Resolving ubiquitous model congruence in phylogenetics and its application for studying macroevolution

Sergei Tarasov[1, 2, 3,*], Josef Uyeda[3]

[1] *Finnish Museum of Natural History, Pohjoinen Raubtatiekatu 13, FI-00014 Helsinki, Finland*
[2] *National Institute for Mathematical and Biological Synthesis (NIMBioS), TN, USA*
[3] *Department of Biological Sciences, Virginia Tech 4076 Derring Hall 926 West Campus Drive Blacksburg, VA 24061, USA*

**Corresponding author, sergei.tarasov@helsinki.fi*

Abstract

This paper investigates model congruence (= asymptotic unidentifiability) in phylogenetics for continuous-time Markov chains (CTMCs) that include models for DNA, protein, discrete trait evolution, and state-dependent diversification. Without exception, all CTMCs have infinite classes of congruent models. Congruent models vary in the number of parameters from one to infinity and may have drastically different evolutionary dynamics, and standard model selection criteria are not applicable to them. We classify the structure of a congruence class and show how the "best" model can be selected based on mathematical and biological reasoning. Thus, congruent models in CTMCs do not suffer from the model selection problem previously identified for the time-dependent diversification process. Moreover, we demonstrate that congruence may serve to explain some evolutionary phenomena, specifically linking macro-and microevolution. We also discuss other types of congruence that may occur in phylogenetics and ways to handle them.

Key words: Markov models, congruent models, model selection, traits, diversification,

microevolution, macroevolution, identifiability

Introduction

Most studies suggest that organismal evolution at different organizational levels, be it DNA, proteins, phenotypes, or species diversification, is driven by heterogeneous processes with rates variable across time and lineages. A central goal of evolutionary biology is to disentangle biotic and abiotic factors that explain this rate variation. A recent study of Louca and Pennell (2020) [hereafter, LP] raised an important issue of model congruence occurring in the birth-death models used to reconstruct diversification. LP showed that any diversification scenario associated with extant timetrees could be explained by alternative scenarios, equally likely to have generated the observed data. In other words, for any observed extant timetree there is an infinite set of congruent models with different dynamics of extinction and speciation but identical likelihood scores. These congruent models are asymptotically unidentifiable (Morlon et al., 2022), which makes them indistinguishable given the observed data. Nevertheless, they suggest different biological interpretations for explaining rate heterogeneity.

Currently, there is no consensus within the phylogenetic community on how to deal with the congruence and correctly treat alternative scenarios (Morlon et al., 2022; O'Meara and Beaulieu, 2021; Helmstetter et al., 2021). LP demonstrated the existence of congruent models for only time-dependent diversification process, where speciation and extinction rates are constant across lineages. While they conjectured that the congruence issue might also arise in State-Speciation and Extinction (SSE) models where the diversification rate is allowed to vary across lineages [e.g., BiSSE, HiSSE, (Maddison et al., 2007; Beaulieu and O'Meara, 2016)], a recent study proposed that SSE models are identifiable (O'Meara and Beaulieu, 2021).

This paper discusses the congruence issue for general continuous-time Markov

chains (CTMCs), not investigated by LP. CTMCs lie at the heart of phylogenetics and embrace the most common models for DNA (e.g, JC, GTR), protein (e.g, Dayhoff), discrete trait evolution, and SSE models for diversification. We show a general and mathematically tractable approach for constructing congruent models using hidden state expansion that relies on the lumpability property of Markov chains. Our approach demonstrates that, without exception, model congruence occurs in all CTMC models, regardless of whether they are modeled on a known phylogenetic tree, or the tree is a parameter to be estimated. Moreover, there is an infinite set of congruent models, called congruence class, for any single CTMC, as was shown for the birth-death models by LP. We find that CTMC models in a congruence class substantially vary in the number of parameters – from one to infinity. Thus, any CTMC, regardless of its level of rate heterogeneity, can be represented by a congruent model with one parameter or vice versa – a model with one parameter can be represented using many parameters. Such transformations come at the cost of expanding the model's state space.

Naturally, congruent models in CTMCs may have drastically different evolutionary dynamics and may support alternative or even incoherent hypotheses in hypothesis-driven research. As noted by LP, this ambiguity imposes a model selection problem. We investigate the structure of a CTMC congruence class by classifying its constituent elements and demonstrate that a biologically and mathematically reasonable approach exists for unambiguously selecting the "best" model within a class. It is called the irreducible model and it is unique in each congruence class. Our selection approach is based on penalizing the model's state complexity. The necessity for penalizing complexity was emphasized earlier for handling congruent models (Morlon et al., 2022). We also provide practical recommendations for it use. Thus, we show that CTMCs do not suffer from a model selection problem in the same way as birth-death models. Additionally, we discuss how a similar approach can be applied to address unidentifiability in time-dependent birth-death models. However, we admit that potential problems of model

selection may arise when comparing models from similar congruence classes, in which models have similar but different likelihoods.

Generally, our findings indicate that model congruence does not pose a problem to phylogenetics. However, additional research is needed to understand relationships between congruence classes and develop general methods to handle congruence across different types of phylogenetic models.

Moreover, congruence might be an exciting tool for studying links between micro-and macroevolution. As we discuss below, it may explain how a constant-rate microevolutionary process generates variable rates at the macroevolutionary level due to the curvature of the adaptive landscape. In the end, we overview other types of congruence, not based on the hidden expansion, which may exist in CTMCs. Thus far, we see no evidence that they should challenge phylogenetic practice, but this requires further investigation. Most of our theoretical findings are supported by R scripts; see "Material and Methods" for more information.

Theoretical Background and Notations

Hidden Markov models.— Hidden Markov models (HMMs) elaborate traditional CTMCs by splitting the model states into two layers – observable and hidden (Tarasov, 2019). The transitions between states are allowed only within the hidden layer that, in essence, represents a traditional CTMC. The observable states exhibit one-to-many mapping with the hidden states. Thus, the probability of an observable state in an HMM is integrated over the probabilities of it belonging to each of its possible hidden states.

Originally, HMMs were proposed to model rate heterogeneity across lineages (Tuffley and Steel, 1998; Beaulieu et al., 2013) and were interpreted as different rate categories influencing character evolution or hidden extrinsic environmental factors. A recent study (Tarasov, 2019) showed that their interpretation is, in fact, multifaceted. The hidden states can be anything of the following: time-heterogeneity, hidden genetic

elements, environmental factors, and subjectivity associated with trait discretization. Thus, interpreting hidden states must be taken with caution and, like in diversification models, will lead to the degeneracy of processes onto particular patterns of trait evolution.

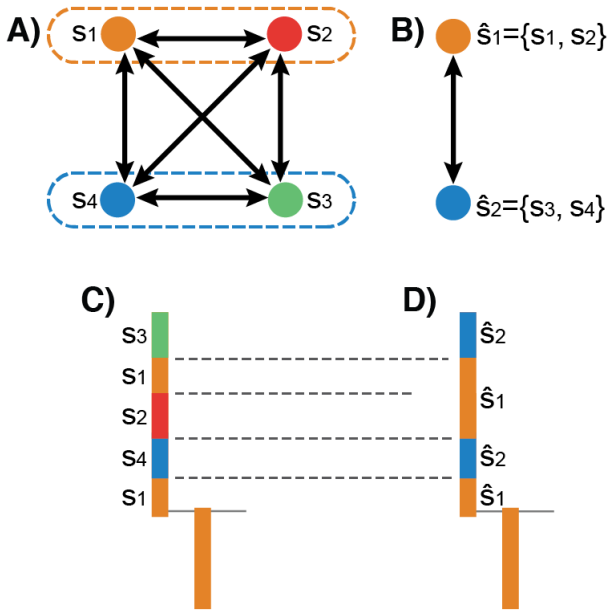

Fig. 1. The lumpability property applied to phylogenetic CTMCs. A-C) The original four-state CTMC from equation (1). B-D) Lumped CTMC from the equation (2). A-B) Topology graphs for CTMC models. C-D) Stochastic realization of CTMC process over tree branch.

Lumpability.— The property of lumpability is central to the present study. Two or more states of an initial CTMC can be aggregated (lumped) into one state that produces the aggregated process with fewer states. If the aggregated process is still Markovian, the initial CTMC is called lumpable for the given state aggregation. If the process is not Markovian, the initial CTMC is called non-lumpable. The aggregation of states is equivalently applicable to three levels of a CTMC representation: rate matrix, Markov model topology, and a stochastic realization of the process. To train an intuition about the aggregation, suppose we have an initial CTMC that characterizes the evolution of a four-state character $\{s_1, s_2, s_3, s_4\}$ and has the following rate matrix Q:

$$Q = \begin{array}{c} \begin{array}{cccc} s_1 & s_2 & s_3 & s_4 \end{array} \\ \left(\begin{array}{cc|cc} - & q_{12} & q_{13} & q_{14} \\ q_{21} & - & q_{23} & q_{24} \\ \hline q_{31} & q_{32} & - & q_{34} \\ q_{41} & q_{42} & q_{43} & - \end{array}\right) \end{array} \begin{array}{c} s_1 \\ s_2 \\ s_3 \\ s_4 \end{array} . \quad (1)$$

The aggregation at the level of the rate matrix means constructing a smaller matrix with fewer states and transitions. Suppose that we are willing to aggregate Q into a two-state process $\hat{Q}$, with states $\{\hat{s}_1, \hat{s}_2\}$, using the following partition of the initial states: $\hat{s}_1 = \{s_1, s_2\}$, $\hat{s}_2 = \{s_3, s_4\}$. This partition can be visualized on Q by splitting the rates into four partition blocks shown with a horizontal and a vertical line within the matrix in the equation (1). Each block includes four initial rates whose function should yield new rates in the aggregated matrix. So, the aggregated matrix should be:

$$\hat{Q} = \begin{array}{c} \begin{array}{cc} \hat{s}_1 & \hat{s}_2 \end{array} \\ \begin{pmatrix} - & \hat{q}_{12} \\ \hat{q}_{21} & - \end{pmatrix} \end{array} \begin{array}{c} \hat{s}_1 \\ \hat{s}_2 \end{array} . \quad (2)$$

Before learning of the relationships between the original and the aggregated rates, let us consider other levels of aggregation. The same aggregation at the level of Markov model topology (Fig. 1A-B) implies collapsing a four-state transition graph of Q into the two-state graph for $\hat{Q}$. Furthermore, the aggregation at the level of a stochastic realization means "recoloring" the realized initial states with fewer colors according to the state's partition (Fig. 1C-D).

The lumpability property guarantees that the transition rates of the aggregated CTMC can be derived from the initial CTMC and thereby modeled using the lumped process. There are several classes of lumpability, of which the "strong lumpability" (Kemeny and Snell, 1960) is the most interesting to us. This is because of its mathematical tractability and the fact that it holds in CTMCs on phylogenetic trees (Tarasov, 2019). Below, we refer to "strong lumpability" as simply "lumpability". This property depends on certain rate symmetries in the rate matrix, which must obey the "row-wise sum rule"

(RWR).

The RWR implies the following: *the original CTMC is lumpable with respect to a given partition of states when the row-wise sum of rates within one partition block in the original Q is the same for all rows within the given partition block, and this property holds for all blocks in Q (Kemeny and Snell, 1960). The rates in the aggregated matrix represent simply the row-wise sums of the original rates.*

For Q in the equation (1), the RWR is maintained if the following two equalities hold: $\hat{q}_{12} = q_{13} + q_{14} = q_{23} + q_{24}$ and $\hat{q}_{21} = q_{31} + q_{32} = q_{41} + q_{42}$. These equalities imply lumpability of Q for the given state partition; $\hat{q}_{12}$ and $\hat{q}_{12}$ represent the new rates in the lumped process $\hat{Q}$ shown in the equation (2). The same aggregation for other levels of the Markov process is shown in Fig. 1B-D. Note, to prove lumpability, it is enough to show that the RWR holds for the off-diagonal blocks because it implies that the main diagonal blocks maintain the RWR too. In the lumped process, the initial probability vector at the tree root (π) should be aggregated, too: by adding up state probabilities from the original process belonging to the same partition subset.

The lumpability property is general and applies to all types of CTMCs. Lumpability does not depend on the time over which a CTMC evolves nor the initial probability vector at the tree root; if a CTMC is lumpable, it is lumpable under any value of the initial vector. If the lumpability property does not hold, then the simple form for the rate matrix for the aggregated process does not exist since it is not Markovian.

We denote the operation of lumping a CTMC, if the RWR holds, as $Lump(\cdot)$. For a lumpable CTMC, the following equality holds:

$$Lump\Big(P(Q, \pi, t)\Big) = P\Big(Lump(Q), Lump(\pi), t\Big), \tag{3}$$

where P is the probability of the Markov process after some time t, that is $P(Q, \pi, t) = \pi e^{Qt}$.

Correlated CTMC evolution.— When dealing with a set of evolving CTMCs, it is sometimes convenient to consider their evolution as a joint process instead of a set of individual processes (Tarasov, 2020). Both such representations are equivalent. To make the joint process, one needs to amalgamate a set of CTMCs into one CTMC by combining their states and rate matrices. The state-space of the amalgamated CTMC consists of all possible state combinations from the individual processes. For example, the amalgamation of two binary CTMCs with the states $\{0, 1\}$ produces a CTMC with four states $\{00, 01, 10, 11\}$. The rate matrix of the amalgamated process has a specific structure where only one-step transitions across state space are allowed while simultaneous transitions are prohibited.

If the initial CTMCs evolve independently, their joint rate matrix has certain symmetries. Specifically, if one partitions its state space with respect to states of some initial CTMC, then this aggregation would be lumpable, and the lumped process would yield the selected initial CTMC. One can construct an amalgamated CTMC that describes the correlated evolution of initial CTMCs in which the transition are dependent. In that case, the rate symmetries are broken, and the amalgamated process is not lumpable with respect to all or some initial chains. The CTMC amalgamation is often used to test (un)correlated character evolution by fitting both correlated and independent matrices and then choosing the best scenario using model selection (Pagel, 1994).

In this study, we consider amalgamated matrices constructed from a set of the initial CTMCs with an equal number of states and identical transition rates, q. We define an initial CTMC as Q^s where s indicates the number of states. We denote the amalgamation of two rate matrices as independently evolving through the Kronecker sum ($\oplus$). Then, the formula for it is $Q_1 \oplus Q_2 = Q_1 \otimes I_{Q_2} + I_{Q_1} \otimes Q_2$, where I_{Q_2} and I_{Q_1} are the identity matrices for the respective Q's, and $\otimes$ denotes the Kronecker product. To amalgamate n identical matrices Q^s as independently evolving, we should apply the Kronecker sum operation successively:

$$\bigoplus^{n} Q^{s} = \bigoplus_{i=1}^{n} Q_i^{s} = Q_1^{s} \oplus Q_2^{s} \oplus ... \oplus Q_n^{s}. \quad (4)$$

Note, below we refer to a rate matrix constructed from n identical and independent CTMCs with s states as $\bigoplus^{n} Q^{s}$.

We can slightly modify the independent amalgamations to make all or some identical Q^{s} evolve in correlation. For this, we can amalgamate them as previously ($\bigoplus^{n} Q^{s}$) and set some allowed transitions (those denoting one-step changes) in that rate matrix to zero. Let us define this operation of the correlated amalgamation via $\odot$. Then, $\bigodot^{n} Q^{s}$ indicates the amalgamation of n CTMCs, each with s states, where all or some CTMCs evolve in correlation.

The number of states in the amalgamated matrices grows exponentially as s^{n}. The processes $\bigoplus^{n} Q^{s}$ and $\bigodot^{n} Q^{s}$ have specific topologies in the state space because only one-step transitions are allowed. We generally refer to those topologies as a space of dimension n on s elements. For example, the space of dimension n on two elements ($\bigoplus^{n} Q^{2}$ and $\bigodot^{n} Q^{2}$) is a hypercube, with dimension $n = 3$ being simply a cube. If $s > 2$, we refer to those amalgamations as a hyperspace on s elements.

Materials and Methods

We support most of our theoretical findings with simulations or computational examples in R using the following packages: *corHMM* (Beaulieu et al., 2013), *diversitree* (FitzJohn, 2012), *hisse* (Beaulieu and O'Meara, 2016), and *phytools* (Revell, 2012). The details on simulations are given, where needed, throughout the text, or in the scripts which are available in Supplementary Material at GitHub `https://github.com/sergeitarasov/Congruent_CTMCs`. The functions to reproduce the congruent transformations on the Markov models discussed in this paper are provided in the R package *rphenoscate* `https://github.com/uyedaj/rphenoscate`.

Results

This section summarizes the main types of Hidden Expansions (HEs) used throughout the paper to build and study congruent models for CTMCs. To show practical aspects of HEs, we use several motivating examples with simple CTMCs. One can think of them as those CTMCs which describe the evolution of some discrete character on a given tree. These simple processes were specifically selected for explanatory purposes. In a separate section, we discuss model congruence in SSE models, which is the same as in other simple CTMCs. Different types of HEs are necessary for understanding the structure of a congruence class, which we also discuss in a separate section. In the Discussion, we show that the structure of the congruence class is crucial for assessing model selection methods.

Hidden Expansion (HE)

Instead of aggregating and thereby constructing rate matrices with fewer states, the lumpability property can be used in the reverse direction. The state-space of any CTMC can be expanded by adding hidden states in a way that preserves lumpability. So, the expanded process can be collapsed back to the original one by lumping. The hidden states should be added to the observable ones without changing the number of observables. Hereafter, we refer to this transformation as hidden expansion (HE).

Consider a two-state CTMC $\{s_1, s_2\}$ defined by the matrix Q_{asy} with asymmetric rates shown in the equation (5). For example, we can think of it as an evolutionary process that generates characters on a known phylogeny. Suppose, we expand s_1 with two hidden states $s_1 = \{s_{11}, s_{12}\}$, then the new matrix Q_{he}, that has two observable and three hidden states, may look like the one shown in right hand side of the equation (5). Obviously, lumping Q_{he} with respect to the partition $\{\{s_{11}, s_{12}\}, s_2\}$ produces Q_{asy} because the transitions $s_{11} \to s_2$ and $s_{12} \to s_2$ maintain the RWR.

$$Q_{asy} = \begin{array}{c} \begin{matrix} s_1 & s_2 \end{matrix} \\ \begin{pmatrix} - & 0.3 \\ 0.2 & - \end{pmatrix} \end{array} \begin{matrix} s_1 \\ s_2 \end{matrix} \quad \cong \quad Q_{he} = \left(\begin{array}{cc|c} - & 0 & 0.3 \\ 0 & - & 0.3 \\ \hline 0.1 & 0.1 & - \end{array}\right) \begin{matrix} s_{11} \\ s_{12} \\ s_2 \end{matrix} \begin{matrix} s_{11} & s_{12} & s_2 \end{matrix} . \tag{5}$$

There are infinite ways to expand Q_{asy} with hidden states and obtain a matrix that would be lumpable back into Q_{asy} because there are infinite ways to maintain RWR. Both Q_{asy} and Q_{he} exhibit congruent processes (denoted by $\cong$). The congruence means that the likelihoods for Q_{asy} and Q_{he} are identical for any set of data because the probabilities for their observable states are the same [the equation (3), Supplementary Material]. In other words, the two processes are phylogenetically indistinguishable – they have identical observable state space and yield identical likelihood scores. The congruence property is symmetric and transitive – if a process X is congruent to Y, and Y is congruent to Z, then X is congruent to Z and Y. The set of congruent processes (or models) forms a congruence class (Louca and Pennell, 2020). Clearly, HE can be applied to any CTMC to produce congruent CTMCs with hidden states because any rate matrix can always be expanded to maintain the RWR. Thus, any particular CTMC process is a member of an infinite congruence class.

Classification of HEs and Parametric Representation

Parametric forms.— To derive parametric forms for congruent processes, we classify HEs into three main types depending on the number of parameters in a HE model compared to the original CTMC used to generate the expanded model. For this classification, we count only parameters located in the off-diagonal blocks (hereafter, off-diagonal parameters) of the HE matrix, but not the total number of the parameters in the model; because the off-diagonal parameters define the parametric form.

The three types of HEs are (1) equivalent expansion, (2) super-expansion, and (3) sub-expansion. They generate models with the same, greater, and fewer off-diagonal parameters, respectively. The parametric form for the equivalent and super-expansion

models derives from the original model by maintaining the RWR, as shown in the Lumpability section above. For the sub-expansion model, the parametric form is conditional on the specific realization of the rate values, and the general form does not exist.

To derive a parametric form, we simply substitute the same values in a numeric form with a unique parameter. For example, the natural parametric form for Q_{asy} process from the equation (5) is a two-parameter model M_{asy} shown in the equation (6). The equivalent expansion of Q_{asy} produces Q_{he} [the equation (5)] that is also a two-parameter model (M_{eq}) shown in the equation (6). Note, the parametric form indicates that the congruence between M_{asy} and M_{eq} holds for any value of α_1 and α_2. The example of a super-expansion for M_{asy} is shown in the equation (6) as well, it is a three-parameter model M_s. There are the following relationships between the original parameters and those in HEs: $\alpha_1 = \beta_1 = \gamma_1, \quad \alpha_2 = 2\beta_2 = \gamma_2 + \gamma_3$. The parametric forms for the sub-expansions are discussed in the following sections.

$$M_{asy} = \begin{array}{c} \begin{matrix} s_1 & s_2 \end{matrix} \\ \begin{pmatrix} - & \alpha_1 \\ \alpha_2 & - \end{pmatrix} \end{array} \begin{matrix} s_1 \\ s_2 \end{matrix} \quad \cong \quad M_{eq} = \left(\begin{array}{cc|c} - & 0 & \beta_1 \\ 0 & - & \beta_1 \\ \hline \beta_2 & \beta_2 & - \end{array}\right) \begin{matrix} s_{11} \\ s_{12} \\ s_2 \end{matrix} \quad \cong \quad M_s = \left(\begin{array}{cc|c} - & 0 & \gamma_1 \\ 0 & - & \gamma_1 \\ \hline \gamma_2 & \gamma_3 & - \end{array}\right) \begin{matrix} s_{11} \\ s_{12} \\ s_2 \end{matrix} . \tag{6}$$

Clearly, there are infinite ways to construct congruent equivalent-and super-expansion models by adding hidden states and/or parameters but maintaining the RWR. Thus, any CTMC model has infinite congruence classes consisting of expanded models.

Waiting-Times Preserving Models.— Also, we define two other categories of HE models and processes that are essential for analyzing the CTMC dynamics. The Waiting-Time Preserving (WTP) hidden expansion does not change the expected waiting times between the original model and HE. Note that the waiting time is indicated by the

negative main-diagonal entities in the rate matrix. Alternatively, the opposite is a non-Waiting-Time Preserving (nWTP) expansion that changes the expected waiting times between the original and HE models. For example, M_{eq} and M_s [the equation (6)] are WTP expansions. However, zeros in those rate matrices can be substituted with a separate parameter. This substitution would maintain the lumpability but would make them nWTP.

Sub-Expansions: Equal Rate Hidden Expansion (EHE)

Let us proceed further with hidden expansion and continue adding hidden states to the observable ones. Eventually, it is possible to reach a point when all allowed transitions have the same rate value. We call this transform Equal Rate Hidden Expansion (EHE). For the process Q_{asy} form the equation (5), it results in the following equal rate matrix (Fig. 2B):

$$Q_{ehe} = 0.1 \cdot \begin{array}{c} \begin{array}{ccccc} s_{11} & s_{12} & s_{21} & s_{22} & s_{23} \end{array} \\ \left(\begin{array}{cc|ccc} - & 0 & 1 & 1 & 1 \\ 0 & - & 1 & 1 & 1 \\ \hline 1 & 1 & - & 0 & 0 \\ 1 & 1 & 0 & - & 0 \\ 1 & 1 & 0 & 0 & - \end{array}\right) \end{array} \begin{array}{c} s_{11} \\ s_{12} \\ s_{21} \\ s_{22} \\ s_{23} \end{array} . \tag{7}$$

The observable states in this particular matrix are represented by two and three hidden states respectively; the transitions in the partition blocks located along the main diagonal are prohibited (i.e., 0), all allowed transitions have the same rate equal to 0.1. Lumping Q_{ehe} with respect to the partition $\{\{s_{11}, s_{12}\}, \{s_{21}, s_{22}, s_{23}\}\}$ produces Q_{asy} from the equation (5). Like before, Q_{asy} and Q_{ehe} are congruent CTMCs (Supplementary Material).

The EHE transform holds for any CTMCs regardless of the observable states and rate values because it stems from the general property of rational numbers. The proof of it is as follows. Imagine some original rate matrix Q whose off-diagonal rates form a set: $\{q_1, q_2, \ldots q_n\}$. These rates are some specific values that we assume to be rational numbers. We approximate them by rational number to any desired precision if they are not. Every

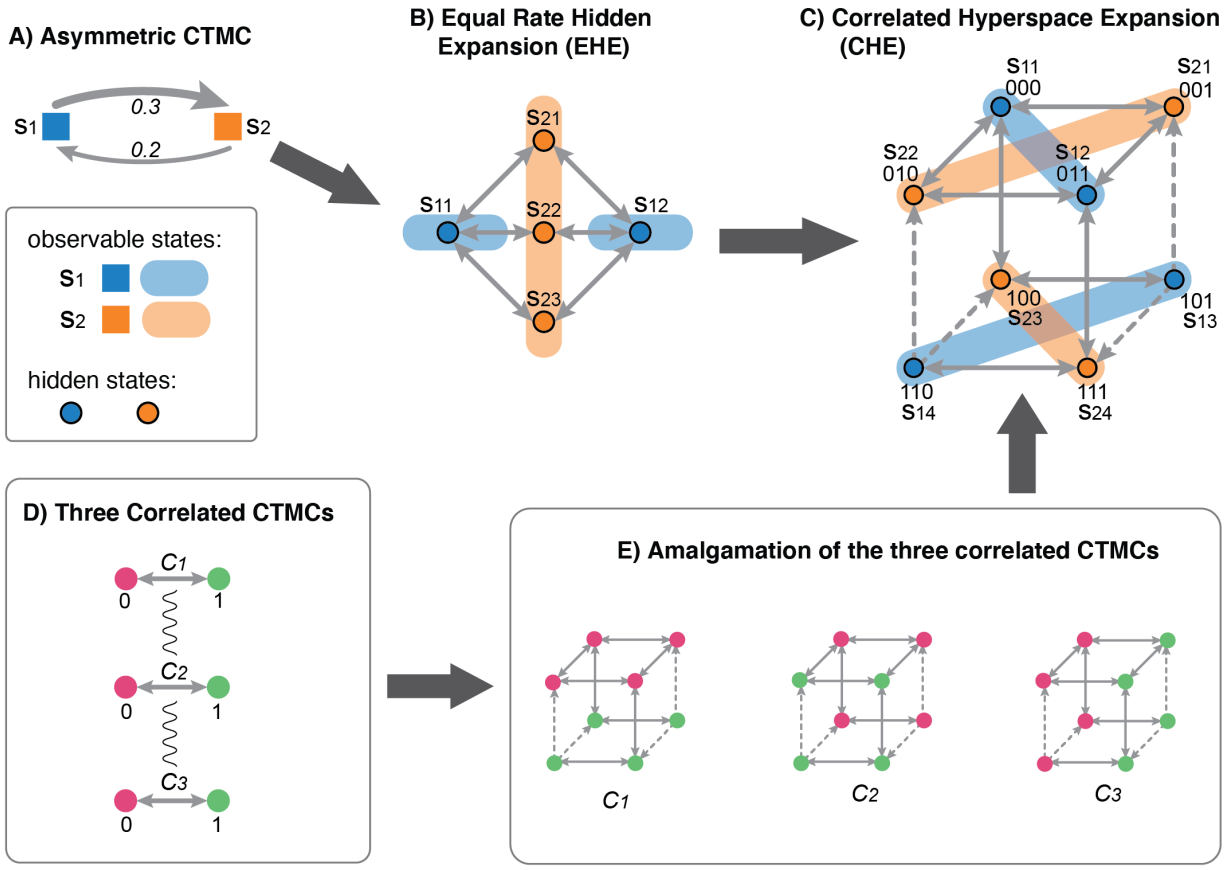

Fig. 2. Examples of Hidden Expansions. A) Original CTMC with asymmetric rates, the equation (5). B) EHE transformation, the equation (7). C) CHE transformation, the equation (9); dashed arrows indicate one-directional transitions. D) The set of three correlated binary characters whose amalgamation produces the graph in (C). E) The states belonging to each binary character from (D) are colored on three instances of the graph from (C).

rational number can be represented as a product of an integer (z), and a rational multiplier (q_{sin}) shared by all elements of the set, which gives $\{q_1 = z_1 q_{sin},\ q_2 = z_2 q_{sin},\ \ldots\ q_n = z_n q_{sin}\}$. The shared multiplier q_{sin} is the equal rate, while integers (z_i) are hidden states in EHE. To calculate them, we should represent those rates as irreducible fractions of two integers and find their least common denominator (lcd). So, the equal rate is $q_{sin} = 1/lcd$. The number of required hidden states for an observable state s_i can be calculated by taking the maximum rate value from the s_i's column in Q and dividing it by q_{sin}. For example, in Q_{asy} from the equation (5), the two rates $\{0.2, 0.3\}$ can be represented as fractions: $2/10$ and $3/10$; their $lcd = 10$, hence

$q_{sin} = 1/lcd = 0.1$. So, the number of hidden states per observable state is $0.2/0.1 = 2$ and $0.3/0.1 = 3$, respectively. Finally, the EHE matrix can be constructed using the calculated number of hidden states and appropriately arranging q_{sin} within the off-diagonal blocks to maintain the RWR [the equation (7)]. The redundant transitions within off-diagonal blocks and the transitions within main diagonal blocks should be set to zero.

The EHE transform exists for any CTMC process because all rate values in Q are rational (or can be approximated by rational numbers with high precision) and thereby have an lcd. The EHE transform for any given Q produces an infinite set of congruent processes because of the infinite number of common denominators for any set of fractions. To construct equal rate expansions, one can choose another common denominator (not lcd) and follow the aforementioned algorithm.

Note that it is not necessary to have zero-valued transitions along the main diagonal blocks to maintain lumpability. Within each such block, zeros can be substituted with any value that will elevate the expected waiting times and make the process nWTP. Yet another congruent nWTP might be constructed by substituting the main diagonal zeros to have equal expected times for all hidden states.

Parametric representation.— The parametric form for Q_{ehe} from the equation (7) can be written by substituting all $1s$ with an unknown parameter (β) which yield a one-parameter model M_{ehe} [the equation (8)]. So, EHE is a sub-expansion. Note, this model is congruent with the original process Q_{asy} [the equation (5)] but not the original model M_{asy} [the equation (6)]. If one takes the original model M_{asy} but different rate values (e.g, $\alpha_1 = 0.3$ and $\alpha_2 = 0.1$), then this setting would require an EHE with a different number of hidden states (e.g., three and one respectively). Thus, the parametric form of EHE or any other sub-expansion is not defined for any original model since the number of hidden states depends on specific rate values. It happens because sub-expansions imply stricter constraints on the rate parameters compared to the original model. For example, if we lump M_{ehe} with respect to the observable states then the

resulting model is M_{lump} [the equation (8)].

$$M_{ehe} = \begin{array}{c} \begin{array}{ccccc} s_{11} & s_{12} & s_{21} & s_{22} & s_{23} \end{array} \\ \left(\begin{array}{cc|ccc} - & 0 & \beta & \beta & \beta \\ 0 & - & \beta & \beta & \beta \\ \hline \beta & \beta & - & 0 & 0 \\ \beta & \beta & 0 & - & 0 \\ \beta & \beta & 0 & 0 & - \end{array}\right) \end{array} \begin{array}{c} s_{11} \\ s_{12} \\ s_{21} \\ s_{22} \\ s_{23} \end{array} \quad \cong \quad M_{lump} = \begin{array}{c} \begin{array}{cc} s_1 & s_2 \end{array} \\ \begin{pmatrix} - & 3\beta \\ 2\beta & - \end{pmatrix} \end{array} \begin{array}{c} s_1 \\ s_2 \end{array} . \tag{8}$$

The model M_{lump} is a one-parameter model where rates are scaled by specific factors, it is a submodel of the original two-parameter model M_{asy} [the equation (5)]. The lack of a defined parametric form and stricter rate constraints is a general property of sub-expansion, where the number of the off-diagonal parameters is fewer than that in the original model. Although the parametric forms are undefined for an original model, they are defined when inferring likelihood since data exhibit the realization of some particular parameter values.

Sub-Expansions: Correlated Hyperspace Expansion (CHE)

We further elaborate the EHE process from the equation (7), by applying our favorite operation – adding hidden states. However, now we add them without altering the rate values. For example, we can add another two and one hidden states to each observable state, respectively. To avoid rows and columns with all rates equal to zero (i.e., prohibited), we rearrange non-zero rates to preserve the RWR. There are many ways to do this, and all of them produce equivalent results for our purpose. One such rearrangement gives the following process (Q_{che}) where each observable state consists of four hidden ones:

$$Q_{che} = 0.1 \cdot \begin{array}{c} \begin{array}{cccc|cccc} s_{11} & s_{12} & s_{13} & s_{14} & s_{21} & s_{22} & s_{23} & s_{24} \end{array} \\ \left(\begin{array}{cccc|cccc} - & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & - & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & - & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & - & 0 & 1 & 1 & 1 \\ \hline 1 & 1 & 0^* & 0 & - & 0 & 0 & 0 \\ 1 & 1 & 0 & 0^* & 0 & - & 0 & 0 \\ 1 & 0 & 1 & 0^* & 0 & 0 & - & 0 \\ 0 & 1 & 0^* & 1 & 0 & 0 & 0 & - \end{array}\right) \end{array} \begin{array}{c} s_{11} \\ s_{12} \\ s_{13} \\ s_{14} \\ \\ s_{21} \\ s_{22} \\ s_{23} \\ s_{24} \end{array} . \quad (9)$$

Lumping all hidden states in this matrix produces the initial CTMC, Q_{asy}, from the equation (5). Clearly, both Q_{asy} and Q_{che} are congruent processes (Supplementary Material).

One may notice that Q_{che} defines a correlated evolution of three binary CTMCs (i.e., $\bigodot^3 Q^2$). To demonstrate it, we need to rearrange the rows and columns in Q_{che} accordingly, which is visualized in Fig. 2C. The correlation with respect to the binary CTMCs arises due to the prohibited transitions marked with "*" in the equation (9). These transitions break the independent evolution between the three CTMCs but are necessary to maintain the lumpability of Q_{che}. If they were allowed (with a rate equal 0.1), it would imply an independent evolution for the binary CTMCs but violate the congruence of CHE with Q_{asy}. Note, each observable state of Q_{che} mismatches the observable states corresponding to the binary CTMCs (Fig. 1C). So, those three CTMCs can be thought of as correlated hidden factors (or characters) whose interaction generates the new observable state space (Fig. 2D-E).

Hereafter, we refer to the proposed transform as Correlated Hyperspace Expansion (CHE). To find CHE for some binary matrix Q, apply EHE to Q, and find an observable state with the maximum number of hidden states; denote this number by N_{max}. Any amalgamation of binary CTMCs ($\bigoplus^n Q^2$), where $n \geqslant N_{max}$, embeds the EHE matrix. Use this $\bigoplus^n Q^2$ by populating allowed transitions with the equal rates found for EHE. Prohibit certain rates in $\bigoplus^n Q^2$ to construct $\bigodot^n Q^2$ and maintain the RWR (as shown above). This yields the desired CHE transform. The derivation of CHE for Q with more than two states

is less trivial.

In fact, any CTMC process transforms via CHE into a congruent and equal-rate matrix that defines the correlated evolution of certain underlying CTMCs (Appendix A1). The number of those underlying CTMC characterizes the dimension in the CHE process and depends on the initial rate values. If the initial CTMC has s states then its CHE transform results in a hyperdimensional CTMC on s elements ($\bigodot^n Q^s$). For example, transforming a three-state CTMC via CHE means amalgamating a certain number of three-state CTMCs ($\bigodot^n Q^3$). One way to interpret CHE is to think of any given CTMC coming from some higher dimensional space where the unobserved hidden factors evolve in a correlated fashion and generate the observable CTMC.

The parametric representation for CHE is analogous to that for EHE since CHE is a sub-expansion.

HEs and State-dependent Speciation and Extinction (SSE) Models

All aforementioned types of HEs apply to CTMCs with infinite state space as in simple birth-death or SSE models, because lumpability is the general property of Markov chains (Appendix A2). Thus, an infinite set of congruent models exists for any SSE model used for reconstructing lineage-specific diversification. The general SSE model consists of two components: (i) a birth-death process that generates extinctions and speciations on a tree; and (ii) a CTMC that generates evolution in observed and hidden traits, which may or may not affect the the birth-death process. The hidden expansions can be constructed for each of these components. The properties of these expansions are the same as described above for the finite state CTMCs. For example, any complex SSE process with variable rates for trait evolution and diversification can be represented by a congruent process with equal rates for both components (i.e., as EHE or CHE, Appendix A2.1). Note, that the hidden expansion for the diversification process requires a dual transition — a simultaneous speciation/extinction event and change of the trait's state. The standard SSE

models currently used in phylogenetics (e.g., BiSSE, HiSSE) do not allow it, but can be extended to include it.

The Structure of a Congruence Class

Every congruence class (a set of all congruent models) generated by HEs, consists of a unique non-lumpable model (M_0) that cannot be lumped further and an infinite set of lumpable models generated by expanding M_0 with hidden states (Fig. 3). We call M_0 the irreducible model and its states the irreducible states; M_0 can be with or without hidden states. Clearly, if M_0 has hidden states, then it characterizes a process that does not have a simpler representation (i.e., without the hidden states). Thus, all other models in the class have more hidden states than M_0. Note that the congruence class can be reconstructed for any given model. For example, if M_0 is known, then the class is all HEs of M_0. If some other model is known, lump it into M_0 and follow the previous step.

The lumpable congruent models fall into three types – equivalent, super-expansion, and sub-expansion – depending on the type of HE that generates them. The parametric forms for the equivalent and super-expansion models can be derived directly from M_0 using the RWR. The total number of parameters in this model is the same (e.g., WTP equivalent type models) or greater (e.g., nWTP equivalent or super-expansion models). Since there is no upper bound to maintain lumpability, the number of parameters in super-expansion models can grow to infinity.

The parametric forms for the sub-expansion models are conditional on the specific parameter values of M_0. It means that every particular point in the parameter space of M_0 has some unique sub-expansion model. The total number of parameters in sub-expansion models might be: (1) less than in M_0 (e.g., WTP models), all the way down to a single parameter (EHE, CHE), indicating that such models are nested within M_0; or (2) the same and greater than in M_0 (e.g., nWTP models).

The irreducible model M_0 is congruent to all equivalent and super-expansion

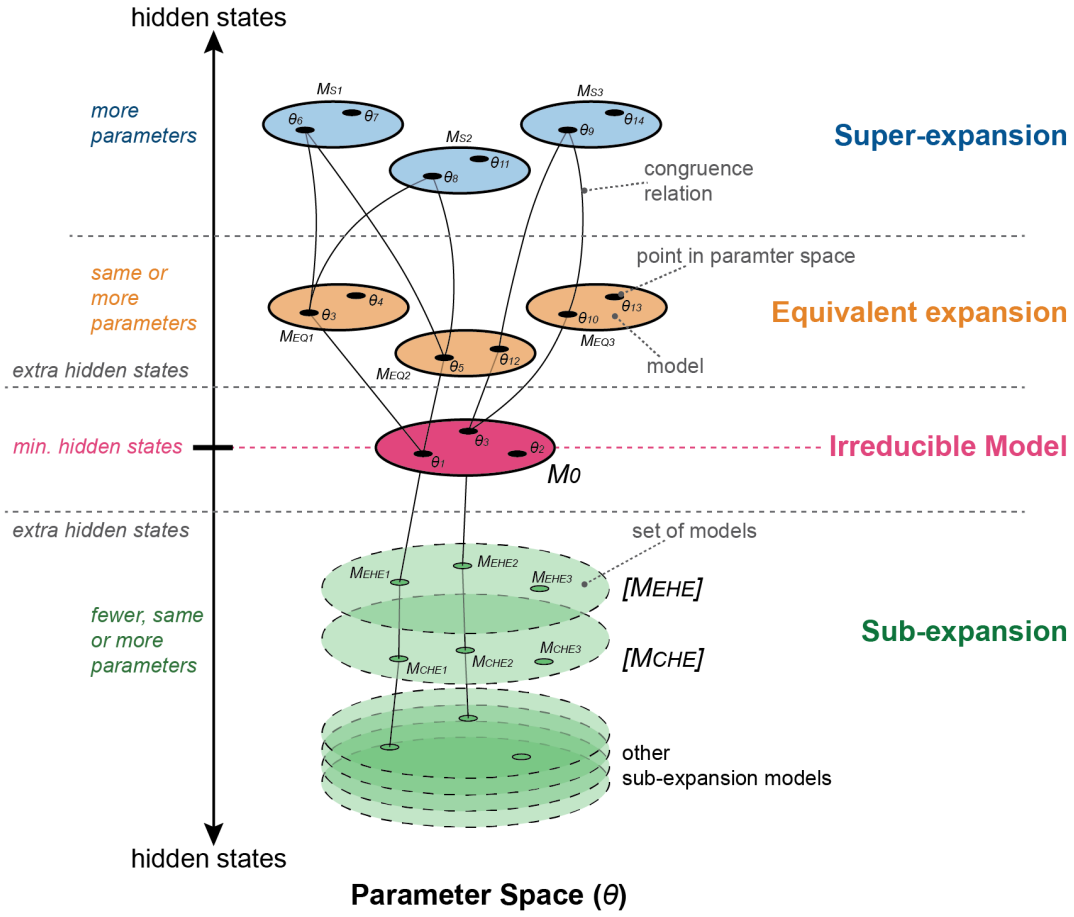

Fig. 3. Structure of Congruence Class generated by hidden expansion. Note, M_0 is congruent to all equivalent and super-expansion models (e.g, $M_0 \cong M_{EQ1} \cong M_{S1}$). Only specific parameter values of M_0 are congruent to sub-expansion models (e.g, $M_0(\theta_1) \cong M_{EHE1}$) because full M_0 is congruent to a set of sub-expansion models (e.g, $M_0 \cong [M_{EHE}] \cong [M_{CHE}]$); see text for details.

models. However, it is not congruent directly with any particular sub-expansion model (e.g., EHE or CHE) because the parameter space of M_0 is larger than that in the sub-expansion models. Instead, M_0 is congruent to some sets of them, for example, the set (denoted by $[\cdot]$) of CHE and EHE models, $M_0 \cong [EHE] \cong [CHE]$. Thus, the congruence class has an infinite number of paths that connect congruent points in parameter space across different models. These points assemble into models whose congruence depends on their type. Each congruence class has an invariant quantity that is constant across all its models – this is the row-wise sum of rates with respect to the irreducible states in M_0 that

obeys the RWR.

Discussion

Model Congruence in Phylogenetics

We have shown the general technique for constructing congruent models by hidden expansion – adding hidden states to a CTMC and maintaining its lumpability property through the RWR. Any CTMC has an infinite set of congruent models, which can be constructed by different types of HEs (equivalent, super-and, sub-expansions). These expansions result in congruent models, in which the number of parameters can vary from one to infinity for any original model. The sub-expansion models have a narrower parameter space, and their parametric form depends on the specific rate values in the original model. The number of the hidden states and the original rate values is confounded in the sub-expansion.

Congruent models are a general property of Markov chains and, thereby, apply to all kinds of CTMCs used in phylogenetics for modeling discrete traits, DNA, proteins, and species diversification (SSE models). Given this generality, the proposed congruence holds in models treating phylogeny as an unknown parameter. Thus, the model congruence discovered in the time-dependent diversification process by Louca and Pennell (2020) also exists in the general CTMCs.

We need to admit that other types of congruence also exist in CTMCs. We discuss them at the end of this paper. They are less tractable mathematically but we did not identify any problems they pose for phylogenetic analysis as in, for example, model selection.

Comparison of HE and LP Congruence.— Based on the current state of knowledge, our congruence classes by HEs seem quite similar to those described by LP. The congruent models result in the same likelihood regardless of the data. They both

contain an infinite set of models. They have an invariant quantity in their rates – pulled speciation and diversification rates in LP class and sums of the transition rates (the RWR) in HE class. Importantly, like diversification models, the dynamics of alternative models within the same class varies substantially (see the next section). Moreover, standard information criteria are not applicable for model selection in both classes (see next section). However, the LP classes are mathematically challenging since they do not have closed-form expressions. Conversely, the HE classes can be systematically constructed by using lumpability and hidden states. Therefore, we do not know whether the general structure of the LP and HE classes is similar enough to apply the same principles to both. For example, whether in LP classes, there exists a unique element that would be analogous to our irreducible M_0. This uncertainty emphasizes that congruence properties require further research in phylogenetics.

Alternative Evolutionary Scenarios

The congruent models are asymptotically unidentifiable but, as noted by LP, may exhibit markedly different evolutionary dynamics, which is also the rule for the HE models. They have two behaviors depending on whether the model preserves the waiting times (WTP) or not (nWTP) compared to the irreducible model (M_0) from its congruence class. We exemplify these dynamics by mapping the evolutionary history of the simple two-state character, Q_{asy} [the equation (5)] on a known phylogeny (Fig. 4). Here, Q_{asy} is the irreducible model. All WTP models within a class have the same number of state changes over a tree compared to M_0, indicating that they and M_0 have the same average evolutionary rate (Fig. 4A-B). However, WTP models have larger state space because of the hidden states. Recoloring the hidden states into the original ones (i.e., lumping) returns the character history for M_0.

The nWTP models have more state transitions, compared to M_0, and thereby a faster evolutionary rate that can be as fast as one desires since there is no limit (Fig. 4C).

Moreover, this rate can be unbalanced between the observable states, for example, slow in one and ultra fast in the other (Fig. 4D). However, recoloring the hidden states also returns the history for M_0. This distinct scenario may indeed lead to drastically different biological interpretations as regards to the underlying evolutionary rates.

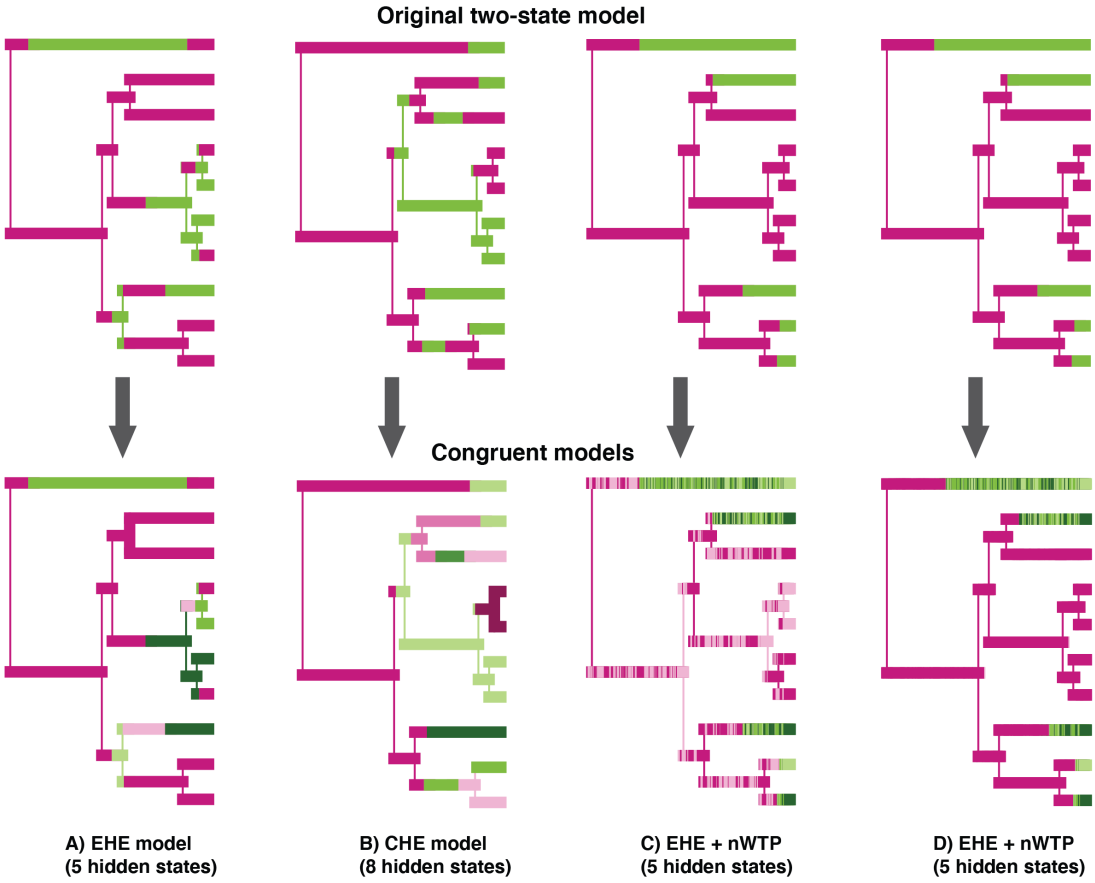

Fig. 4. Alternative evolutionary scenarios for congruent models. The top row shows stochastic maps for the original binary character (green and purple states) from the equation (5). The bottom row shows alternative scenarios for models congruent to the original character. A) EHE transformation, from equation (7). B) CHE transformation, from equation (9). C) CHE with equally elevated waiting time between the observable states (i.e, nWTP model). C) CHE with elevated waiting time in only one state (green).

Analogous behavior would hold for SSE models and diversification rates as they can be similarly mapped onto phylogeny with time intervals on tree branches colored according to speciation and extinction rates. However, the available SSE software does not allow dual transitions; hence, full HEs cannot be applied to SSE models for this demonstration.

Another type of congruence between alternative scenarios may mislead hypotheses

testing. In SSE models, two alternative hypotheses are usually tested. One implies that diversification does not depend on the observable trait, known as the character-independent (CID) model (Beaulieu and O'Meara, 2016). The other implies that diversification is character-dependent. Interestingly, the CID model is congruent with some (but not all) character-dependent models, and they have the same number of parameters (Appendix A2.2 and Supplementary Material). Thus, the standard model selection procedure cannot distinguish the two hypotheses.

Selecting among congruent models

LP suggested that the models within the same congruence class are not identifiable and equally possible, precluding reliable inference of the diversification dynamics in time-dependent birth-death models.

Given a similar indentifiability issue in CTMCs, can we select among the alternative models in a HE congruence classes? We believe that we can, but not using the standard model selection. The standard model selection criteria as the Akaike Information Criterion [AIC, (Akaike, 1981)] or the Bayesian Information Criterion [BIC, (Schwarz, 1978)] do not work for this problem. They would always select sub-expansion models because these models have fewer parameters in a congruence class. Although these models look simple (e.g., one-parameter EHE or CHE), their parameter space is very narrow, and the state space is complex, rendering them biased and having little explanatory power.

If the EHE or CHE is selected based on AIC or BIC, we would have a situation where a separate model would describe every single dataset with an almost unique and large number of hidden states. If, we were to follow this route, then it would require we discuss why one binary character has a hundred hidden states, and another character has four hundred hidden states. Such questions may not make sense biologically unless we have access to alternative means for studying the underlying state space. Instead, explaining this process using variable transition rates in a binary character (without hidden states)

will likely represent much more interpretable evolutionary information. Therefore, we can rule out the sub-expansion models from the candidate list.

Using similar logic, we can straightforwardly remove the equivalent and super-expansion models from the selected candidates list because they have more parameters or hidden states than M_0. Thus, the only reasonable candidate model becomes M_0 itself.

The same principles apply to hypotheses testing in SSE. We may prefer selecting the CID model (that is M_0) since it provides a simpler explanation of data instead of selecting the congruent character-dependent model. Thus, following the parsimony principle that embraces the penalty on the number of hidden states and parameters, we can unambiguously select M_0 – the unique irreducible model in a congruence class. This parsimony principle is also backed by the biological reasoning outlined above.

Our main recommendation is to always use irreducible models for model comparison. When comparing models between different congruent classes, one needs to ensure that they are irreducible. It is trivial if a model has no hidden states because all such models are irreducible. However, one should always check if a smaller irreducible model may exist for the models with hidden states using the RWR.

In fact, irreducible models have been used in phylogenetics for decades (i.e., our standard models) and seem to often have been quite successful in explaining the observed data. A priori selection of an irreducible model means that even if the true process was evolving under some other model (non M_0), we would never be able to identify it. But do we really need it? Possibly not; if following parsimony, we assume that other models are less likely to occur in nature (Morlon et al., 2022).

General considerations on model penalization and statistical regularization for congruent models are overviewed in Morlon et al. (2022). We agree that this is the way forward to handle the congruence issue described by LP.

Solving LP congruence with the HE principles? — If we assume that the HE congruence is similar enough to the LP congruence, what the analogous parsimony principle would look like for the LP problem?

Imagine that the rate matrices for our congruent models are unknown. We even do not know how to write them on paper. However, we have an algorithm that can reconstruct mapped evolutionary histories for our congruent models given data, as shown in Fig. 4. We observe that congruent histories have different dynamics but the same likelihood. In this case, the complexity penalty would refer to the total number of states and the frequency of state changes in a mapped history, which is the same as the penalty derived using the rate matrices. Thus, the analogous metric of complexity for a time-dependent birth-death process in LP might be some kind of curvature for the extinction and speciation functions. This metric would select a model with the smoothest functions in a congruence class [see, Magee et al. (2020) for one implementation of this philosophy].

Curvature of Adaptive Landscape Explains Macroevolutionary Rate Variation

We believe that the provided hidden expansions may open up new avenues for studying evolutionary processes. For example, CHE transformation may explain rate variation at the macroevolutionary scale through a prism of microevolutionary processes, data, and biological knowledge. The "holey adaptive landscape" model (Gavrilets, 1997; Gavrilets and Gravner, 1997) suggests that genotypic space is a high dimensional graph where viable genotypes form clusters connected by a chain of one-step mutations and surrounded by "holes" – volumes of the space with low fitness (Fig. 5A). Evolution might be seen as a random walk of a species or population on such multidimensional graphs over time. Topologically, these graphs are identical to those that result from CHE transformation [e.g., Fig. 2C and Gavrilets and Gravner (1997)]. Interpreting hidden factors in CHE as individual genetic elements, while its prohibited transitions as holes, suggests that CHE exhibits an analog of a holey adaptive landscape. Although the

"mutation" rate in CHE is constant, lumping it for the observable states returns a heterogeneous Markov process. It happens because the evolutionary walk of species or population between the hidden factors proceeds at a constant rate. However, this walk should constantly go around the holes (i.e., prohibited transitions), rendering its trajectory a hyperdimensional curve. The observable states exhibit a projection of the high-dimensional hidden space into our low-dimensional world. Projecting the curved walk onto the observable states forms a process that evolves at variable rates (Polly, 2008).

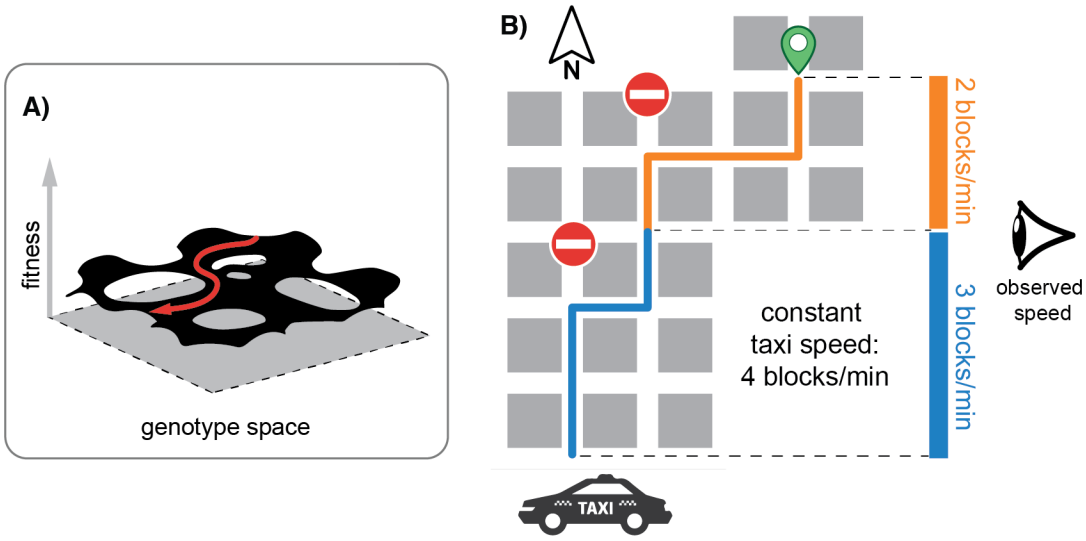

Fig. 5. A) The conceptual figure of the holey adaptive landscape, from Gavrilets (2003); the red arrow shows the curved population's path. B) A taxi driving around downtown at constant speed; its path observed from the south-north direction gives an illusion of speed shift. Thus, macroevolutionary rate heterogeneity can be recast as curvature of the state-space resulting from the underlying topology, or macroevolutionary architecture, of trait space.

A simple metaphor would help train an intuition of the curved walk. Imagine a taxicab driving around downtown at a constant speed (Fig. 5B). Due to road constructions (= holes), the taxicab can only travel on certain streets to reach the final destination. Its movement within the two-dimensional downtown landscape is constant by our assumption. However, if one observes its movement in one dimension alone – south-north direction – then it appears to happen at different speeds due to the curved driving path. This is similar to what CHE transform implies. Thus, the constant-rate dynamics on the holey

adaptive landscape at the microevolutionary scale may generate rate heterogeneity over larger time intervals at the macroevolutionary scale. Given the generality of CHE, this phenomenon equally may apply to molecular evolution, traits, and diversification.

In the literature it is common to explain macroevolutionary rate variation by temporal changes in biotic and abiotic factors. Our theoretical finding indicates that rate variation at macroevolutionary scale might be explained by the topology or "evolutionary architecture" of adaptive landscape alone.

At the macroevolutionary scale, multiple factors will change over time – the selective environment, genetic architecture, development, mutation rates, life history, etc. Each of these may map onto additional hidden factors that may alter transition rates between states or expand the state space of evolutionary change in our framework. As the environment changes, so too does the "holey adaptive landscape,"–expanding the dimensionality of the state space further. Over millions of years of macroevolutionary change, evolutionary walks traverse this vast "macroevolutionary landscape" during trait evolution, shedding light on its larger topology and accessibility. When using only macroevolutionary data, many of these different factors–including the observational process by the researcher's construct characters [see, the Two-Scientist Paradox in Tarasov (2019)] are confounded, and many of them may lead to the appearance of heterogeneity in rate. However, integrating knowledge of "microevolutionary topologies" known through developmental biology, mutation screening, trait variation within and among populations, and selection experiments, for example, can begin to constrain congruence classes to the set of macroevolutionary topologies that are consistent with the observed data. We believe this will provide a pathway for integrating microevolutionary data into macroevolutionary studies by using micro and macroevolutionary data jointly to "learn" the overarching topology of macroevolutionary landscapes. If macroevolutionary dynamics change, for example, with rate shifts in particular clades–candidate factors can be investigated within groups to identify plausible causes that may further expand the state space of

macroevolution. Such causes can range from trivial (improperly lumping distinct states into a single state, e.g., the wings of birds and insects) to elucidating dramatic changes in the evolvability of a character due to developmental reorganization.

Thus, we see two primary advantages to recognizing and exploring the congruence classes of Markov models. First, we can better understand the limits to macroevolutionary inference and enumerate the large family of candidate processes that can explain a given macroevolutionary pattern by exhaustively or heuristically exploring higher dimensional state spaces [as in Höhna et al. (2022)]. But second, within those congruence classes can be found topologies that map onto specific biological hypotheses for the structure of evolutionary change, each of which may make specific predictions, for example, for how traits are constructed developmentally or evolve in response to selection at the microevolutionary scale. Thus, while the apparent widespread existence of congruence classes across comparative models is a substantial challenge to macroevolutionary inference, full recognition and understanding of this higher dimensional state space may nevertheless provide the key insights needed to connect macroevolutionary models to the causes and underlying processes of evolutionary change.

Are there other congruence classes?

CTMCs have other types of asymptotically unidentifiable congruent models besides those reviewed here, based on the lumpability property. Some of such models can be constructed given that $e^X e^Y = e^{(X+Y)}$, if the matrices commute: $XY = YX$. For example, imagine a homogeneous CTMC that we denote as $\hat{Q}$ with all transitions equal to one (i.e, $\hat{q} = 1$); it evolves over a tree branch of length τ, with probability $P(\hat{Q}, \tau) = e^{\hat{Q}\tau}$. This CTMC is congruent to a heterogeneous CTMC consisting of two rate matrices Q_1 and Q_2, in which the parameters (q_1 and q_2, respectively) are different between the matrices but equal within them following this relationship $q_1 + q_2 = 2$; for example, the rates can be: $q_1 = 0.5$ and $q_1 = 1.5$; apparently, Q_1 and Q_2 commute. The heterogeneous CTMC evolves

under Q_1 from the start ($\tau = 0$) to the half-time ($t = \tau/2$), and under Q_2 from the half-time to τ; its probability is $P(Q_1, Q_2, \tau) = e^{(Q_1+Q_2)\frac{\tau}{2}}$. Clearly, the probabilities are the same for these two processes $P(\hat{Q}, \tau) = P(Q_1, Q_2, \tau)$, and hence the likelihoods are the same too. There are infinite ways to construct a congruent heterogeneous process from a homogeneous one, following the properties of matrix commutativity. Obviously, in such trivial cases, the heterogeneous CTMC will always have more parameters than the homogeneous CTMC, which does not pose any problem of model selection since the homogeneous model can always be chosen due to its simplicity. Even if the true process was heterogeneous, we would never be able to detect it. We must assume that the homogeneous CTMC represents the average rate of the true process that we consider the simplest description. Therefore, this trivial congruence is less interesting than the HE congruence at this moment.

Things become more complicated if the matrices do not commute (i.e., $XY \neq YX$). Then, the homogeneous rate matrix ($\hat{Q}$) may or may not exist. This issue is related to the multiplicative closure of Markov models (Sumner et al., 2012b) that has been studied using interesting methods of Lie group theory (Sumner et al., 2012a; Woodhams et al., 2015). These methods might be promising for studying model congruence in general. In this type of congruence, the problem of model selection may arise if it is possible to show that some homogeneous process ($\hat{Q}$) can be represented by a set of rate matrices that have fewer parameters in total than $\hat{Q}$. In that case, this would make the model selection ambiguous. We do not know any such examples, but this issue requires further research.

Another class of congruent models for general CTMCs can be constructed using weak lumpability that holds under specific values of the initial probability vector (Kemeny and Snell, 1960). The earlier study (Tarasov, 2019) showed that the weak lumpability does not generally hold in the CTMCs operating on phylogenetic trees. However, weak lumpability is less tractable mathematically, and it is possible, but unlikely, that some specific kind of it may apply to phylogenetics. Thus, model congruence needs additional

research to identify cases that may pose ambiguity for model selection or prove that such cases do not exist.

Future Perspectives

This paper showed that congruent models exist in all types of CTMCs used in modeling various phylogenetic phenomena. The HE can be used to construct infinite congruence classes that include models with a varying number of parameters. However, we show that the model selection within the same congruence class is not particularly problematic by penalizing model complexity based on mathematical and biological reasoning. We also believe that similar methods that penalize for complexity, use priors, or other regularization (Morlon et al., 2022; Magee et al., 2020) techniques should be adopted for the congruence occurring in the time-dependent diversification process discussed in LP. Thus, model congruence does not generally pose a problem but a challenge that helps us better understand statistical models and underlying structure of evolutionary change. It can be potentially used as a tool for understanding evolutionary processes, as we demonstrated using the holey adaptive landscape evolution. Thus, we call for additional research in congruent models to better understand relationships between congruence classes and classify other types of congruence that may occur in CTMCs.

Funding

ST was supported by the Academy of Finland grant: 339576, and the three-year grant from University of Helsinki; ST partially conducted this work while a Postdoctoral Fellow at the National Institute for Mathematical and Biological Synthesis sponsored by the NSF Award DBI-1300426.

JCU was funded by NSF-DEB-1942717 and NSF-DBI-1661516.

Appendix

A1. Proofs for CHE

In the proofs below, consider thinking of any CTMC as of a graph with states (nodes) and transitions between them (edges). Every such a graph is specified by a rate matrix. Coloring graph states corresponds to splitting the states into groups which refer to the observable states (different colors means different observable states). We say that a graph can be colored (or coloring is possible) if no two adjacent vertices (i.e., those connected by the same edge) are of the same color. We provide two proofs for the existence of CHE, one in Q^2 (trivial) and another in Q^s.

A1.1 Proving CHE for a two-state CTMC.— Any two-state CTMC with asymmetric and defined rates (Q_0^2) can be represented by a hyperdimensional CTMC with hidden states ($\bigodot^n Q^2$). The hyperdimensional CTMC characterizes correlated evolution of n two-state CTMCs (Q^2). Both Q_0^2 and $\bigodot^n Q^2$ are congruent processes.

Proof.— Suppose $\bigoplus^n Q^2$ describes evolution of n independent CTMCs with equal transition rate q. The state space of $\bigoplus^n Q^2$ is a hypercube with dimension n. The number of states in $\bigoplus^n Q^2$ is 2^n, every state is labeled by the combination of states (i.e., 0s and 1s) from n initial Q^2. For example, for $\bigoplus^2 Q^2$, the state labels are $\{00, 01, 10, 11\}$.

Every hypercube graph can be colored with two colors because all hypercube graphs are bipartite. The colors are two observable states. This coloring may be constructed by giving one color to those labels that sum up to an even number and the other color to those that sum up to an odd number. Arranging the rows and columns in the rate matrix $\bigoplus^2 Q^2$ by colors splits the matrix into four blocks that correspond to transitions between the observable states (i.e., colors). Main diagonal blocks in the arranged $\bigoplus^2 Q^2$ has all transitions equal to 0s (by the coloring condition). The transitions between the hidden states are allowed only in the off-diagonal blocks, the number of

possible transitions per hidden sate equals to the dimension n. Obviously, $\bigoplus^2 Q^2$ is lumpable, with the respect to the two colors, into a binary CTMC with symmetric transition rate nq (by the RWR). It is always possible to modify one off-diagonal partition block in $\bigoplus^2 Q^2$ by prohibiting certain allowed transitions (i.e. making them 0) but, at the same, time maintaining the RWR within this block; this implies setting the number of allowed transitions for each hidden state in the modified block to k, $(k < n)$. Such a transformation results into $\bigodot^n Q^2$ that describes evolution of n-correlated binary CTMCs. Obviously, $\bigodot^n Q^2$ is also lumpable (by the RWR) with the respect to the two colors and produces some binary CTMC, Q_0^2, with asymmetric rates nq and kq. Apparently, the construction of $\bigodot^n Q^2$ is flexible: the variables n, q and k can be chosen so that the rates nq and kq become any desired rational numbers. Moreover this construction is not unique for Q_2. This implies that any Q_0^2 is congruent to $\bigoplus^n Q^2$.

A1.2 Proving CHE for a CTMC with any number of states.— An s-state CTMC with variable and defined rates (Q_0^s) can be represented by a hyperdimensional CTMC with hidden states ($\bigodot^n Q^s$). The hyperdimensional CTMC characterizes correlated evolution of n s-state CTMCs (Q^s). Both Q_0^s and $\bigodot^n Q^s$ are congruent processes.

Proof.— The proof is similar to above. Suppose, there is $\bigoplus^n Q^s$ that describes evolution of n independent CTMCs (Q^s) with equal transition rate q; the number of states in $\bigoplus^n Q^s$ is s^n. The main idea is to show that amalgamated $\bigoplus^n Q^s$ can be colored by s colors that correspond to observable s states. Since the coloring is possible (see the next section) then arranging the rate matrix of $\bigoplus^n Q^s$ by colors results in s^2 matrix blocks. The main diagonal blocks have all transitions equal to 0s (except for the negative transitions); the off-diagonal blocks have equal number of transitions per each hidden state that is n. Lumping with the respect to them produces s-state rate matrix with equal transition rate qn. Obviously, $\bigoplus^n Q^s$ can be transformed into $\bigodot^n Q^s$ by prohibiting some transitions to match the rate values of Q_0^s and maintain the RWR (as in the previous proof). Thus,

$\bigodot^n Q^s$ can lumped into Q_0^s. This implies that any Q_0^s is congruent to $\bigodot^n Q^s$.

Proof of the existence of coloring for $\bigoplus^n Q^s$.— Constructing $\bigoplus^n Q^s$ means amalgamating Q^s with itself $n-1$ times (by the definition of $\bigoplus^n Q^s$) that can be done recursively. To make $\bigoplus^2 Q^s$, take Q^s and uniquely enumerate each state with one of s natural numbers $\{1, 2, \ldots, s\}$. This classifies states in enumerated classes. Make $s-1$ copies of Q^s, which in total gives s copies of Q^s and s copies of states within each enumerated class. To complete the amalgamation, create edges (i.e., transitions with rate q) between all states within the same enumerated class (because all state transitions are allowed in Q^s). This gives $\bigoplus^2 Q^s$.

Now, we show that coloring exists for this amalgamation. Color each copy of Q^s identically by s colors $\{c_1, c_2, \ldots c_s\}$; this coloring maps the set $\{c_1, c_2, \ldots c_s\}$ to the enumerated state classes $\{1, 2, \ldots, s\}$. Leave one copy of Q^s intact and permute colors in the remaining $s-1$ copies. For this, construct a permutation $\sigma = (c_1, c_2, \ldots c_s)$ that sends c_1 to c_2, $\ldots$, c_{s-1} to c_s, c_s to c_1. Note, permutation power (σ^k) means applying σ sequentially k times. For example, applying σ and σ^2 to the set of the three colors $\{c_1, c_2, c_3\}$ gives the following reordered sets: $\{c_2, c_3, c_1\}$ and $\{c_3, c_1, c_2\}$. The permutation σ has a cycle of length s indicating that σ^s returns the identity set. Thus, applying σ^k while incrementing k from $k = 1$ to $k = s-1$, produces $s-1$ sets with unique (up to permutation) color order. Now, apply the set of permutations $\{\sigma, \sigma^2, \ldots, \sigma^{s-1}\}$ to the set of the remaining copies of Q^s (there are $s-1$ copies). The first copy (left intact) will preserve the original coloring, while the remaining $s-1$ copies will have colors permuted. This will result in each of s copies of Q^s having unique order of colors, which also implies that the states within the same enumerated class (across the copies) are differently colored. Note, in the amalgamated $\bigoplus^2 Q^s$ the states are adjacent only if they are within the same copy of Q^s or if they belong to the same enumerated class. All states within the same copy have different colors (by the initial condition) and all states within the same enumerated class have different colors too, due to the permutation. Thus, no two adjacent states have

the same color and coloring is possible.

Using the same recursive technique can construct higher dimensional amalgamations. For $\bigoplus^{d+1} Q^s$, take $\bigoplus^{d} Q^s$, create a new enumeration for its s^d states, create $s-1$ copies of $\bigoplus^{d} Q^s$, link edges within the same enumerated class, use the coloring obtained for $\bigoplus^{d} Q^s$ but permute it for $s-1$ copies using the powers of σ. This will yield the appropriately colored $\bigoplus^{d+1} Q^s$. Since the permutation σ always exists for any $\bigoplus^{n} Q^s$, it implies that coloring of $\bigoplus^{n} Q^s$ is always possible.

A2. Hidden Expansions in State-Speciation and Extinction Process (SSE)

A2.1. Hidden Expansions.— We define a birth-death process (Q_{bd}) for lineage speciation (λ) and extinction (μ), and a binary CTMC for trait evolution (Q_1) as:

$$Q_{bd} = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{pmatrix} - & \lambda & 0 & \cdots \\ \mu & - & \lambda & \ddots \\ 0 & \mu & - & \ddots \\ \vdots & \ddots & \ddots & - \end{pmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix}, \quad Q_1 = \begin{array}{c} \begin{matrix} 0 & 1 \end{matrix} \\ \begin{pmatrix} - & q_1 \\ q_2 & - \end{pmatrix} \end{array} \begin{matrix} 0 \\ 1 \end{matrix}. \tag{10}$$

The simplest joint process for Q_{bd} and Q_1, known as BiSSE model (Q_{bi}), is shown in the equation (11). A more complex model, HiSSE, that includes hidden traits is given in the equation (12). The capital integers $\{1, 2, 3, ...\}$ in the state names indicate number of born species, the numerical subscripts $\{0, 1\}$ denote states in the observed character, and letters $\{a, b\}$ denote hidden states.

$$Q_{bi} = \begin{array}{c} \begin{array}{ccccccc} 1_0 & 1_1 & 2_0 & 2_1 & 3_0 & 3_1 & 4_0 \end{array} \\ \left(\begin{array}{cc|cc|cc|c} - & q_1 & \lambda_1 & 0 & 0 & 0 & \cdots \\ q_2 & - & 0 & \lambda_2 & 0 & 0 & \ddots \\ \hline \mu_1 & 0 & - & q_1 & \lambda_1 & 0 & \ddots \\ 0 & \mu_2 & q_2 & - & 0 & \lambda_2 & \ddots \\ \hline 0 & 0 & \mu_1 & 0 & - & q_1 & \ddots \\ 0 & 0 & 0 & \mu_2 & q_2 & - & \ddots \\ \hline \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & - \end{array}\right) \end{array} \begin{array}{c} 1_0 \\ 1_1 \\ 2_0 \\ 2_1 \\ 3_0 \\ 3_1 \\ 4_0 \end{array} . \tag{11}$$

$$Q_{hi} = \begin{array}{c} \begin{array}{cccccccccc} 1_{0a} & 1_{0b} & 1_{1a} & 1_{1b} & 2_{0a} & 2_{0b} & 2_{1a} & 2_{1b} & 3_{0a} \end{array} \\ \left(\begin{array}{cccc|cccc|c} - & q_{12} & q_{13} & q_{14} & \lambda_1 & 0 & 0 & 0 & \cdots \\ q_{21} & - & q_{23} & q_{24} & 0 & \lambda_2 & 0 & 0 & \ddots \\ q_{31} & q_{32} & - & q_{34} & 0 & 0 & \lambda_3 & 0 & \ddots \\ q_{41} & q_{42} & q_{43} & - & 0 & 0 & 0 & \lambda_4 & \ddots \\ \hline \mu_1 & 0 & 0 & 0 & - & q_{12} & q_{13} & q_{14} & \cdots \\ 0 & \mu_2 & 0 & 0 & q_{21} & - & q_{23} & q_{24} & \ddots \\ 0 & 0 & \mu_3 & 0 & q_{31} & q_{32} & - & q_{34} & \ddots \\ 0 & 0 & 0 & \mu_4 & q_{41} & q_{42} & q_{43} & - & \ddots \\ \hline \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{array}\right) \end{array} \begin{array}{c} 1_{0a} \\ 1_{0b} \\ 1_{1a} \\ 1_{1b} \\ 2_{0a} \\ 2_{0b} \\ 2_{1a} \\ 2_{1b} \\ 3_{0a} \end{array} . \tag{12}$$

These are the conventional SSE models used in phylogenetics. Note, in both models every trait's state is associated with only one speciation and extinction event. The hidden expansions for traits can be constructed by adding extra hidden states to them. However, the hidden expansions for diversification process cannot be constructed on those conventional models. To allow it, the conventional SSE models should be extended to permit a dual transition – speciation/extinction event and change in trait's state. The simplest of such models with dual transitions is:

$$Q_{bi^+} = \begin{array}{c} \begin{array}{cccccccc} 1_0 & 1_1 & 2_0 & 2_1 & 3_0 & 3_1 & 4_0 & \end{array} \\ \left(\begin{array}{cc|cc|cc|c} - & q_1 & \lambda_{11} & \lambda_{12} & 0 & 0 & \cdots & 1_0 \\ q_2 & - & \lambda_{21} & \lambda_{22} & 0 & 0 & \ddots & 1_1 \\ \hline \mu_1 & 0 & - & q_1 & \lambda_{11} & \lambda_{12} & \ddots & 2_0 \\ 0 & \mu_2 & q_2 & - & \lambda_{21} & \lambda_{22} & \ddots & 2_1 \\ \hline 0 & 0 & \mu_1 & 0 & - & q_1 & \ddots & 3_0 \\ 0 & 0 & 0 & \mu_2 & q_2 & - & \ddots & 3_0 \\ \hline \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & - & 4_0 \end{array}\right) \end{array} . \tag{13}$$

So, the birth-death dual transitions can be used to construct any HE. For example, suppose there is a BiSSE process with variable diversification: $\lambda_1 = 0.1, \lambda_2 = 0.2, \mu = 0.1$, and asymmetric transitions between the trait's states: $q_1 = 0.2, q_2 = 0.1$, shown on the left hand side of the equation (14). The EHE of this process is shown on right hand side of the equation (14); it is characterized by the equal rate value $q_{sin} = 0.1$. Note that the EHE is lumpable by the RWR with respect to the aggregation of states $\{\{1_{0a}\}, \{1_{1a}, 1_{1b}\}, \{2_{0a}\}, \{2_{1a}, 2_{1b}\}, ...\{S_{0a}\}, \{S_{1a}, S_{1b}\}\}$ that produces the original BiSSE matrix. The same logic of allowing dual transitions applies to other hidden expansions of SSE processes.

$$0.1\cdot \begin{array}{c} \begin{array}{ccccccc} 1_0 & 1_1 & 2_0 & 2_1 & 3_0 & & \end{array} \\ \left(\begin{array}{cc|cc|c} - & 2 & 1 & 0 & \cdots & 1_0 \\ 1 & - & 0 & 2 & \ddots & 1_1 \\ \hline 1 & 0 & - & 2 & \cdots & 2_0 \\ 0 & 1 & 1 & - & \ddots & 2_1 \\ \hline \vdots & \ddots & \ddots & \ddots & \ddots & 3_0 \end{array}\right) \end{array} \cong 0.1\cdot \begin{array}{c} \begin{array}{ccccccc} 1_{0a} & 1_{1a} & 1_{1b} & 2_{0a} & 2_{1a} & 2_{1b} & 3_{0a} \end{array} \\ \left(\begin{array}{ccc|ccc|c} - & 1 & 1 & 1 & 0 & 0 & \cdots & 1_{0a} \\ 1 & - & 0 & 0 & 1 & 1 & \ddots & 1_{1a} \\ 1 & 0 & - & 0 & 1 & 1 & \ddots & 1_{1b} \\ \hline 1 & 0 & 0 & - & 1 & 1 & \cdots & 2_{0a} \\ 0 & 1 & 0 & 1 & - & 0 & \ddots & 2_{1a} \\ 0 & 0 & 1 & 1 & 0 & - & \ddots & 2_{1b} \\ \hline \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & 3_{0a} \end{array}\right) \end{array} . \tag{14}$$

A2.2. Congruence between correlated and independent models.— Consider the character independent (CID) model where diversification process is dependent on the

hidden trait but independent for the observable trait. The CID model implies the following constraints between the rates in the equation (12) to hold simultaneously: (1) $q_{14} = q_{23} = q_{32} = q_{41} = 0$; (2) $q_{13} = q_{24}, q_{12} = q_{34}, q_{21} = q_{43}, q_{31} = q_{42}$; (3) $\lambda_1 = \lambda_3, \lambda_2 = \lambda_4$; and (4) $\mu_1 = \mu_3, \mu_2 = \mu_4$. The submatrix ($B_1$) describing trait evolution in the CID model is shown in the equation (15); the submatrix notation is used herein for brevity. Note, it lumpable with respect to the observed trait (Q_{ob}) by the following aggregation $\{\{1_{0a}, 1_{0b}\}, \{2_{1a}, 2_{1b}\}\}$, which means that Q_{ob} can be modeled separately without any information on the hidden trait and diversification process. Obviously, there are many ways to maintain lumpability in B_1 with the respect to the observed trait. For example, a lumpable submatrix B_2 can be constructed by allowing dual transition in top-right block shown in the equation (15). Note, B_2 breaks the independence constraints of the CID model. Thus, it implies the observed and hidden traits evolve in correlation and affect the diversification process; so the latter is dependent on the observed trait too. Both models, B_1 and B_2 are congruent and have the same number of parameters but imply drastically different evolutionary interpretations. These two models cannot be distinguished phylogenetically. Apparently any CID model can be modified in similar way by allowing dual transitions to result in a congruent model that implies state dependent diversification. Thus, any CID model is congruent with the state-dependent model.

$$B_1 = \left|\begin{array}{cc|cc} 1_{0a} & 1_{0b} & 1_{1a} & 1_{1b} \\ - & q_{12} & q_{13} & 0 \\ q_{21} & - & 0 & q_{13} \\ \hline q_{31} & 0 & - & q_{12} \\ 0 & q_{31} & q_{21} & - \end{array}\right| \cong B_2 = \left|\begin{array}{cc|cc} 1_{0a} & 1_{0b} & 1_{1a} & 1_{1b} \\ - & q_{12} & q_{13} & q_{13} \\ q_{21} & - & q_{13} & q_{13} \\ \hline q_{31} & 0 & - & q_{12} \\ 0 & q_{31} & q_{21} & - \end{array}\right| \cong Q_{ob} = \begin{array}{cc} 0 & 1 \\ \left(\begin{array}{cc} - & q_{13} \\ q_{31} & - \end{array}\right) & \end{array} . \tag{15}$$

807 # References

808 H. Akaike. Likelihood of a model and information criteria. *Journal of econometrics*, 16(1):
809 3–14, 1981.

810 J. M. Beaulieu and B. C. O'Meara. Detecting hidden diversification shifts in models of
811 trait-dependent speciation and extinction. *Systematic Biology*, 65(4):583–601, 2016.
812 URL `https://academic.oup.com/sysbio/article/65/4/583/1753616`.

813 J. M. Beaulieu, B. C. O'meara, and M. J. Donoghue. Identifying hidden rate changes in
814 the evolution of a binary morphological character: the evolution of plant habit in
815 campanulid angiosperms. *Systematic biology*, 62(5):725–737, 2013.

816 R. G. FitzJohn. Diversitree: Comparative phylogenetic analyses of diversification in r.
817 *Methods in Ecology and Evolution*, in press, 2012. doi:
818 10.1111/j.2041-210X.2012.00234.x.

819 S. Gavrilets. Evolution and speciation on holey adaptive landscapes. *Trends in ecology &*
820 *evolution*, 12(8):307–312, 1997.

821 S. Gavrilets. Perspective: models of speciation: what have we learned in 40 years?
822 *Evolution*, 57(10):2197–2215, 2003.

823 S. Gavrilets and J. Gravner. Percolation on the fitness hypercube and the evolution of
824 reproductive isolation. *Journal of theoretical biology*, 184(1):51–64, 1997.

825 A. J. Helmstetter, S. Glemin, J. Kafer, R. Zenil-Ferguson, H. Sauquet, H. de Boer, L.-P. M.
826 Dagallier, N. Mazet, E. L. Reboud, T. L. Couvreur, et al. Pulled diversification rates,
827 lineage-through-time plots and modern macroevolutionary modelling. *bioRxiv*, 2021.

828 S. Höhna, B. T. Kopperud, and A. F. Magee. Acdc: Analysis of congruent diversification
829 classes. *bioRxiv*, 2022.

830 J. Kemeny and J. Snell. Finite markov chains. princeton (nj): van nostrand, 1960.

831 S. Louca and M. W. Pennell. Extant timetrees are consistent with a myriad of
832 diversification histories. *Nature*, 580(7804):502–505, 2020.

833 W. P. Maddison, P. E. Midford, and S. P. Otto. Estimating a binary character's effect on
834 speciation and extinction. *Systematic biology*, 56(5):701–710, 2007.

835 A. F. Magee, S. Höhna, T. I. Vasylyeva, A. D. Leaché, and V. N. Minin. Locally adaptive
836 bayesian birth-death model successfully detects slow and rapid rate shifts. *PLoS*
837 *computational biology*, 16(10):e1007999, 2020.

838 H. Morlon, S. Robin, and F. Hartig. Studying speciation and extinction dynamics from
839 phylogenies: addressing identifiability issues. *Trends in Ecology & Evolution*, 2022.

840 B. O'Meara and J. Beaulieu. Potential survival of some, but not all, diversification
841 methods. 2021.

842 M. Pagel. Detecting correlated evolution on phylogenies: a general method for the
843 comparative analysis of discrete characters. *Proc. R. Soc. Lond. B*, 255(1342):37–45,
844 1994.

845 P. D. Polly. Developmental dynamics and g-matrices: Can morphometric spaces be used to
846 model phenotypic evolution? *Evolutionary biology*, 35(2):83–96, 2008.

847 L. J. Revell. phytools: An r package for phylogenetic comparative biology (and other
848 things). *Methods in Ecology and Evolution*, 3:217–223, 2012.

849 G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464,
850 1978.

851 J. G. Sumner, J. Fernández-Sánchez, and P. D. Jarvis. Lie markov models. *Journal of*
852 *theoretical biology*, 298:16–31, 2012a.

853 J. G. Sumner, P. D. Jarvis, J. Fernández-Sánchez, J. Ferńandez-Sánchez, B. T. Kaine,
854 M. D. Woodhams, and B. R. Holland. Is the general time-reversible model bad for
855 molecular phylogenetics? *Systematic biology*, 61(6):1069–1074, 2012b.

856 S. Tarasov. Integration of anatomy ontologies and evo-devo using structured markov
857 models suggests a new framework for modeling discrete phenotypic traits. *Systematic*
858 *biology*, 68(5):698–716, 2019.

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

861 C. Tuffley and M. Steel. Modeling the covarion hypothesis of nucleotide substitution.
862 *Mathematical biosciences*, 147(1):63–91, 1998.

863 M. D. Woodhams, J. Fernández-Sánchez, and J. G. Sumner. A new hierarchy of
864 phylogenetic models consistent with heterogeneous substitution rates. *Systematic*
865 *biology*, 64(4):638–650, 2015.