New phylogenetic Markov models for inapplicable morphological characters

Sergei Tarasov

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

Email: sergei.tarasov@helsinki.fi

Abstract

This paper describes new Markov models for phylogenetic inference with inapplicable (anatomically dependent) morphological characters. It shows that the new solutions fall into two categories depending on the type of character hierarchy and require use of different Markov models. A combination of the new models and previous solutions is capable to model all general types of dependencies that may exists between morphological traits. The performance of the new models is assessed using theory and simulations. Bayesian phylogenetic inference with the new models and their construction for a hierarchy of custom complexity is demonstrated using RevBayes and the R package *rphenoscate* respectively.

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

Recent publications have revitalized the 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 likelihood-based methods (Tarasov, 2019). The character inapplicability refers to the hierarchical dependency of traits due to anatomical dependency of their 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 state(s) of the controlling character – when the tail is present. Hereafter, I call this as the "embedded dependency" (ED) to distinguish it from other types of dependencies that may exist between traits.

The study of Goloboff et al. (2021) has criticized the likelihood-based solution using structured Markov models (SMMs) with hidden states in Tarasov (2019) for inappropriately treating the ED. Additionally, Goloboff et al. (2021) claimed that the ED would require new computational algorithms for likelihood estimation 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 SMMs do not explicitly imply the ED but disagree that they treat ED incorrectly; SMMs can model various dependencies and behave appropriately with respect to the 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 different models.

Given the recent interest, it is the best time as ever to provide another solution for the inapplicable characters, which is based on a new class of Markov models that explicitly model the ED. Below, I demonstrate that these models can be straightforwardly constructed using the techniques of SMMs for character amalgamation, aggregation and lumpability given in Tarasov (2019, 2020). These techniques are central to modeling morphological characters since they equip them with a mathematically consistent closure

property – 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 models come in two flavours depending on the type of character hierarchy. They can be represented using single rate matrix Q and do not need new algorithms as mentioned in Goloboff et al. (2021). I run simulations to evaluate the new ED models in the context of the tail color problem (Maddison, 1993) and, its modified version, "tail armor case" which well-exemplify anatomical dependencies and were investigated by me earlier with SMMs (Tarasov, 2019). Additionally, I describe an algorithm for constructing any custom character hierarchy and give practical recommendations for using the new models in phylogenetic inference with RevBayes (Höhna et al., 2016). The functions for constructing custom hierarchies are provided in the R package *rphenoscate*.

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

Preliminaries.— Below, I begin with deriving rate matrices for general ED models and then consider their possible parametrizations for phylogenetic inference. There are two fundamental categories of characters: (i) an absence-presence (a/p) character denotes (in)existence of an anatomical structure (e.g., tail), while (ii) a qualitative character describes qualities of a structure (e.g., shape or color). In ED characters interact that renders the ED to include two components: a controlling and dependent character. Depending on which category these components belong to, there are two main types of EDs: (1) *qualitative type* (ED_{ql}) where all states of a qualitative character depend on a/p of a controlling character character (e.g., tail color depends on tail presence); and (2) *absence-presence type* (ED_{ap}) where a/p of a dependent character depends on the a/p of a controlling character (e.g., presence of tail armor depends on tail presence). These two types require different Markov models for phylogenetic inference, which can be derived using the SMM techniques of state aggregation, amalgamation and lumpability. These

techniques are extensively used throughout the paper and I would suggest consulting my earlier work for details (Tarasov, 2019). In a nutshell, two or more individual characters can be merged into one character by the operation of amalgamation; hereafter, I refer to the amalgamation of independently evolving characters as "SMM amalgamation" [the equation (2) in Tarasov (2019)]. The lumpability property occurs when state aggregation of a Markov model (character) produces another model (with fewer states) that preserves the Markovian property [page 6 in Tarasov (2019)]; this property guarantees that the aggregated character can be unbiasedly modeled regardless of the original state space. In this study, the lumpability is mostly used to investigate dependencies between the controlling and dependent characters in ED. In the mathematical notations below, bold font is used for vectors, while capitals are for matrices.

Necessary and sufficient conditions.— Let us consider the simplest scenario of ED – the tail color case (TC) that seeks to optimally model traits in species with no tails, blue tails, and red tails (Maddison, 1993). This ED_{ql} implies two characters: T — tail presence $\{absent(a), present(p)\}$, and C — tail color $\{red(r), blue(b)\}$. Suppose, these two characters evolve over a tree branch; in the language of Markov models, this means that the Markov chain for the dependent character C, is allowed to evolve only within the time interval when the Markov chain for the controlling character T is in the state p; if the controlling character is in the state a then C is absent too (Fig. 1A). Such an interaction between the chains implies the following necessary and sufficient conditions to hold: (1) the controlling character is independent of the dependent character; (2) the dependent character is dependent on the controlling one; (3) after certain time, the probability of states in the dependent character depends only on the last realized state in the controlling character (Appendix 1). The last condition indicates that observing r or b depends solely on the last a/p state of the tail but not the previous tail occurrences. Hereafter, I call these conditions as the "ED conditions", they apply to all types of EDs.

Below, I show that the dependent and controlling chains can be represented as one

Markov chain that satisfies the aforementioned ED conditions and hence exhibits the solution to inapplicable characters. This chain can be constructed using the technique of "ED amalgamation" that is similar but different from the SMM amalgamation mentioned above. At the end of this section, I demonstrate how the combination of ED and SMM amalgamations can be used to model any character hierarchy.

General ED models

Qualitative model, ED_{ql}: Tail Color Case.— Suppose that C and T evolve under the ED 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:

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

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

$$ED_{ql}(T, C) = \begin{array}{c} \begin{array}{ccc} a & r & b \end{array} \\ \begin{pmatrix} - & \lambda_1 & \lambda_2 \\ \beta & - & \gamma_1 \\ \beta & \gamma_2 & - \end{pmatrix} \end{array} \begin{array}{c} a \\ r \\ b \end{array}, \tag{0.1}$$

this matrix has the initial vector $\boldsymbol{\pi_{ql}} = (1 - \pi_{T_2}, \pi_{T_2}\pi_{C_1}, \pi_{T_2}\pi_{C_2})$. Its rates λ_1 and λ_2 consist of two components: the transition rate α from T, and the embedded vector $\boldsymbol{\phi_C} = (\phi_{C1}, \phi_{C2})$, so that $\lambda_1 = \phi_{C_1}\alpha$ and $\lambda_2 = \phi_{C_2}\alpha$. The embedded vector $\boldsymbol{\phi_C}$ specifies the initial probabilities for the states in C when this character is allowed to evolve (i.e., when transiting to "tail" from "no tail").

I refer to this operation of combining T and C as the ED_{ql} amalgamation. Note, $ED_{ql}(T, C)$ fulfils all three ED 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 (the condition 1). In turn, C has no lumpable aggregations indicating its dependency on T (the condition 2). The fulfilment of the condition 3 is given in Appendix 1. Character evolution defined by this ED_{ql} rate matrix is exactly the one shown in Fig. 1A.

Absence-Presence model, ED_{ap}: Tail Armor Case.— Consider the tail armor (TA) 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 the matrix A indicates presence $\{absent(a_1), present(p_1)\}$ of the tail armor:

$$T = \begin{array}{cc} \begin{array}{cc} a & p \end{array} & \\ \begin{pmatrix} - & \alpha \\ \beta & - \end{pmatrix} & \begin{array}{c} a \\ p \end{array} \end{array}, A = \begin{array}{cc} \begin{array}{cc} a_1 & p_1 \end{array} & \\ \begin{pmatrix} - & \alpha_1 \\ \beta_1 & - \end{pmatrix} & \begin{array}{c} a_1 \\ p_1 \end{array} \end{array}, C = \begin{array}{cc} \begin{array}{cc} r & b \end{array} & \\ \begin{pmatrix} - & \gamma_1 \\ \gamma_2 & - \end{pmatrix} & \begin{array}{c} r \\ b \end{array} \end{array}. \tag{0.2}$$

It is logical to assume that in an amalgamated matrix for T and A, the change from "no tail" (aa_1) to "armor and tail present" (pp_1) should go though a stepping-stone transition "tail present, no armor" (pa_1), to allow gradual grow of the hierarchical complexity and permit only one change during an infinitesimal time interval. Thus, the amalgamation for the two a/p characters (hereafter denoted as ED_{ap}) should be different from the qualitative amalgamation ED_{ql}; it yields the following rate matrix for T and A:

$$ED_{ap}(T, A) = \begin{array}{cc} \begin{array}{ccc} aa_1 & pa_1 & pp_1 \end{array} & \\ \begin{pmatrix} - & \alpha & 0 \\ \beta & - & \alpha_1 \\ \beta & \beta_1 & - \end{pmatrix} & \begin{array}{c} aa_1 \\ pa_1 \\ pp_1 \end{array} \end{array}. \tag{0.3}$$

this matrix is similar to the ED_{ql} in the equation (0.1), but the embedded vector of initial probabilities $\boldsymbol{\phi_A}$ for A is $\boldsymbol{\phi_A} = (1, 0)$, which gives the following relationships between this matrix and the λ's in the equation (0.1): $\lambda_1 = \alpha_1$, $\lambda_2 = 0$; herein $\boldsymbol{\phi_A}$ penalizes the direct change from aa_1 to pp_1. Note, ED_{ap} also fulfils the three ED conditions given above.

To construct the rate matrix for their entire tail armor case, the ED_{ql} and ED_{ap}

amalgamations should be repeated successively:

$$ED_{ap}(T, ED_{ql}(A, C)) = \begin{array}{c} \begin{array}{cccc} aa_1 & pa_1 & pp_1r & pp_1b \end{array} \\ \begin{pmatrix} - & \alpha & 0 & 0 \\ \beta & - & \lambda_1 & \lambda_2 \\ \beta & \beta_1 & - & \gamma_1 \\ \beta & \beta_1 & \gamma_2 & - \end{pmatrix} \end{array} \begin{array}{c} \\ aa_1 \\ pa_1 \\ pp_1r \\ pp_1b \end{array}, \tag{0.4}$$

this matrix has the initial vector $\boldsymbol{\pi_{TA}} = (1 - \pi_{T_2}, \pi_{T_2}\pi_{A_1}, \pi_{T_2}\pi_{A_2}\pi_{C_1}, \pi_{T_2}\pi_{A_2}\pi_{C_2})$; its $\lambda_1 + \lambda_2$ equals to α_1 from the equation (0.3).

General ED model.— If the controlling character (Q_c) is the same as T in the equation (0.2) then the general equation for ED amalgamation is:

$$ED(Q_c, Q_d) = \begin{array}{c} \begin{array}{cc} a & p \end{array} \\ \begin{pmatrix} -\alpha & \boldsymbol{\phi_d}\alpha \\ \boldsymbol{\beta^T} & Q_d - I\beta \end{pmatrix} \end{array}, \tag{0.5}$$

where Q_d is the dependent characters, I is the identity matrix, $\boldsymbol{\beta}$ is the vector of β rates, superscript $\boldsymbol{T}$ is an array transpose and $\boldsymbol{\phi_d}$ is the appropriate type of embedded vector for the initial probabilities of Q_d when it is allowed to evolve . This equation can be applied to Q_d with any number of states. The dimensionality of the amalgamated matrix equals to the number of states in Q_d plus one. Note, this equations indicates that the the type of the ED rate matrix solely depends on the embedded initial vector $\boldsymbol{\phi_d}$.

Custom Character Hierarchy.— Anatomical dependencies between body parts can generate character hierarchies of varying complexity. If a hierarchy includes several characters it might be challenging to keep the right track on the sequence of character amalgamations. To facilitate the construction of amalgamated rate matrices, I propose using dependency diagrams and DDA (Dependency Diagram Amalgamation) algorithm.

The dependency diagram (Fig. 1D-J) schematically depicts relationships in character hierarchies and consists of three building blocks (Fig. 1B): (i) a/p character; (ii) qualitative character; and (iii) dependency arrow. The a/p character includes two

subblocks: the $a/p - node$ and $p - node$, which are the character itself and its state p respectively. In the diagram, the a/p and qualitative characters are interlinked by means of a dependency arrow to reflect hierarchical relationships; specifically, the dependency arrow links the $p - node$ of a controlling character with a dependent character (that can be either from the a/p or qualitative category). The DDA algorithm traverses along the dependency diagram by amalgamating characters at its nodes and produces the required rate matrix at the end.

For example, consider the complex hierarchy of three a/p and four qualitative tail characters shown in Fig. 1C, for which we wish to construct a rate matrix. The DDA algorithm works as follows: it takes the required dependency diagram (Fig. 1D) and traverses it from top to root in the topologically ordered manner. At each $a/p-$ or $p - node$, it combines all node's child characters into one matrix via SMM, ED_{ql} or ED_{ap} amalgamations; the type of amalgamation depends on the node and its children (see the Appendix 2 for details). The amalgamation at the root of the diagram yields the final rate matrix. For our complex hierarchy, the diagram traversing requires six steps and returns the final matrix of dimension=31 (Fig. 1D).

The dependency diagram for other aforementioned tail color and armor cases are shown in Fig. 1E-J. The vignette and functions implementing DDA algorithm for all considered examples are given in the package *rphenoscate*.

Phylogenetic inference with ED models

The general ED models from 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 models should be appropriately parameterized for tree inference to avoid overfitting. For example, the ED_{ql} for the tail color case [the equation (0.1)] can have as many as five parameters that might be overwhelming given the limited number of data points in morphological datasets. To maintain the balance between model complexity

and its biological interpretation, I propose several parametrizations for ED_{ql} and ED_{ap}.

Tail color case.— The tail color ED_{ql} [the equation (0.1)] can be straightforwardly converted into the Mk-like model (Lewis, 2001) by constraining it to only one free parameter (that is usually interpreted as a branch length), which gives the following rate matrix:

$$Q_{TC_1} = \begin{array}{c} \begin{array}{ccc} a & r & b \end{array} \\ \begin{pmatrix} - & 1 & 1 \\ 1 & - & 1 \\ 1 & 1 & - \end{pmatrix} \end{array} \begin{array}{c} a \\ r \\ b \end{array} . \tag{0.6}$$

Given that ED_{ql} includes different types of transitions, it is biologically reasonable to assume that they may occur at different rates. This suggests that the transitions for tail gain, loss and changes between the colors can be assigned to different categories as follows:

$$Q_{TC_2} = \begin{array}{c} \begin{array}{ccc} a & r & b \end{array} \\ \begin{pmatrix} - & \lambda & \lambda \\ \beta & - & \gamma \\ \beta & \gamma & - \end{pmatrix} \end{array} \begin{array}{c} a \\ r \\ b \end{array} . \tag{0.7}$$

The number of parameters in this rate matrix can be reduced to two by constraining the sum of the rates to 1 as $\beta + \lambda + \gamma = 1$ that gives $\gamma = 1 - \beta - \lambda$.

Tail armor case.— Unlike the previous case, the tail armor one [the equation (0.4)] might be converted into the Mk-like model by two different approaches. This duality happens because its rate matrix contains zero elements. The first type of the Mk parametrization implies that all non-zero rates are equal:

$$Q_{TA_1} = \begin{array}{c} \begin{array}{cccc} aa_1 & pa_1 & pp_1r & pp_1b \end{array} \\ \begin{pmatrix} -1 & 1 & 0 & 0 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{pmatrix} \end{array} \begin{array}{c} aa_1 \\ pa_1 \\ pp_1r \\ pp_1b \end{array} . \tag{0.8}$$

In the second type of the Mk parametrization, the main diagonal elements can be set to be equal that makes the off-diagonal rates uniformly rescaled within each row:

$$Q_{TA_2} = \begin{array}{c} \begin{array}{cccc} aa_1 & pa_1 & pp_1r & pp_1b \end{array} \\ \begin{pmatrix} -1 & 1 & 0 & 0 \\ 1/3 & -1 & 1/3 & 1/3 \\ 1/3 & 1/3 & -1 & 1/3 \\ 1/3 & 1/3 & 1/3 & -1 \end{pmatrix} \end{array} \begin{array}{c} \\ aa_1 \\ pa_1 \\ pp_1r \\ pp_1b \end{array} . \tag{0.9}$$

Although these two models have the same number of parameters (one), they may have different biological interpretations. The model Q_{TA_2} assumes the equal amount of expected waiting time that is spent in each state before leaving it, while Q_{TA_1} imposes longer waiting time to spent in the state aa_1 than in the three other states. Thus, these models imply different evolutionary dynamics for the "no tail, no armor" sate (aa_1). One may think that Q_{TA_2} can be biologically more reasonable than Q_{TA_1} if there are no prior evidence to believe why the absence tail should evolve slower that its presence.

The third way to parametrize the general tail armor model is to assign its rates into different categories which are of three main types – the rates responsible for the grow and loss of the hierarchy and changes between the qualitative states:

$$Q_{TA_3} = \begin{array}{c} \begin{array}{cccc} aa_1 & pa_1 & pp_1r & pp_1b \end{array} \\ \begin{pmatrix} - & q_2 & 0 & 0 \\ q_1 & - & \frac{1}{2}q_2 & \frac{1}{2}q_2 \\ q_1 & q_1 & - & q_3 \\ q_1 & q_1 & q_3 & - \end{pmatrix} \end{array} \begin{array}{c} \\ aa_1 \\ pa_1 \\ pp_1r \\ pp_1b \end{array} . \tag{0.10}$$

The number of the parameters is this model can be further reduced to two by constraining the sum of the rates to 1, that gives $q_3 = 1 - q_1 - q_2$.

In practice, three types of initial vectors can be used for the proposed models: equal, equilibrium and inferred state frequencies at the root of a tree. Note, all proposed models are generally time-irreversible due to their asymmetries in rates or presence of zero-valued rates. This means that their likelihood should be calculated on a rooted tree (Felsenstein, 1981; Yang et al., 2006). In this regard, Bayesian dating analysis represents a

convenient framework for their implementation. The only model in this set, that is trivially time-reversible, is Q_{TC_1} when its initial vector has equal or equilibrium state frequencies (both are the same); thus, the phylogenetic inference using this model does not require rooted tree. The tree search on unrooted topologies is the settings common for many maximum likelihood software.

Last but not least, morphological datasets usually contain character hierarchies which are coded using a set of characters, for the tail color case it would be two characters (one coded as inapplicable), while for the tail armor case it would be three (two coded as inapplicables). Note, all the ED models require representing these sets of characters as one single character whose states match those in the respective ED rate matrix. Thus, the set of characters defining a hierarchy should be recoded into once character prior to running analysis.

Materials and Methods

Availability

The functions and vignette for implementing the DDA algorithm and cosntructing character hierarchies are provided in the R package *rphenoscate* `https://github.com/sergeitarasov/rphenoscate` [Note, the package will be submitted to CRAN prior to acceptance of the MS]. The scripts to reproduce the simulations below are available at GitHub repository `https://github.com/sergeitarasov/MorphoModels` and in Supplementary Material (Supplemental Material is available on Dryad).

Simulations

To assess the ED models, I run three series of simulations in the Bayesian framework using RevBayes. (1) The fist series tests the effect of various parametrizations, proposed for the ED models, on tree topology and model performance. (2) The second series compares performance between the ED models and SMMs which were earlier studied

by me for modeling inapplicable characters (Tarasov, 2019). I use the tail armor case for these two series. (3) The third series tests the behaviour of the ED models in the original formulation of the tail color problem that is commonly used for assessing inapplicable characters (Maddison, 1993; Hawkins et al., 1997). Further details on these three series of simulations are given below.

Since generally the ED models are time irreversible and require a rooted tree, I run all analyses in the dating framework using global molecular clock model (Zuckerkandl and Pauling, 1962) with the branch rate prior $\tau \sim Exponential(10)$, Yule prior distribution over the tree topologies, and the tree root age set to 1.0. In all simulations, I use datasets with fifty identical characters referring to the tail armor case and tail color problem respectively; this number of characters was shown to contain sufficient information content for model assessment (Tarasov, 2019). The model likelihoods were conditioned on observing only variable characters in the data (Lewis, 2001). Given the simplicity of my toy datasets, I use 200K generations in Markov Chain Monte Carlo algorithm and two independent runs for each simulation.

The relative model performance was evaluated using Bayes factors (BF) which compares the ratio of marginal likelihoods (Mln) of two candidate models (estimated as Mln of the best model minus Mln of the focal model). The marginal likelihood was calculated using the stepping-stone algorithm (Xie et al., 2011) implemented in RevBayes by averaging over the two runs.

1. The performance of the ED models.— I tested eight different ED models (ED1-ED8) using the tail armor case (Table 1). They were constructed as combinations of the three rate matrices Q_{TA1}, Q_{TA2} and Q_{TA3} from the equations (0.8-0.10) and the three types of the initial vectors at tree root with equal, equilibrium and inferred probabilities. The tested models range in the number of parameters from 1 to 4. This number refers exclusively to the components of Markov model of character evolution that is calculated as the sum of free parameters in $Q + \pi + 1$ (one stands for the branch rate τ). The

combinations of Q_{TA3} and the inferred initial vector were not included in the simulations due to excessive number of parameters in them.

2. ED vs. SMM.— In this series, I compare two SMMs (Mk-SMM-ind and Mk-SMM-sw) with the selected ED models using BF to assess their relative performance (Table 1). Both Mk-SMM-ind and Mk-SMM-sw are one-parameter models that have eight hidden and four observable states [page 3 in Tarasov (2019)]. The conventional representation of the respective ED models has only four observable states. For the sake of model comparison the data that are fed into different models should be the same, which might be violated by varying state space. To avoid it, I expanded the state-space of the ED models to match that of the SMMs. This expansion creates the eight-state rate matrix that, if lumped, collapses back to the original four-state one as described in Tarasov (2019).

3. ED and Tail Color Problem.— In this series of simulations, I test if the ED models can appropriately model the tail color problem (TCP). I follow the same simulation setup as before (Tarasov, 2019). The TCP considers a tree of 14 species where the tailed species (with red and blue tails) are nested within the left and right clades of the tailless species (Maddison, 1993). The tree (Fig. 2C) is assumed to be fully resolved except for the relationships of the left tailed clade (LTC); all resolved clades are supported by, at least, one binary character to avoid zero-length branches. The goal of these simulations is to assess possible resolutions of the LTC. For this purpose, I use one-parameter (Q_{TC_1}, the equation 0.6) and three-parameter (Q_{TC_2}, the equation 0.7) ED models with equal state frequencies at the tree root. For each of these models, I ran two simulations with varying number of synapomorphies supporting the LTC (one or eleven characters) to assess the effect of branch length on the topology.

Results

1. The performance of the ED models.— The ED models (Table 1) produce two distinct tree topologies for the tail armor case. One topology suggests a gradual appearance of the tail and tail armor from no tail state (Fig. 2A); it is recovered by the models ED1, ED3-5, Mk-SMM-sw, Mk-SMM-ind on maximum clade credibility (MCC) tree. Majority rule (50%) consensus topology for these models is the same as the MCC one, except for ED1,5 where "tail" and "no tail" taxa collapse into a polytomy. Another topology groups together species with and without the armor (Fig. 2B); it is recovered by the models ED2, ED6-8 on both MCC and majority rule (50%) consensus trees.

The marginal likelihood values significantly vary across the ED models (Table 1). The best fit models are ED5 and ED3 that reflect the gradual appearance of the tail from no tail (Fig. 2A). These two models are characterized by the Mk-like rate matrix with equal transition rates (Q_{TA_1}, the equation 0.8). The worst performing models are ED1 and ED2 with the two-parameter rate matrix (Q_{TA_3}, the equation 0.10). The BF value between the best fit model (ED5) and the worst fit model (ED2) is 21.13 that points to a substantial overperformance of ED5 over ED2.

2. ED vs. SMM.— Based on the criteria of simplicity and performance, I selected two one-parameter models ED3 and ED6 from the previous series of simulations to assess their fit in comparison to one-parameter SMMs for inapplicable characters: Mk-SMM-ind and Mk-SMM-sw. The selected ED models performed decisively better than the respective SMMs in these toy simulations ($BF >> 10$, Table 1).

3. ED and Tail Color Problem.— The ED models show the appropriate behavior with respect to the TCP. Both Q_{TC_1} and Q_{TC_2} demonstrate the desirable behavior in resolving the LTC – they support the monophyly of the red and blue tailed species with the posterior probability 1.0 (Fig. 2C). The same topology is recovered regardless of the number of synapomorpies (one or eleven) used in the simulations.

Discussion

Assessing the proposed ED models using the simulations

Coding the tail color characters as inapplicables, in the context of the TCP, was shown to suffer from the undesirable behavior in parsimony analysis by incorrectly resolving the species relationships within the LTC (Maddison, 1993). Previously, I showed that, unlike parsimony, SMMs (e.g., Mk-SMM-ind and Mk-SMM-sw) provide the solution to modeling the inapplicables (Tarasov, 2019). In this study, I have demonstrated that the ED models, likewise SMMs, also solve this problem – they correctly resolve the LTC and, thereby, do not misbehave. This result indicates that the three-state Mk model [i.e., Q_{TC_1} from the equation (0.6)] does, in fact, represent one of the solutions to the TCP. This is in contrast to my previous conclusion (Tarasov, 2019) that now should be considered invalid.

Even for a simple dataset, one can construct several ED models, as shown for the tail armor and tail color cases, which may differ in rate matrices and initial vectors. The first series of simulations demonstrates that the fit and topology of different models can vary significantly because different rate matrix parametrizations and initial vectors naturally imply different assumptions for character evolution. Specifically, the ED models recover two typologies – one assuming gradual evolution of the tail and tail armor, and the other grouping "tail" and "no tail" species (Fig. 2A-B). This result tends to be associated with the parametrization of the rate matrix: the matrix Q_{TA_1} with equal transition rates usually produces the tree in Fig. 2A, while the matrix Q_{TA_2} with equal diagonal elements recovers the tree in Fig. 2B. Note, the models favoring the gradual tail evolution are capable to correctly recover the expected root of the tree – the species with no tail (Fig. 2A).

Both, the parametrization of Q and initial vector affect model marginal likelihood. In the presented toy simulations, the ED models outperform SMMs, and the best identified model (ED5 with four parameters) comes to support the gradual evolution of the tail and armor. However, this results may differ depending on dataset. In practice, one should

always compare the set of ED models and SMMs using statistical methods for model selection (e.g., Akaike information criterion, BF) as shown in the first series of simulations.

Modeling Character Hierarchies

The ED exhibits a widespread phenomenon in phylogenetics since organismal body parts often depend on each other. It occurs when one character (Markov chain) is allowed to evolve only within certain state(s) of another character (chain). In practice, the ED has been widely coded using inapplicable characters and then analyzed under parsimony or likelihood-based methods. However, in parsimony, the inapplicable coding was shown to be inappropriate (Maddison, 1993), which encouraged development of new parsimony-based methods to handle the inapplicables (Brazeau et al., 2019; Hopkins and St John, 2021; Goloboff et al., 2021).

In this paper, I used theoretical derivations to show that certain types of Markov models can be straightforwardly adopted to explicitly model the ED. These models are defined using single rate matrices via the appropriate amalgamation of the involved characters [the equation (0.5)]. They do not require special algorithms for likelihood computations and can be run in any software that allows custom specification of Q's and inference on rooted typologies (since many models are time-irreversible). Specifically, I demonstrated how the Bayesian tree inference with the new models can be conducted in RevBayes. The ED models come in two flavors – ED_{ql} and ED_{ap} – depending on the type of character dependency. They differ from SMMs, in which inapplicables are modeled by means of hidden states, that does not explicitly imply the ED. Note, the performance of the ED models and SMMs can be data-dependent and the best model for a specific character hierarchy should be selected using model selection criteria.

In practice, anatomical dependencies between characters may be complex. The proposed DDA algorithm and the packages *rphenoscate* can be used to construct a rate matrix that defines any custom character hierarchy for its use in downstream phylogenetic

analysis. The DDA algorithm is based on the combination of ED and SMM amalgamations. For phylogenetic inference, all characters associated with anatomically dependent data should be recoded to match the state space of the respective ED or SMM rate matrices. The recoded characters can be placed in a single data partition subset to which a selected Markov model is assigned.

The likelihood-based phylogenetics has only a limited set of models available for discrete traits (Lewis, 2001; Klopfstein et al., 2015; Pyron, 2017). This imposes certain constraints for tree inference when trying to account for a diverse array of processes that generate trait evolution. The new ED models expand the current arsenal of modeling approaches for morphological characters. The flexibility of ED and SMM in comparison to parsimony lays in their natural ability to model the trait-generating process and, thereby, dependencies. The combination of ED models and SMMs can be used to potentially account for any type of dependency that may exist between traits.

Supplementary Material

Data available from the Dryad Digital Repository: `http://dx.doi.org/dryad.[NNNN]`.

Funding

This work was supported by the Academy of Finland grant: 339576.

Appendix

Appendix 1: Proof of the Condition 3

In the context of tail color case the ED condition 3 means that the probability of observing C (r or b state) at time T (Fig. 1A) depends only on the last time interval $T - t$ of having the tail absent/present but not the previous tail occurrences, which is

$Pr(C,T) = Pr(C,T-t)$.

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 $\boldsymbol{\pi} e^{ED_{ql}(T,C)}$ (the formula is not given here for brevity). The equality of the two approaches would suggest that the condition 3 is true. For tractability suppose that Q_c is time-reversible:

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

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:

$$\begin{aligned} Pr(\boldsymbol{\pi} Q_d, T \mid \vec{ap}) = & \int_0^T Pr\Big(\vec{aa}(t) \mid N(t)>0\Big) Pr\Big(\vec{pp}(T-t) \mid N(T-t)=0\Big) \\ & \times Pr\Big(\boldsymbol{\pi} Q_d, T-t\Big) dt \\ = & \int_0^T \Big(\frac{1}{2} + \frac{1}{2} e^{-2at}\Big)(ae^{-a(T-t)})(\boldsymbol{\pi} e^{Q_d(T-t)}) dt, \end{aligned} \tag{0.11}$$

$$\begin{aligned} Pr(\boldsymbol{\pi} Q_d, T \mid \vec{pp}) = & Pr\Big(\vec{pp}(T) \mid N(t)=0\Big) Pr\Big(\boldsymbol{\pi} Q_d, T\Big) \\ & + \int_0^T Pr\Big(\vec{pa}(t) \mid N(t)>0\Big) Pr\Big(\vec{pp}(T-t) \mid N(t)=0\Big) \\ & \times Pr\Big(\boldsymbol{\pi} Q_d, T-t\Big) dt \\ = & e^{-aT}\Big(\boldsymbol{\pi} e^{Q_d T}\Big) + \int_0^T \Big(\frac{1}{2} - \frac{1}{2} e^{-2at}\Big)(ae^{-a(T-t)})(\boldsymbol{\pi} e^{Q_d(T-t)}) dt. \end{aligned} \tag{0.12}$$

Now averaging over the initial vector:

$$Pr(Q_d, T \mid Q_c) = \pi_a Pr(\boldsymbol{\pi} Q_d, T \mid \vec{ap}) + \pi_p Pr(\boldsymbol{\pi} Q_d, T \mid \vec{pp}) \tag{0.13}$$
$$= \Big(\frac{(\gamma_1 - \gamma_2) e^{T(-(a+\gamma_1+\gamma_2))} + a + 2\gamma_2}{4(a+\gamma_1+\gamma_2)}, \frac{(\gamma_2 - \gamma_1) e^{T(-(a+\gamma_1+\gamma_2))} + a + 2\gamma_1}{4(a+\gamma_1+\gamma_2)} \Big).$$

Note, $Pr(Q_d, T \mid Q_c)$ in the euqation (0.13) is the same as elements π_2 and π_3 in $\boldsymbol{\pi} e^{ED_{ql}(T,C)}$ that proofs the aforementioned condition 3.

Appendix 2: The Dependency Diagram Amalgamation (DDA) algorithm

As an input, the DDA algorithm requires a dependency diagram (e.e., Fig. 1D). The algorithm works by taking the diagram and traversing it from tips to root in the topologically ordered manner; during the traversing the DDA amalgamates characters. The amalgamations occur only at $a/p-$ or $p-nodes$ using the following rules:

1. At each $p-node$ combine all child characters via SMM amalgamation [the equation (2) in Tarasov (2019)].
2. At each $a/p-node$ conditionally combine all child characters via:
   (a) ED_{ql} amalgamation if the children are qualitative characters (e.g., as in the equation 0.1).
   (b) ED_{ap} amalgamation if the children are a/p characters (e.g., as in the equation 0.3).
   (c) The mixture of ED_{ql} and ED_{ap} amalgamations if the children are a/p and qualitative characters; this requires appropriate construction of $\boldsymbol{\phi_d}$ to match the dependencies between states (the equation 0.5).
3. The amalgamation a the root of the dependency diagram yields the desired final rate matrix.

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.

S. Klopfstein, L. Vilhelmsen, and F. Ronquist. A nonstationary markov model detects directional evolution in hymenopteran morphology. *Systematic biology*, 64(6):1089–1103, 2015.

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.

R. A. Pyron. Novel approaches for phylogenetic inference from morphological data and total-evidence dating in squamate reptiles (lizards, snakes, and amphisbaenians). *Systematic Biology*, 66(1):38–56, 2017.

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.

W. Xie, P. O. Lewis, Y. Fan, L. Kuo, and M.-H. Chen. Improving marginal likelihood estimation for bayesian phylogenetic model selection. *Systematic biology*, 60(2):150–160, 2011.

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.

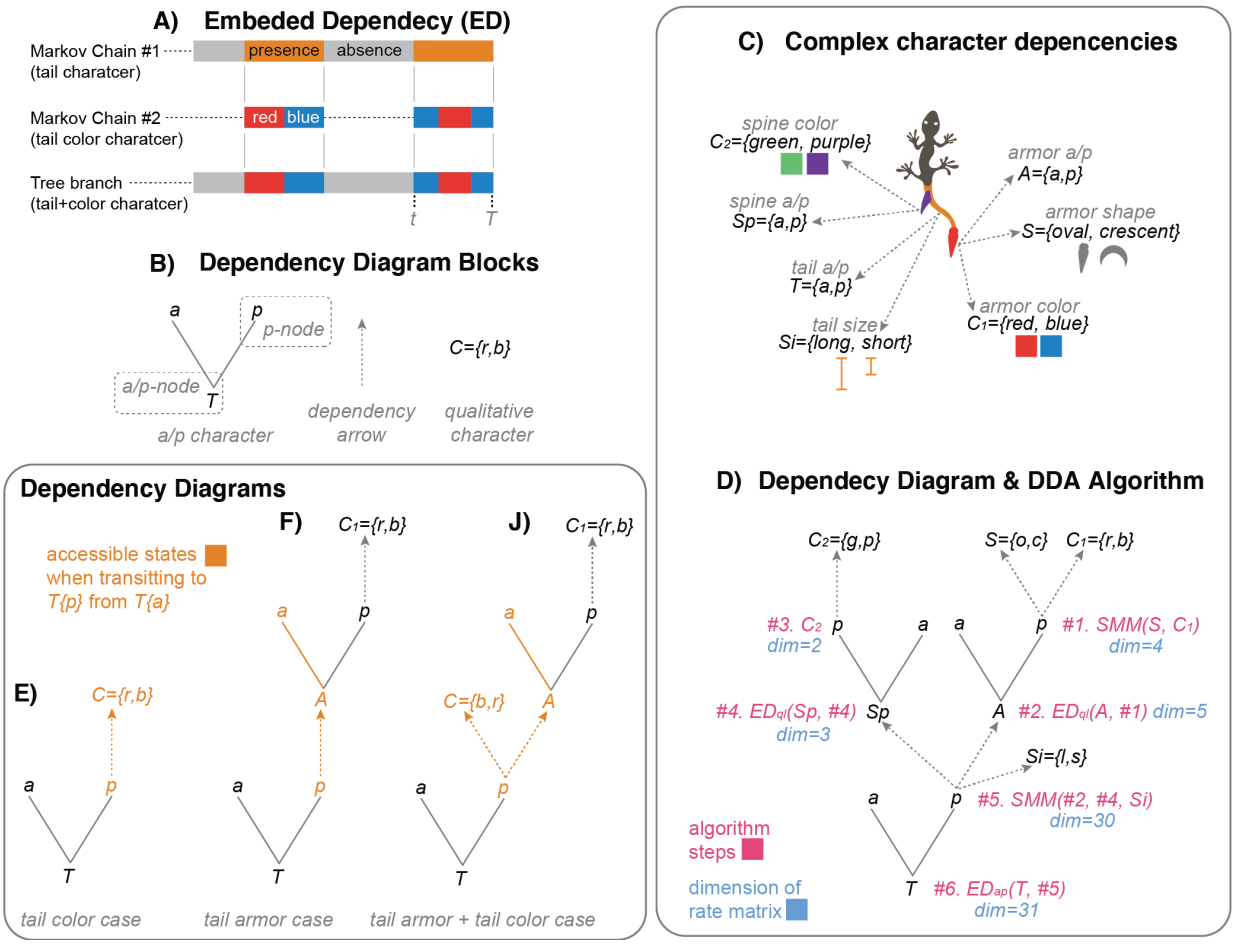

Fig. 1. The Embedded Dependency (ED). A) Visualizing the ED using two Markov chains (characters) – #1 and #2 – evolving along the tree branch; ED means that the two states in the chain #2 are allowed to evolve only within the state "present" of the chain #1. B) The building blocks of a dependency diagram. C) Complex character hierarchy for the hypothetical tail traits, which consists of three a/p characters and four qualitative ones. D) The dependency diagram for the characters in (C); red indicates steps of the DDA algorithm, while blue refers to the dimension of rate matrix; ED_{ql}, ED_{ap}, SMM refer to the respective amalgamations. E-J) The dependency diagrams for the discussed character hierarchies and their combination.

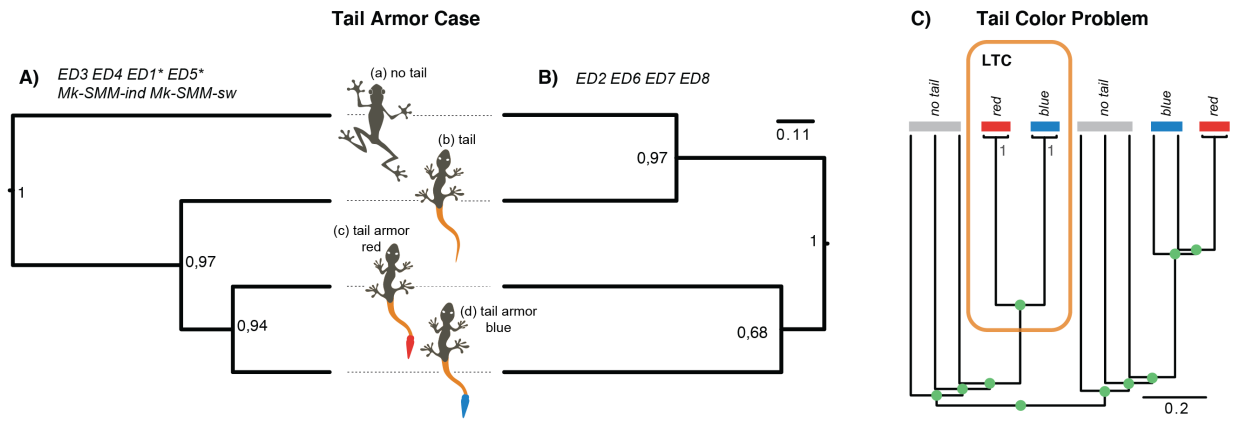

Fig. 2. Assessing the ED models. A-B) The two majority rule (50%) consensus trees recovered by eight ED models and two SMMs for the tail armor case; the values at tree nodes indicate posterior probabilities from the analyses with ED4 (A) and ED8 (B) models (they are selected for the demonstrative purpose only); "*" indicates that "tail" and "no tail" taxa collapse into a polytomy on the majority rule (50%) consensus trees but are resolved on MCC tree that is the same as the tree in (A). C) The tree that exemplifies the tail color problem where the tailed species are nested within the two major clades of the tailless species; all relationships are assumed to be resolved (green balls) except those in the left tailed clade (LTC) that includes two species with red tail and two species with blue tail respectively; the values in the LCT indicate posterior probabilities.

Table 1. *The Performance of the ED models. The "Topology" column indicates the tree A or B from Fig. 2.*

	Model	Parameters	Marginal Ln	BF	Topology
	The first series of simulations:				
1	ED5 ($Q_{TA_1}, \pi = inferred, \tau$)	4	-292.29	0.00	A
2	ED3 ($Q_{TA_1}, \pi = equal, \tau$)	1	-296.13	3.85	A
3	ED8 ($Q_{TA_2}, \pi = inferred, \tau$)	4	-299.10	6.82	B
4	ED4 ($Q_{TA_1}, \pi = equilibrium, \tau$)	1	-300.50	8.22	A
5	ED6 ($Q_{TA_2}, \pi = equal, \tau$)	1	-303.65	11.36	B
6	ED7 ($Q_{TA_2}, \pi = equilibrium, \tau$)	1	-304.98	12.69	B
7	ED1 ($Q_{TA_3}, \pi = equal, \tau$)	3	-308.24	15.95	A
8	ED2 ($Q_{TA_3}, \pi = equilibrium, \tau$)	3	-313.42	21.13	B
	The second series of simulations:				
1	ED3 ($Q_{TA_1}, \pi = equal, \tau$)	1	-303.94	0.00	A
2	ED6 ($Q_{TA_2}, \pi = equal, \tau$)	1	-309.75	5.81	B
3	Mk-SMM-sw ($\pi = equal, \tau$)	1	-328.07	24.13	A
4	Mk-SMM-ind ($\pi = equal, \tau$)	1	-334.71	30.77	A