Integration of anatomy ontologies and Evo-Devo using structured Markov chains suggests a new framework for modeling discrete phenotypic traits

Sergei Tarasov

National Institute for Mathematical and Biological Synthesis, University of Tennessee, Knoxville, TN 37996, USA

E-mail: sergxf@yandex.ru

The author declares no conflict of interest.

ABSTRACT

Modeling discrete phenotypic traits for either ancestral character state reconstruction or morphology-based phylogenetic inference suffers from ambiguities of character coding, homology assessment, dependencies, and selection of adequate models. These drawbacks occur because trait evolution is driven by the two key processes – hierarchical and hidden – which are not accommodated simultaneously by the available phylogenetic methods. The hierarchical process refers to the dependencies between anatomical body parts, while the hidden process refers to the evolution of gene regulatory networks (GRNs) underlying trait development. Herein, I demonstrate that these processes can be efficiently modeled using structured Markov chains equipped with hidden states, which resolves the majority of the problems associated with discrete traits. Integration of structured Markov chains with anatomy ontologies adequately incorporates the hierarchical dependencies, while use of the hidden states accommodates hidden GRN evolution and mutation rate heterogeneity. This model is insensitive to alternative coding approaches which is shown by solving the Maddison's tail color problem. Additionally, this model provides new insight into character concept and homology assessment. The practical considerations for implementing this model in phylogenetic inference and comparative methods are discussed.

KEY WORDS:

Discrete trait, character, morphology, homology, anatomy ontology, structured Markov chain, hidden Markov chain, lumpability, gene regulatory networks

Understanding the processes underlying changes in a phenotype during the course of evolution is one of the fundamental challenges in biology (Prud'homme et al. 2007; Dececchi et al. 2015). The study of these processes promises to advance our knowledge of the dynamics of evolutionary radiations (Price et al. 2010; Van Bocxlaer et al. 2010; Tobias et al. 2014), complexity, and novelties (Moczek 2008; Ramirez and Michalik 2014) , as well as to enhance our understanding of the relationships between genotype and phenotype (Houle et al. 2010; Hiller et al. 2012; McCune and Schimenti 2012; Manda et al. 2015). Despite the numerous methods available for analyzing discrete morphological characters [reviewed in O'Meara (2012)], the lack of repeatable and agreed approaches (similar to those existing in DNA alignment) for primary homology assessment and character coding generates ambiguity during the *character construction* phase of analysis. This process of encoding traits into characters (see the definitions in Box 1) consists of a two-step procedure: (1) delimitation of trait within phenotype and its primary homology assessment across species, as well as (2) trait encoding into character vector or matrix (Wiens 2001). As a result, for the same trait one may propose different hypotheses of primary homology (Ramirez 2007; Agnarsson and Coddington 2008) and different ways of coding the same hypothesis of homology into character [reviewed in Brazeau (2011)]. The analysis of competing characters formulated for the same trait will naturally produce different and largely incomparable results, which can mislead the understanding of trait evolution.

All sources of ambiguity during character construction can be traced to a single root – the complex nature and organization of phenotypic traits (Wiens 2001; Houle et al. 2010; Burleigh et al. 2013). This complexity arises due to the two key processes – hidden and hierarchical – driving trait evolution. The hidden process refers to the evolution of gene regulatory networks (GRNs) which underlay trait development (Wagner 2007; Carroll 2008; Houle et al. 2010). It

implies that the actual trait evolution is hidden behind a "curtain" from the direct observation of morphology. In most cases, the observer (i.e. scientist) has no clue of how this process is operating, and only knows what is going on behind the "curtain" from the outcome of the process (i.e. morphological traits). The hierarchical process refers to the hierarchical relationships between traits that arise due to hierarchical dependencies between anatomical parts. For example, digits and characters associated with them are located on limbs; loss of limbs during evolution simultaneously causes the loss of digits. Additionally, hidden processes of GRN evolution – through interacting cascades of genes – can also result in hidden dependencies among observable morphological traits.

The hidden and hierarchical processes are not accommodated by available phylogenetic methods simultaneously regardless the approach used for trait analyses, be it parsimony (Lee and Bryant 1999; Fitzhugh 2006; Brazeau 2011) or traditional Markov models. By proposing a new framework, I will demonstrate that the simultaneous inclusion of these processes, to a large extent, eliminates ambiguities associated with trait modeling. This framework is the extension of the traditional Markov model approach commonly used for modeling traits (Lewis 2001; O'Meara 2012). The most common version, hereafter referred as simple Markov chains (SMC), implies that a discrete character is a continuous-time Markov chain that moves sequentially from one character state to another over the course of evolution. Such a Markov chain is defined by the transition rate matrix containing infinitesimal rates of change between the states, and a base frequency vector specifying the initial probabilities of states at the root of a phylogenetic tree [see e.g., Huelsenbeck et al. (2003) for details].

The new framework extends SMC using the theory of structured Markov chains [StMC, (Nodelman et al. 2002)] and hidden Markov chains [HMC, (Beaulieu et al. 2013)] to

accommodate complex evolutionary space and anatomical dependencies among traits. To justify the proposed framework, I will show how the anatomical dependencies can be efficiently incorporated into model using structured Markov chains and anatomy ontologies. This integration provides the solution for the well-known Maddison's tail color problem (Maddison 1993; Hawkins et al. 1997). Next, I will discuss how the correspondence between traits and their GRNs can be modeled using HMC. Finally, the unified framework for character modeling will be proposed and practical considerations on trait modeling and phylogenetic inference will be given. The central focus of this paper is morphological traits; however, the presented results can be directly extended to other discrete traits of phenotype, such as behavior.

MODELING HIERARCHICAL PROCESS USING STRUCTURED MARKOV CHAINS AND ANATOMY ONTOLOGIES

Dependencies and coding schemes

Encoding traits is a crucial step in constructing discrete morphological characters. Although for some traits this can be straightforward, many real-life situations lack an unambiguous coding approach. In practice, this means selecting between (1) ordered or unordered characters, (2) coding schemes (a set of binary characters, one multistate character or a mixture of both), and (3) a way to code inapplicable observations (using reductive coding "?" versus a separate state). Each of the alternatives have their own pros and cons but none of them, due to the lack of consensus in the published studies (Maddison 1993; Pleijel 1995; Hawkins et al. 1997; Strong and Lipscomb 1999; Forey and Kitching 2000; Brazeau 2011), can be regarded as the ideal one. Additionally, there is no general consensus on the distinction and validity between character and character state. Some studies insist that the distinction is important (Pinna

1991; Hawkins et al. 1997; Wagner 2015), some suggest that both concepts are the same (Patterson 1982), while the others (Sereno 2007) view character states as mutually exclusive observations that have to be combined to perform analysis. The use of different coding approaches drastically affects the interpretation of results (Brazeau 2011).

The incorporation of anatomical dependencies between traits using structured Markov models resolves the conundrum associated with coding schemes and yields an invariance under alternative coding approaches. This invariance implies that any coding scheme produces the same result. I demonstrate the resiliency of my modeling framework by revisiting the exemplar Maddison's tail color problem (Maddison 1993; Hawkins et al. 1997) that seeks the optimal scheme for scoring tail traits in species which can have a tail with either blue or red color, or have no tail at all (Figs. 1, 2). The Maddison's problem exemplifies a common situation of scoring complex traits with dependencies between traits, and any solution to this problem can be extrapolated to the majority of ambiguous coding cases encountered in practice. The core ingredient of the proposed approach is the explicit incorporation of anatomical dependencies which can be inferred from ontological structure of anatomy (Fig. 1), therefore I refer to this approach as "structured Markov models informed by anatomy ontology". In this chapter, I will begin with the reviewing the general properties of StMC, then I will demonstrate how various types of dependencies can be accommodated and how StMC can be used to solve the Maddison's problem. Finally, I will give some consideration for retrieving dependency data from anatomy ontologies.

Incorporating character dependencies using StMC

Structured Markov Models. – This class of models (Nodelman et al. 2002), also known as continuous-time Bayesian Networks (Shelton and Ciardo 2014), arises from simple Markov

chains. The only difference between the two is that the structured Markov models are equipped with a specific parametrization of the rate matrix that accommodates conditional dependencies between characters.

The structuring of Markov chains allows combining two or more initial characters into a single character. The initial characters can be combined either as independently or dependently evolving; the latter enables modeling correlated evolution across states between different initial characters. The states of the initial characters will be modified states in the combined character. The combination of characters does not affect their mathematical characteristics. This means that in a Markov chain model there is no distinction between character and character state as both are equipped with a scale-free property in respect to each other. This property provides a flexible way to deal with characters, as characters can be combined or decomposed into several smaller ones. Additionally, the scale-free property can be used to test the correlated evolution between characters (Pagel 1994). The approaches summarized below can be used to incorporate all types of dependencies that, to my knowledge, exist among morphological characters.

Two characters: independent evolution. – This case occurs when independently evolving characters are combined. Suppose there are two initial two-state characters: X {with states: $\boldsymbol{x_1}$, $\boldsymbol{x_2}$} and Y {with states: $\boldsymbol{y_1}$, $\boldsymbol{y_2}$} which we wish to combine into one single character Z. The initial characters X and Y are defined by the transition rate matrices **X** and **Y** respectively:

$$\boldsymbol{X} = \begin{matrix} \\ \boldsymbol{x_1} \\ \boldsymbol{x_2} \end{matrix} \begin{matrix} \begin{matrix} \boldsymbol{x_1} & \boldsymbol{x_2} \end{matrix} \\ \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix} \end{matrix}; \quad \boldsymbol{Y} = \begin{matrix} \\ \boldsymbol{y_1} \\ \boldsymbol{y_2} \end{matrix} \begin{matrix} \begin{matrix} \boldsymbol{y_1} & \boldsymbol{y_2} \end{matrix} \\ \begin{pmatrix} -\gamma & \gamma \\ \theta & -\theta \end{pmatrix} \end{matrix}. \quad (1)$$

The rate matrix of the combined character is constructed out of the matrices **X** and **Y** using the equation (2) by merging the two matrices in a mathematically valid way (Supplementary Material, section S1):

$$\boldsymbol{Z} = \boldsymbol{X} \otimes \boldsymbol{I_Y} + \boldsymbol{I_X} \otimes \boldsymbol{Y}, \quad (2)$$

where $\boldsymbol{I_Y}$ and $\boldsymbol{I_X}$ are the identity matrices of the same dimension as the matrices **Y** and **X** respectively, and ⊗ denotes the Kronecker product. Given this, **Z** is:

$$\mathbf{Z} = \begin{array}{c} \\ \boldsymbol{x_1y_1} \\ \boldsymbol{x_1y_2} \\ \boldsymbol{x_2y_1} \\ \boldsymbol{x_2y_2} \end{array} \begin{array}{c} \begin{array}{cccc} \boldsymbol{x_1y_1} & \boldsymbol{x_1y_2} & \boldsymbol{x_2y_1} & \boldsymbol{x_2y_2} \end{array} \\ \begin{pmatrix} -\alpha-\gamma & \gamma & \alpha & 0 \\ \theta & -\alpha-\theta & 0 & \alpha \\ \beta & 0 & -\beta-\gamma & \gamma \\ 0 & \beta & \theta & -\beta-\theta \end{pmatrix} \end{array}. \quad (3)$$

The combined matrix **Z** defines the four-state combined character *Z*. The four states of the character Z correspond to all possible permutations of states in the initial characters *X* and *Y*, which are {$\boldsymbol{x_1y_1}$, $\boldsymbol{x_2y_1}$, $\boldsymbol{x_1y_2}$, $\boldsymbol{x_2y_2}$}. The right-diagonal cells of this matrix are populated with zeros indicating that only one state of the initial characters can change during the infinitesimal time interval. The independent evolution of the initial characters constrains this matrix to possess some rate symmetries that do not hold otherwise (see the next section).

Two characters: general case of correlation. – The dependent evolution of the two characters *X* and *Y* implies that some states in one character are correlated with some states in the other. In this case, the combined character *Z* is characterized by the same states {$\boldsymbol{x_1y_1}$, $\boldsymbol{x_2y_1}$, $\boldsymbol{x_1y_2}$, $\boldsymbol{x_2y_2}$} as in the independent case but the rate symmetries in the combined matrix **Z** are different. In character *Z*, each state consists of the two elements: the first corresponding to a state of *X*, the second to a state of *Y*. Let us denote by * any element in a combined state. For example, notation $\boldsymbol{x_1}$* indicates either state $\boldsymbol{x_1y_1}$ or $\boldsymbol{x_1y_2}$ of *Z*. Independent evolution implies that the transition rates $\boldsymbol{x_1}$*→$\boldsymbol{x_2}$* in *Z* must be the same as the transition rate $\boldsymbol{x_1}$→$\boldsymbol{x_2}$ in the initial chain *X* (i.e., rate α); the same should apply for the rates $\boldsymbol{x_2}$*→$\boldsymbol{x_1}$* that must be equal to the rate $\boldsymbol{x_2}$→$\boldsymbol{x_1}$ in *X* (rate β); and analogous symmetries must hold for pairs *$\boldsymbol{y_1}$→*$\boldsymbol{y_2}$ and *$\boldsymbol{y_2}$ →*$\boldsymbol{y_1}$ whose rates have to be equal to $\boldsymbol{y_1}$→$\boldsymbol{y_2}$ and $\boldsymbol{y_2}$ →$\boldsymbol{y_1}$ in the initial chain *Y* respectively. If these equalities do not hold simultaneously, then the evolution of the two initial chains is correlated. So, in the most

sophisticated case of the correlated evolution, the rate matrix **Z** has all rates different except for the right-diagonal ones that are set to zeros as in the independent case:

$$\boldsymbol{Z} = \begin{pmatrix} -\alpha - \gamma_1 & \gamma_1 & \alpha & 0 \\ \theta_1 & -\alpha_1 - \theta_1 & 0 & \alpha_1 \\ \beta & 0 & -\beta - \gamma & \gamma \\ 0 & \beta_1 & \theta & -\theta - \beta_1 \end{pmatrix}. \quad (4)$$

The analytical derivation of these matrix is given in the Supplementary Material (section S2).

Two characters: "switch-off" case of correlation. – Beside the general case of correlated evolution, the StMC can be used to incorporate "switch-off" dependencies between states which arise when hierarchically upstream state switches-off the downstream one. Suppose that during the course of evolution both states of character *Y* can appear only if the character *X* is in the state $\boldsymbol{x_2}$; if *X* is in the state $\boldsymbol{x_1}$ then the character *Y* is "switched-off" (in practice this often means using inapplicable coding for *Y*). The modified version of the equation (2) can be used to construct the transition matrix **Z** of such correlated character (Supplementary Material, section S3) which gives:

$$\boldsymbol{Z} = \begin{pmatrix} -\alpha & 0 & \alpha & 0 \\ 0 & -\alpha & 0 & \alpha \\ \beta & 0 & -\beta - \gamma & \gamma \\ 0 & \beta & \theta & -\beta - \theta \end{pmatrix}. \quad (5)$$

In this matrix, the transitions $\boldsymbol{x_1y_1} \rightarrow \boldsymbol{x_1y_2}$ and $\boldsymbol{x_1y_2} \rightarrow \boldsymbol{x_1y_1}$ are prohibited and equal to zero due to this particular type of dependency.

Two characters: synchronous changes. – This dependency refers to the case when some states belonging to two different characters always change simultaneously. It can be observed when two characters, produced by decomposing one single character, are combined. Suppose there is a two-state character *X* $\{\boldsymbol{x_1}, \boldsymbol{x_2}\}$ specified by the transition rate matrix **X** as in (1); we can decompose the character *X* into two separate characters X_1 and X_2 which denote presence or

absence of the initial states of X. So, each separate character is interpreted as follows: X_1 $\{\boldsymbol{x_{1\ absent}}, \boldsymbol{x_{1\ present}}\}$, and X_2 $\{\boldsymbol{x_{2\ absent}}, \boldsymbol{x_{2\ present}}\}$. Since states $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ are mutually exclusive as they are the parts of the same initial character, the transitions in X_1 and X_2 occur synchronously: for example, the transition $\boldsymbol{x_{1\ absent} \rightarrow x_{1\ present}}$ in X_1 immediately causes the compliment transition $\boldsymbol{x_{2\ present} \rightarrow x_{2\ absent}}$ in X_2. The scale-free property of character suggests that the reverse operation – combining X_1 and X_2 into one character – is possible. Let us denote this combination of X_1 and X_2 by Z. The character Z is supposed to have four states which are permutations of the original states: $\{\boldsymbol{x_{1\ absent}\ x_{2\ absent}}, \boldsymbol{x_{1\ present}\ x_{2\ absent}}, \boldsymbol{x_{1\ absent}\ x_{2\ present}}, \boldsymbol{x_{1\ present}\ x_{2\ present}}\}$. The synchronous transitions in X_1 and X_2 precludes using the previous approaches to combine X_1 and X_2. This happens because the structure of the previous rate matrices allows only one change over the infinitesimal time interval. In contrast, the synchronous evolution assumes the opposite – one change transitions have to be prohibited, while the two change transitions must be allowed. Additionally, any transitions associated with the states $\{\boldsymbol{x_{1\ absent}\ x_{2\ absent}}\}$ and $\{\boldsymbol{x_{1\ present}\ x_{2\ present}}\}$ must be also prohibited as these states cannot exists given the initial condition. This yields the following transition matrix of the character Z:

$$\boldsymbol{Z} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -\alpha & \alpha & 0 \\ 0 & \beta & -\beta & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \quad (6)$$

Further, without loose of generality this combined matrix can be reduced to that of the initial character X (1). This property of StMC provides an invariance for character decomposition and subsequent merge.

Combining arbitrary number of characters. – The techniques shown above can be extrapolated to construct combined rate matrices for any arbitrary number of initial characters and character states. To combine n independently evolving characters one needs to successively

repeat the equation (2) *n-1* times until all initial characters are combined. For example, in the case of the three initial characters *A, B, C,* the first step constructs combined matrix for characters *A* and *B* (i.e. *AB*) and the second steps uses matrices of *AB* and *C* to construct the final combined character. In equation form this can be expressed as:

$$(\boldsymbol{A}\otimes\boldsymbol{I_B} + \boldsymbol{I_A}\otimes\boldsymbol{B})\otimes\boldsymbol{I_C} + \boldsymbol{I_{AB}}\otimes\boldsymbol{C}. \quad (7)$$

The order at which matrices are combined does not matter. In the case of the correlated character evolution, the final matrix can be constructed as that for the independent characters and then modified to accommodate the desired pattern of correlation.

The combined matrices of coevolving characters have peculiar symmetries; although such matrices can be enormous, the vast majority of their cells are zeros, while the transition rates are located along the secondary diagonals (Fig. 3a-g). For a chain of *n* coevolving characters with the equal number of states ω the proportion of non-zero elements in the large matrix is:

$$\frac{1+n(\omega-1)}{\omega^n}. \quad (8)$$

The number and pattern of secondary diagonals populated with transition rates increases as the total number of states grows (Fig. 3a-g). If all elementary characters share equal quantity of states, the total number of secondary diagonals is $n(\omega-1)$, (Fig. 3a,b,f,g). Each secondary diagonal encompasses rates from a single elementary character; asymmetries in the rates within a character split diagonals into rate groups (Fig. 3c-d). If elementary characters are correlated, then the number of rate categories increases rapidly, culminating at the extreme case when all cells along secondary diagonals get populated with different rate values (Fig. 3d).

The Maddison's problem

Maddison's problem has been intensively discussed in the framework of parsimony but the coding consensus is still lacking (Brazeau 2011). I reduce this problem to assesses the alternative

schemes used for coding tail traits in the three species with: (1) absent tail, (2) blue tail, and (3) red tail (Fig. 1). Traditionally, there exist three main schemes (Hawkins et al. 1997) of scoring these tail traits (Fig. 2).

The *first scheme* (Fig. 2a) uses two characters: (i) tail presence with two states, and (ii) tail color with three states, to encode the observations. In this scheme, the character (ii) can be reduced to only two states (blue and red) if its state "absent" is coded as inapplicable observation using "?". In the current context, the distinction between these two flavors is irrelevant.

The *second scheme* (Fig. 2b) employs three binary characters: (i) tail present, (ii) blue tail present, and (iii) red tail present. This scheme can also have an alternative version that uses only characters (ii) and (iii) to encode the observations (Hawkins et al. 1997).

Finally, the *third scheme* (Fig. 2e) uses one multistate character with three states (absent, blue, red) to simultaneously summarize the observations. In phylogenetic inference, one of the prevailing ways to deal with this problem is to use the scheme #2 with inapplicable coding (Maddison 1993; Hawkins et al. 1997; Strong and Lipscomb 1999). However, this coding scheme was shown to be flawed (Maddison 1993; Strong and Lipscomb 1999). In comparative phylogenetics, the preferable coding scheme is selected to best fit the needs of an analysis.

Solution using StMC. – All schemes become invariant if anatomical dependencies between characters are incorporated using the StMC. These ontology-informed dependencies imply that tail color (either blue or red) depends on the tail presence; if tail is absent, then color trait is "switched-off". Below, I successively incorporate these dependencies using the alternative coding schemes which, in all instances, produces the scheme #3 with the special parametrization of the rate matrix.

First, let us consider the scheme #1 that treats tail traits as two binary characters. Given the invariance under character decomposition, the synchronous dependency between *tail* {***absent***} and *tail color* {**absent**} (Fig. 2a) is a redundant observation that can be omitted without loose of generality. So, each character can be represented by a two-state rate matrix:

$$\boldsymbol{TL} = \begin{matrix} \\ \boldsymbol{a} \\ \boldsymbol{p} \end{matrix} \begin{matrix} \begin{matrix} \boldsymbol{a} & \boldsymbol{p} \end{matrix} \\ \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix} \end{matrix}, \quad \boldsymbol{CR} = \begin{matrix} \\ \boldsymbol{r} \\ \boldsymbol{b} \end{matrix} \begin{matrix} \begin{matrix} \boldsymbol{r} & \boldsymbol{b} \end{matrix} \\ \begin{pmatrix} -\gamma & \gamma \\ \theta & -\theta \end{pmatrix} \end{matrix},$$

where **TL** is a matrix for tail presence {***a*** – absent, ***p*** – present}, and **CR** is a matrix for tail color {***r*** – red, ***b*** – blue}. The dependencies between these two characters (Fig. 2c) can be incorporated using the "switch-off" correlation, which results in the following combined rate matrix:

$$\begin{matrix} & \boldsymbol{ar} & \boldsymbol{ab} & \boldsymbol{pr} & \boldsymbol{pb} \\ \boldsymbol{ar} & -\alpha & 0 & \alpha & 0 \\ \boldsymbol{ab} & 0 & -\alpha & 0 & \alpha \\ \boldsymbol{pr} & \beta & 0 & -\beta-\gamma & \gamma \\ \boldsymbol{pb} & 0 & \beta & \theta & -\beta-\theta \end{matrix}. \quad (9)$$

This matrix contains redundancy: the states ***ar*** and ***ab*** correspond to the same observation specifying absence of tail since tail color cannot be observed when the tail is absent. Interestingly, this matrix can be reduced by aggregating states ***ar*** and ***ab*** using the rule of strong lumpability (see the "Cases of lumpable chains" section) that gives the three-state matrix:

$$\begin{matrix} & \boldsymbol{a} & \boldsymbol{pr} & \boldsymbol{pb} \\ \boldsymbol{a} & -\alpha & \frac{\alpha}{2} & \frac{\alpha}{2} \\ \boldsymbol{pr} & \beta & -\beta-\gamma & \gamma \\ \boldsymbol{pb} & \beta & \theta & -\beta-\theta \end{matrix}. \quad (10)$$

This final matrix has specific symmetries between the rate parameters characterizing the dependencies between the states. So, when anatomical relationships are incorporated, the scheme #1 of the two characters, collapses to scheme #3 of one character, and the latter becomes equipped with the special parametrization of the matrix. In fact, the parametrization of such matrix should not be necessarily restricted to that in matrix (10). For example, equipping

transitions $\mathbf{pr} \rightarrow \mathbf{a}$, and $\mathbf{pb} \rightarrow \mathbf{a}$ with different rates allows modeling a more complex correlation pattern.

The same collapse happens for the coding scheme #2 that uses three binary characters to encode tail traits. In respect to anatomy, two characters in this scheme (red tail presence and blue tail presence) are dependent on the presence of the tail. There are two synchronous changes in these characters: (i) the tail absence causes absence of the blue and red color, while (ii) the absence of the red color causes presence of the blue color (Fig. 2b). Based on the properties of the synchronous changes, these dependencies can be straightforwardly reduced to those of the scheme #1 without loose of generality. In turn, the scheme #1 can be further collapsed, as shown above, to the rate matrix (10).

To sum up, the incorporation of ontological information from anatomy using StMC produces coding invariance regardless of the scheme used. This invariance eliminates ambiguity of character coding. If the given alternative schemes are modelled onto an existing phylogenetic tree, the result is expected to be the same. So, these alternatives are just different ways of representing the same morphological observations. Moreover, incorporation of dependencies does not require using inapplicable coding, thus avoiding uncertainty associated with it. At the same time, the incorporated dependencies have biologically meaningful interpretations as they are imposed by the structure of organismal anatomy.

Anatomy ontology and dependencies

As shown above, the ontological knowledge is important for constructing realistic models of trait evolution (Fig. 1). The importance of integrating anatomy ontologies for character analysis has been recently emphasized and discussed (Vogt 2016, 2017a, 2017b). At present, a computer-assisted way to formalize ontologies is turning them into a promising tool for

arranging and managing knowledge of organismal anatomies (Deans et al. 2015). The formalized ontologies are available for several taxa [e.g., Mungall et al. (2012) and Yoder et al. (2010)] providing an opportunity to directly link anatomy ontologies with character matrices. Presently, this linking can be performed using the specialized software (Balhoff et al. 2010), while the dependency data can be automatically extracted from semantic descriptions using anatomy ontologies (Dececchi et al. 2016). Also, the incorporation of the dependencies can be done by simply using a scientist's own knowledge of organismal anatomy. Integration of StMC with anatomy ontologies enables reconstruction of the ancestral anatomy ontologies in a way similar to the parsimony-based method proposed by Ramirez and Michalik (2014).

MODELING HIDDEN PROCESS: INTEGRATING HIDDEN MARKOV CHAINS AND GRN EVOLUTION

GRNs and morphology

Morphology is a realization of a complex GRN over spatiotemporal scales of embryo development. The consequence of this complex process is the discordance between homology of morphological traits and the homology of underlying GRNs (Abouheif 1999; Hall 2003; Moczek 2008; McCune and Schimenti 2012; Wagner 2015). Therefore, it is essential to summarize the main mechanisms of GRN evolution and their effect on morphological traits before proceeding to modelling their correspondence.

At global scale, GRN consists of modules; each module is a smaller GRN exhibiting a cluster of interacting components whose interactions are relatively autonomous in respect to other modules (Oakley; Babu et al. 2004; Longabaugh et al. 2005; Kuratani 2009; McDougall et al. 2011; Siegal 2013; Rebeiz et al. 2015; Voordeckers et al. 2015). Although the mechanisms of

GRN evolution are multifaceted, they can be summarized into three general principles. *(1) First*, it is a birth of new functional GRN that occurs by co-option of pre-existent module into a new body place (Babu et al. 2004; Wagner 2007; Erwin and Davidson 2009; Monteiro 2012; Siegal 2013; Hinman and Cheatle Jarvela 2014; McKeown et al. 2014; Glassford et al. 2015; Rebeiz et al. 2015) or by integration of two or more pre-existent modules (Clark-Hachtel et al. 2013; Arendt et al. 2016). *(2) Second*, this is a transformation of pre-existent module from one state to another state mediated by reorganization of existing regulatory linkage among genes (Abouheif 1999; Erkenbrack and Davidson 2015). *(3) Third*, it is an inactivation of GRN module by mutation(s) in upstream regulatory modules that disables its realization (Shapiro et al. 2006; Shbailat and Abouheif 2013; Held 2014). The limits between the aforementioned mechanisms are somewhat arbitrary and it is logical to believe that GRN evolution occurs under a mixture of the listed mechanisms especially at the global time scales.

Markov chains is a convenient way to formalize GRN evolution mathematically. In this formalization each step in evolutionary path of GRN represents a state of a Markov chain. The complete evolutionary path of GRN evolution involves birth of GRN modules and transition between GRN states that culminates at the death of GRN. Morphological traits are realizations of some modules of global organismal GRN. In this respect, the construction of a discrete character can be viewed as a mapping of Markov chain state(s) characterizing GRN evolution to another set of states corresponding to the states of the discrete character. This mapping yields one of the three types of correspondence between GRN and trait primary homology (Fig. 4).

Type 1: one-to-one correspondence. – This is a case when there is one-to-one correspondence between GRN states and those of a discrete character (Fig. 4a) indicating that primary homology hypotheses correctly identify underlying GRN states. This case is ideal but it

is far from being realistic, as morphological states are likely to have a complex unobservable GRN space. Nevertheless, this case is possible when changes in morphology are controlled by a single gene.

Type 2: many-to-one correspondence. – It is the case when GRN space is larger than morphological space meaning that one morphological state consists of many GRN states (Fig. 4b). This case seems to be very likely as the majority of morphological traits are realizations of the complex mechanisms controlled by multiple genetic factors (Rebeiz et al. 2015) which are largely unknown to the researcher. Numerous Evo-Devo studies confirm this type of the correspondence. For instance, some males of *Drosophila* have a pigmented spot on the wing which they use in courtship display. Given that the spot shapes are similar across many species, it would be logical to encode this trait as a character containing two states: "spot present" and "spot absent". The study of Prud'Homme et al. (2006) demonstrates that in the clade of 29 *Drosophila* species this spot has been gained and lost multiple times. In all analyzed species, the pigmented spots were products of the expression of gene *yellow*. Interestingly, the loses of pigmentation were caused by parallel inactivation of the same *cis*-regulatory element controlling *yellow* expression. Contrary to that, the independent gains of spot were caused by co-option of different *cis*-regulatory elements associated with the *yellow* gene. This unequivocally suggests that evolution of such simple trait as wing spot is combinatorial at the underlying GRN level, and the GRN state space is larger than the observable one.

The type #2 correspondence may also arise when the space of morphological observations is underestimated due to the precision of morphological examination. This may happen when external structures are examined without referencing to skeletal structures (in vertebrates), or when external skeletal structures are examined without referencing to the underlying architecture

of muscles (in invertebrates). For example, body elongation in salamanders is a result of convergent evolution that occurs by addition and elongation of vertebrae (Wake et al. 2011). This may lead to similarly elongated body shapes by modifying different individual vertebrae. If body shape is studied without referencing to skeletal structure, then similarly elongated bodies must be considered to be homologues and coded with the same state. However, a thorough examination of skeleton will eventually find this coding misleading as the real space of observable morphological evolution is undoubtedly larger. In the model formalism, the salamanders and *Drosophila* cases are the same as they both assume the underestimation of the underlying evolutionary space.

Type 3: partial matching. – This type embraces cases when one complex trait governed by a single GRN is thought to represent two or more independent characters (Fig. 4c). So, different states of the same GRN are mapped onto states in different morphological characters which are separately analyzed suggesting that a focal discrete character does not necessarily exhibit an independent identity. In nature, this case seems to be common as it arises when evolution of states between separate characters is correlated. For instance, mouthpart traits in insects can undergo synchronous evolutionary changes when species get adapted to a new feeding substrate. Considering an anatomical element (e.g., mandibles) of the mouthparts as a separate character without reference to all traits of the mouthparts exhibits this type of correspondence since such character is correlated with other mouthpart traits. In some cases, the existence of correlation is obvious and can be retrieved from anatomy; however, cases with unobservable correlation, which cannot be inferred from the structure of anatomy or species biology, seem to be widespread.

Modeling correspondence between GRN and morphology

All abovementioned cases except type #1 correspondence suggest that the original state space of GRN evolution is larger than the observable trait space and that such pattern is common in biology. This means that one inevitably aggregates different GRN states together when delimiting the observable character state; thus, making morphological state to be a mixture of the underlying GRN states. In Markov chain terms this means that the construction of a discrete morphological character is a substitution of the original GRN Markov chain with a large number of states (Fig. 5b) by an aggregated morphological chain with the reduced number of states (Fig. 5c). A natural question which can be raised is when such aggregation is mathematically valid for modelling observable character if the underlying processes is unknown? If the aggregation is valid, then the original chain is called lumpable and its behavior can be modeled without errors by the aggregated morphological Markov chain. The study of Vera-Ruiz et al. (2014) confirms that chain aggregation under the lumpability condition is appropriate in phylogenetic inference with molecular data. This justifies, to certain a extent, the use of simple Markov chains to model traits without referencing to the underlying genetic processes. However, if the aggregation is invalid, then the original chain is not lumpable and its approximation using the aggregated chain is biased. The lumpability of chain depends on certain strict symmetries in the transition matrix and base frequency vector. These symmetries are unlikely to be observed in real-life examples suggesting that the majority of chains occurring in phylogenetics have to be non-lumpable (see the next section). In the cases, when the conditions of the lumpability are not fulfilled, the aggregated chain would erroneously approximate the original one, and the inferred parameters would be largely meaningless suggesting that the aggregated chain is a biased proxy of the original process. The range of error, in this case, depends on the rate values and rate ratios in the original transition matrix. Figure 6b exemplifies the error in rates estimate when the original

three-state chain is aggregated into the two-state chain under scenarios when the lumpability is fulfilled (*rate.diff.=1*) and not (*rate.diff.≠1*). The increase in rate values along with the differences between rates increases the error in estimating characteristics of the original chain. Thus, the use of simple Markov models, can be misleading. However, the problem of invalid aggregation can be overcome using Markov chains with hidden states (Beaulieu et al. 2013; Beaulieu and O'Meara 2014), thereby directly allowing modeling the hidden process of GRN evolution without aggregation of the original states (see the "Hidden Markov chains and traits" section).

Cases of lumpable chains

The aggregation of the original state space can be viewed as partitioning of the original transition rate matrix into partition blocks which correspond to the transition rates in the aggregated chain. Here, in characterizing the conditions under which aggregation is possible, the term "Markov chain" corresponds to an irreducible continuous-time and time-homogeneous Markov chain unless otherwise specified. There are two main types of lumpability – strong and weak (Kemeny and Snell 1960; Rubino and Sericola 2014). Additionally, there is a case of nearly lumpable chains that occurs under certain conditions in large matrices describing correlated evolution of multiple characters thereby allowing to lump large matrices with insignificant error.

Strong lumpability. – The strong lumpability implies that the original chain can be lumped with respect to some partitioning scheme, under any possible values of the base frequency vector. Suppose, there is a four state $S = \{s_1, s_2, s_3, s_4\}$ Markov chain that is defined by the base frequency vector π and for-by-four matrix $\mathbf{Q}$ whose entities are rates $q_{i,j}$ (Fig. 6a). Let us assume, the aggregated chain is constructed by partitioning the state space S into two groups denoted by

the partitioning scheme $B = \{\{s_1, s_2\}, \{s_3, s_4\}\}$, so the total number of states in the aggregated chain, defined by matrix $\widehat{\mathbf{Q}}$, is two $F = \{f_1, f_2\}$. The analogous notification procedure can be extrapolated to any arbitrary Markov chain.

The sufficient and necessary condition for Markov chain to be strongly lumpable in respect to the given partitioning scheme and any base vector is that the row-wise sum of rates within one partition block of rate matrix must be the same for all rows within given partition block, and this property must hold for all blocks in the rate matrix (Kemeny and Snell 1960; Rubino and Sericola 2014). This row-wise sum of rates constitutes the new transition $\hat{q}_{k,l}$ rates in the lumped chain. Specifically, for the matrix **Q** to be lumpable under the partitioning scheme *B* this implies that:

$$q_{13} + q_{14} = q_{23} + q_{24} = \hat{q}_{1,2} \quad (11a)$$

$$q_{31} + q_{32} = q_{41} + q_{42} = \hat{q}_{2,1} \quad (11b)$$

where $\hat{q}_{1,2}$ and $\hat{q}_{2,1}$ are the transition rates in the aggregated matrix $\widehat{\mathbf{Q}}$ (Fig. 6a).

Weak lumpability. – The weak lumpability allows lumping an original chain only under particular values of base vector (Kemeny and Snell 1960; Rubino and Sericola 2014). In other words, the weak lumpability imposes stricter dependencies between the base frequency vector and rate matrix. There is no straightforward way to find all possible dependencies between any arbitrary rate matrix and base vector that fulfill the conditions of the weak lumpability (Rubino and Sericola 2014); however, there exist a finite algorithm for elucidating the base frequency vector satisfying the conditions of weak lumpability given rate matrix and partitioning scheme (Rubino and Sericola 1993, 2014). Some sufficient conditions for weak lumpability are given in Kemeny and Snell (1960). An example of weakly lumpable chain is given in the Supplementary Material (section S4).

Interesting case of weak lumpability arises when the initial vector of Markov chain coincides with the stationary distribution of that chain (meaning that the process is stationary). Such chain is weakly lumpable in respect to any possible partitioning scheme (Supplementary Material, section S5). This special case of weak lumpability have theoretical implications for modelling character evolution (discussed below).

Nearly lumpable Markov chains. – This type of lumpability occurs in large matrices describing correlated evolution of multiple characters thereby allowing lumping large matrices with insignificant error. Biologically such chains can be seen as many elementary characters coevolving together either dependently or independently (see the "Incorporating character dependencies using StMC" section). Aggregation in such matrices can be thought of as linking evolutionary processes occurring at the level of DNA sites or numerous GRNs with their realization at the phenotypic level.

For example, consider some DNA locus of 1000 sites; each site is a four-state character (four nucleotides). Based on the scale-free property of character, all elementary characters (sites) can be combined into one large Markov chain, where the number of states is $\prod_{i=1}^{n} \omega_i$, given that ω_i is a number of states in the *i*th character and *n* is the total number of characters. So, for the DNA locus example the number of states in the combined Markov chain is 4^{1000}. The properties for combining coevolving characters into one matrix suggest that this combined character has the peculiar symmetries of the rate matrix in which over 99% of cells are equal to zero (the equation [8]). The construction of a phenotypic character is the aggregation of the molecular state-space. Since the molecular states space is significantly larger, we can assume that the aggregated phenotypic chain is composed of a few states and each state comprises large number of the molecular states.

In the trivial case of equal evolutionary rates across all sites, the strong lumpability condition can be satisfied for numerous partitioning schemes that can be applied to the molecular rate matrix (Fig. 3b,g). Interestingly, if all elementary characters are correlated, and all transition rates as well as probabilities of the base frequency vector are different but are identically and independently distributed (*i.i.d.*) then the combined matrix can be nearly lumpable under any possible partitioning scheme. Obviously, such chains drastically violate the condition of strong lumpability. Nevertheless, aggregation of states would produce a nearly lumpable chain whose error, in approximating the original rates, is insignificant and decreasing as the number of original states increases (Supplementary Material, section S6).

Lumpability and reality. – Apparently, the exact conditions for strong and weak lumpability are unlikely to be encountered in nature due to their symmetry constraints in the rates and base frequency vector. Moreover, the weak lumpability is generally not relevant for modelling characters on phylogenetic trees as character reconstruction is largely insensitive to the base frequency vector (Yang 2006). Nevertheless, the case of weak lumpability when the original chain is stationary may occur on large trees. The stationary distribution might be an approximation to the character distribution on branches located closer to tips if the process is time-homogeneous. Since, these branches predominate on large trees, the stationary state of Markov chain will occupy majority of branches on this tree. In this respect, the stationary state will be a good approximation of the global dynamics, thereby allowing lumping the chain under any possible partitioning scheme. There are also no evidence supporting frequent occurrence of nearly lumpable chains as violations of *i.i.d.* conditions can be widespread. Overall, this suggests that lumpable chains are very unlikely to be encountered in nature; nevertheless, as in the case of

nearly lumpable chains, the original process can be sometimes efficiently approximated by the aggregated simple Markov chains.

Hidden Markov chains and traits

Use of HMC for character modelling avoids errors associated with non-lumpable Markov chains and, at the same time, provides new insight into the process driving trait evolution. The HMC model consists of the two layers (Fig. 5a): the observable corresponding to trait states and the hidden one corresponding to the GRN states; the transitions between states are allowed only within the hidden layer. The observable layer represents a mapping of the GRN states onto the observable trait states, thus perfectly matching the concept of phenotypic character construction. The structure of HMC avoids aggregation of the GRN states and thereby does not suffer if lumpability conditions are not fulfilled. This allows the realistic modeling of the Type #2 and Type #3 correspondence between GRNs and traits. Additionally, the properties of hidden Markov chains offer a somewhat different interpretation for such focal concepts of phylogenetics as character, rate heterogeneity and homology.

Rate heterogeneity and GRN states. – Initially, HMC were mainly proposed to accommodate the heterogeneity evolutionary rate across time (Tuffley and Steel 1998; Beaulieu et al. 2013). It is worth noting, however, that both time-heterogeneity and complexity of the underlying GRN space are confounded in HMC. The HMC does not separate switches between hidden rate categories and hidden GRN states: when a hidden transition happens it can be either of those or both simultaneously. So, HMC accommodates both the complexity of GRN space and rate variation at the same time.

Trait homology and HMC. – In phylogenetics, every character statement is a hypothesis of primary homology (Hawkins et al. 1997). The “real” assessment of whether an observation in

one species is homologous or homoplasious in respect to the other species can be done only through phylogenetic inference or reconstruction of character evolution (secondary homology). It is considered that a thoughtful statement of primary homology is a prerequisite for the successful comparative analysis and phylogenetic inference. Obviously, in simple Markov models and parsimony, the secondary homology is dependent on primary homology statement (Agnarsson and Coddington 2008; Brazeau 2011) since the construction of a discrete character is a subjective procedure. The discordance between primary and secondary homology, to a large extent, occurs due to the discordance between the morphological traits and GRNs. The HMC can directly account for this discordance. Theoretically, this means that the quality of primary homology statement should not matter as the underlying hidden space of evolution can be automatically adjusted by the HMC. The practical aspects of this adjustment need further research.

IMPLEMENTING STRUCTURED AND HIDDEN MARKOV MODELS

Modelling character onto a known phylogenetic tree

The HMC provides the ability to accommodate hidden evolutionary space and rate heterogeneity, whereas structured Markov chains can efficiently incorporate hierarchical dependencies. These two approaches can be integrated by equipping structured Markov chains with hidden states, thus allowing the model to account for all aspects of trait evolution simultaneously. The provided theoretical ground suggests that this eliminates subjectivity in character coding and homology assessment. Moreover, this integration makes morphological character scale-free. This means that a part of phenotype can be a character and the entire set of all phenotypic traits can be one complex character as well (see the “Combining arbitrary number

of characters" section). So, structured Markov chains equipped with hidden states offer a flexible approach for constructing discrete characters from morphological observations.

If a trait is modeled onto a known phylogenetic tree to reconstruct ancestral states and evolutionary rates, then the incorporation of hierarchical and hidden processes is straightforward. The hierarchical processes can be incorporated by any software which allow specification of user-defined rate matrices. These rate matrices have to be parametrized using approaches summarized above to reflect the dependencies between anatomical parts. These dependencies can be retrieved from anatomy ontologies or prior knowledge of organismal anatomies. The flexible incorporation of hidden process, in the context of this study, is presently provided by only two software packages *corHMM* (Beaulieu et al. 2013; Beaulieu and O'Meara 2014) and *RevBayes* (Höhna et al. 2016).

CorHMM. – This likelihood-based method assumes that every observable state consists of a specified number of hidden states. The hidden states were originally proposed to model rate heterogeneity among the observable states (Beaulieu et al. 2013); however the hidden states can be equally interpreted as different GRN states. This method offers full flexibility to accommodate all necessary correspondences (type #2 and #3) between morphology and GRNs. Although, it requires *a priori* specification of the number of hidden states and their mapping with the observable states, it allows testing different models by manually varying the number and mapping of the hidden states; so the best parametrization of the matrix can be selected using e.g., Akaike information criterion (Akaike 1974). Thus, the potential space of possible models can be explored. This flexibility is extremely important as the structure of transitions between hidden states as well as their number is unknown *a priori*, and therefore have to be tested during the analysis.

RevBayes. – This software provides a broad flexibility for developing new phylogenetic models using the graphical model concept and *Rev* language. The *RevBayes* supports implementation of hidden Markov models using *expandCharacters("number of hidden states")* method as well as specification of the user-defined rate matrices with *fnFreeK()* function. These two features are sufficient to incorporate the hierarchical and hidden process reviewed in this paper. Since the modelling capabilities of *RevBayes* are enormous and require programming in the *Rev* language, I direct interested reader to the original paper (Höhna et al. 2016) for the additional information.

HMC and data. – Testing between simple Markov chains and HMC to select the best model must be a prerequisite of any comparative analysis. However, it is worth mentioning that the performance of HMC depends on the number of taxa and thereby small trees are not likely to favor HMC over SMC (Beaulieu et al. 2013). On small trees, SMC can be an efficient approximation of the underlying complex process. Even if data is sufficient and favor HMC, one should not expect that the selected HMC perfectly reconstructs the underlying GRN space as it merely performs the best approximation of the underlying hidden space.

Phylogenetic inference using morphology

Unlike DNA data, where states are the same across different sites, morphological data are challenging to parametrize. This occurs because morphological states are not aligned across characters in a sense that ***"state 1"*** in one character is not the same as ***"state 1"*** in another. This issue poses a general barrier for modelling rate heterogeneity (Lewis 2001) as well as hidden processes in phylogenetic inference. Theoretically, HMC models can be used in phylogenetic inference with morphology but their implementation will be limited; it will likely require assignment of the same HMM with *a priori* specified hidden states to all characters of the same

partition. This approach substantially restricts the possibility for exploring hidden space of evolution. However, the ontology-informed characters using structured Markov models can be implemented in inference. Presently, this can be done using the *RevBayes* software that allows using user-defined rate matrices with *fnFreeK()* function. In this respect, characters which are known to be interdependent have to be merged in one character and assigned to a separate partition. This approach does not require using inapplicable coding. Next, a user-defined rate matrix characterizing dependencies within such characters can be applied to this partition.

CONCLUSIONS

This paper lays out the theoretical consideration for improving the modeling of discrete morphological characters in phylogenetic context. The main suggestion of the paper from the theoretical perspective is to use structured Markov models equipped with hidden states to accommodate the complexity of processes driving the evolution of traits. The approaches summarized here can be used for developing new methods aiming at reconstructing correlated trait evolution and ancestral ontologies. Since the considerations summarized here are mainly theoretical, further empirical research is needed to understand the behavior of the proposed models.

SUPPLEMENTARY MATERIAL

The supplementary file is available from the BioRxiv (www.biorxiv.org).

ACKNOWLEDGEMENTS

This work was conducted while a Postdoctoral Fellow at the National Institute for Mathematical and Biological Synthesis, an Institute sponsored by the National Science Foundation through NSF Award #DBI-1300426, with additional support from The University of Tennessee, Knoxville. Thanks are due to Brian O'Meara (University of Tennessee), Sergey Gavrilets (University of Tennessee), Sarah Flanagan (National Institute for Mathematical and Biological Synthesis, University of Tennessee), and Dimitar Dimitrov (University of Copenhagen, Natural History Museum of Denmark) for useful suggestions on the text of the manuscript.

REFERENCES

Abouheif E. 1999. Establishing homology criteria for regulatory gene networks: prospects and challenges. Novartis Foundation symposium. 222:207–21; discussion 222–5.

Agnarsson I., Coddington J.A. 2008. Quantitative tests of primary homology. Cladistics. 24:51–61.

Akaike H. 1974. A new look at the statistical model identification. IEEE transactions on automatic control. 19:716–723.

Arendt D., Musser J.M., H Baker C.V., Bergman A., Cepko C., Erwin D.H., Pavlicev M., Schlosser G., Widder S., Laubichler M.D., Wagner G.P. 2016. The origin and evolution of cell types. Nature Publishing Group. 17.

Babu M.M., Luscombe N.M., Aravind L., Gerstein M., Teichmann S.A. 2004. Structure and evolution of transcriptional regulatory networks. Current Opinion in Structural Biology. 14:283–291.

Balhoff J.P., Dahdul W.M., Kothari C.R., Lapp H., Lundberg J.G., Mabee P., Midford P.E., Westerfield M., Vision T.J. 2010. Phenex: ontological annotation of phenotypic diversity. PLoS One. 5:e10500.

Beaulieu J.M., O'Meara B.C., Donoghue M.J. 2013. Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in campanulid angiosperms. Systematic biology. 62:725–37.

Beaulieu J.M., O'Meara B.C. 2014. Hidden Markov models for studying the evolution of binary morphological characters. Modern Phylogenetic Comparative Methods and their Application in Evolutionary Biology. Springer. p. 395–408.

Van Bocxlaer I., Loader S.P., Roelants K., Biju S., Menegon M., Bossuyt F. 2010. Gradual adaptation toward a range-expansion phenotype initiated the global radiation of toads. Science. 327:679–682.

Brazeau M.D. 2011. Problematic character coding methods in morphology and their effects. Biological Journal of the Linnean Society. 104:489–498.

Burleigh G., Alphonse K., Alverson A.J., Bik H.M., Blank C., Cirranello A.L., Cui H., Daly M., Dietterich T.G., Gasparich G., Irvine J., Julius M., Kaufman S., Law E., Liu J., Moore L., O'Leary M.A., Passarotti M., Ranade S., Simmons N.B., Stevenson D.W., Thacker R.W., Theriot E.C., Todorovic S., Velazco P.M., Walls R.L., Wolfe J.M., Yu M. 2013. Next-generation phenomics for the Tree of Life. PLoS Currents.

Carroll S.B. 2008. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. Cell. 134:25–36.

Clark-Hachtel C.M., Linz D.M., Tomoyasu Y. 2013. Insights into insect wing origin provided by functional analysis of vestigial in the red flour beetle, Tribolium castaneum. Proceedings of the National Academy of Sciences. 110:16951–16956.

Deans A.R., Lewis S.E., Huala E., Anzaldo S.S., Ashburner M., Balhoff J.P., Blackburn D.C., Blake J.A., Burleigh J.G., Chanet B., others. 2015. Finding our way through phenotypes. PLoS Biol. 13:e1002033.

Dececchi T.A., Balhoff J.P., Lapp H., Mabee P.M. 2015. Toward Synthesizing Our Knowledge of Morphology: Using Ontologies and Machine Reasoning to Extract Presence/Absence Evolutionary Phenotypes across Studies. Systematic biology. 64:936–52.

Dececchi T.A., Mabee P.M., Blackburn D.C., Strier K., Altmann J., Brockman D., Bronikowski A., Cords M., Fedigan L., Tacutu R., Craig T., Budovsky A., Wuttke D., Lehmann G., Taranukha D., Smith C., Finger J., Hayamizu T., McCright I., Xu J., Berghout J., Sprague J., Clements D., Conlin T., Edwards P., Frazer K., Schaper K., Parr C., Schulz K., Hammock J., Leary P., Hammock J., Rice J., O'Leary M., Kaufman S., Mabee P., Balhoff J., Dahdul W., Lapp H., Midford P., Vision T., Deans A., Lewis S., Huala E., Anzaldo S., Ashburner M., Balhoff J., Mabee P., Ashburner M., Cronk Q., Gkoutos G., Haendel M., Segerdell E., Dahdul W., Balhoff J., Engeman J., Grande T., Hilton E., Kothari C., Vogt L., Nickel M., Jenner R., Deans A., Dececchi T., Balhoff J., Lapp H., Mabee P., Deans A., Yoder M., Balhoff J., Vogt L., Bartolomaeus T., Giribet G., Göpel T., Richter S., Richter S., Wirkner C., Poe S., Wiens J., Wiens J., O'Keefe F., Wagner P., Wake D., Wake D., Wake D., Wake M., Specht C., West-Eberhard M., Daeschler E., Shubin N., Thomson K., Amaral W., Shubin N., Daeschler E., Coates M., Mabee P., Edmunds R., Su B., Balhoff J., Eames B., Dahdul W., Lapp H., Boisvert C., Boisvert C., Garvey J., Johanson Z., Warren A., Shubin N., Daeschler E., Jr F.J., Shubin N., Daeschler E., Jr F.J., Boisvert C., Mark-Kurik E., Ahlberg P., Coates M., Daeschler E., Shubin N., Jr F.J., Mungall C., Gkoutos G., Washington N., Lewis S., Mungall C., Gkoutos G., Smith C., Haendel M., Lewis S., Ashburner M., Balhoff J., Dahdul W., Kothari C., Haendel M., Lewis S., Ashburner M., Mungall C., Torniai C., Gkoutos G., Lewis S., Haendel M., Haendel M., Balhoff J., Bastian F., Blackburn D., Blake J., Bradford Y., Dahdul W., Lundberg J., Midford P., Balhoff J., Lapp H., Vision T., Dahdul W., Balhoff J., Blackburn D., Diehl A., Haendel M., Hall B., Bastian F., Parmentier G., Roux J., Moretti S., Laudet V., Robinson-Rechavi M., Bairoch A., Cohen-Boulakia S., Froidevaux C., Niknejad A., Comte A., Parmentier G., Roux J., Bastian F., Robinson-Rechavi M., Gkoutos G., Green E., Mallon A.-M., Hancock J., Davidson D., Midford P., Dececchi T., Balhoff J., Dahdul W., Ibrahim N., Lapp H., Jarvik E., Jarvik E., Lebedev O., Coates M., Dahdul W., Dececchi T., Ibrahim N., Lapp H., Mabee P., Swartz B., Sereno P., Kazakov Y., Krötzsch M., Simancík F., Clack J., Ahlberg P., Blom H., Finney S., Ruta M. 2016. Data Sources for Trait Databases: Comparing the Phenomic Content of Monographs and Evolutionary Matrices. PLOS ONE. 11:e0155680.

Erkenbrack E.M., Davidson E.H. 2015. Evolutionary rewiring of gene regulatory network linkages at divergence of the echinoid subclasses. Proceedings of the National Academy of Sciences. 112:E4075–E4084.

Erwin D.H., Davidson E.H. 2009. The evolution of hierarchical gene regulatory networks. Nature Reviews Genetics. 10:141–148.

Fitzhugh K. 2006. The philosophical basis of character coding for the inference of phylogenetic hypotheses. Zoologica Scripta. 35:261–286.

Forey P.L., Kitching I. 2000. Experiments in coding multistate characters. Systematics Association Special Volume. 58:54–80.

Glassford W.J., Johnson W.C., Dall N.R., Smith S.J., Liu Y., Boll W., Noll M., Rebeiz M. 2015. Co-option of an Ancestral Hox-Regulated Network Underlies a Recently Evolved Morphological Novelty. Developmental Cell. 34:520–531.

Hall B.K. 2003. Descent with modification: the unity underlying homology and homoplasy as seen through an analysis of development and evolution. Biological Reviews. 78:S1464793102006097.

Hawkins J.A., Hughes C.E., Scotland R.W. 1997. Primary Homology Assessment, Characters and Character States. Cladistics. 13:275–283.

Held L.I. 2014. How the snake lost its legs: curious tales from the frontier of evo-devo. Cambridge University Press.

Hiller M., Schaar B.T., Indjeian V.B., Kingsley D.M., Hagey L.R., Bejerano G. 2012. A “forward genomics□? approach links genotype to phenotype using independent phenotypic losses among related species. Cell reports. 2:817–823.

Hinman V.F., Cheatle Jarvela A.M. 2014. Developmental gene regulatory network evolution: Insights from comparative studies in echinoderms. genesis. 52:193–207.

Höhna S., Landis M.J., Heath T.A., Boussau B., Lartillot N., Moore B.R., Huelsenbeck J.P., Ronquist F. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Systematic biology.syw021.

Houle D., Govindaraju D.R., Omholt S. 2010. Phenomics: the next challenge. Nature Reviews Genetics. 11:855–866.

Huelsenbeck J.P., Nielsen R., Bollback J.P. 2003. Stochastic mapping of morphological characters. Systematic Biology. 52:131–158.

Kemeny J.G., Snell J.L. 1960. Finite markov chains. van Nostrand Princeton, NJ.

Kuratani S. 2009. Modularity, comparative embryology and evo-devo: Developmental dissection of evolving body plans. Developmental Biology. 332:61–69.

Lee D.-C., Bryant H.N. 1999. A Reconsideration of the Coding of Inapplicable Characters: Assumptions and Problems. Cladistics. 15:373–378.

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

Longabaugh W.J.R., Davidson E.H., Bolouri H. 2005. Computational representation of developmental genetic regulatory networks. Developmental Biology. 283:1–16.

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

Manda P., Balhoff J.P., Lapp H., Mabee P., Vision T.J. 2015. Using the Phenoscape Knowledgebase to relate genetic perturbations to phenotypic evolution. genesis. 53:561–571.

McCune A.R., Schimenti J.C. 2012. Using genetic networks and homology to understand the evolution of phenotypic traits. Current genomics. 13:74–84.

McDougall C., Degnan B.M., Lowe C., Wu M., Salic A., Evans L., Lander E., Stange-Thomann N., Gruber C., Gerhart J., Kirschner M., Yankura K., Martik M., Jennings C., Hinman V., Denes A., Jékely G., Steinmetz P., Raible F., Snyman H., Prud B., Wawersik S., Maas R., Nielsen C., Degnan S., Degnan B., Jackson D., Meyer N., Seaver E., Pang K., McDougall C., Moy V., Gordon K., Degnan B., Martindale M., Burke R., Peterson K., Dunn E., Moy V., Angerer L., Angerer R., Morris R., Peterson K., Collin R., Raff R., Byrne M. 2011. Modularity of gene-regulatory networks revealed in sea-star development. BMC Biology. 9:6.

McKeown A.N., Bridgham J.T., Anderson D.W., Murphy M.N., Ortlund E.A., Thornton J.W. 2014. Evolution of DNA Specificity in a Transcription Factor Family Produced a New Gene Regulatory Module. Cell. 159:58–68.

Moczek A.P. 2008. On the origins of novelty in development and evolution. BioEssays. 30:432–447.

Monteiro A. 2012. Gene regulatory networks reused to build novel traits. BioEssays. 34:181–186.

Mungall C.J., Torniai C., Gkoutos G.V., Lewis S.E., Haendel M.A. 2012. Uberon, an integrative multi-species anatomy ontology. Genome biology. 13:R5.

Nodelman U., Shelton C.R., Koller D. 2002. Continuous time Bayesian networks. Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence.378–387.

O'Meara B.C. 2012. Evolutionary Inferences from Phylogenies: A Review of Methods. http://dx.doi.org/10.1146/annurev-ecolsys-110411-160331.

Oakley T.H. The eye as a replicating and diverging, modular developmental unit. .

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

Patterson C. 1982. Morphological characters and homology. Problems of phylogenetic reconstruction. 21:21–74.

Pinna M.C. 1991. Concepts and tests of homology in the cladistic paradigm. Cladistics. 7:367–394.

Pleijel F. 1995. On character coding for phylogeny reconstruction. Cladistics. 11:309–315.

Price S.A., Wainwright P.C., Bellwood D.R., Kazancioglu E., Collar D.C., Near T.J. 2010. Functional innovations and morphological diversification in parrotfish. Evolution. 64:3057–3068.

Prud 'homme B., Gompel N., Rokas A., Kassner V.A., Williams T.M., Yeh S.-D., True J.R., Carroll S.B. Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. .

Prud'homme B., Gompel N., Carroll S.B. 2007. Emerging principles of regulatory evolution. Proceedings of the National Academy of Sciences of the United States of America.8605–12.

Prud'Homme B., Gompel N., Rokas A., Kassner V.A., Williams T.M., Yeh S.-D., True J.R., Carroll S.B. 2006. Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. Nature. 440:1050–1053.

Ramirez M.J., Michalik P. 2014. Calculating structural complexity in phylogenies using ancestral ontologies. Cladistics. 30:635–649.

Ramirez M.J. 2007. Homology as a parsimony problem: a dynamic homology approach for morphological data. Cladistics. 23:588–612.

Rebeiz M., Patel N.H., Hinman V.F. 2015. GG16CH05-Rebeiz Unraveling the Tangled Skein: The Evolution of Transcriptional Regulatory Networks in Development. Annu. Rev. Genomics Hum. Genet. 16:103–31.

Rubino G., Sericola B. 1993. A finite characterization of weak lumpable Markov processes. Part II: The continuous time case. Stochastic processes and their applications. 45:115–125.

Rubino G., Sericola B. 2014. Markov chains and dependability theory. Cambridge University Press.

Sereno P.C. 2007. Logical basis for morphological characters in phylogenetics. Cladistics. 0:070907095847001–???

Shapiro M.D., Bell M.A., Kingsley D.M. 2006. Parallel genetic origins of pelvic reduction in vertebrates. Proceedings of the National Academy of Sciences. 103:13753–13758.

Shbailat S.J., Abouheif E. 2013. The Wing-Patterning Network in the Wingless Castes of Myrmicine and Formicine Ant Species Is a Mix of Evolutionarily Labile and Non-Labile Genes. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution. 320:74–83.

Shelton C.R., Ciardo G. 2014. Tutorial on Structured Continuous-Time Markov Processes. J. Artif. Intell. Res.(JAIR). 51:725–778.

Siegal M.L. 2013. Evolution of molecular networks. The Princeton guide to evolution. Princeton University Press, Princeton.428–436.

Strong E.E., Lipscomb D. 1999. Character Coding and Inapplicable Data. Cladistics. 15:363–371.

Tobias J.A., Cornwallis C.K., Derryberry E.P., Claramunt S., Brumfield R.T., Seddon N. 2014. Species coexistence and the dynamics of phenotypic evolution in adaptive radiation. Nature. 506:359–363.

Tuffley C., Steel M. 1998. Modeling the covarion hypothesis of nucleotide substitution. Mathematical biosciences. 147:63–91.

Vera-Ruiz V.A., Lau K.W., Robinson J., Jermiin L.S. 2014. Statistical tests to identify appropriate types of nucleotide sequence recoding in molecular phylogenetics. BMC bioinformatics. 15:S8.

Vogt L. 2016. Assessing similarity: on homology, characters and the need for a semantic approach to non-evolutionary comparative homology. Cladistics.

Vogt L. 2017a. The logical basis for coding ontologically dependent characters. Cladistics.

Vogt L. 2017b. Towards a semantic approach to numerical tree inference in phylogenetics. Cladistics.

Voordeckers K., Pougach K., Verstrepen K.J. 2015. How do regulatory networks evolve and expand throughout evolution? Current Opinion in Biotechnology. 34:180–188.

Wagner G.P. 2007. The developmental genetics of homology. Nature Reviews Genetics. 8:473–479.

Wagner G.P. 2015. Homology in the Age of Developmental Genomics. Evolutionary Developmental Biology of Invertebrates 1. Vienna: Springer Vienna. p. 25–43.

Wake D.B., Wake M.H., Specht C.D., Jablonka E., Raz G., Wake M.H., Wake D.B., Larson A., Wake D.B., Hall B.K., Hall B.K., Diogo R., Arendt J., Reznick D.N., Arendt J., Reznick D.N., Leander B.S., Davidson E.H., Donoghue M.J., Ree R.H., Shapiro M.D., Albert A.Y., Chan Y.F., Schluter D., Marchinko K.B., Barrett

R.D.H., Rogers S.M., King M.-C., Wilson A.C., Nachman M.W., Hoekstra H.E., D'Agostino S.L., Manceau M., Domingues V.S., Linnen D.R., Rosenblum E.B., Hoekstra H.E., Zanis M.J., Soltis P.S., Qiu Y.L., Zimmer E., Soltis D.E., Craene L.P.R.D., Irish V.F., Gregory T.R., Alberch P., Gale E., Roth G., Nishikawa K.C., Wake D.B., Shubin N., Tabin C., Carroll S., Shubin N., Tabin C., Carroll S., Carroll S.B., Piatigorsky J., Kirchoff B.K., Lagomarsino L.P., Newman W.H., Bartlett M.E., Specht C.P., Kramer E.M., Abouheif E., Preston J.C., Hileman L.C., Collin R., Miglietta M.P., Kohlsdorf T., Wagner G.P., Goldberg E.E., Igic B., Galis F., Arntzen J.W., Lande R., Marshall C.R., Raff E.C., Raff R.A., Kohlsdorf T., Lynch V.J., Rodrigues M.T., Brandley M.C., Wagner G.P., Zufall R.A., Rausher M.D., Igic B., Lande R., Kohn J.R., Hoekstra H.E., Coyne J.A., Mueller R.L., Macey J.R., Jaekel M., Wake D.B., Boore J.L., Parra-Olea G., Wake D.B., Wake M.H., Gower D.J., Giri V., Dharne M.S., Shouche Y.S., Coen E.S., Meyerowitz E.M., Ditta G., Pinyopich A., Robles P., Pelaz S., Yanofsky M.F. 2011. Homoplasy: from detecting pattern to determining process and mechanism of evolution. Science (New York, N.Y.). 331:1032–5.

Wiens J.J. 2001. Character Analysis in Morphological Phylogenetics: Problems and Solutions. Syst. Biol. 50:689–699.

Yang Z. 2006. Computational molecular evolution. Oxford University Press.

Yoder M.J., Miko I., Seltmann K.C., Bertone M.A., Deans A.R. 2010. A gross anatomy ontology for Hymenoptera. PloS one. 5:e15991.

BOXES

Trait	An observation of some feature(s) of phenotype.
Character	A formalized coding of a trait (observation) into a string or matrix (i.e. character) that consists of two or more entities called "character states."
Phenotype	A set of all traits of an organism

Box 1. Definitions of the terms employed in this paper. Alternative definitions are reviewed by Sereno (2007).

LEGENDS TO FIGURES

Figure 1. Ontology-informed character.

The exemplar case of dependence between two characters: (i) tail presence and (ii) tail color. The character (ii) is dependent on the state present of character (i). This dependence is imposed by the ontological relationships between body parts. The green links show various types of ontological relationships between characters and between entities of UBERON anatomy ontology (Mungall et al. 2012).

Figure 2. Coding schemes.

Three alternative schemes of coding tail traits (presence and color) from Fig. 1. Scheme #1 **(a)** uses two characters, scheme #2 **(b)** uses three binary characters, and scheme #3 **(e)** uses one multistate character to encode tail traits. The traits states are indicated by orange balls which are explained in **(d)**; the latter also explains species used in coding traits. All three coding schemes (**a**, **b** and **e**) imply identical relationships between characters captured by the dependency graph **(c)**. This graph is described by the structured Markov chain shown in **(e)**.

Figure 3. Dependent and independent coevolution of several characters.

This plot exemplifies combined matrices and their rate symmetries characterizing the coevolution of several characters. Different rates within matrix are differently colored. Grey lines dividing the combined matrices exemplify their partitioning. The matrices are composed of the following elementary characters: **(a, b, d)** ten two-state characters; **(c)** six two-state characters; **(e)** three three-state characters; **(f)** four five-state characters; **(g)** three ten-state characters. The rate symmetries and dependencies between elementary characters are: **(a, b, f g)** independently coevolving characters with equal rates across and within characters; **(c)** independently coevolving characters with different rates across characters; **(e)** independently coevolving characters with different rates across and within characters; **(d)** dependently coevolving characters with all rates different. **(b)** an **(g)** exemplify balanced portioning schemes preserving the strong lumpability. **(a)** and **(f)** exemplify partitioning scheme violating strong lumpability; *P* indicates the partitioning block that violates the row-wise sum rule of strong lumpability.

Figure 4. Correspondence between morphological and GRN states.

Trait of tail color is divided into two states (shown with rectangles): blue and red. Hypothetical GRN states producing the states of the trait are shown with balls. The correspondence between GRN states and trait states can be of three types. **(a)** Type #1, one-to-one correspondence: each GRN state corresponds to its own trait state. **(b)** Type #2, many-to-one correspondence: each trait state is composed of several GRN states. **(c)** Type #3, partial correspondence: only part of GRN states correspond to the trait states; the remaining GRN state may be involved in production of other traits.

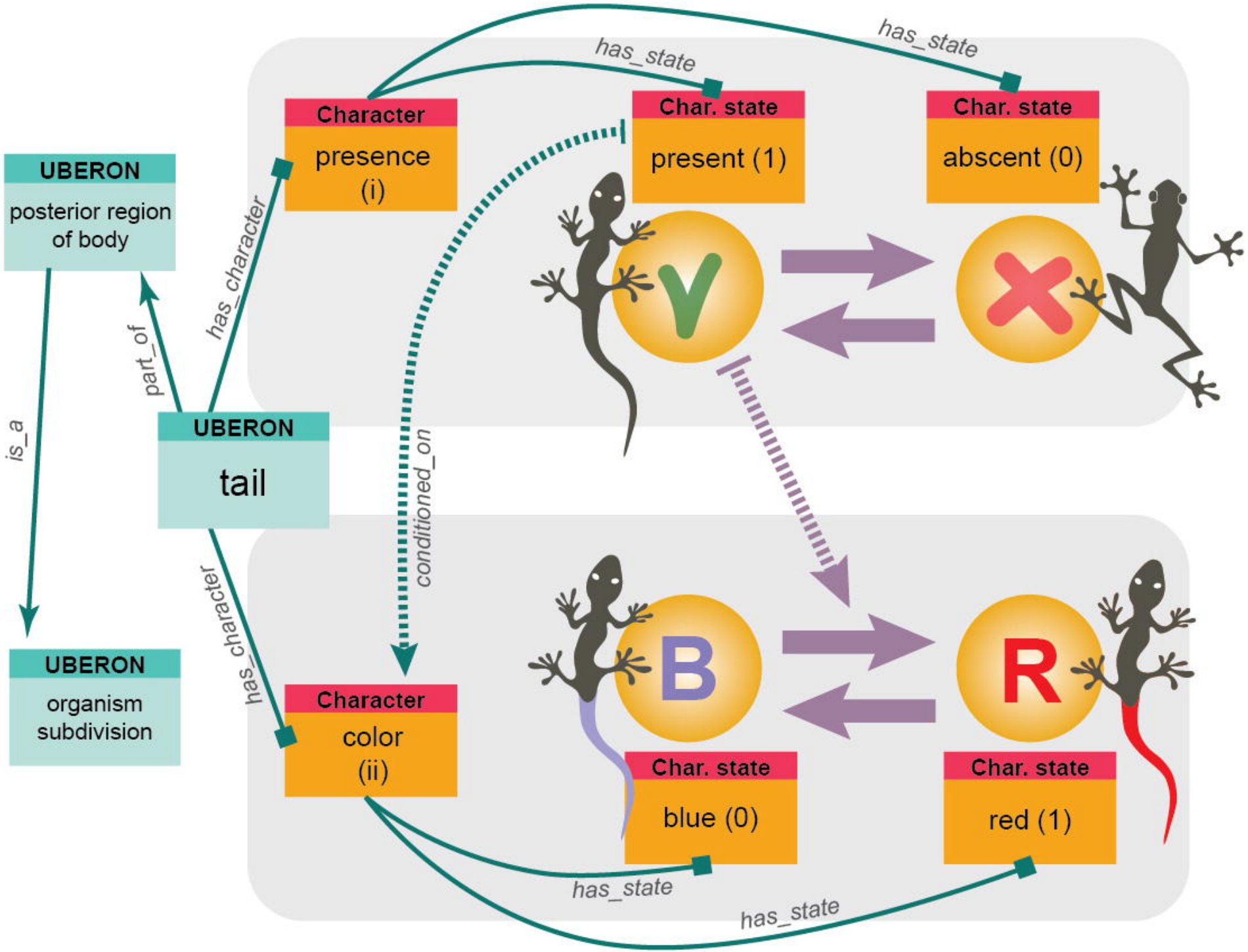
Figure 5. Hidden Markov chains and lumpability.

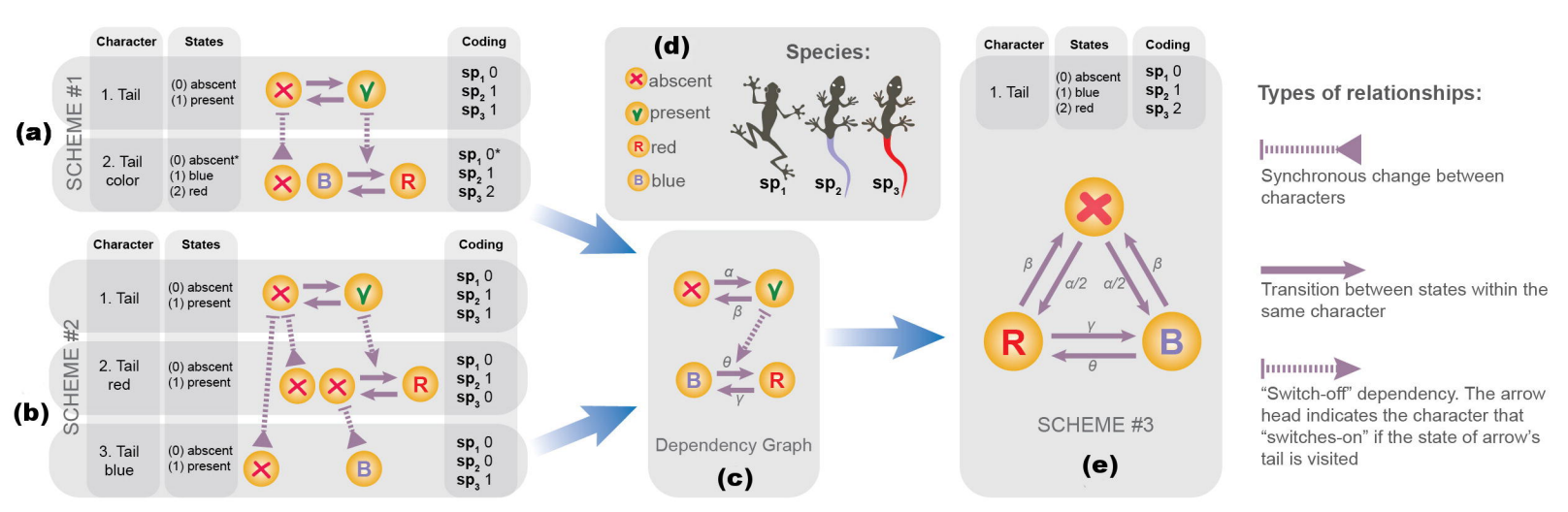
This figure exemplifies approaches by which the original three-state chain **(b)** can be substituted with a two-state chain. The reduction of state number is assumed to happen by aggregating the states *B* and *C*. One approach is to lump the original chain; this is only possible if transition rates between the states $C \rightarrow A$ and $B \rightarrow A$ are identical ($=\alpha$) which results in the aggregated chain **(c)**, where the transition rate between $C+B$ and A is equal to α**.** If transition rates $C \rightarrow A$ and $B \rightarrow A$ are different, then the original chain is not lumpable. In this case the reduction of states can be only done by substituting the original chain with a hidden Markov chain **(a)**; the latter consists of the hidden layer (three states) mapped onto the observable layer (two states); the hidden state space has the same topology as that of the original chain **(b)**.

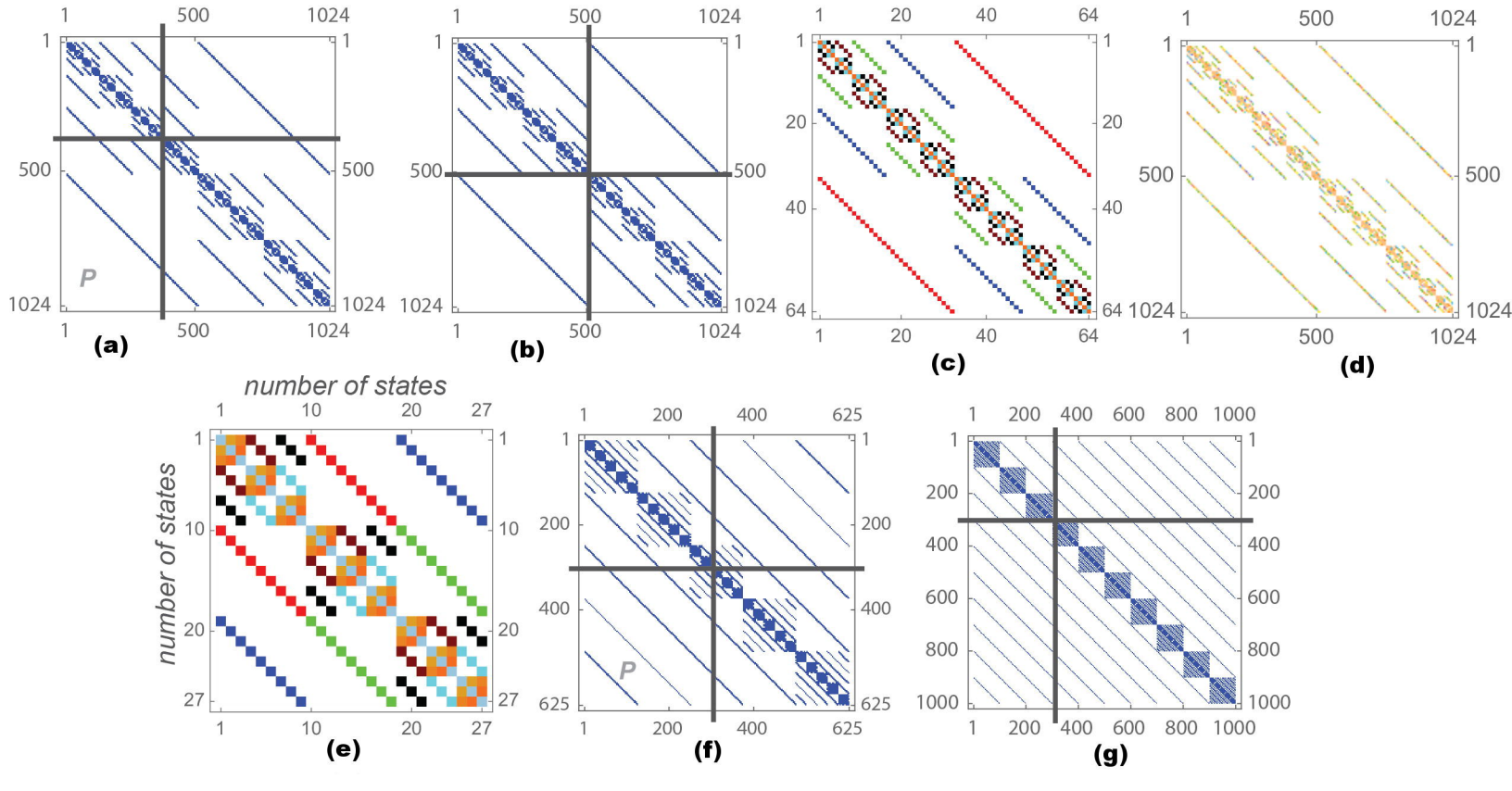
Figure 6. Lumpable Markov chain and estimation error.

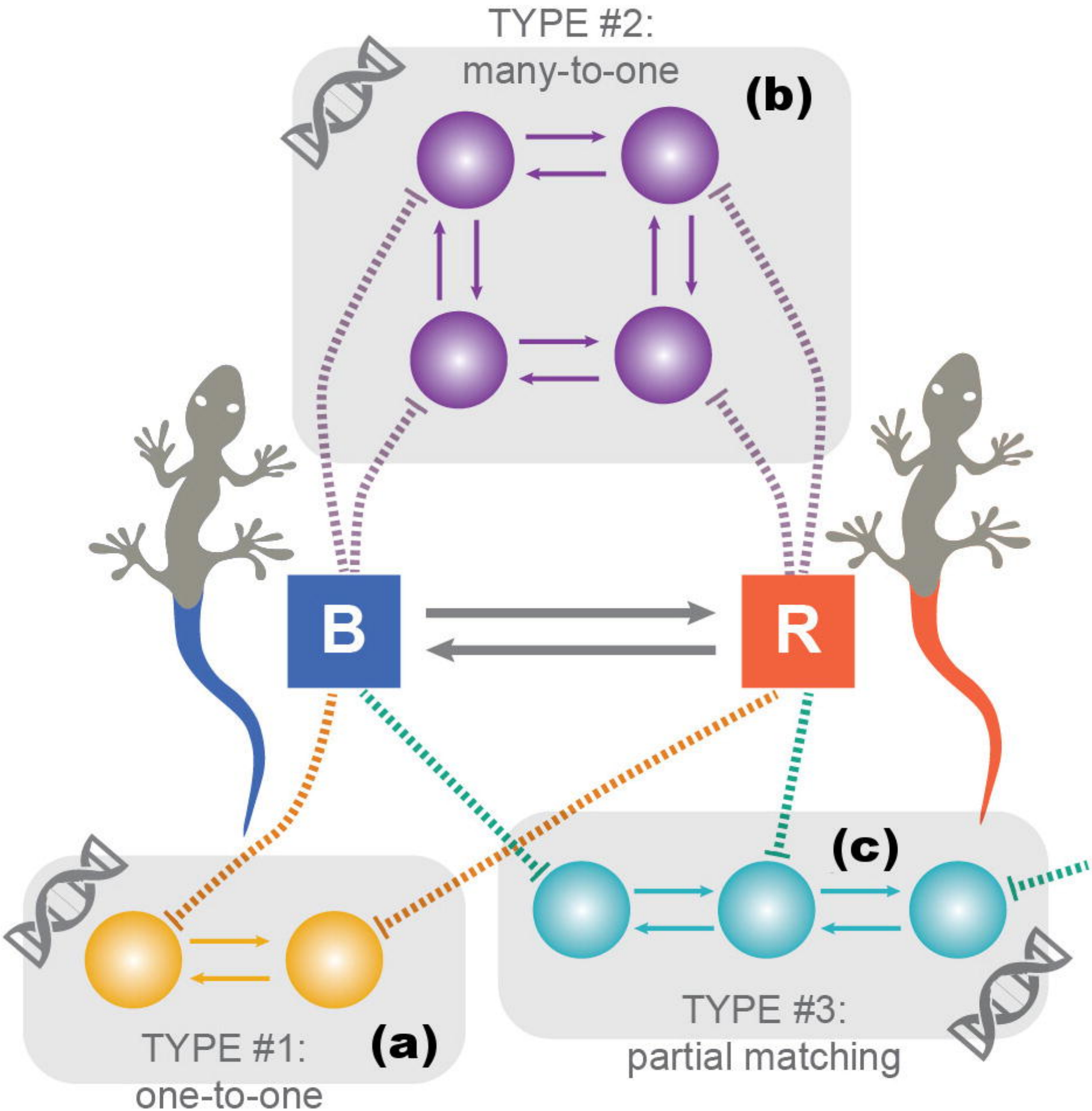
(a) Rate symmetries of the transition matrix preserving the property of strong lumpability (see the "Cases of lumpable chains" section). Partitioning of the original transition matrix (shown in purple color) yields an aggregated matrix (shown in grey color). The original matrix consists of four states $\{s_1, s_2, s_3, s_4\}$ and rate parameters q_{ij}. Its states are partitioned into groups $\{\{s_1, s_2\}, \{s_3, s_4\}\}$ yielding the aggregated matrix with the two states $\{f_1, f_2\}$ and rate parameters $\hat{q}_{ij}$. The original matrix is lumpable if the following equalities hold $\hat{q}_{12} = q_{13} + q_{14} = q_{23} + q_{24}$ and $\hat{q}_{21} = q_{31} + q_{32} = q_{41} + q_{42}$. **(b)** Error in rate estimation when strong lumpability does not hold. The plot shows an error (*y* axis) for approximating an original three-state Markov chain with an aggregated two-state Markov chain. The rates in the original chain were identical except one aggregated rate that was divided by the scaling parameter (*rate diff.*) specifying the rate ratio (*x* axis) between this and the remaining rates. The scaling parameter was used to violate strong lumpability property; the value of *rate diff.* =1 (meaning that all rates were identical) corresponds

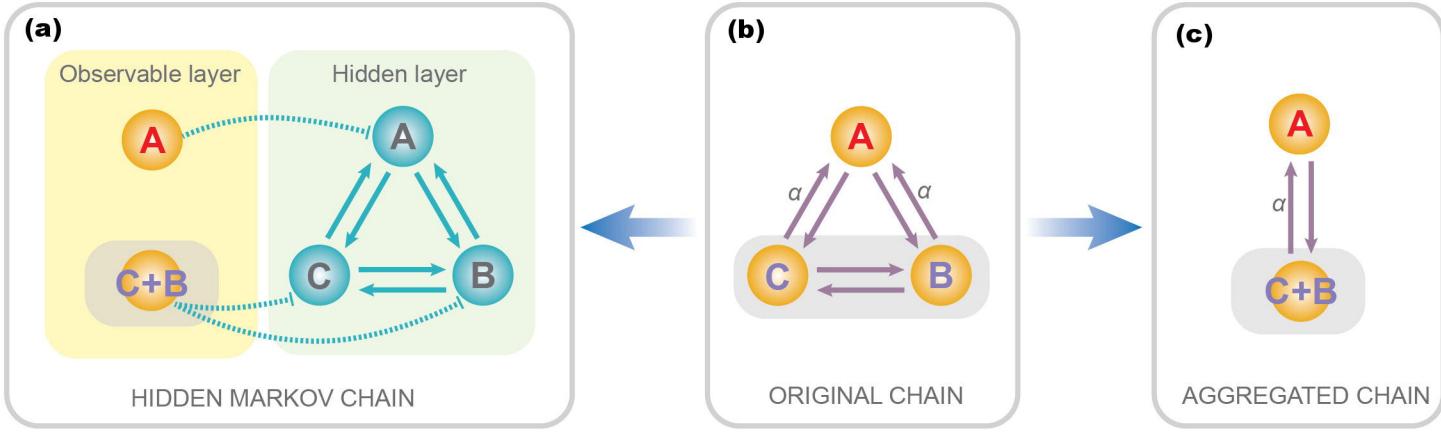
to the only instance of strong lumpability. The results are shown for a range of the transition rates $q = \{0.1, 0.5, 1, 2\}$ in the original chain (Supplementary Material, section S7).

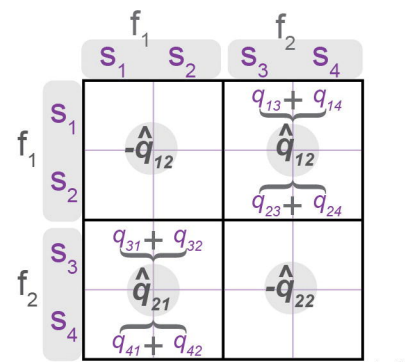

UBERON
posterior region of body
is_a
UBERON
organism subdivision
part_of
UBERON
tail
has_character
has_character
Character
presence (i)
has_state
has_state
Char. state
present (1)
Char. state
abscent (0)
conditioned_on
Character
color (ii)
has_state
has_state
Char. state
blue (0)
Char. state
red (1)
B
R


(a)
SCHEME #1
Character
States
Coding
1. Tail
(0) abscent
(1) present
sp1 0
sp2 1
sp3 1
2. Tail color
(0) abscent*
(1) blue
(2) red
sp1 0*
sp2 1
sp3 2
(b)
SCHEME #2
Character
States
Coding
1. Tail
(0) abscent
(1) present
sp1 0
sp2 1
sp3 1
2. Tail red
(0) abscent
(1) present
sp1 0
sp2 1
sp3 0
3. Tail blue
(0) abscent
(1) present
sp1 0
sp2 0
sp3 1
(d)
Species:
abscent
present
red
blue
sp1
sp2
sp3
Dependency Graph
(c)
α
β
θ
γ
Character
States
Coding
1. Tail
(0) abscent
(1) blue
(2) red
sp1 0
sp2 1
sp3 2
α/2
SCHEME #3
(e)
Types of relationships:
Synchronous change between characters
Transition between states within the same character
“Switch-off” dependency. The arrow head indicates the character that “switches-on” if the state of arrow’s tail is visited

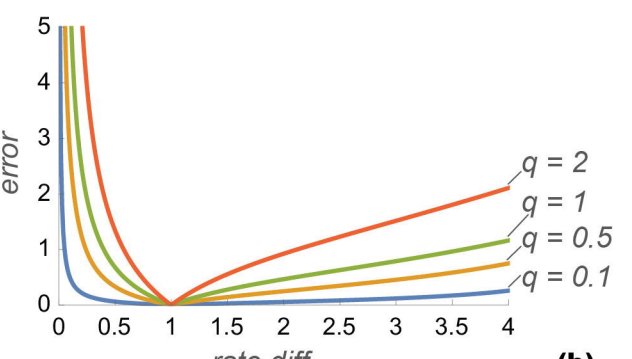

P
(a)
(b)
(c)
(d)
number of states
number of states
(e)
P
(f)
(g)


TYPE #2:
many-to-one
(b)
B
R
(c)
TYPE #3:
partial matching
TYPE #1:
one-to-one
(a)


(a)
Observable layer
Hidden layer
A
A
C+B
C
B
HIDDEN MARKOV CHAIN
(b)
A
α
α
C
B
ORIGINAL CHAIN
(c)
A
α
C+B
AGGREGATED CHAIN


f1
f2
s1
s2
s3
s4
f1
s1
s2
f2
s3
s4
-q̂12
q13+ q14
q̂12
q23+ q24
q31+ q32
q̂21
q41+ q42
-q̂22

(a)


error
rate diff.
q = 2
q = 1
q = 0.5
q = 0.1
0
1
2
3
4
5
0
0.5
1
1.5
2
2.5
3
3.5
4

(b)