Insect Systematics and Diversity, (2019) 3(6): 1; 1–7
doi: 10.1093/isd/ixz009
Method

Morphology and Ontology

PARAMO: A Pipeline for Reconstructing Ancestral Anatomies Using Ontologies and Stochastic Mapping

Sergei Tarasov,[1,2,3,6] István Mikó,[4] Matthew Jon Yoder,[5] and Josef C. Uyeda[2]

[1]Finnish Museum of Natural History, Pohjoinen Rautatiekatu 13, FI-00014 Helsinki, Finland, [2]Department of Biological Sciences, Virginia Tech, 4076 Derring Hall, 926 West Campus Drive, Blacksburg, VA 24061, [3]National Institute for Mathematical and Biological Synthesis, University of Tennessee, Knoxville, TN 37996, [4]University of New Hampshire, Durham, NH, [5]Illinois Natural History Survey, Champaign, IL, and [6]Corresponding author, e-mail: sergxf@yandex.ru

Subject Editor: Brendon Boudinot

Received 17 February, 2019; Editorial decision 13 June, 2019

Abstract

Comparative phylogenetics has been largely lacking a method for reconstructing the evolution of phenotypic entities that consist of ensembles of multiple discrete traits—entire organismal anatomies or organismal body regions. In this study, we provide a new approach named *PARAMO* (***P**hylogenetic **A**ncestral **R**econstruction of **A**natomy by **M**apping **O**ntologies*) that appropriately models anatomical dependencies and uses ontology-informed amalgamation of stochastic maps to reconstruct phenotypic evolution at different levels of anatomical hierarchy including entire phenotypes. This approach provides new opportunities for tracking phenotypic radiations and evolution of organismal anatomies.

Key words: ontology, stochastic mapping, morphology, anatomy, evolution, ancestral character state reconstruction

Ancestral character state reconstruction has been long used to gain insight into the evolution of individual traits in organisms (Pagel 1999). However, organismal anatomies (=entire phenotypes [EPs]) are not merely ensembles of individual traits, rather they are complex systems where traits interact with each other due to anatomical dependencies and developmental constraints. Individual trait approaches substantially simplify the full picture of phenotypic evolution by reducing it to only a single feature at a time, which can potentially hinder the discovery of new evolutionary patterns or even reconstruct logically impossible evolutionary scenarios. Only a handful of studies have been focused on reconstructing evolution of EPs by treating them as sets of all available traits (e.g., Sauquet et al. 2017, O'Leary et al. 2013, Peters et al. 2014). Nevertheless, even these studies still employ individual character approaches, which remains the predominant paradigm in comparative phylogenetics due to a lack of methods for modeling an EP (or its parts) as a single complex character. These limitations thereby prevent researchers from reconstructing entire organismal anatomies. To our knowledge, the only approach that attempts to overcome this problem is the parsimony-based method of Ramírez and Michalik (2014).

In this article, we propose a new pipeline called *PARAMO* (*Phylogenetic Ancestral Reconstruction of Anatomy by Mapping Ontologies*) that takes into account anatomical dependencies and uses stochastic mapping (Huelsenbeck et al. 2003) along with anatomy ontologies to reconstruct the evolution of entire organismal anatomies; this pipeline can be implemented in likelihood or Bayesian frameworks. Our approach treats the EP or its component body regions (BRs) as single complex characters and allows exploring and comparing phenotypic evolution at different levels of anatomical hierarchy. These complex characters are constructed by ontology-informed amalgamation of elementary characters (i.e., those coded in a character matrix) using stochastic maps. In our approach, characters are linked with the terms from an anatomy ontology, which allows viewing them not just as an ensemble of character state tokens but as entities that have their own biological meaning provided by the ontology.

The goal of this article is to give the description of *PARAMO* pipeline and *R* (R Core Team 2018) scripts that can be used to run it. Additionally, we use a Hymenopteran dataset to demonstrate the workflow of the pipeline. At the end of the article, we discuss biological questions that can be addressed using our method. We believe that reconstructing evolutionary dynamics of EPs and their major parts opens up new perspectives for comparative morphology and phylogenetics, which, in turn, allows tracking phenotypic radiations across time and phylogeny.

Methodological Background

The Core Ingredient: Character and Character State Invariance

At the core of our method lies the property of character and character state invariance that exists in Markov models of discrete trait evolution (Tarasov 2018, 2019). This property removes the distinction between character and character state, meaning that multiple

© The Author(s) 2019. Published by Oxford University Press on behalf of Entomological Society of America.
All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

Downloaded from https://academic.oup.com/isd/article/3/6/1/5584145 by guest on 02 September 2020

individual characters can be represented as a single character and vice versa, which makes the two concepts equivalent. In other words, the invariance property preserves the result of inference if the same trait is coded using several or single character(s) if the appropriate model of character evolution is chosen. In this article, this property is used to construct larger characters from ensembles of elementary characters by using the operation of character amalgamation, in which the states of the larger character are created from the combinations of the states of the elementary characters (see the next section). Character amalgamation is crucial for reconstructing ancestral anatomies because it offers a convenient way to incorporate anatomical dependencies and reconstruct simultaneous evolution of traits at different levels of anatomical hierarchy (= levels of amalgamation).

In the present study, we consider three major levels of amalgamation: (1) the level of anatomical dependencies (AD), (2) BRs and (3) EP. The AD level implies that anatomically dependent traits have to be amalgamated into a single character to appropriately model anatomical dependencies (see Step 2 in the *PARAMO* description). The amalgamation at the BR level implies that all individual characters associated with a particular BR become combined into a single character for comparative analysis. For example, amalgamation of all characters associated with the 'head' produces a character that describes evolution of this BR; the same can be done for legs and other BRs of interest. Construction of these characters facilitates comparison of BR evolution across phylogeny. For example, BR amalgamation can be used to address questions of whether different BRs change over the same or different branches on a phylogeny. Amalgamation of all characters in a dataset produces one gigantic character at the EP level that describes the evolution of the entire anatomy. The character amalgamation has to be performed in a mathematically consistent way, as discussed in the next section.

Character Amalgamation Using Stochastic Maps

In probabilistic models of phylogenetics, a 'character' represents a Markov process that sequentially moves from one state to another over time. The realization of this process at tips of a phylogenetic tree generates the observed character states. Discrete characters can be represented as a discrete state Markov process that is defined by a transition rate matrix containing infinitesimal rates of change between states, and an initial vector of probabilities at the root of the phylogenetic tree. Any number of individual characters can be amalgamated into one character through amalgamating their rate matrices (Tarasov 2019) that defines the joint evolution of the initial characters. In the present article, we assume that initial characters, if they are not dependent anatomically (see Step 2 in the *PARAMO* description), are independent entities that have to be independently amalgamated. Suppose there are two characters C_1{with states: *0, 1*} and C_2{with states: *0, 1*} defined by:

$$C_1 = \begin{pmatrix} 0 & 1 \\ -\alpha_1 & \alpha_1 \\ \beta_1 & -\beta_1 \end{pmatrix} \begin{matrix} 0 \\ 1, \end{matrix} \quad C_2 = \begin{pmatrix} 0 & 1 \\ -\alpha_2 & \alpha_2 \\ \beta_2 & -\beta_2 \end{pmatrix} \begin{matrix} 0 \\ 1, \end{matrix} \tag{1}$$

where α_1, α_2, β_1, and β_2 are rate parameters. Their independent amalgamation, herein denoted by $\oplus$ (the Kronecker sum), results in the following character $C_{1,2}$ with four states:

$$C_{1,2} = C_1 \oplus C_2 = \begin{matrix} 00 \\ 01 \\ 10 \\ 11 \end{matrix} \begin{pmatrix} -\alpha_2 - \alpha_1 & \alpha_2 & \alpha_1 & 0 \\ \beta_2 & -\beta_2 - \alpha_1 & 0 & \alpha_1 \\ \beta_1 & 0 & -\beta_1 - \alpha_2 & \alpha_2 \\ 0 & \beta_1 & \beta_2 & -\beta_1 - \beta_2 \end{pmatrix} \begin{matrix} 00 \\ 01 \\ 10 \\ 11 \end{matrix} \tag{2}$$

The full formula for character amalgamation can be written as $C_1 \oplus C_2 = C_1 \otimes I_{C_2} + I_{C_1} \otimes C_2$ where I_{C_2} and I_{C_1} are the identity matrices for C_1 and C_2, and $\otimes$ denotes the Kronecker product. The amalgamated character $C_{1,2}${*00,01,10,11*} is constructed by forming its states using the combinations of states from the characters C_1 and C_2. Unfortunately, the rate matrix amalgamation has a shortcoming—the number of its states grows exponentially (as 2^n for n binary characters) resulting in an enormous rate matrix that makes computations infeasible even if it is constructed from a few dozens of initial characters. Here, we propose an approach that bypasses this issue by using stochastic maps.

A stochastic map (S) is a phylogenetic tree with an instance of mapped evolutionary history of a character (i.e., state transitions) conditional on data at the tips and a Markov model used for ancestral state reconstruction (Huelsenbeck et al. 2003). This tree is divided into segments; each segment corresponds to time spent in a particular state. Thus, stochastic mapping is a function Sm that converts realization of a character rate matrix C and data to the corresponding stochastic map(s) [i.e., $S = Sm(C)$]. Both definitions of a character—using rate matrix or stochastic map(s)—are equivalent as they can be converted into each other.

Interestingly, character amalgamation can be performed by only using stochastic maps of the initial characters. Suppose, that S_1 and S_2 are the stochastic maps obtained from the realizations of the characters (=Markov processes) C_1 and C_2 over some phylogeny respectively. The amalgamation of the maps S_1 and S_2 implies construction of a joint stochastic map $S_{1,2} = S_1 \oplus S_2$ by forming new segments from the combinations of the segments in S_1 and S_2 as shown in (Fig. 1); the map $S_{1,2}$ defines the character $C_{1,2}$. In other words, the stochastic mapping performed directly on character $C_{1,2}$ is identical to the amalgamation of the stochastic maps obtained for C_1 and C_2 separately:

$$Sm(C_1 \oplus C_2) = Sm(C_1) \oplus Sm(C_2). \tag{3}$$

The amalgamation using stochastic maps is computationally cheap as it avoids gigantic rate matrices, for which matrix exponentiation is computationally challenging, and can be virtually applied to any number of elementary characters in a dataset. Thus, the invariant property necessary for reconstructing ancestral anatomies can be feasibly maintained. This approach of stochastic map amalgamation is employed in this article. Note, the proposed amalgamation technique does not allow modeling correlated character evolution, see the discussion section below for further considerations on this issue.

Querying Characters Using Ontologies

Ontologies are graphs that describe relationships (edges) among entities (nodes) from a domain of knowledge under interest. In the present study, we are specifically interested in linking morphological

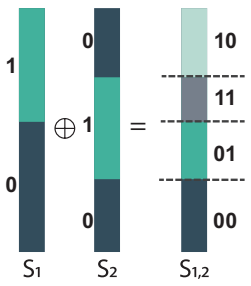

Fig. 1. Amalgamation of stochastic maps. Vertical bars are tree branches, their segments are mapped character states. The amalgamation of the stochastic map S_1{0,1} and S_2{0,1} yields the map $S_{1,2}${00,01,11,10}.

Downloaded from https://academic.oup.com/isd/article/3/6/1/5584145 by guest on 02 September 2020

characters with anatomy ontologies to be able to query and retrieve all characters associated with a particular ontology term (as, e.g., 'head', 'wing', 'legs', etc.). Herein, we call this query '*Retrieve all characters*' (*RAC*) and use it to construct character amalgamations for the different levels of the anatomical hierarchy.

In an anatomy ontology, the nodes are the anatomical entities, while edges are their relationships. Linking a character with an ontology, in the context employed here, means assigning a link between ontology term(s) and the character's ID from a character matrix. Technically, this implies that a character becomes a node in the ontology graph connected with the ontology term(s) it is linked to. Two fundamental types of edges—*is_a* and *part_of*—occur in almost all anatomy ontologies. The relationship *A is_a B* indicates that *A* is a subtype of *B*, and the relationship *A part_of B* indicates that every instance of *A* is, on the instance level, a part of some instance of *B* (Haendel et al. 2008). Computationally, *RAC* works by taking an input term and traversing the ontology graph using *is_a* and *part_of* edges to retrieve all characters that are descendant nodes of the input term. In our case, the input is an ontology term that corresponds to a BR or EP. For example, if *RAC* takes the term 'head', it returns all characters associated with this BR (e.g., 'shape of eyes', 'length of antennae', etc.). Thus, ontologies offer a convenient way to automatically query character matrices. The implementation of *RAC* is discussed in the Step 4 section of the pipeline description below.

Description of the Pipeline

Our pipeline requires three initial pieces of data: a character matrix, a dated phylogeny, and an anatomy ontology. To demonstrate the workflow, we use a modified subset of nine characters (Table 1) and 87 species from a large-scale phylogeny of Hymenoptera (Sharkey et al. 2012). The character matrix is sketched in Fig. 2A, a detailed description is given in the Supp. Material (online only). Note, the two pairs of characters in the matrix $\{C_2, C_3\}$ and $\{C_5, C_6\}$ are subject to anatomical dependencies (Table 1). For reconstructing character histories, we use the dated phylogeny of Klopfstein et al (2013), and for linking characters to the ontology, we use the Hymenoptera Anatomy Ontology (HAO; Yoder et al. 2010). In this demonstration, we are interested in constructing the amalgamated characters for the AD, BR, and EP levels of anatomical hierarchy. At the BR level, three main BRs are considered—'head', 'legs', and 'wings' (Fig. 2B–C).

The *PARAMO* pipeline includes five steps as shown in Fig. 2 and described below. In the Supp. Material (online only), we provide a set of *R* functions that can be used to implement this pipeline in practice and the tutorial (Supp. Files [online only]: *PARAMO_pipeline.pdf* or *PARAMO_pipeline.Rmd*). The newest version of the pipeline and tutorial is also available on *GitHub* https://github.com/sergeitarasov/PARAMO.

Step 1: Initial Character Matrix

Workflow

The first step requires getting or constructing an initial character matrix that codes a set of characters for a set of species. In our case, this matrix is shown in Fig. 2A, and the character report is given in Table 1.

Software

Any software for building character matrices can be used at this step, e.g., the popular software *Mesquite* (Maddison and Maddison 2018).

Step 2: Incorporating Anatomical Dependencies—Constructing Amalgamations at the AD Level

Workflow

The structure of organismal anatomies imposes anatomical dependencies among traits (i.e., the presence of digits is dependent on the presence of limbs; Fig. 2B). Coding anatomical dependencies has been a subjective procedure because different experts have different views of how to code dependent trait(s) into a character(s). Traditionally, three main coding approaches have been proposed to deal with dependencies: 1) one multistate character, 2) the presence/absence approach, or 3) the inapplicable approach. Obviously, AD traits have to be coded appropriately to avoid undesirable bias and incompatible evolutionary states that can negatively affect downstream analyses. The appropriate treatment of anatomical dependencies is discussed in Tarasov (2019) and is followed in the present pipeline. Thus, at this step, we suggest recoding the miscoded AD characters from the initial character matrix obtained in Step 1. There are two main types of dependencies—hierarchical and synchronous—that appear frequently miscoded in character matrices. Their proper treatment requires the use of different coding approaches (Tarasov 2019).

A *hierarchical dependency* occurs when a hierarchically upstream character controls a downstream one. For example, the state *present* (1) of the character C_3(*Labrum*) controls the presence of both states in C_3(*Position of labrum*), which are inapplicable otherwise (Table 1). The proper way to model this dependency is to use structured Markov models with hidden states that can be constructed by amalgamating the two characters into one as, for instance, shown in equation 2; this amalgamation results in the character $C_{2,3}\{00,01,10,11\}$ that can be represented using four states as $C_{2,3}\{0,1,2,3\}$; in $C_{2,3}$ the states $\{0, 1\}$ are hidden and correspond to the observable state *labrum absent*, while the states $\{2, 3\}$ have the same meaning as those in C_2, respectively. In the character matrix, the hidden states can be scored using polymorphic coding as $\{0 \ \& \ 1\}$ (Fig. 2B). Note,

Table 1. Initial characters of Hymenoptera used in demonstration

ID	Character statement	State 0	State 1	Dependency
C_1	Notch on medial margin of eye	Absent	Present	–
C_2	Position of labrum	Anterior	Posterior	$C_2\{0,1\} < C_3\{1\}$
C_3	Labrum	Absent	Present	$C_3\{1\} > C_2\{0,1\}$
C_4	Forewing costal and radial vein fusion	Not fused	Fused along their lengths	–
C_5	Hind wing subcostal vein, absent	No	Yes	$C_5 <> C_6$
C_6	Hind wing subcostal vein, present	Yes	No	$C_5 <> C_6$
C_7	Inner posterior mesotibial spur	Simple	Modified into a calcar	–
C_8	Foretibial apical sensillum	Present	Absent	–
C_9	Metatibial apical sensillum	Present	Absent	–

The symbols > and < indicate the direction of a hierarchical dependency; the symbol <> indicates synchronous dependency (see Step 2 of the pipeline description).

Downloaded from https://academic.oup.com/isd/article/3/6/1/5584145 by guest on 02 September 2020

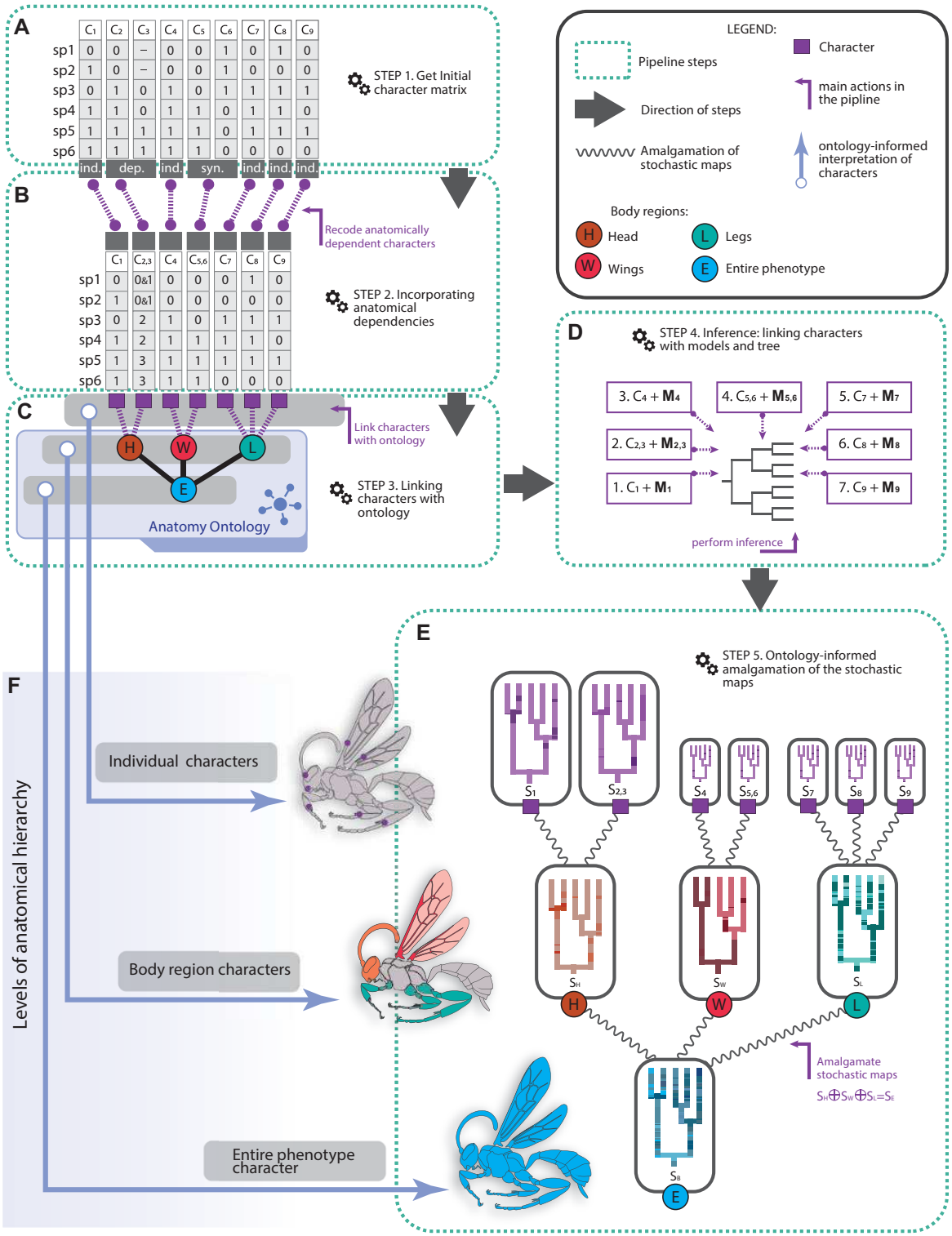

Fig. 2. *PARAMO* pipeline. The panels A–E represent the five steps of the pipeline (see the text). (E) The size of the stochastic maps S_4–S_9 is reduced for the illustrative purpose. (F) Three levels of anatomical hierarchy. Abbreviations: *C*: character, *S*: stochastic map, *ind.*: independent character, *dep.* and *syn.*: hierarchically and synchronously dependent characters, respectively.

equation 2 uses the independent amalgamation, other more complex models can be used to model AD characters as well (Tarasov 2019).

A *synchronous dependency* usually occurs when a trait is redundantly scored using a binary coding scheme. For example, the characters C_5 and C_6 (Table 1) code the same trait *presence of hind wing subcostal vein* and their character states depend on each other simultaneously: C_5 {0} and {1} occur when C_6 is {1} and {0}, respectively. This synchronous dependency has to be eliminated by combining the two characters into a single character $C_{5,6}$ without changing the state pattern (Fig. 2B).

Downloaded from https://academic.oup.com/isd/article/3/6/1/5584145 by guest on 02 September 2020

The recoding of dependent characters constructs the amalgamated characters at the AD level. If a character does not display any dependencies then we treat it as correctly amalgamated at the AD level by default. In our demonstrative example, we place the new AD characters in a separate character matrix (Fig. 2B) that is used for the downstream steps in the pipeline.

Software

Currently, there is no software that is capable of automating the recoding of the AD characters. Obviously, the AD characters can be recoded manually using any software for viewing and editing character matrices. If the number of the miscoded characters is large, manual recoding must be taken with caution, as it may result in errors.

Step 3: Linking Characters to an Ontology

Workflow

In the next step, characters from the previous step must be linked to their respective term(s) from an anatomy ontology (Fig. 2C). Depending on the scope of a study, the same character might be linked with one or more ontology terms; several ontology terms can be used when a character refers to several BRs (e.g., color of head and legs, see also Step 5). For the needs of the *PARAMO* approach, the linking requires assigning one or more ontology terms to a respective character as shown in Table 2. The ontology terms have to be selected to best fit a character statement and the scope of a study. In the demonstration here, we are specifically interested in linking the initial characters in a way that facilitates the construction of BD and EP characters.

Software

For *PARAMO*, character-ontology linking is straightforward methodologically and can be done manually by, e.g., constructing a table with the two columns—character ID and an ontology term's ID (Table 2). The linking can be facilitated using *R* package *ontoFAST* (https://github.com/sergeitarasov/ontoFAST) that provides a graphical interface for selecting terms by navigating through the ontology. Various text processing tools like the Hymenoptera Anatomy Ontologie's 'URI Table' creator (Seltmann et al. 2012) are also available.

Character-ontology linking, sensu this article, falls into a general area of bioinformatics that focuses on annotating phenotypes with ontologies. The recent developments in this area offer comprehensive methods for constructing detailed annotations of phenotypes and characters (Dahdul et al. 2018, Balhoff et al. 2013, Dececchi et al. 2015, Cui et al. 2016). Such detailed annotations can be constructed in *Phenex* (Balhoff et al. 2010) and used in *PARAMO* as well.

Step 4: Inference—Linking Characters With Models and Tree

Workflow

The goal of this step is to obtain stochastic maps for the characters amalgamated at the AD level (Step 2) (Fig. 2D). To perform the inference, the characters have to be associated with a dated phylogeny and the respective Markov models of trait evolution. Note, that the anatomically dependent characters require structured Markov models with hidden states (see Step 2); thus, these models have to be appropriately assigned to the characters with such dependencies. In our dataset, the only character that requires such model is $C_{2,3}$. The stochastic maps can be obtained in likelihood and Bayesian frameworks. The use of the latter is preferable as it provides a convenient way for sampling the stochastic maps from the posterior distribution of character histories, which also incorporates uncertainty.

Software

Technically, this step requires creating a data object file(s) for each character that includes the character data, model and tree in an appropriate format that can be read by the software used in inference. In the present tutorial, we use *RevBayes* (version 1.0.7) (Höhna et al. 2016) to perform character inference and generate stochastic maps. The creation of the data files for *RevBayes* is automatized using *R* scripts (Supp. Material [online only]).

Step 5: Ontology-Informed Amalgamation of the Stochastic Maps for AD, BR, and EP Levels

Workflow

Our ultimate goal is to construct the amalgamated characters for the AD, BR, and EP levels of the anatomical hierarchy (Fig. 2E). The stochastic maps generated at the previous step are the individual characters of the AD level. The construction of the BR level characters implies the ontology-informed amalgamation of the AD level maps that can be done using a *RAC* query and the characters linked with the ontology in Step 3. In our example, for each of the focal BR terms ('head', 'legs', 'wings'), the *RAC* query returns a set of the associated initial characters and their stochastic maps. Next, the stochastic maps are used to produce the amalgamated BR characters. The construction of the EP level character is similar to that of BR, but requires amalgamation of all available initial characters. The amalgamated leg character from the Hymenoptera phylogeny is demonstrated in Fig. 3.

If a character (C_x) is linked to several ontology terms that refer to different BRs (e.g., BR_1 and BR_2; see also Step 3), then two-copies of C_x can be separately used for producing the amalgamated BR_1 and BR_2 characters. However, if there is a BR_3 character that includes BR_1 and BR_2 (in this case BR_3 may correspond to EP), then only one copy of C_x should be used for creating amalgamated BR_3 character.

As soon as the amalgamations are done, this step culminates the pipeline. In the next section, we discuss the use of the amalgamations for addressing various biological questions.

Software

The *RAC* query is implemented in *R* using *OntologyIndex* package (Greene et al. 2017) and a set of *PARAMO* functions. The *paramo*()

Table 2. Hymenoptera characters linked with HAO terms

ID	Character statement	HAO ID	HAO ID name
C_1	Notch on medial margin of eye	HAO:0000234	Cranium
$C_{3,2}$	Labrum + Position of labrum	HAO:0000639	Mouthparts
C_4	Forewing costal and radial vein fusion	HAO:0000351	Fore wing
$C_{5,6}$	Hind wing subcostal vein, present	HAO:0000400	Hind wing
C_7	Inner posterior mesotibial spur	HAO:0001351	Mesotibia
C_8	Foretibial apical sensillum	HAO:0000350	Fore tibia
C_9	Metatibial apical sensillum	HAO:0000631	Metatibia

Downloaded from https://academic.oup.com/isd/article/3/6/1/5584145 by guest on 02 September 2020

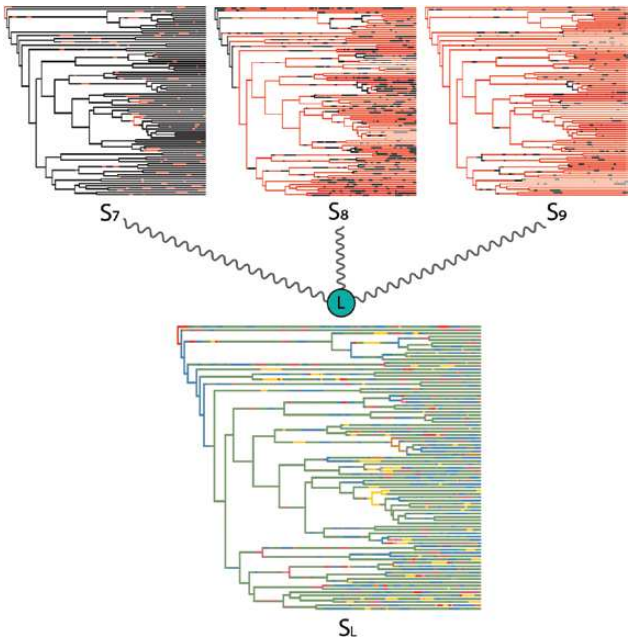

Fig. 3. Amalgamation of stochastic maps corresponding to the characters of legs from Hymenoptera phylogeny (S_7, S_8, S_9) into one 'leg character' (S_L); see also Fig. 2.

function in the provided *PARAMO* scripts performs the ontology-informed amalgamation of the stochastic maps.

Discussion

The *PARAMO* pipeline allows users to appropriately incorporate anatomical dependencies and construct characters for phenotypic entities that consist of ensembles of discrete traits. This is achieved through ontology-informed amalgamation of stochastic maps, which allows tracking of evolution at different levels of the anatomical hierarchy—individual characters, BRs, and EP (Fig. 2F). Our approach can be applied to a dataset of virtually any size, e.g., one with hundreds or thousands of characters. Ontology-informed amalgamation at the BR and EP levels represents each entity, usually described by numerous individual characters, as a single multistate character (=single-character representation).

In this article, we assume that initial characters evolve independently (except those which are anatomically dependent) and hence use independent amalgamation of their stochastic maps. It is known that evolution of morphological characters might be subjected to correlations or more complex scenarios of state changes due to hidden intrinsic or extrinsic factors, which cannot be modeled by the independent amalgamation. In this case, more complex models of trait evolution that incorporate non-anatomical correlations (Pagel 1994) or hidden states—where one observed state consists of two or more hidden ones—(Beaulieu et al. 2013, Tarasov 2019) can also be used in *PARAMO*. Similar to modeling anatomical dependencies, these models require appropriately structured rate matrices constructed for a focal set of initial characters at Steps 2 and 4. The application of hidden state models is straightforward in *PARAMO*—a separate hidden model can be assigned to each individual character in the same way the traditional models were used in the provided demonstrative example (Tarasov 2019). In contrast, modeling non-anatomical correlation has limitations—it would work well if the focal sets of characters are relatively small that allows using computationally manageable amalgamated rate matrices. However, if the focal sets are large then the amalgamations produce gigantic and computationally intractable matrices thereby precluding the use of the character correlations models. Thus, we emphasize that our method does not resolve the problem of modeling character correlations in morphological dataset to full extent, and further research is needed to develop appropriate methods.

Like traditional ancestral state reconstruction, the ancestral states can also be inferred for entire organismal anatomy or its parts through single-character representation. In this case, the states in a large composite BR or EP character correspond to a combination of states from the initial characters that form BR or EP. Note, such integrative reconstruction is similar to the traditional character-by-character reconstruction because *PARAMO* uses an independent amalgamation of stochastic maps.

The advantage of *PARAMO* approach is that the single-character representation of BR or EP opens up new avenues for comparing and assessing the dynamics of phenotypic radiations and diversifications. As a response to novel extrinsic or intrinsic factors, a phenotypic radiation may occur by rapidly diversifying an adaptive trait, inherited from a common ancestor, into a diversity of new forms in the ancestor's descendants. A well-known example of this radiation is Darwin's finches, which evolved a remarkable diversity of beak shape and functionality. Almost always, such a radiation represents an ensemble of characters that are located on the same or functionally similar BRs. Other BRs (in the same or different species) may also undergo radiations triggered by different factors and coded by their own ensembles of characters. Apparently, the varying coding of these ensembles preclude a consistent identification and assessment of phenotypic radiations. In contrast, single-character representation of BRs avoids this problem and may provide insight into the timing, location (clades) and number of phenotypic radiations occurring across a phylogeny. In this respect, each BR character is a stochastic map showing state changes in a tree where the number of changes over a branch (or time interval) reflects the evolutionary rate of the BR character in that branch—in other words, the more changes the faster the rate. The per-branch rate estimates can be used to determine rate shifts in the BR character and identify the timing of an evolutionary radiation. The same approach applied to a set of BRs can be used to map different phenotypic radiations onto the phylogeny. Obviously, any organism has many BRs that are hierarchically structured due to the nature of the anatomy. Ontology-informed amalgamation can generate characters for all potential BRs, thereby allowing to study phenotypic radiations hierarchically (Slater and Friscia 2019). Additionally, the single-character representation of EP can be used to identify rate shifts in the entire organismal anatomy and address questions on tempo and mode of its evolution. Thus, *PARAMO* can be used to disentangle evolution of phenotypes across tree and BRs. Unfortunately, so far, identification of rate shifts using amalgamated stochastic maps requires new statistical methods that would be built upon *PARAMO*. Their development is beyond the focus of the present study but we anticipate their emergence in near future.

Decades of systematists' effort have generated thousands of datasets (see *Phenoscape* [Mabee et al. 2012] and *Morphobank* [O'Leary and Kaufman 2011]) that score morphological characters for numerous clades across the Tree of Life. Frequently, these morphological data become forgotten shortly after publishing a phylogenetic tree they were used to construct. The *PARAMO* approach and anticipated further development in this area provide a new dimension for analyzing these data, which, as we believe, will aid understanding of how phenotypes evolve.

Downloaded from https://academic.oup.com/isd/article/3/6/1/5584145 by guest on 02 September 2020

Supplementary Data

Supplementary data are available at *Insect Systematics and Diversity* online.

Acknowledgments

S.T. was supported by Postdoctoral Fellowships at the National Institute for Mathematical and Biological Synthesis, an Institute sponsored by the National Science Foundation through NSF Award DBI-1300426 and through NSF-ABI award DBI-1661529 to J.C.U., with additional support from The University of Tennessee, Knoxville. We are grateful to Brian O'Meara (University of Tennessee) for helping to obtain the dated tree of Hymenoptera.

References Cited

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

Balhoff, J. P., I. Mikó, M. J. Yoder, P. L. Mullins, and A. R. Deans. 2013. A semantic model for species description applied to the ensign wasps (Hymenoptera: Evaniidae) of new Caledonia. Syst. Biol. 62: 639–659.

Beaulieu, J. M., B. C. O'meara, and M. J. Donoghue. 2013. Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in campanulid angiosperms. Syst. Biol. 62: 725–737.

Cui, H., D. Xu, S. S. Chong, M. Ramirez, T. Rodenhausen, J. A. Macklin, B. Ludäscher, R. A. Morris, E. M. Soto, and N. M. Koch. 2016. Introducing explorer of taxon concepts with a case study on spider measurement matrix building. BMC Bioinf. 17: 471.

Dahdul, W., P. Manda, H. Cui, J. P. Balhoff, A. Dececchi, N. Ibrahim, H. Lapp, T. Vision, and P. M. Mabee. 2018. Annotation of phenotypes using ontologies: a gold standard for the training and evaluation of natural language processing systems. Database. 2018.

Dececchi, T. A., J. P. Balhoff, H. Lapp, and P. M. Mabee. 2015. Toward synthesizing our knowledge of morphology: using ontologies and machine reasoning to extract presence/absence evolutionary phenotypes across studies. Syst. Biol. 64: 936–952.

Greene, D., S. Richardson, and E. Turro. 2017. ontologyx: a suite of r packages for working with ontological data. Bioinformatics. 33: 1104–1106.

Haendel, M. A., F. Neuhaus, D. Osumi-Sutherland, P. M. Mabee, J. L Mejino, C. J. Mungall, and B. Smith. 2008. Caro–the common anatomy reference ontology, pp. 327–349. *In Anatomy ontologies for bioinformatics*. Springer.

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

Huelsenbeck, J. P., R. Nielsen, and J. P. Bollback. 2003. Stochastic mapping of morphological characters. Syst. Biol. 52: 131–158.

Klopfstein, S., L. Vilhelmsen, J. M. Heraty, M. Sharkey, and F. Ronquist. 2013. The hymenopteran tree of life: evidence from protein-coding genes and objectively aligned ribosomal data. PLoS One 8: e69344.

Mabee, P., J. P. Balhoff, W. M. Dahdul, H. Lapp, P. E. Midford, T. J. Vision, and M. Westerfield. 2012. 500,000 fish phenotypes: the new informatics landscape for evolutionary and developmental biology of the vertebrate skeleton. J. Appl. Ichthyol. 28: 300–305.

Maddison, W. and D. Maddison. 2018. Mesquite: a modular system for evolutionary analysis. version 3.40. http://mesquiteproject.org.

O'Leary, M. A., J. I. Bloch, J. J. Flynn, T. J. Gaudin, A. Giallombardo, N. P. Giannini, S. L. Goldberg, B. P. Kraatz, Z-X. Luo, J. Meng, et al. 2013. The placental mammal ancestor and the post–k-pg radiation of placentals. Science. 339: 662–667.

O'Leary, M. A., and S. Kaufman. 2011. Morphobank: phylophenomics in the 'cloud'. Cladistics. 27: 529–537.

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

Pagel, M. 1999. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. Syst. Biol. 48: 612–622.

Peters, R. S., K. Meusemann, M. Petersen, C. Mayer, J. Wilbrandt, T. Ziesmann, A. Donath, K. M. Kjer, U. Aspöck, H. Aspöck. et al. 2014. The evolutionary history of holometabolous insects inferred from transcriptome-based phylogeny and comprehensive morphological data. BMC Evol. Biol. 14: 52.

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

R Core Team. 2018. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Sauquet, H., M. von Balthazar, S. Magallón, J. A. Doyle, P. K. Endress, E. J. Bailes, E. B. de Morais, K. Bull-Hereñu, L. Carrive, M. Chartier, et al. 2017. The ancestral flower of angiosperms and its early diversification. Nat. Commun. 8: 16047.

Seltmann, K. C., M. J. Yoder, I. Mikó, M. Forshage, M. A. Bertone, D. Agosti, A. D. Austin, J. P. Balhoff, M. L. Borowiec, S. G. Brady, et al. 2012. A hymenopterists' guide to the Hymenoptera Anatomy Ontology: utility, clarification, and future directions. J. Hymenopt. Res. 27: 67.

Sharkey, M. J., J. M. Carpenter, L. Vilhelmsen, J. Heraty, J. Liljeblad, A. P. Dowling, S. Schulmeister, D. Murray, A. R. Deans, F. Ronquist, et al. 2012. Phylogenetic relationships among superfamilies of hymenoptera. Cladistics. 28: 80–112.

Slater, G. J., and A. R. Friscia. 2019. Hierarchy in adaptive radiation: a case study using the carnivora (mammalia). Evolution. 73: 524–539.

Tarasov, S. 2018. The invariant nature of a morphological character and character state: insights from gene regulatory networks. bioRxiv. page 420471.

Tarasov, S. 2019. Integration of anatomy ontologies and evo-devo using structured Markov models suggests a new framework for modeling discrete phenotypic traits. Syst. Biol.

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

Downloaded from https://academic.oup.com/isd/article/3/6/1/5584145 by guest on 02 September 2020