ones – expand the arsenal of approaches for modeling evolution of morphological characters. Together with the other SMM models they can be used to potentially account for any type of character dependency. Any model for a complex character hierarchy can be constructed using the properties of character lumpability, state aggregation and amalgamation (Tarasov, 2019, 2020), which are crucial for modeling and analyzing morphological characters. I have showed that the complex hierarchies can be expressed using rate matrices and, thereby, do not require expensive computations as suggested previously (Goloboff et al., 2021). Following these receipts, any custom models can be straightforwardly implemented in RevBayes for statistical inference.

The ED models and other types of SMMs can be compared using statistical methods for model selection (e.g., Akaike information criterion, Bayes factor, etc.). Based on my recent experience, the SMM solutions for the tail color problem in (Tarasov, 2019) and the one given here $A_{ql}(T, C)$ may behave similar because the SMM solution may imply $A_{ql}(T, C)$ if the SMM one is weakly lumpable. Although, the weak lumpability (Tarasov, 2019) seem to not hold exactly in the likelihood computations, it apparently can be approximated with the high degree of precision. I also showed that the three-state Mk model represents an additional solution to the the tail color case. This is in contrast to my

previous conclusion (Tarasov, 2019) that should be now considered invalid. Thus, Markov models in the likelihood-based framework provide a natural way for modeling dependencies and offer different solution to the tail color and similar problems.

Appendix

Proving the Condition 3

The probability of observing r or b states depends only on the last time interval of having the tail present but not the previous tail occurrences.

I will derive the joint distribution of the controlling Q_c and dependent Q_d characters using the two approaches: (i) the conditional approach $Pr(\boldsymbol{\pi} Q_d, T \mid \vec{pp})$ that directly implies the condition 3; and (ii) the direct calculation using matrix exponential $\boldsymbol{\pi} e^{A_{ql}(T,C)}$. The equality of the two approaches would suggest that the condition 3 is true. For tractability suppose that Q_c is time-reversible:

$$Q_c = \begin{array}{cc} a & p \\ \end{array} \begin{pmatrix} - & \alpha \\ \alpha & - \end{pmatrix} \begin{array}{c} a \\ p \end{array} .$$

The probability of last p in time-reversible Q_c is hence exponential $Pr\Big(\vec{pp}(T-t) \mid N(T-t)=0\Big) = ae^{-a(T-t)}$. Herein, the notation $x_1\vec{x}_2$ indicates a transition from the state x_1 to the state x_2; $N(t)$ counts the number of transitions to different states. Two scenarios are possible:

$$\begin{aligned} Pr(\boldsymbol{\pi} Q_d, T \mid \vec{ap}) = & \int_0^T Pr\Big(\vec{aa}(t) \mid N(t) > 0\Big) Pr\Big(\vec{pp}(T-t) \mid N(T-t)=0\Big) \\ & \times Pr\Big(\boldsymbol{\pi} Q_d, T-t\Big) dt \\ = & \int_0^T \Big(\frac{1}{2} + \frac{1}{2} e^{-2at}\Big)(ae^{-a(T-t)})(\boldsymbol{\pi} e^{Q_d(T-t)}) dt, \end{aligned} \tag{0.2}$$

$$
\begin{aligned}
Pr(\boldsymbol{\pi} Q_d, T \mid \vec{pp}) &= Pr\Big(\vec{pp}(T) \mid N(t) = 0\Big) Pr\Big(\boldsymbol{\pi} Q_d, T\Big) \\
&\quad + \int_0^T Pr\Big(\vec{pa}(t) \mid N(t) > 0\Big) Pr\Big(\vec{pp}(T-t) \mid N(t) = 0\Big) \\
&\quad \times Pr\Big(\boldsymbol{\pi} Q_d, T-t\Big) dt \\
&= e^{-aT}\Big(\boldsymbol{\pi} e^{Q_d T}\Big) + \int_0^T \Big(\frac{1}{2} - \frac{1}{2} e^{-2at}\Big)(a e^{-a(T-t)})(\boldsymbol{\pi} e^{Q_d (T-t)}) dt.
\end{aligned} \tag{0.3}
$$

Now averaging over the initial vector:

$$
\begin{aligned}
Pr(Q_d, T \mid Q_c) &= \pi_a Pr(\boldsymbol{\pi} Q_d, T \mid \vec{ap}) + \pi_p Pr(\boldsymbol{\pi} Q_d, T \mid \vec{pp}) \\
&= \Big(\frac{(\gamma_1 - \gamma_2) e^{T(-(a+\gamma_1+\gamma_2))} + a + 2\gamma_2}{4(a+\gamma_1+\gamma_2)}, \frac{(\gamma_2 - \gamma_1) e^{T(-(a+\gamma_1+\gamma_2))} + a + 2\gamma_1}{4(a+\gamma_1+\gamma_2)}\Big).
\end{aligned} \tag{0.4}
$$

Note, $Pr(Q_d, T \mid Q_c)$ is the same as elements π_2 and π_3 in $\boldsymbol{\pi} e^{A_{ql}(T,C)}$ that proofs the aforementioned condition.

References

M. D. Brazeau, T. Guillerme, and M. R. Smith. An algorithm for morphological phylogenetic analysis with inapplicable data. *Systematic biology*, 68(4):619–631, 2019.

J. Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376, 1981.

P. A. Goloboff, J. De Laet, D. Ríos-Tamayo, and C. A. Szumik. A reconsideration of inapplicable characters, and an approximation with step-matrix recoding. *Cladistics*, 2021.

J. A. Hawkins, C. E. Hughes, and R. W. Scotland. Primary homology assessment, characters and character states. *Cladistics*, 13(3):275–283, 1997.

S. Höhna, M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist. Revbayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology*, 65 (4):726–736, 2016.

M. J. Hopkins and K. St John. Incorporating hierarchical characters into phylogenetic analysis. *Systematic Biology*, 2021.

P. O. Lewis. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic biology*, 50(6):913–925, 2001.

W. P. Maddison. Missing data versus missing characters in phylogenetic analysis. *Systematic Biology*, 42(4):576–581, 1993.

S. Tarasov. Integration of Anatomy Ontologies and Evo-Devo Using Structured Markov Models Suggests a New Framework for Modeling Discrete Phenotypic Traits. *Systematic Biology*, 01 2019. ISSN 1063-5157. doi: 10.1093/sysbio/syz005. URL `https://doi.org/10.1093/sysbio/syz005`.

S. Tarasov. The invariant nature of a morphological character and character state: insights from gene regulatory networks. *Systematic biology*, 69(2):392–400, 2020.

Z. Yang et al. *Computational molecular evolution*, volume 284. Oxford University Press Oxford, 2006.

E. Zuckerkandl and L. Pauling. Molecular disease, evolution, and genetic heterogeneity. *Horizons in biochemistry*, pages 189–225, 1962.

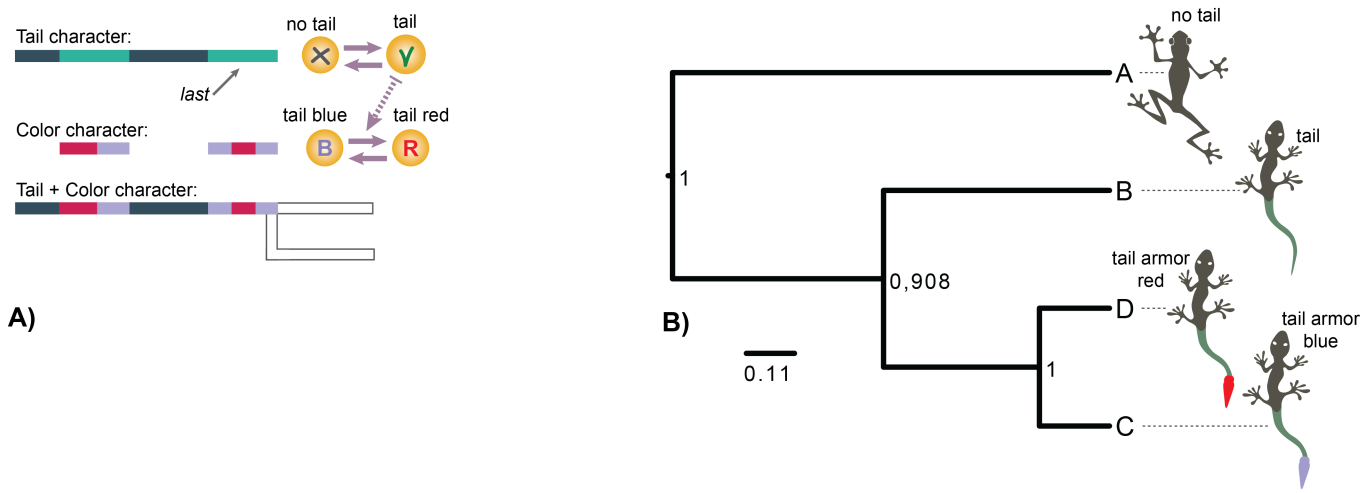

Fig. 1. a) The Embedded Dependency for the tail color case; "last" indicates that the probability of observing r or b state depends only on the last time interval of having the tail present. b) A phylogenetic tree for the tail armor case inferred with Mk_{armor} model in RevBayes; values above nodes indicate their posterior probabilities.