COLLÈGE DE FRANCE
1530

Ecole Doctorale 227 : Sciences de la Nature et de l'Homme. Numéro attribué par la bibliothèque :

THÈSE
pour l'obtention du grade de
DOCTEUR DU MUSÉUM NATIONAL D'HISTOIRE NATURELLE
Spécialité : Écologie et Évolution

Prospecting for unconventional hypotheses in biodiversity macroevolution modeling.

Présentée par
Marc Manceau

Sous la direction de

Mr. Amaury Lambert Professeur au laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université.

Mme. Hélène Morlon Directeur de recherche CNRS à l'Institut de Biologie de l'École Normale Supérieure.

La thèse sera soutenue publiquement le 21 Juin 2018
devant le jury composé de

Amaury Lambert	Professeur Sorbonne université	Directeur de thèse
Hélène Morlon	Directeur de recherche CNRS	Directeur de thèse
Nicolas Galtier	Directeur de recherche CNRS	Rapporteur
L. Lacey Knowles	Collegiate Professor University Michigan	Rapporteur
Sandrine Pavoine	Maître de conférence MNHN	
Nicolas Lartillot	Directeur de recherche CNRS	
Stéphane Robin	Directeur de recherche INRA	

Résumé

La macroévolution est la branche de la biologie évolutive qui étudie l'évolution des organismes sur des échelles de temps suffisamment longues pour que des espèces nouvelles apparaissent et disparaissent. Dans cette thèse, nous nous attachons à modéliser, à l'aide d'outils probabilistes, différents phénomènes évolutifs observables sur ces longues échelles de temps.

Le premier chapitre de la thèse est une introduction générale du domaine de la modélisation en macroévolution qui insiste plus particulièrement sur les méthodes et les modèles probabilistes les plus utilisés. Les hypothèses communément admises, parmi lesquelles certaines que les chapitres suivant s'attacheront à relâcher, sont d'abord exposées. Les méthodes statistiques permettant l'inférence des paramètres des modèles, qui seront employées en particulier dans les quatre chapitres suivants, sont également présentées. Le chapitre se clôt enfin sur une vision générale du type de questions que permettent de résoudre ces modèles.

Les quatre chapitres suivants présentent notre travail de recherche. Ils permettent d'évaluer la pertinence de nouvelles hypothèses, peu représentées dans la littérature, prenant en particulier en compte l'effet de processus écologiques en macroévolution.

Le chapitre 2 présente une étude critique de la modélisation des espèces dans un type de modèle individu-centré très employé dans le domaine, considérant des individus à reproduction clonale, se différenciant via la superposition d'un processus de mutations ponctuelles. Motivés par la difficulté à reconstruire la phylogénie des espèces sous les hypothèses présentes dans la littérature, nous présentons deux définitions d'espèces, permettant de faciliter le passage de la généalogie des individus à la phylogénie des espèces. Ces définitions d'espèces découlent naturellement de la considération de propriétés, désirables d'un point de vue biologique, que nous formalisons dans un cadre mathématique.

Le chapitre 3 propose une application de l'une des deux définitions d'espèces présentées dans le chapitre 2. Il repose sur un modèle individu-centré de diversification, permettant d'analyser le rythme auquel les événements de spéciation et extinction surviennent. En sus de la définition d'espèce, nous considérons une seconde hypothèse non conventionnelle, selon laquelle la dynamique de la métapopulation d'individus est donnée par un processus de naissance-mort. Nous décrivons un moyen efficace de simuler des phylogénies reconstruites sous ce modèle, ainsi qu'un moyen de calculer la densité de probabilité d'une phylogénie reconstruite sous ce modèle. Nous montrons enfin que ces hypothèses permettent de reproduire des patrons de forme d'arbres phylogénétiques empiriques.

Dans notre chapitre 4, nous nous attachons à un second type de questions posées en macroévolution, concernant l'évolution de caractères continus parmi un ensemble de taxa liés par des relations phylogénétiques. Alors que les approches comparatives communément employées font l'hypothèse que les traits évoluent indépendamment sur différentes branches d'une phylogénie, nous proposons au contraire un moyen de modéliser des interactions entre des traits portés par des lignées différentes. Ce type de modélisation permettrait d'étudier l'impact conjoint des relations phylogénétiques, et d'interactions écologiques, sur la distribution des phénotypes.

Le chapitre 5 présente un second modèle d'évolution des traits au cours du temps, pour un trait moléculaire cette fois. Là où les modèles de datation moléculaire reposent sur une hypothèse gradualiste, anagénétique, de modification des séquences, nous envisageons un modèle qui, à un fond de mutations anagénétiques à taux constant, superpose un processus d'évolution ponctuelle cladogénétique. Ce travail étant encore inachevé, nous présentons uniquement le modèle, une ébauche de procédure d'inférence, ainsi que les directions futures envisagées.

Enfin, le dernier chapitre 6 est l'occasion de synthétiser et mettre en lien les chapitres précédents. Nous terminons par quelques perspectives sur les défis que différents types de travaux de modélisation en macroévolution devront résoudre dans les années futures.

Abstract

Macroevolution consists in the study of biological evolution over timescales large enough that distinct species can appear and disappear. In this PhD thesis, we are interested in modeling, in a probabilistic framework, various evolutionary processes acting over these large timescales and responsible for the biodiversity patterns observed today.

The first chapter of this thesis is a background introduction to the field of macroevolution modeling putting particular emphasis on the most used models and methods. Conventional hypotheses are first exposed, including the very ones that we will try to relax in the following chapters. We present the statistical methods enabling to infer model parameters, which will be in particular used in the next four chapters. This chapter ends with a general overview of the type of questions that all these models aim to address.

The next four chapters expose our research work. They assess the relevance of new sets of hypotheses, either never or under-considered in the literature. They attempt in particular to better incorporate the effects of ecological processes in macroevolution.

Chapter 2 presents a critical survey of species definitions in a very popular individual-based modeling framework considering clonally reproducing organisms, which phenotype is given by the superimposition of a process of point mutations. We present two new species definitions, which naturally come out from the consideration of a set of biologically desirable properties of species. These new, unconventional, definitions, allow us to more clearly link the genealogy of individuals with the phylogeny of species.

Chapter 3 then presents an application of one of these species definitions. An individual-based model of diversification is built, allowing us to study the tempo at which speciations and extinctions occur. Additionally to the species definition, we consider a second unconventional hypothesis: the metapopulation dynamics is given by a birth-death process. We describe an efficient algorithm enabling us to simulate reconstructed phylogenies, as well as a second algorithm to compute the probability density of a reconstructed phylogeny, under this model. We finally show that this set of hypotheses produces phylogenies quite in agreement with empirical ones.

In our chapter 4, we are interested in a second broad type of questions addressed in macroevolution modeling, which concerns the evolution of continuous phenotypes among a set of phylogenetically related organisms. While commonly used comparative approaches assume that traits evolve independently from one another on distinct branches of a phylogeny, we instead propose a way to model interactions between traits present on distinct lineages. This modeling holds the promise to study the joint impact of phylogenetic relatedness and ecological interactions on the present-day distribution of phenotypes.

Chapter 5 presents a second model of trait evolution through time, but for a molecular trait instead of a continuous one. While molecular clocks used in tree dating studies generally assume that sequences evolve through the gradual accumulation of anagenetic mutations, we assume that both gradual anagenetic mutations and punctual cladogenetic mutations happen along the phylogeny. This chapter presents work-in-progress. We describe the model, then sketch an inference procedure, and finally provide future directions.

Last, we summarize our work and attempt to link the various chapters of this thesis in the concluding chapter 6. We end with a presentation of the challenges that different types of modeling approaches will have to tackle in future years.

Remerciements

Quatre années d'aventures humaines et scientifiques s'achèvent symboliquement avec cette soutenance de thèse. Je saisis ma chance de remercier tous ceux qui m'ont accompagné et m'ont aidé à mener ce projet à terme. Et puisque la fin de cette thèse signifie également la fin de ma scolarité, j'en profite pour adresser un remerciement "diffus" à tous les êtres humains qui ont permis mon éducation scientifique. En premier lieu bien sûr, le système éducatif français et ses acteurs : j'ai eu la chance de rencontrer d'excellents profs tout au long de ma scolarité, qui ont su éveiller puis nourrir mon envie d'apprendre. Enfant gâté du milieu académique, je mesure le privilège qui m'a ensuite été accordé d'étudier si longtemps, en toute liberté, au gré de mes envies et sans aucun souci matériel. J'ai pris énormément de plaisir à côtoyer durant cette thèse toute la constellation de femmes et d'hommes qui œuvrent pour cet idéal de recherche et de diffusion de la connaissance.

Parmi ces hommes et ces femmes, je souhaite en premier lieu remercier chaleureusement les membres de mon jury de thèse. Merci à Nicolas Galtier et Lacey Knowles, qui ont relu avec attention et évalué le présent manuscrit. Merci à Sandrine Pavoine et Stéphane Robin, qui se sont rendus disponibles pour prendre part à ce jury. Le monde académique est définitivement un monde à part, où il est possible de réunir sans aucune contre partie des chercheurs reconnus du domaine pour évaluer le travail d'un doctorant qui leur était encore anonyme hier. Merci de votre générosité. Merci à Nicolas Lartillot, que j'ai eu le plaisir de rencontrer dès mon stage de master, et qui s'est toujours rendu accessible pour répondre à mes questions. Merci de m'avoir fait l'honneur de participer à la fois au suivi de cette thèse et au jury de soutenance.

Un très grand merci à Amaury et Hélène, qui, sous l'étiquette de directeurs de thèse, ont joué pour moi les rôles très variés de profs, conseillers d'orientation et modèles scientifiques. Les deux cours de 'maths pour biologistes' dispensés par Amaury, suivis du stage de master que vous avez co-encadré, ont été à l'origine de mon choix de suivre un master de maths appliquées, en rêvant déjà à revenir faire une thèse avec vous. Merci pour votre patience, vos conseils, vos encouragements, votre optimisme à toute épreuve. Merci pour votre oreille toujours attentive à mes questionnements scientifiques autant que personnels. Ces années passées avec vous ont été extrêmement enrichissantes, et je ne peux imaginer de meilleur choix ni *a priori* ni *a posteriori*.

Un certain niveau d'exigence dans les remerciements de thèse semble se perpétuer dans l'équipe *SMILE* ces dernières années. J'ai voulu initialement prétexter éviter toute surenchère pour faciliter l'écriture de mes co-thésards. Je fais finalement confiance à votre bienveillance générale pour pardonner mon manque d'originalité et ma pudeur: je n'écrirai ni roman ni poème, je ne présenterai aucune oeuvre d'art, mais je n'en pense pas moins. Merci donc aux co-bureaux, co-thésards, ou plus largement co-équipiers, des deux côtés du Panthéon, pour la bonne humeur, la gentillesse, et l'attention que vous portez à chacun. Merci pour les cafés réguliers où l'on refait la science, pour les petits goûters de 16h, pour les week-ends organisés d'une main de maître par monts et par vaux. Merci pour les traditionnels chouquettes ou gâteaux d'anniversaire (avec chanson, s'il vous plaît !), le réconfort d'un jus de fruit et le plaisir d'un apéro sur la terrasse, ou simplement pour les petites remarques attentionnées du quotidien. Merci à tous d'avoir créé tous les jours cet environnement de travail exceptionnel.

Aussi agréable que ce soient mes deux équipes d'adoption, mes années de thèse auraient été également moins belles sans les amis extérieurs que j'ai eu la chance de rencontrer. Deux associations sportives ont joué un rôle essentiel en offrant régulièrement leur concours pour me permettre de m'évader de la grisaille parisienne. Un grand merci à tous les copains de l'ASP6 et du GUMS, qui m'ont formé respectivement à la plongée et à 'la montagne', autour de valeurs de partage, d'entraide, et de respect de

l'environnement. Merci à tous les amis avec qui j'ai pu partager les petites frustrations, merci à ceux qui font relativiser et sourire dans les moments de doute. Aux cinq plus ou moins brefs colocs du K-B, qui m'ont sorti parfois, nourri souvent, et aéré l'esprit au quotidien, un grand merci ! Je ne sais pas où j'atterirai l'année prochaine, mais soyez déjà certains d'y être invités.

A mes parents, ma soeur, mes grands-parents et ma famille en général, merci pour votre soutien, votre écoute, votre regard rassurant. Une bonne partie de ce manuscrit a été rédigée dans le cadre familier et apaisant de la maison de mes grands-parents, que je peux remercier pour leur accueil, au moins une fois, formellement.

Enfin, à Claire, évidemment, merci.

Table of Contents

Chapter 1

Introduction

This first chapter presents an overview of what macroevolution is about. We present the broad questions asked in this field, both for the study of diversification and for the study of phenotypic evolution. We try to do so with a special focus on the formalism and methodologies that constitute the foundation of comparative methods.

It is intended to serve as an introductory course on probabilistic modeling in macroevolution. More specific and modern references will be discussed in the separate introduction of each chapter.

Contents of the chapter

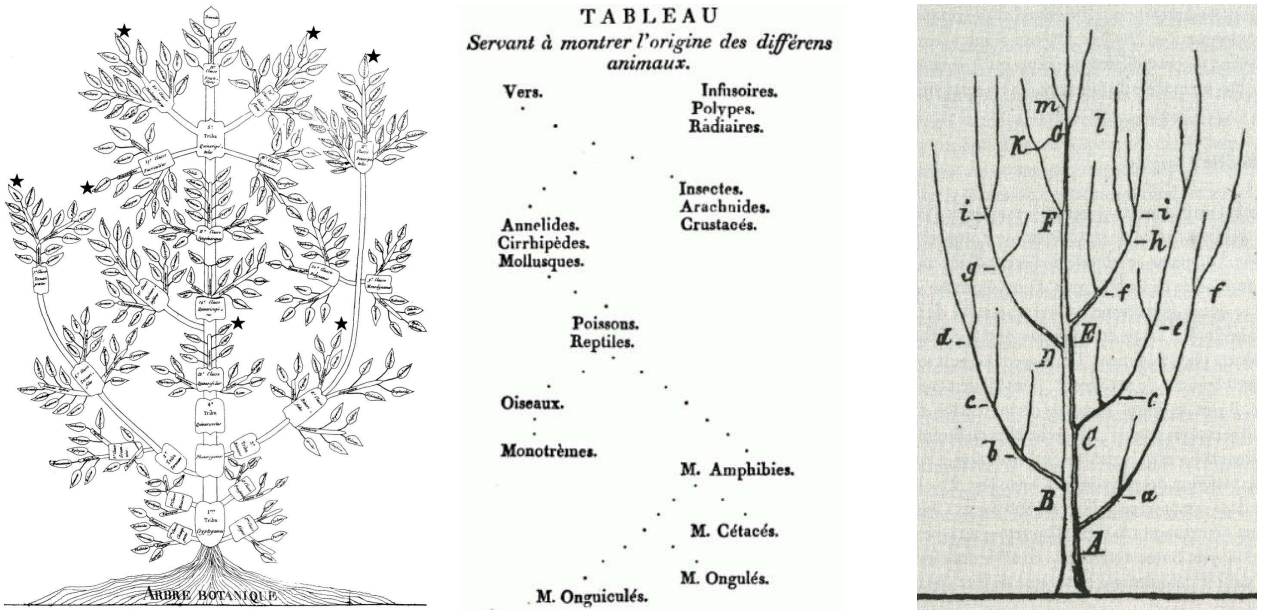
In spite of the many creative efforts deployed by human societies to degrade it, the Earth is still covered with a staggering biodiversity inspiring lots of very challenging, yet concrete, questions.
Why are there so many beetle species ? Why do some beetles possess wings while others do not ? What determines their size, their colour ? How do we classify them into distinct species ? What did the first beetle of all beetles look like ?

These questions are so challenging, indeed, that you may end up answering that *you are still too young to understand* or, paraphrasing J.B.S. Haldane, that *the Creator has an inordinate fondness for beetles* to the 8 years-old child who dares questioning you during your traditional family Sunday walk in a forest.

These are, in fact, long-standing questions in evolutionary biology. Addressing them scientifically requires the use of a precise framework to represent genealogical relationships between organisms, together with sound modeling tools adequate to make rigorous statistical inferences from the observation of the natural world.

Tree thinking

While trees were already used extensively to represent pedigrees of human families, the first records of trees of life in the biological literature all arise in the 19$^{\text{th}}$ century. Several representations pre-date Darwin (1859)'s most read book, *On the Origins of species by means of natural selection*, representing species relationships on a tree without proposing the mechanism responsible for bifurcations (Archibald, 2009).

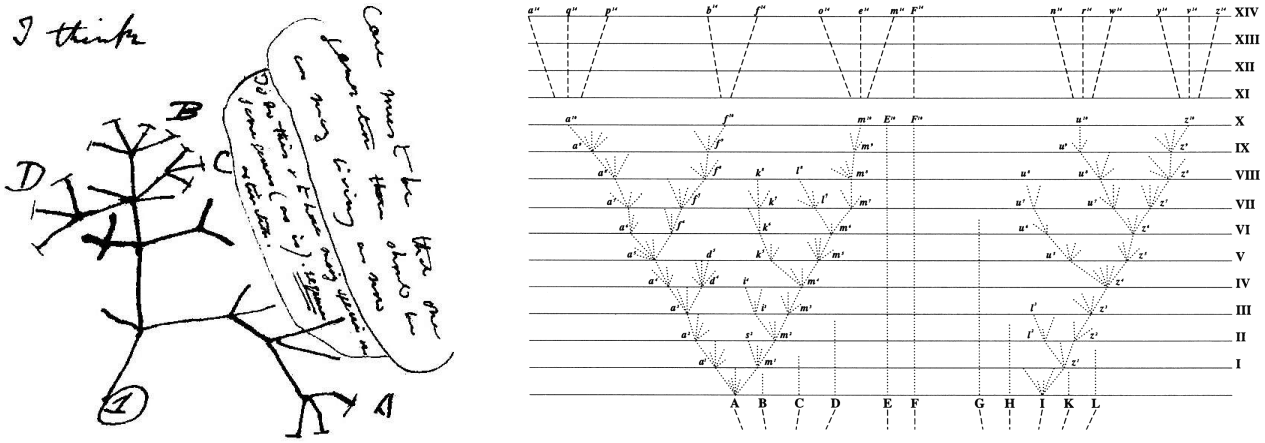


(a) Augustin Augier, 1801. (b) Jean-Baptiste Lamarck, 1809. (c) Heinrich Georg Bronn, 1858.

Figure 1.1 – Some pre-Darwin representations of a tree of life.

Because he is the first to associate the diagram to the mechanism of evolution, Darwin (1859)'s two tree representations have achieved posterity in the field (see Fig. 1.2). The first one is a small diagram quickly drawn in the margin of his notebook, while the second one is the only figure on the whole book *On the origin of species*.

Post-Darwin tree of life representations are much more numerous. Those drawn by Ernst Haeckel in the late 19$^{\text{th}}$ century, notably, are still a vivid memory in the field (see Fig. 1.3). Many more tree of life representations will be exposed in the rest of the thesis, with branches heading upward (as is popular in systematics), downward (as is popular in mathematics), to the left or right (as is popular in



(a) On a notebook.

(b) Only figure in Darwin (1859).

Figure 1.2 – Trees of life sketched by Darwin.

population genetics and macroevolution modeling), or radially extending from the root (as is popular in science popularization pictures).

Whatever their appearance, trees represent today the main ground hypothesis about, and representation of, genealogical relationships among species. They will be omnipresent in particular throughout this thesis, as the central component upon which to build models of evolution, to address questions on the past history of organisms.

Mathematical modeling in micro- an macro- evolution

Biology in general, and evolutionary biology in particular, are very recent scientific fields. Yet proportionately, mathematical modeling has a quite long history in evolutionary biology, starting right in the beginning of the 20$^{\text{th}}$ century with the work of population geneticists Fisher, Haldane and Wright. Building on the recently described concepts of *genetic laws* and *natural selection*, they propose the first mathematical models explaining diversity change through time in a population of individuals. Their work paved the way to modern population genetics, introducing e.g. the concepts of *adaptive landscape* or *genetic drift*. This pioneering modeling work was mainly focusing on processes happening at the scale of populations of interbreeding individuals, among one species.

Around the same period, Dobzhansky (1937) initiated a synthesis of this work linked to experimental evolution studies, where he popularized the term *macroevolution*, to refer to the study of biological evolution over time periods long enough to see distinct species appear and go extinct among wide groups of organisms. Macroevolution is thus opposed to *microevolution*, which refers to the study of the evolution of organisms over shorter timescales, of the order of few generations, generally among one species. The very simple scale distinction introduced by these two terms nonetheless hides a quite fundamental difference for researchers. While microevolution relies on processes that act at the individual-level and are, as such, observed directly by biologists, macroevolution relies on processes acting at the species level, thus preventing their direct observation. The birth, death, and reproduction of individuals can be observed and studied through time, even manipulated by scientists when needed. In contrast, the origination (but unfortunately, not extinction) of species requires more than human life to be manipulated. The challenge of macroevolution is thus to imagine what the most important processes could be over long timescales, and from them to reconstruct what likely happened in the past.

Evolutionary biologists interested in macroevolution nonetheless have access to a wealth of data. All these data are simply empirical observations of *phenotypes* of organisms. We will use the term phenotype throughout this thesis in its most extensive acceptance, referring to all observable, measurable,

characteristics of individuals. In particular, the phenotype is not opposed to the genotype, and may include here morphological, cellular, as well as genetic characteristics of individuals. Paleontological data give access to the phenotypes of ancient organisms well conserved, by chance, in the fossil record. But the main source of reliable data consists in the observation of phenotypes of present-day individuals.

Making sense out of these data requires (i) to be able to propose distinct scenarios that could have led to the same observations, and (ii) be able to compare these scenarios. The field has thus been particularly attractive to mathematical modelers (i) helping to formalize the questions and (ii) bringing them into a probabilistic framework well suited for statistical inference.

While modeling work was mainly interested in modeling microevolution processes in the beginning of the 20$^{\text{th}}$ century, some other pioneers began to show interest in macro patterns too. Yule (1925), for example, introduced the famous *Yule model* (see section 1.1.4) while trying to explain the surprising heterogeneity in the number of species in distinct genera. It laid the groundwork to Kendall (1948), who used and extended the study of birth-death processes in a context of population dynamics modeling.

A drastic inflation of modeling work in evolutionary biology happened after the discovery of the DNA as the support of genetic material in all living forms, followed by the description of its structure in the mid-20$^{\text{th}}$ century. Zuckerkandl and Pauling (1962) proposed soon after the concept of molecular clock, lighting the way to tree reconstruction and tree dating using the abundant genetic data instead of only morphological characters. Models of molecular evolution (Jukes and Cantor, 1969), and continuous trait evolution (Felsenstein, 1973), on a tree structure followed, introducing the idea of tree reconstruction in a maximum likelihood framework.

The continuous amelioration of sequencing technologies, together with the inflation of computing power, led to a boom of modeling work in evolutionary biology. This in turn allowed researchers to address very diverse questions on the past history of organisms:

tree reconstruction What are the most likely ancestral relationships between a given set of organisms ?

trait evolution What did ancestors of contemporary species look like ? What are the most labile or conserved traits in the history of some clade ?

diversification What is the tempo of speciation/extinction events in distinct clades ? How could we explain their heterogeneity ?

and many more questions which will be skimmed through in section 1.4.

Background presentation

Because the next chapters address questions on three quite distinct facets of life evolution over long timescales - diversification, phenotypic evolution, and molecular evolution - we intend to provide first an accessible and self-contained introduction to mathematical modeling in macroevolution.

Although we mention some key historical papers in the process, as well as some examples of applications in the end, it is not intended to be a comprehensive review of the field. Prospecting for unconventional hypotheses in the field requires mainly to know the ground hypotheses of classic models, on which we direct the spotlight in this introduction. This will allow us to contrast what we did in the next chapters with the most common methodologies.

We start this opening chapter with a general overview of the tree formalism and of probabilistic models of trees used in macroevolution. We then expose the most commonly used probabilistic models of trait evolution, describing the change of characteristics through time, and along a tree. Last, we present the statistical framework, as well as the questions that these models aim to address. We provide some examples of applications on empirical data.

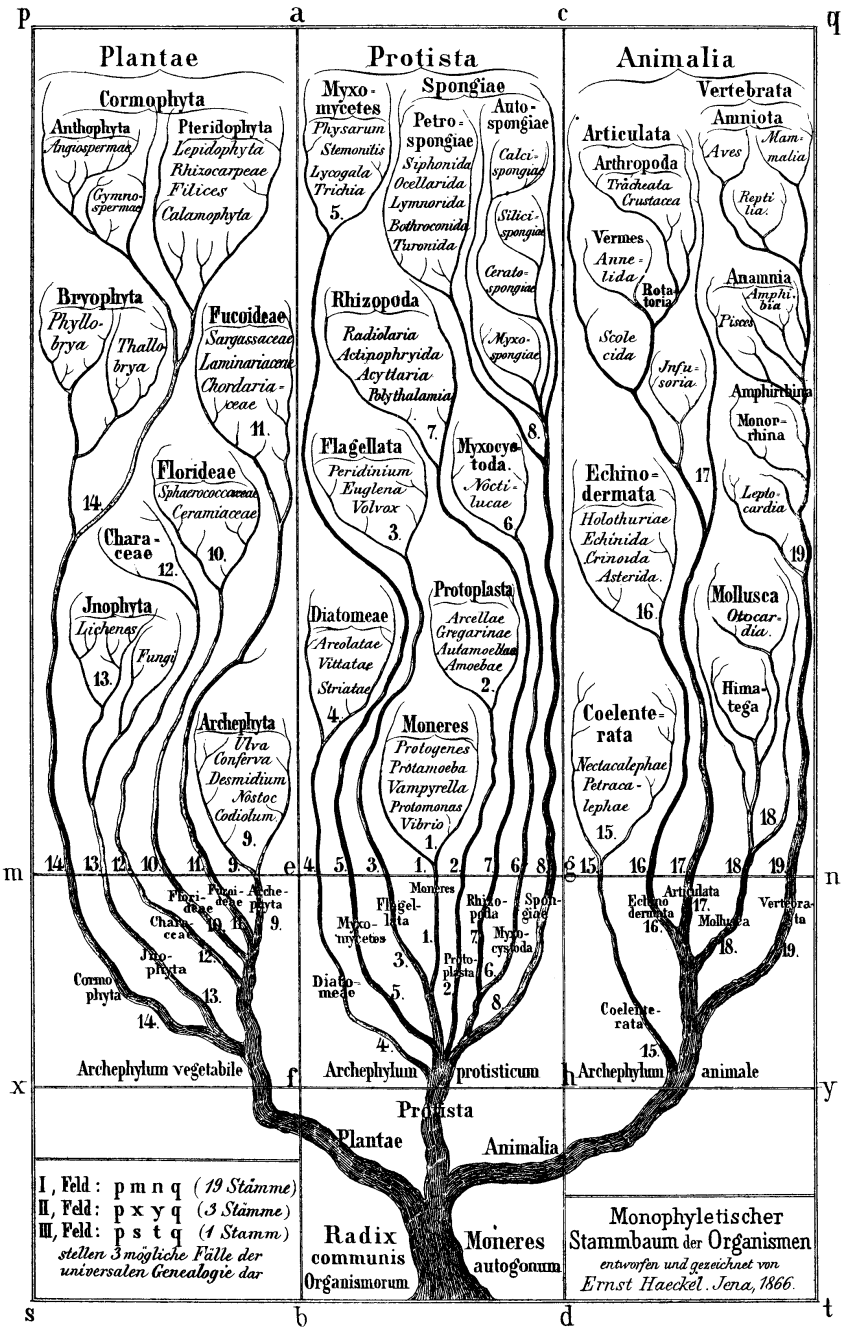


Figure 1.3 – Tree of life representation by Ernst Haeckel, 1866.

1.1 Formalizing parental relationships with trees

1.1.1 Tree formalism

In this section we expose the formalism of graphs and trees, that we will need in our applications. Our introduction strongly relies on the book of Semple and Steel (2003). We mainly stick to their definitions and terminology, and alert the reader in the few cases where we don't.

The graph nomenclature

We start with some graph definitions, carefully chosen to introduce trees.

Definition 1 *We say that $G = (V, E)$ is a* graph *if V is a non-empty set of vertices and E is a set of edges with $E \subset \{\{v_1, v_2\} : v_1, v_2 \in V^2\}$.*

We further need some terminology to describe the subclass of graphs known as trees.
If $e = \{v_1, v_2\} \in E$, we say that e is *incident* with v_1 and v_2, and that v_1 and v_2 are *adjacent*. We call *degree* of a vertex v the number of edges incident with v. The following special structures in graphs need also be mentioned:

a loop is an edge $\{v, v\}$ incident with only one vertex;

parallel edges are two edges incident with the same two vertices;

a path is a sequence of distinct vertices $(v_1, v_2, ..., v_k)$ such that $\forall i \in \{1, 2, ..., k-1\},\ \{v_i, v_{i+1}\} \in E$;

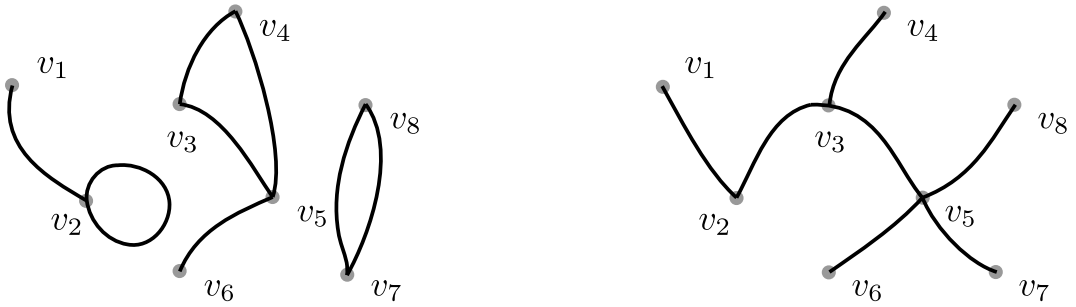
a cycle is a path $(v_1, v_2, ..., v_k)$ with v_1 and v_k adjacent;

a connected graph is a graph in which any two vertices can be linked through a path.

This now allows us to define what is called a tree. This object is ubiquitous in evolutionary biology and will be central throughout this thesis.

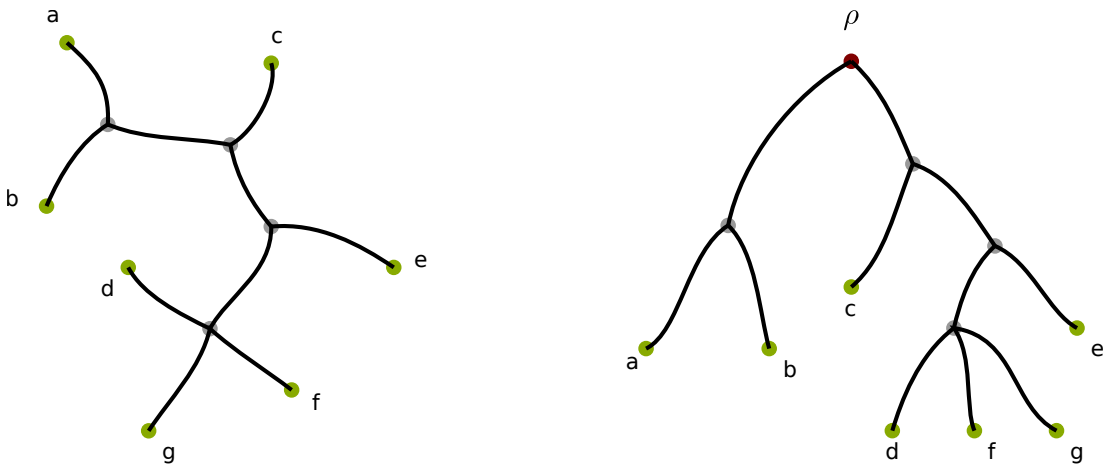
Definition 2 *A* tree *is a connected graph without loop, parallel edges and cycle.*

Extending the botanical analogy, vertices of degree one are called *tips* or *leaves*, while others are called *interior nodes*. Edges are also referred to as *branches*, either internal (incident with two interior nodes) or external (incident with one leaf).



(a) A graph. (b) A tree.

Figure 1.4 – Two graphs with set of vertices $V = \{v_1, v_2, ..., v_8\}$. In (a) $\{v_2, v_2\}$ is a loop. Two parallel edges are incident with v_7 and v_8. The path (v_3, v_4, v_5) is a cycle. Moreover, the graph is not connected, because their is no path, e.g. between v_2 and v_3. In (b), the graph is connected and has no loop, parallel edges and cycle: it is a tree.



(a) Unrooted X-tree.

(b) Rooted X-tree.

Figure 1.5 – Two X-trees with $X = \{a, b, c, d, e, f, g\}$. Leaves are colored in green, interior nodes are colored in gray, and the root ρ is in red. Note that we can transform (a) into (b) by choosing an edge, cutting it and placing a new root vertex in the middle.

Rooted trees

The trees that we just defined are also referred to as *unrooted trees* in evolutionary biology. They are used as representations of proximity between distinct biological entities (e.g. genes), without making any assumption on the ancestor-descendant relationship. However, evolutionary biologists being essentially interested in representing the history of biological entities, they need to describe a directed ancestral relationship between adjacent vertices. This is exactly what the following *rooted trees* are meant to.

Definition 3 *A* rooted tree T_ρ *is a tree where a special vertex* ρ *is marked as the* root *of the tree.*

The simple addition of a root allows us to define the following order on vertices:
If v_1 lies in the path from ρ to v_2, then we say that v_2 is a *descendant* of v_1, and v_1 is an *ancestor* of v_2.
Furthermore, if v_1 and v_2 are adjacent, then v_1 is called the *mother* of v_2 and v_2 is a *child* of v_1.
Last, the *is an ancestor* relationship is a partial order on V that we denote $\preceq_T$.

Labelled trees

Rooted trees are commonly used representations of ancestral relationships between biological entities, such as genes, individuals, species. Whatever the application, biologists consider trees which leaves are associated to labels. This construction can be formalized as follows:

Definition 4 *An X-tree is an ordered pair $(T; \phi)$, where T is a tree, and ϕ is a bijection from X into the set of leaves of T.*

Note that we took the liberty to call this an X-tree, whereas X-trees according to Semple and Steel (2003) are more general. Our definition would correspond to their *phylogenetic X-trees*, but we prefer to keep the term *phylogeny* closer to its biology meaning.

Depending on their set of labels X, X-trees have indeed distinct names for biologists. When X is a set of gene labels, the resulting X-tree is called a *gene tree*. It is called a *genealogy* when X is a set of individuals, and *phylogeny* when X is a set of species.

Dated trees

Last, we often need to consider X-trees inducing a distance concept on X to make our description of history more precise. To do so, we associate edges with a positive edge-length, as follows.

Definition 5 *Let $((V,E);\phi)$ be an X-tree, and w be a function with $w : E \to \mathbb{R}^+$.*
We say that $((V,E);\phi;w)$ is an edge-weighted X-tree.

An edge-weighted tree T naturally induces a standard distance d_T on V. For $(v_1, v_n) \in V^2$, we define:

$$\text{if } v_1 = v_n, \;\; d_T(v_1, v_n) = 0$$

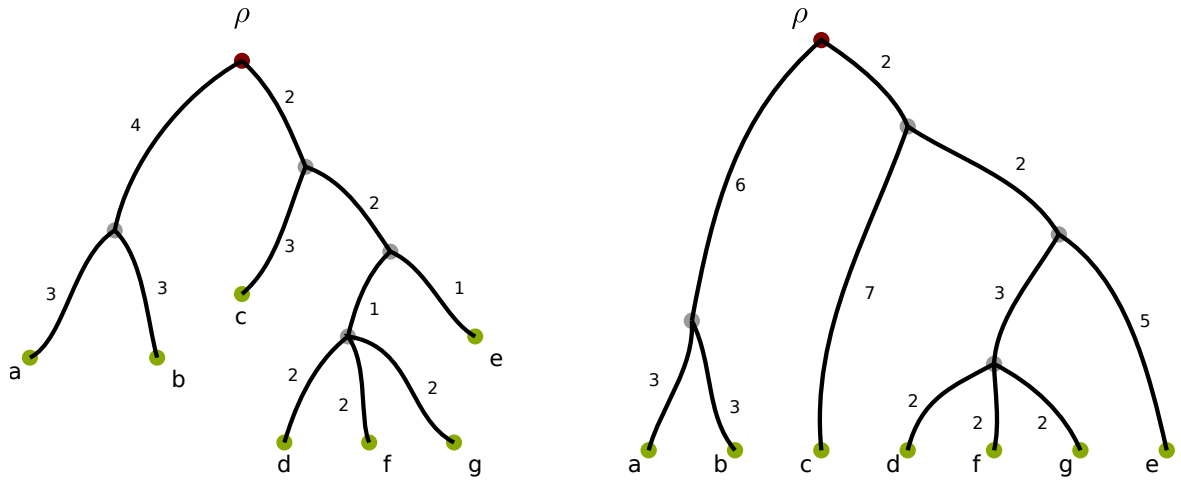
$$\text{if } v_1 \neq v_n, \;\; d_T(v_1, v_n) = \sum_{k=1}^{n-1} w(\{v_k, v_{k+1}\}) \quad \text{where } (v_1, v_2, ..., v_n) \text{ is the unique path from } v_1 \text{ to } v_n$$

Moreover, $d_T \circ \phi$ is a distance on X.
When all edges have a weight equal to 1, this distance is called *topological distance*.

Edge weightings may have very distinct interpretations depending on the context. On an unrooted gene tree, it could for instance represent an average number of substitutions. On a rooted phylogeny T_ρ, it would more presumably represent time intervals. When it does so, the tree is also called *dated tree*, and a *living time* can be associated to each vertex in a backward or forward in time manner.

Forward in time, we consider that the root lives at time 0. Each child v_2 of a vertex v_1 living at time t_1 is then considered to live at time $t_1 + w(\{v_1, v_2\})$. This rule recursively allows us to date all vertices in the tree.

Backward in time, we consider that the most distant leaf from the root lives at time 0. The mother v_1 of a child v_2 living at time t_2 is then considered to live at time $t_2 + w(\{v_1, v_2\})$. This convention, together with the previous one, allows us to recursively date all vertices in the tree.



(a) Edge-weighted X-tree. (b) Ultrametric X-tree.

Figure 1.6 – Two rooted edge-weighted X-trees with $X = \{a, b, c, d, e, f, g\}$. Leaves are colored in green, interior nodes are colored in gray, and the root ρ is in red. Numbers correspond to edge weights.

When all elements in X are entities sampled at present time, then all leaves should live at the same time, i.e. at equal distance from the root:

$$\forall x, y \in X^2,\ d_T(\rho, \phi(x)) = d_T(\rho, \phi(y))$$

An edge-weighted rooted tree T_ρ is called *ultrametric tree* when this property is verified. Non-ultrametric dated trees display leaves that do not live at present. They are called *fossil leaves* or simply *fossils*.

Unless otherwise stated, all trees in this thesis can be considered to be edge-weighted, rooted, X-trees, where the weighting represents time.

Common ancestry on trees

The tree of life is now the most common hypothesis about, and representation of, species history in macroevolution. It is viewed as a rooted, timed, X-tree $((V, E); \phi)$, where X is a set of species without descent. Each tip is thus labelled by a species name, being either extant (if the branch reaches present) or extinct/fossil (if it does not). Internal vertices correspond to hypothetical (not observed) common ancestors to the tips it subtends. Edges between vertices represent ancestral relationships, directed from the root to the leaf, and may be interpreted as a succession of ancestors through time, also called a *lineage*.

Recall that a partial order on V noted $\preceq_T$ convey the 'is an ancestor' relationship. We use it now to define two operations.

Definition 6 *We call descent of a vertex $v \in V$, and we write $desc(v)$ the set $\{w \in V, v \preceq_T w\}$.*
We call most recent common ancestor *of a subset $W \subset V$, and we write $mrca(W)$ the greatest lower bound of the set W for the $\preceq_T$ relationship.*

Depending on ancestral relationships among them, subsets of vertices of T might be called *monophyletic groups* or *paraphyletic groups*.

Definition 7 *Let $T = (V, E)$ be a rooted tree. Let $W \subset V$.*
W is said to be monophyletic if it includes one common ancestor and all its descent, i.e. if:

$$desc(mrca(W)) = W$$

Otherwise, W is said to be paraphyletic.

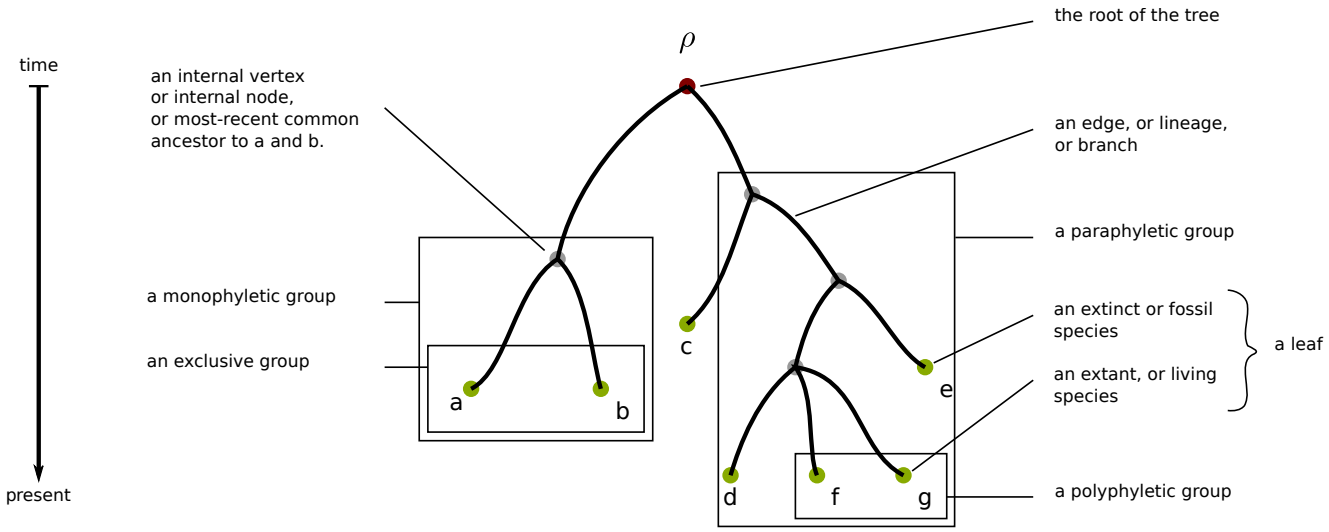


Figure 1.7 – Tree of life terminology in evolutionary biology.

The biological literature further introduces a classification of subsets Y of X depending on whether or not the species in Y are more closely related together than to any other species. These notions are very similar to the monophyly/paraphyly dichotomy, and are often mixed up.

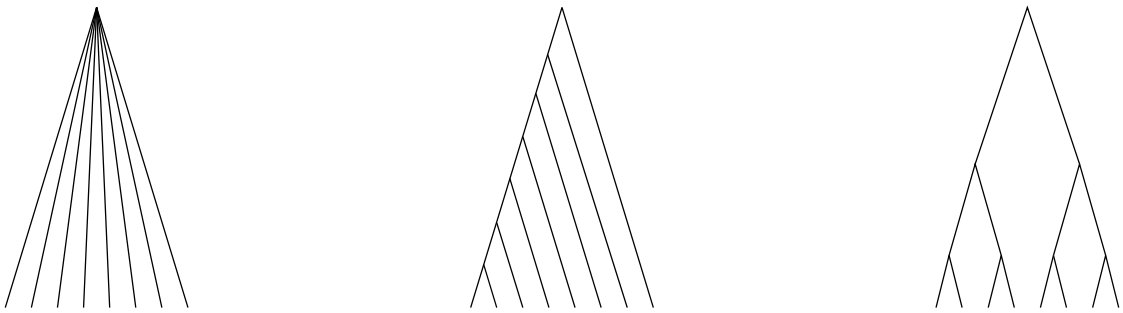
Definition 8 *We say that $Y \subset X$ is* exclusive *if*

$$\forall y, y' \in Y^2, \forall x \in X \backslash Y,\ mrca(\phi(x), \phi(y)) \preceq_T mrca(\phi(y), \phi(y'))$$

If not, we say that Y is polyphyletic.

These 4 notions are illustrated in Figure 1.7. They convey the same underlying idea, of trying to highlight groups of organisms more closely related to one another than to any organism outside the group. While the monophyly/paraphyly focuses on vertices, the exclusivity/polyphyly focuses on leaves only. An exclusive group Y is indeed a set of labels such that $\phi(Y)$ is the intersection of a monophyletic group with the set of leaf vertices. A common slight misuse of language consists in only using the most popular terms of monophyly/ paraphyly in all situations (see chapter 2). These terms have been brought to the forefront by Hennig (1965). They stand at the root of the *cladistics* approach, which considers that only those groups that are monophyletic should appear in a biological classification of life.
A summary of the biological nomenclature surrounding phylogenetic trees can be found in Figure 1.7.

Last, particular trees illustrated in Figure 1.8 play key roles in macroevolutionary thinking. We call a *star tree* the rooted tree in which vertices are either the root or a leaf. The star tree represents the absence of any knowledge on ancestral relationships between species. At the other end of a continuum of tree refinements stand *binary trees*, also referred to as *ternary trees*, *totally bifurcating trees*, or still *totally resolved trees*. These are trees with vertices of degree at most 3 (explaining the 'ternary' nomenclature). They represent the most precise scenarios of ancestral relationships that we can hypothesize for our set of species.



(a) Star tree. (b) Caterpillar tree. (c) Balanced tree.

Figure 1.8 – Particular tree shapes in evolutionary biology. The caterpillar tree (b) is also called 'completely unbalanced' tree. The number of leaves of a balanced tree (c) is a power of 2.

1.1.2 Tree manipulation

In this section, we aim to describe constructions that enable us to represent and explore trees, both conceptually and numerically. We focus only on rooted trees, but unrooted trees are commonly represented in the same way, by specifying an arbitrary root position.

A set representation

We first present a set representation of rooted X-trees discussed at great length in Semple and Steel (2003), and known as X-hierarchies.

Definition 9 *An X-hierarchy H is a set of subsets of X satisfying the following 3 properties:*

1. $\forall A, B \in H^2,\ A \cap B \in \{A, B, \emptyset\}$.
2. $\forall x \in X,\ \{x\} \in H$.
3. $X \in H$.

It can be shown that X-hierarchies and X-trees are in one-to-one correspondence. We illustrate this correspondence in Figure 1.9b and detail it below.

To build a tree from a X-hierarchy H, first consider that each element of H is a vertex. Then, draw an edge incident with A and B if $A \subset B$ and if there is no $C \in H$ such that $A \subset C \subset B$. This construction, which is also called *cover graph* of H, is a rooted X-tree, with root the vertex identified by the element X.

From a rooted X-tree T, we can build its associated hierarchy by taking, for each vertex v of T, the set of labels of the leaves descending from v.

Note that another set representation exists for unrooted trees, known under the name of X-splits, which we will not use in the present manuscript. More details can be found in the book of Semple and Steel (2003).

We will make use of the concept of hierarchy in chapter 2, when we will present some formal definitions of species based on the knowledge of the genealogy of individuals. However, this is not a standard way of representing trees. We now turn to these representations that are commonly used, in practice, in macroevolution.

Recursive decomposition

Rooted trees have a natural recursive representation, also called *standard decomposition*, that proves useful both for theoretical and programming considerations. Let T_ρ be a rooted tree with k edges $(e_i)_{i=1}^k$, where each e_i is incident with ρ and a vertex ν_i. Cutting these k edges, we isolate ρ, and a list of trees $(T_i)_{i=1}^k$ each rooted at vertex ν_i. Each of these rooted trees can in turn be decomposed, and so on recursively until reaching the tips of the tree.

This decomposition can be applied to X-trees, in which case each subtree is an X'-tree, with $X' \subset X$. It can also be applied to edge-weighted trees, in which case each vertex is stored along with the edge-weight incident with the vertex and its mother. We illustrate the decomposition in Figure 1.9c.

Newick representation

The recursive representation mentioned in the last paragraph enables us to devise a string representation of trees known as a *parenthesis representation* or *Newick representation* in evolutionary biology. It consists in only one string object, and is thus well adapted for storage and distribution.

If we call `Newick`(v) the Newick representation of a vertex v, we have:

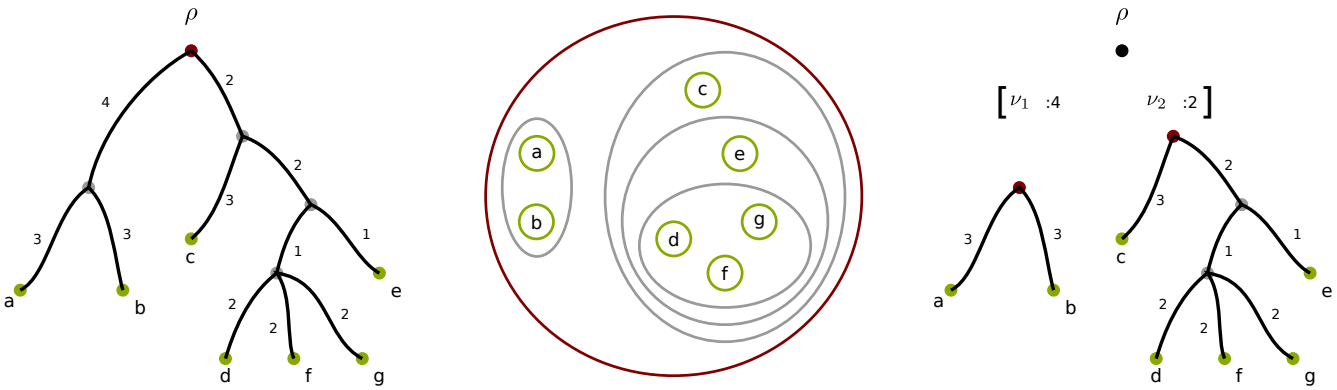
— if v is a tip, `Newick`(v) is an empty string.

— otherwise v has child vertices $v_1, ..., v_k$ and recursively apply:
`Newick`(v) = '(`Newick`(v_1), `Newick`(v_2), ..., `Newick`(v_n))'.

The Newick representation of a rooted tree T_ρ is then `Newick`(ρ).

Additional informations, such as edge-weights, or tip labels, can be incorporated. The most commonly followed conventions would lead the tree represented in Figure 1.9a to be stored as:

'((a:3,b:3):4,(c:3,((d:2,f:2,g:2):1,e:1):2):2);'



(a) A rooted X-tree. (b) A X-hierarchy. (c) Recursive decomposition.

Figure 1.9 – A rooted edge-weighted $\{a, b, c, d, e\}$-tree is represented (a), together with some alternative representations. In (b), the associated $\{a, b, c, d, e\}$-hierarchy where ellipses represent elements of the hierarchy, and the colors correspond to colored vertices on (a). In (c), the principle of recursive tree decomposition is illustrated at the first vertex.

List of edges

Rooted trees might also be represented formally as a list of edges, each specifying its two incident vertices (with, for instance the mother first and her child afterwards). This is the choice made, e.g. in the popular R package 'ape' (Paradis et al., 2004). When dealing with an edge-weighted tree, one only needs to store along each edge its weight. For an X-tree, another list is needed, to match each tip identifier to its label.
The tree in Figure 1.9a could be represented as:

$$\begin{pmatrix} \text{mother:} & \nu_1 & \nu_1 & \nu_2 & \nu_4 & \nu_4 & \nu_4 & \nu_3 & \nu_3 & \nu_2 & \rho & \rho \\ \text{child:} & a & b & c & d & f & g & \nu_4 & e & \nu_3 & \nu_1 & \nu_2 \\ \text{length:} & 3 & 3 & 1 & 2 & 2 & 2 & 1 & 1 & 2 & 4 & 2 \end{pmatrix}$$

While this tree representation does not depend on the order of edges in the list, some particular orders, like pre- and post-orders that we discuss below, are more popular than others.

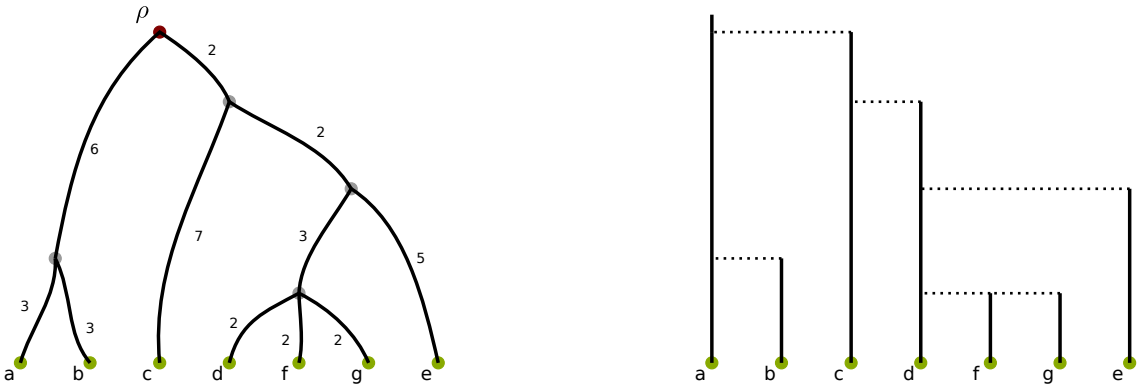
Comb representation for ultrametric trees

Because all tips of an ultrametric tree reach the same exact time position, information on its topology and edge-weights can be stored in a very efficient way, as an ordered list of the internal vertex depths.

Consider one fixed planar representation of a rooted, edge-weighted, X-tree with n tips ordered from left to right labelled as $x_0, x_1, ..., x_{n-1}$. We can associate to this representation a list of vertex depths $(h_i)_{0 \leq i \leq n-1}$, defined through:

$$h_0 := \infty$$
$$\forall i > 0,\ h_i := \frac{1}{2} d_T(\phi(x_i), \phi(x_{i+1}))$$

From the ordered list of (h_i), one can draw back the initial tree by lining up segments of length h_i starting at present. Each segment is then joined to its left neighbours by an horizontal line stopped at the first vertical segment. The procedure is illustrated in Figure 1.10. Note that the tree in this figure would be represented as:



(a) A rooted ultrametric X-tree.

(b) Vertex depths of the tree.

Figure 1.10 – Representation of an ultrametric tree (a) through the list of its internal vertex depths (b).

$$\begin{pmatrix} \infty & 3 & 9 & 7 & 2 & 2 & 5 \\ a & b & c & d & f & g & e \end{pmatrix}$$

The representation is not unique, for each tree can be embedded in the plane by considering any ordering of subtrees at internal vertices. The same tree could as well be represented, e.g. as:

$$\begin{pmatrix} \infty & 3 & 9 & 5 & 2 & 2 & 7 \\ b & a & e & d & g & f & c \end{pmatrix}$$

We will see in section 1.1.4 that this construction is fruitful for simulating reconstructed phylogenies under some probabilistic models.

Tree traversal

A tree traversal consists in exploring each vertex of a tree exactly once, thus defining a total order on the set V, that we will denote here $\leq$.

The traversal is called *breadth-first*, when the order in which vertices are explored is such that,

$$\forall v_1, v_2 \in V^2,\ d_T(v_1, \rho) \leq d_T(v_2, \rho) \ \Rightarrow\ v_1 \leq v_2$$

where d_T refers to the topological distance, i.e. the number of edges in the path between two vertices. The successive *corollas* of vertices at the same topological distance from the root are explored sequentially.

In contrast, the traversal is called *depth-first* when the order in which vertices are explored is such that:

$$\begin{aligned} \text{either (i)} \quad & \forall v_1, v_2 \in V^2,\ v_1 \preceq_T v_2 \ \Rightarrow\ v_1 \leq v_2 \\ \text{or (ii)} \quad & \forall v_1, v_2 \in V^2,\ v_1 \preceq_T v_2 \ \Rightarrow\ v_2 \leq v_1 \end{aligned}$$

When the order satisfies (i), it is called a *pre-order*, while if it satisfies (ii) it is called *post-order*. In the field of macroevolution, (ii) is also called a *pruning order*, because it is the order used in the *pruning algorithm* proposed by Felsenstein (1981) to compute the likelihood of tip data under some models of phenotypic evolution (see algorithm 3 in section 1.2).

Depth-first traversal is much more common in macroevolution, for it allows one to explore all vertices before their descent (pre-order) or all vertices before their parent (post-order). Both procedures will prove very useful when studying the evolution of a heritable trait in section 1.2.

We now take the most of our tree formalization, and seize the opportunity to describe probability distributions on trees and dated trees. We discuss those models that we consider parts of the shared foundation for *diversification* studies, i.e. for studying the tempo at which species appear and go extinct through time. Applications of these models will be discussed in section 1.4.

1.1.3 Probabilistic models of trees

We start with an overview of three probabilistic models for rooted, binary, X-trees on a set X comprising n elements. These three models represent a reference for evolutionary biologists because they allow generating binary trees with distinct *balance*. We say that a tree is *balanced* when each branching event tends to split the number of leaves subtended by the node into equal parts. This characteristic of empirically reconstructed phylogenies has been compared to expectations under the three models. The third one has led to a widely used index of tree balance, that we also use in chapter 3.

Yule-Harding model

The Yule-Harding model is best described by its random generating procedure. It consists in iteratively picking elements from the set X of leaf-labels and growing the tree by adding one edge at a time:

1. Pick two elements in X and build a rooted binary tree with these two leaves.
2. While the tree has less than n leaves:
   (a) Pick uniformly an element $x \in X$ never picked before.
   (b) Select uniformly a leaf labeled y.
   (c) Subdivide the edge incident with y, by creating a new vertex leading to x and y.

This procedure is illustrated in Figure 1.11.

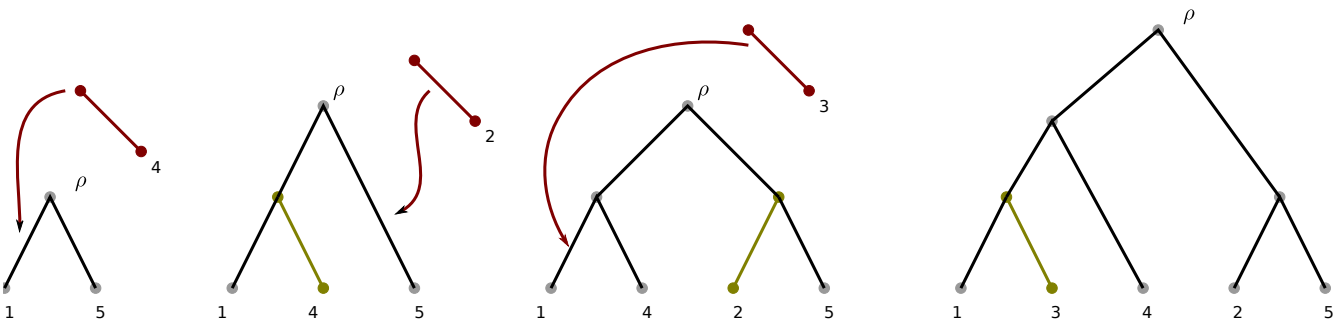


Figure 1.11 – An example of a step-by-step Yule-Harding simulation, generating a rooted binary $\{1, 2, 3, 4, 5\}$-tree.

Uniform or PDA model

The uniform model consists in assigning to all distinct rooted, binary, X-trees the same probability. At this point, it might be interesting to have an idea of the number of distinct rooted, binary, X-trees on a set X of n elements.

Let us call r_n the number of distinct rooted binary $\{1, 2, ..., n\}$ trees, and e_n^r the number of edges on a rooted binary $\{1, 2, ..., n\}$-tree. We can find by induction that:

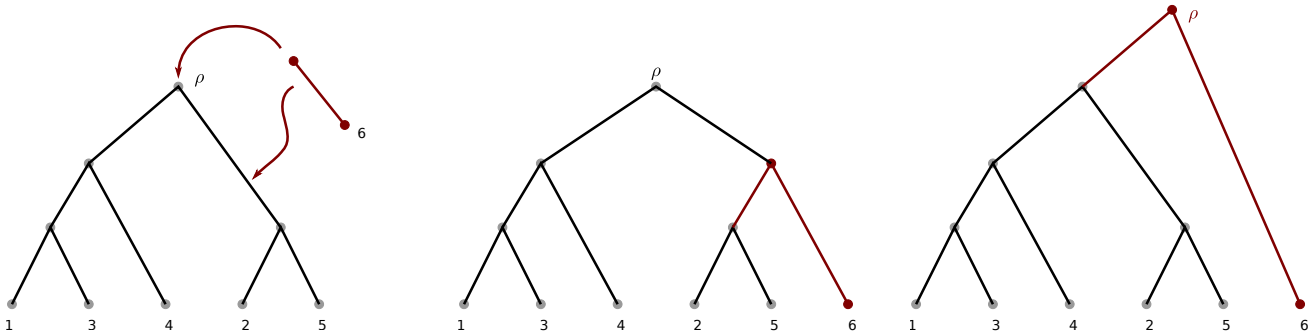
$$r_n = 1 \times 3 \times 5 \times ... \times (2n-3) =: (2n-3)!!$$
$$e_n^r = 2n - 2$$

On a rooted tree with two leaves, we indeed have $r_2 = 1$ and $e_2^r = 2$.

Now, suppose we have a rooted, binary, $\{1, 2, ..., n-1\}$-tree T. We can get a rooted, binary, $\{1, 2, ..., n\}$-tree by either:

1. selecting one of the $2n - 4$ edges of T, subdividing it with a new vertex leading to leaf n (see Fig. 1.12b);
2. or, adding an edge e between the root ρ and the new leaf n, before creating a new root in the middle of e (see Fig. 1.12c).

We can thus get $2n - 3$ distinct binary rooted $\{1, 2, ..., n\}$-tree from each of the possible binary rooted $\{1, 2, ..., n - 1\}$-trees. Moreover, the new tree with n leaves has 2 additional edges compared to the tree with $n - 1$ leaves.



(a) A $\{1, 2, ..., 5\}$-tree. (b) Branching on an existing edge. (c) Adding an outgroup.

Figure 1.12 – Two operations to build a rooted binary $\{1, 2, ..., 6\}$-tree from a rooted binary $\{1, 2, ..., 5\}$-tree (as illustrated in (a)), depending on the place of the new leaf labeled n.

Steel (2014) discusses some differences between the Yule-Harding distribution and the uniform distribution on binary rooted X-trees, with $|X| = n$. The Yule-Harding probability distribution tends to favour more balanced trees than the uniform one. This can be seen through two characteristics of these distributions. First, the expected number of edges between the root and leaves under the Yule-Harding distribution grows as $\log n$, whereas it grows as $\sqrt{n}$ under the uniform distribution. Second, there is a single leaf adjacent to the root with probability $2/(n-1)$ under the Yule-Harding distribution, whereas it is much more likely, with a probability $n/(2n-3)$ under the uniform distribution.

Beta-splitting model

The two previous distributions are two distinct points in a continuum of distributions that is well described by Aldous (1996). The β-splitting model is intended to provide a distribution on binary rooted X-trees, parametrized by a single parameter $-2 \leq \beta \leq \infty$.

The procedure for generating a random tree under the β-splitting distribution is illustrated in Figure 1.13 and described now for $\beta > -1$:

1. place n points labeled by the n elements of X uniformly on the unit segment.
2. while there exist segments with more than one points,
   (a) select all segments with at least two points.
   (b) rescale each of them to unit length.
   (c) split independently each interval at a random point chosen from the density f on $(0, 1)$ with $f(x) \propto x^\beta (1 - x)^\beta$.

Note that for $-2 < \beta \leq -1$ the function f is not a density anymore. We consider a similar recursive procedure, where each split of a set of n elements has a probability $q_n(i)$ to have i elements in the left subset and $n - i$ in the right subset, with:

$$q_n(i) \propto \binom{n}{i} \int_0^1 x^i (1-x)^{n-i} f(x) dx$$

Each of the $\binom{n}{i}$ choices of elements is then equiprobable. This procedure makes sense for $\beta \in (-2, -1]$ because the normalizing constant is finite.

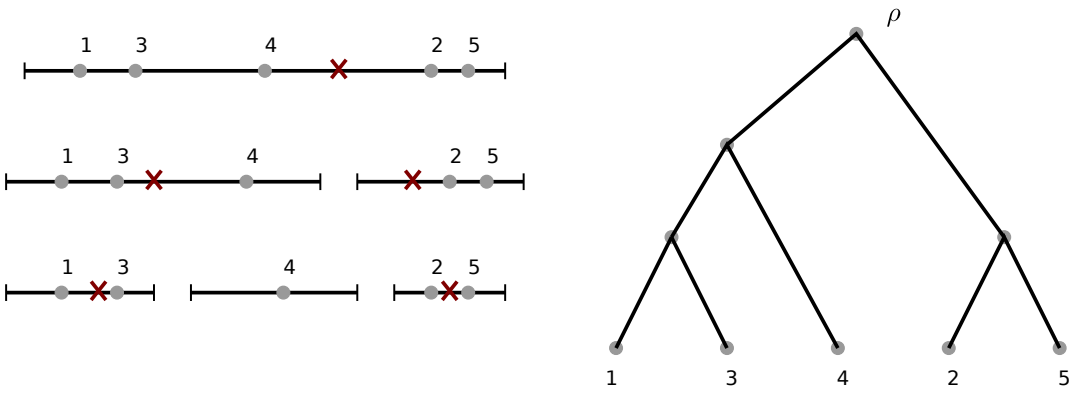


Figure 1.13 – An example of a step-by-step β-splitting simulation, generating a rooted binary $\{1, 2, 3, 4, 5\}$-tree (adapted from Aldous (2001)).

When $\beta = -2$, we consider that the distribution puts non-zero probabilities only on completely unbalanced trees. In contrast, when $\beta = \infty$ the interval $(0, 1)$ is always splitted in its middle, thus leading to more balanced trees. In between these two extremes, one can recover the uniform distribution when $\beta = -1.5$ and the Yule-Harding distribution when $\beta = 0$.

The β-splitting model has been widely used as a way to get an index of a tree balance. This index is $\hat{\beta}$, the maximum likelihood estimator of the β parameter of the model. Empirical phylogenies typically have $\hat{\beta}$ values close to -1. This will be discussed further in chapter 3.

These probabilistic models allow us to compare the shape of empirical phylogenies to those obtained from some theoretical distributions, with the ambition to get insights into the processes that have shaped phylogenies. However, evolutionary biologists are mainly interested in trees displaying an additional time information. We will turn now to probabilistic models aimed at generating such dated trees.

1.1.4 Probabilistic models of dated trees

We start with the well-known Kingman coalescent, that is in fact more used in population genetics than in macroevolution. We then expose some results on birth-death processes, which are intensively used in macroevolution.

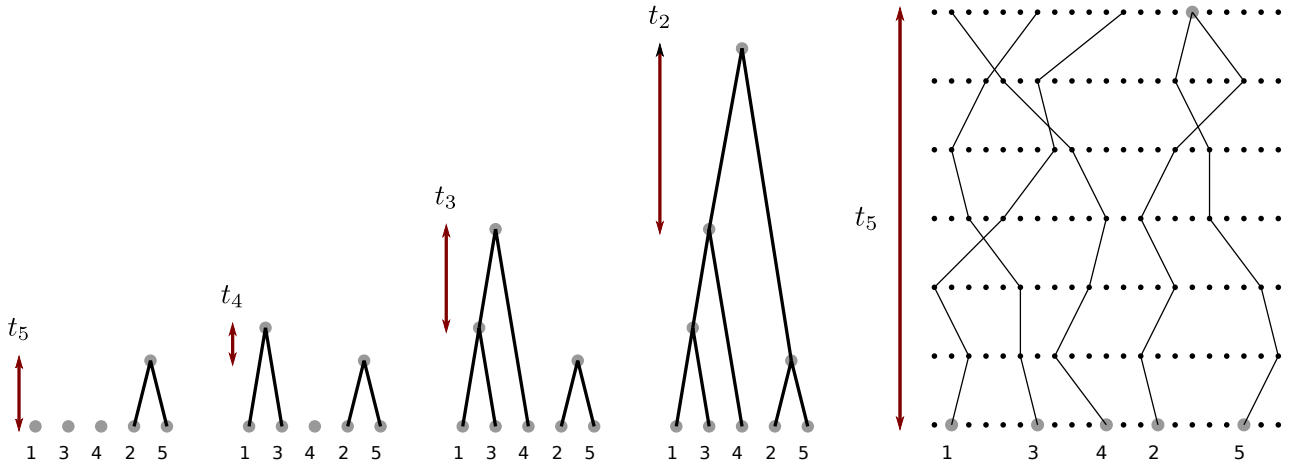
Kingman coalescent

The Kingman coalescent is used to generate rooted, binary, dated X-trees under the procedure illustrated in Figure 1.14a and detailed below.

1. Start with a set of n *active vertices* labeled by the elements of X.
2. While there are $k \geq 2$ active vertices in the process, recursively do the following:
   (a) draw a realisation t_k of the next random time when a *coalescent* event happen: $T_k \sim \mathcal{E}(k(k-1)/2)$.
   (b) select two active vertices uniformly and merge them at time t_k into one active vertex.

Interestingly, this model is obtained as a limit of two well-known discrete processes known under the names of *Moran* model and *Wright-Fisher* model. We explain now the Wright-Fisher model (also illustrated in Fig. 1.14b), and we refer to Durrett (2008) for more details about these two models, the Kingman coalescent, and applications to population genetics.

We consider a population of fixed size, comprising N particles having discrete and synchronized generations. The Wright-Fisher model then randomly assigns parental relationships between particles. At each generation $t+1$, each individual of the generation chooses its mother uniformly among the individuals of generation t. The Kingman coalescent is the distribution of the genealogy of a sample of n individuals under this process, when time is accelerated by a factor N and $N \to \infty$.



(a) Kingman coalescent. (b) Wright-Fisher model.

Figure 1.14 – In (a), an example of a simulation of a Kingman coalescent with 5 leaves. In (b), a zoom on the t_5 time interval, illustrating the link with the Wright-Fisher model.

Birth-death model

The birth-death process that we describe below will provide us with another way to generate random rooted, binary, dated trees. It is parametrized by two parameters, called the *birth rate* b and the *death rate* d. The forward-in-time description of the process is the following, where each particle can be represented as a growing edge of a tree.

1. start with one vertex, initiating one particle at time 0.
2. while there are living particles, draw recursively their history:
   (a) each particle gives birth to another, independent, one, at rate b (i.e. after an exponentially distributed time with parameter b).
   Create a vertex at this point. Two independent particles start here.
   (b) each particle dies at rate d (i.e. after an exponentially distributed time with parameter d).
   Create a leaf vertex ending the particle growing edge.

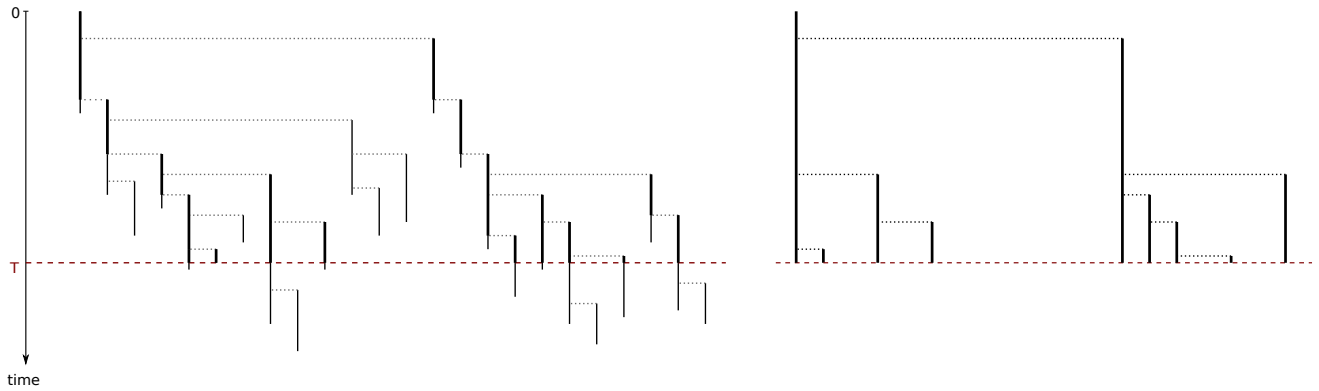
Note that this process is called *pure birth* or *Yule process* with parameter b when $d = 0$. The extinction probability before time t of a process originating from one particle at time 0 is a quantity of central importance in this process. We can show that it is:

$$u(t) = \frac{1 + \frac{d}{b-d} e^{(b-d)t}}{1 + \frac{b}{b-d} e^{(b-d)t}}$$

The quantity $b - d$ is called the growth rate, or *diversification* rate in the context of phylogenies. When $b - d = 0$, the process is called *critical*. As time increases to ∞, the process goes extinct with

probability 1. When $b - d < 0$, the process is called *subcritical* and the process also goes extinct almost surely. Last, when $b - d > 0$, the process is called *supercritical* and the extinction probability tends to d/b as time tends to ∞.

From the description above, one can simulate a tree conditioned on having a total height less than T by simply stopping the simulation at time T. When $d = 0$, the resulting tree is ultrametric. However, when $d \neq 0$, the tree might have extinct branches, i.e. leaves that do not live at time T. In the context of diversification studies, a lot of taxa do not have a good fossil record, and the trees that are studied often contain only extant taxa. Researchers thus compare empirical ultrametric trees to *reconstructed trees*, i.e. trees simulated under a birth-death process, with erased extinct branches (see Fig. 1.15).



(a) Tree with extinct branches.

(b) Reconstructed tree.

Figure 1.15 – Tree generation under a birth-death process (adapted from Lambert and Stadler (2013)). In (a), the whole history of the process is represented. Each segment represents the lifetime of one particle, horizontal dotted lines link to the birth of a child to the right. In (b), the process is stopped at some time (interpreted as present time) and extinct lineages are removed. It is called a *reconstructed tree*, and can be directly simulated using theory on CPPs.

A fancier, an faster, way to directly simulate reconstructed trees is offered by theory on *coalescent point processes* that we now discuss.

Coalescent Point Processes

The following short presentation is extracted from Lambert and Stadler (2013). A comprehensive survey of the topic can be found in this paper.

Definition 10 *A* coalescent point process *(CPP) with stem age T is a random ultrametric tree with height T (in the sense that all tip points are at the same distance T from the stem age), whose node depths are characterized by independent draws from the same distribution H, until a value larger than T is drawn.*

The authors show under which conditions a birth-death process is a CPP. In particular, a constant rate birth-death process is a CPP, and the law H of node depths can be derived. Reconstructed trees can thus be simulated directly without extinct parts, by growing the tree *horizontally* through iterative simulations of node depths under the law H. Moreover, it allows one to compute the density probability of any reconstructed tree as a product of the node depth densities. Figure 1.15b illustrates the simulation of a CPP.

Lambert and Stadler (2013) also discuss general extensions to the previously described constant-rate birth-death process. Rates have been proposed to vary with time (Nee et al., 1994), with the number of species (Schluter, 2000), with a non-heritable trait value (Condamine et al., 2013), or with a heritable

trait value (Maddison et al., 2007). They expose under which conditions the resulting dated tree is a CPP, and under which conditions the resulting (non-dated) tree follows a Yule-Harding distribution. These results are key to the applications presented in section 1.4.

Trees are central structures in macroevolution studies. They first need to be reconstructed, usually using molecular data. Once reconstructed, they might also be used to further investigate the evolution of phenotypic traits among related organisms. Both of these applications, that will be detailed in section 1.4, also strongly rely on probabilistic models of trait evolution unfolding on a tree, that we now turn to discuss.

1.2 Probabilistic models of trait evolution

In this section, we aim to introduce the methods used to model the phenotypic evolution of organisms on phylogenies. We first state some generalities on the probabilistic tools that we need, and then present the usual hypotheses made in the field. We provide an idea of what models are, how to simulate them, and how to compute the probability of a final, observed, state. Empirical applications will be discussed later in section 1.4.

1.2.1 Markov processes on a discrete state space

We start with a description of Markov chains, useful to model a random walk with discrete time steps on a discrete state space. We then extend it to Markov processes, with steps happening in a continuous time. We then provide examples of models for molecular evolution.

Mathematical description

Definition 11 *Let I be a countable set. Let ν be a probability distribution on I.*
We say that $P = (p_{ij})_{i,j \in I^2}$ is a stochastic matrix if $\forall i, j \in I^2,\ p_{ij} \geq 0$ and $\sum_j p_{ij} = 1$.
A sequence $(X_k)_{k \in \mathbb{N}}$ of random variables is a Markov chain *with initial distribution ν and transition probabilities P if:*

$$\forall i,\ \mathbb{P}(X_0 = i) = \nu_i$$
$$\forall n > 0,\ \forall (i_0, i_1, ..., i_n) \in I^{n+1},\ \mathbb{P}(X_n = i_n \mid X_0 = i_0,\ X_1 = i_1,\ ...,\ X_{n-1} = i_{n-1}) = p_{i_{n-1} i_n}$$

This definition allows us to model a random process with a state evolving through time, where the index of the random variables X_n is interpreted as a discrete time. However, it is much more intuitive and useful in macroevolution to model the evolution of characters in a continuous in time manner. Intuitively, we should not be surprised that a similar Markovian property with a continuous time could rely on the *memoryless* property of the exponential distribution. This is exactly how *Markov processes* are constructed.

Definition 12 *Let I be a countable set. Let ν be a probability distribution on I.*
We say that $Q = (q_{ij})_{i,j \in I^2}$ is a rate matrix if $\forall i \in I,\ q_{ii} \leq 0,\ \ \forall i \neq j,\ q_{ij} \geq 0$ and $\sum_j q_{ij} = 0$.
The jumping times of the process are the sequence of (J_n) defined through:

$$J_0 = 0$$
$$J_{n+1} = \inf\{t > J_n,\ X_t \neq X_{J_n}\}$$

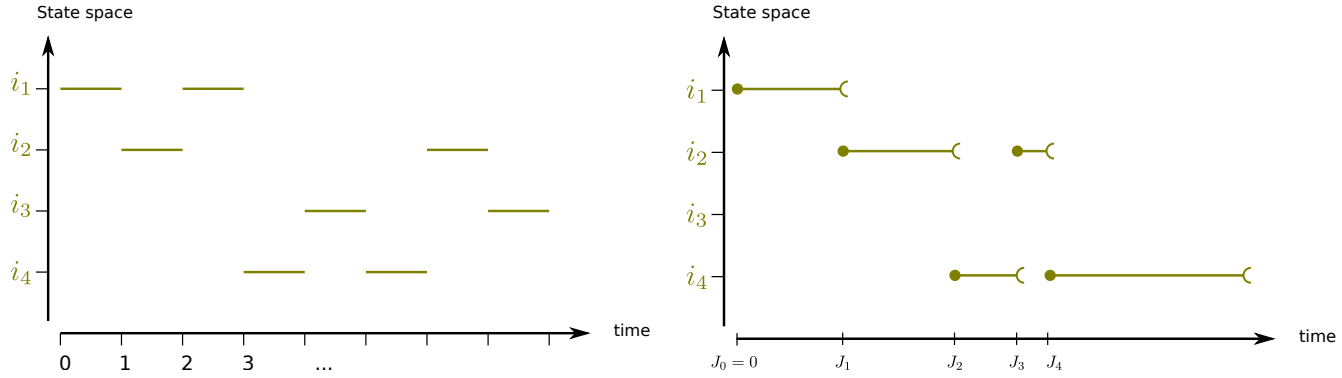
Further, let P be the stochastic matrix with elements:

$$\forall i \in I, \quad \text{if } q_{ii} = 0,\ p_{ii} = 1, \quad \text{and } \forall j \neq i,\ p_{ij} = 0$$
$$\text{if } q_{ii} \neq 0,\ p_{ii} = 0, \quad \text{and } \forall j \neq i,\ p_{ij} = \frac{q_{ij}}{-q_{ii}}$$

A sequence $(X_t)_{t\in\mathbb{R}}$ *of random variables is a* Markov process *with initial distribution* ν *and rate matrix* Q *if:*

1. $(X_{J_n})_{n\in\mathbb{N}}$ *is a Markov chain, called* embedded chain *with initial distribution* ν *and transition probabilities* P*.*
2. $\forall n > 0,\ (J_n - J_{n-1}) \sim \mathcal{E}(-q_{X_{J_{n-1}} X_{J_{n-1}}})$*.*

Figure 1.16 illustrates definitions 11 and 12. It conveys the basic intuition that we need remembering in order to manipulate Markov processes: the future of the process depends on what is known in the past only through its last known state. When the process is in state $i \in I$, the time before the next jump is exponentially distributed with parameter $-q_{ii} \neq 0$. When this 'exponential clock' rings, the process jumps to another state j with a probability given by the ratio $-q_{ij}/q_{ii}$. If $q_{ii} = 0$, then the process is trapped in state i. This definition thus gives us a practical way to simulate numerically such a process, provided we know ν and Q.



(a) Markov chain. (b) Markov process.

Figure 1.16 – A discrete time Markov chain (a) and a continuous time Markov process (b) trajectories on the same state space $I = \{i_1, i_2, i_3, i_4\}$.

Kolmogorov equations

The theory of Markov processes further allows us to compute the transition probabilities between any two states, after a time t. The following, very fruitful, result, is known as a *Kolmogorov equation*. If we call $P(t)$ the matrix of transition probabilities with elements:

$$P(t) \ := \ (\mathbb{P}(X_t = j \mid X_0 = i))_{i,j\in I^2}$$

Then we have $P(0)$ that is the identity matrix, and $P'(t) = P(t)Q$. This further implies that

$$P(t) = e^{tQ}$$

The latter equation typically allows us to simulate the end state of a Markov process, after a time t, without simulating the whole trajectory. It nonetheless requires to be able to compute, either analytically or numerically, the exponential of a matrix.

Time reversibility and stationarity

Two properties of Markov processes, known as *time reversibility* and *stationarity*, turn out to be central hypotheses for the applications we are interested in. We provide here the definitions of the terms.

Definition 13 *Let ν be a probability distribution on a state I.*
We say that a Markov process has a stationary distribution ν *if* $\forall t \in \mathbb{R}$, $\nu P(t) = \nu$.

Note that this is also equivalent to finding a probability distribution ν verifying $\nu Q = 0$. Two consequences will prove effective for the applications we are interested in.

First, the probability distribution of a process starting in its stationary distribution remains, after any time, in the same distribution. Second, when a process admits a unique stationary state, with, $\forall i \in I,\ \nu_i > 0$, then it is said to be *ergodic*, which means that the distribution of the process after a large amount of time will be ν, i.e.

$$\lim_{t\to\infty} p_{ij}(t) = \nu_j$$

This is the rationale for almost always considering stationary processes in the literature. Moreover, stationary processes are almost always considered to start in their stationary distribution, for we suppose that they have been running for a long time before.

The second definition will provide a way to define such stationary processes.

Definition 14 *A Markov process on I with initial distribution ν and rate matrix Q is said to be* time reversible *if* $\forall i, j \in I^2,\ \nu_i q_{ij} = \nu_j q_{ji}$.

This property intuitively means that the probability of trajectories forward or backward in time are the same, i.e. that the law of $(X_s)_{0\leq s\leq t}$ is the same as the law of $(X_{(t-s)-})_{0\leq s\leq t}$.
It further implies that ν is a stationary distribution of the Markov chain and that the process is stationary.

Biological example: nucleotide evolution

Models of nucleotide or amino-acid evolution supply a wealth of examples for Markov processes (Galtier et al., 2005). A given nucleotide referenced by its precise position along a DNA sequence, for example, may exist in four different states: adenine, thymine, cytosine, guanine. The state space is thus denoted $I = \{A, T, C, G\}$.

A model for a single nucleotide evolution can be proposed by specifying an initial distribution ν on I, and a rate matrix Q. Many such models have been proposed, historically ranging from simple parameterizations to more and more complex ones. The most general time reversible model is (indeed) called *Generalized Time-Reversible* model (GTR, Lanave et al. (1984)). Its 9 parameters are $\pi_A, \pi_T, \pi_C, a, b, c, d, e, f$ and, if we denote $\pi_G = 1 - \pi_A - \pi_T - \pi_C$, the model is defined through:

$$\nu = (\pi_A, \pi_T, \pi_C, \pi_G)$$

$$Q = \begin{pmatrix} -(a\pi_T + b\pi_C + c\pi_G) & a\pi_T & b\pi_C & c\pi_G \\ a\pi_A & -(a\pi_A + d\pi_C + e\pi_G) & d\pi_C & e\pi_G \\ b\pi_A & d\pi_T & -(b\pi_A + d\pi_T + f\pi_G) & f\pi_G \\ c\pi_A & e\pi_T & f\pi_C & -(c\pi_A + e\pi_T + f\pi_C) \end{pmatrix}$$

More simple, popular, models include:

the Jukes-Cantor (JC69) model assuming that all nucleotide proportions are equal, and that all substitutions are equiprobable.

the Kimura (K80) model assuming that all nucleotide proportions are equal, but that two substitution rates exist. A first one concerns transition ($A \leftrightarrow G$, $C \leftrightarrow T$) while a second one concerns transversions (all other substitutions).

the Tamura (T92) model building on K80 by relaxing the hypothesis that all nucleotide proportions are equal. Instead, two groups of nucleotide are proposed: (A and T) and (C and G). They have the same proportions, within the group, but it may differ from the other group.

the Tamura and Nei (T93) model proposing that all nucleotide proportions are distinct, and introducing three substitution rates: one for $A \leftrightarrow G$, one for $C \leftrightarrow T$ and one for all other transversions.

Their definition can be expressed as special nested cases of the GTR model, as described in Figure 1.17.

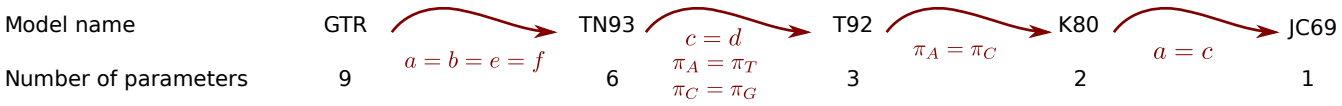


Figure 1.17 – Popular models of nucleotide evolution seen as nested models (adapted from Perrière and Brochier-Armanet (2010)).

1.2.2 Diffusion processes on a continuous state space

In the previous section, we first mentioned Markov chains, that are well adapted tools to describe the evolution of a discrete trait in discrete time. We then turned to Markov processes, allowing us to model discrete traits in continuous time. In this section, we will introduce the basics that we need to model continuous traits in continuous time. These processes are known as *diffusion processes*.

Introducing the Brownian Motion

The Brownian Motion is the first diffusion that one needs to study, for all others rely on its definition. It can be obtained as the limit of the random walk described below.

We consider a sequence of random independent increments $(\xi_i)_{i\in\mathbb{N}}$ with, $\forall i$, $\mathbb{P}(\xi_i = 1) = \mathbb{P}(\xi_i = -1) = \frac{1}{2}$. From these, we can build a random walk on $\mathbb{Z}$, with the state of the chain at step n given by:

$$S_n = \sum_{i=1}^{n} \xi_i$$

This random walk is represented in Figure 1.18a. We are now interested in finding the appropriate scale at which to look at this walk, in order to get a non-trivial continuous in time process.

Let's zoom out on the time scale by a factor n, and a bit less on the state space, with an order $\sqrt{n}$, as in Figure 1.18b. The central limit theorem gives us the convergence in law:

$$\frac{1}{\sqrt{n}} S_{[nt]} = \frac{\sqrt{[nt]}}{\sqrt{[n]}} \frac{1}{\sqrt{[nt]}} \sum_{i=1}^{[nt]} \xi_i \longrightarrow \mathcal{N}(0,t)$$

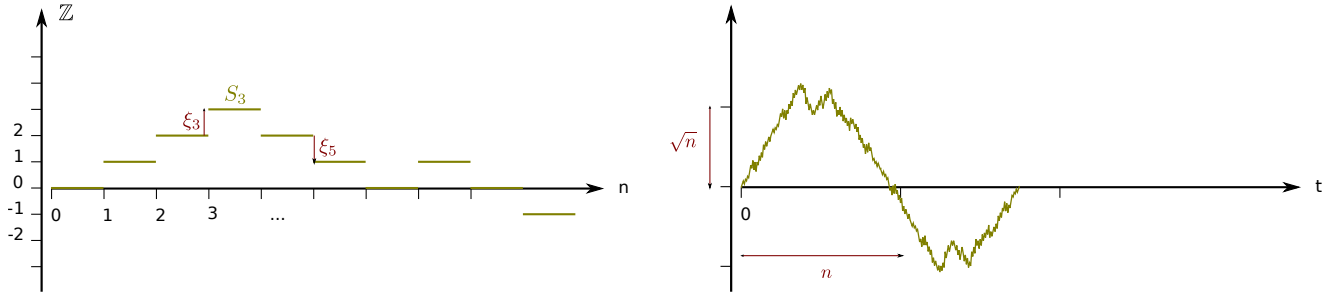
This limit of a random walk really gives us the intuition of what the Brownian motion is, together with a numerical way to simulate it. Here is now the appropriate mathematical definition of the Brownian motion:

Definition 15 *We say that $(B_t)_{t\geq 0}$ is a standard Brownian motion, if:*

1. *$t \mapsto B_t$ is continuous*
2. *$\forall s, t$, the law of $(B_{t+s} - B_t)$ does not depend on t*

3. $(B_s)_{s\leq t}$ *and* $(B_{t+s}-B_t)_{s\geq 0}$ *are independent*
4. $B_0 = 0$
5. $B_1 \sim \mathcal{N}(0,1)$.

In what follows, we need only remember that it is a continuous process, with random, Gaussian increments between any two time points. This, by the way, gives us another useful way to simulate a Brownian trajectory on an interval $[0,T]$. We can first discretize the interval into n small pieces with length T/n, and then simulate iteratively small random Gaussian increments with parameters $(0,T/n)$ (see algorithm 1 in what follows, for more details).



(a) Random walk on $\mathbb{Z}$.

(b) Brownian Motion.

Figure 1.18 – From a random walk on $\mathbb{Z}$ (a) to the Brownian motion (b). Changing the scale by zooming out by a factor n in time and $\sqrt{n}$ in space, then letting $n \to \infty$, leads to a non-trivial process, with Gaussian, independent, increments.

Diffusion processes

Diffusion processes are built upon the Brownian Motion. Defining them properly requires a lot of theory that we do not want to introduce here. Instead, we will try to convey the intuition one needs in order to use them for modeling purposes only.

One convenient way to get the intuition on diffusion processes consists in providing the *Stochastic Differential Equation* (SDE) it verifies. If the reader is familiar with the meaning of (deterministic) ordinary differential equations, there is not much to add to get the meaning of their stochastic counterparts.

$$dx_t = f(x_t,t)dt \tag{1.1}$$
$$dX_t = f(X_t,t)dt + \sigma(X_t,t)dB_t \tag{1.2}$$

Equation (1.1) governs the evolution of a small increase in the trajectory of x at time t. This first version is entirely deterministic, and provided it is coupled with an initial condition for x_0, it fully determines the value of x_t.

In contrast, equation (1.2) governs the stochastic evolution of a random variable X_t. The small increment dX_t at time t depends both on its value, and on a small random Brownian increment dB_t, that can be thought of as a *noise term* added to the trajectory. Provided it is coupled with an initial condition giving the law of X_0, this equation fully determines the distribution of X_t.

For the applications we have in mind, the best way to grasp intuition on SDE is probably to discuss how we can simulate random trajectories from them. This is the aim of algorithm 1.

Algorithm 1 *(diffusion simulation through an Euler-Maruyama scheme)*
We aim at simulating an approximation of a random trajectory on $[0,T]$ *of a diffusion satisfying the SDE:*

$$dX_t = f(X_t, t)dt + \sigma(X_t, t)dB_t$$
$$X_0 \sim \mathcal{L}_0$$

The Euler-Maruyama scheme is the adaptation to SDE of the Euler method used to numerically solve an ordinary differential equation.

1. *Discretize* $[0, T]$ *into* n *small pieces of length* $\Delta_t = T/n$.
2. *Initialize the trajectory by drawing a realisation* x_0 *under the law* $\mathcal{L}_0$. *Fix* $t_0 = 0$ *and* $i = 1$.
3. *While* $i \leq n$:
   (a) *Fix* $t_i := iT/n$.
   (b) *Draw a realisation* s_i *under the law* $\mathcal{N}\left(0,\ \sigma(x_{i-1}, t_{i-1})^2\Delta_t\right)$.
   (c) $x_i := x_{i-1} + f(x_{i-1}, t_{i-1})\Delta_t + s_i$
   (d) *Increment* $i \leftarrow i + 1$.

This algorithm emphasizes the two components that drive the evolution of X_t: the small deterministic part $f(x_{i-1}, t_{i-1})\Delta_t$ and the small stochastic part that is a random Gaussian increment centered on 0. As a result, any two random trajectories will be distinct from one another.

In practice, researchers using diffusion processes for modeling purposes tend to choose specifically well studied diffusions. Typically, a lot of those used to study trait evolution through time are such that we can, analytically, derive the exact distribution of X_t at any time t. We provide now examples gleaned from the literature.

Biological example: body mass evolution

The body mass of organisms is probably the continuous trait that has been most studied, for this is a trait that can be measured fairly easily for a wide range of organisms. Many distinct models have been proposed, but we detail here the simplest ones only. The very first observation that we must make is that body mass is a trait that lives in $(0, \infty)$.

One possibility that has been explored is thus to propose a model for the logarithm of the trait, that lives in $\mathbb{R}$. If X_t denotes the logarithm of the trait, then we can propose that:

$$dX_t = adt + \sigma dB_t$$

This first model, known as *drifted Brownian motion* allows one to look for a directional trend in the evolution of the body mass, with drift strength a. The second parameter σ is usually called *diffusion coefficient*.

However, under this model, even if $a = 0$, the expectation of the body mass increases through time. Indeed, by Jensen's inequality, we have $\mathbb{E}(\exp(\sigma B_t)) > \exp(\mathbb{E}(\sigma B_t)) = 0$. It is further possible to show that:

$$\mathbb{E}(e^{\sigma B_t}) = e^{\frac{\sigma^2}{2}t}$$

This increase in expectation through time can be suppressed by considering a process called *Geometric BM*, with the following expression:

$$X_t = X_0\ e^{(a - \frac{\sigma^2}{2})t + \sigma B_t}$$

This process follows the stochastic differential equation: $dX_t = aX_t dt + \sigma X_t dB_t$.

Another very popular model that has been applied to body mass evolution is the *Ornstein-Uhlenbeck process* (OU). Its stochastic differential equation can be written as:

$$dX_t = \lambda(X_t - \theta)dt + \sigma dB_t$$

It is intended to model a trait that evolves under selective pressure, with a strength λ pushing back the trait value toward the optimum θ as the trait value moves away from θ. The third parameter σ is still the diffusion coefficient.

This model is closely linked to another one, named *Early Burst* (EB). It is defined through the following stochastic differential equation:

$$dX_t = \sigma_0 e^{-rt}\, dB_t$$

The EB model has been originally introduced as a way to model the fast evolution in the beginning of an adaptive radiation, followed by a slowdown as ecological niches are progressively filled. However, under some particular initial conditions, and when applied on an ultrametric tree, it has been shown that it leads to exactly the same distribution of traits at present than the OU model (Uyeda et al., 2015).

Random trajectories drawn using three basic models are reproduced in Figure 1.19. The legend correspond to the following stochastic differential equations:

$$\begin{aligned} \text{drifted BM} \quad & dX_t = dt + dB_t \\ \text{BM} \quad & dX_t = dB_t \\ \text{OU} \quad & dX_t = (5 - X_t)dt + dB_t \end{aligned}$$

Last, some authors have proposed to directly model the body mass instead of its logarithm. In this case, it has been proposed that the body mass would evolve as a *reflected BM*, i.e. a BM that is reflected upwards when it reaches zero (Boucher and Démery, 2016). An example of such a trajectory is also drawn in Figure 1.19.

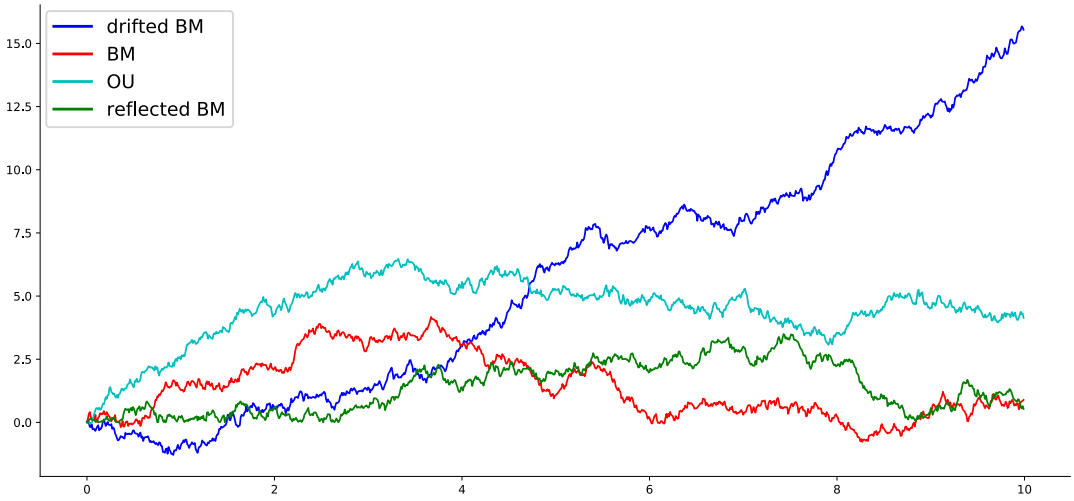


Figure 1.19 – Four random trajectories drawn under distinct models of trait evolution on 10 time units. Parameters of the distinct processes are provided in the main text.

1.2.3 Trait evolution on a fixed tree

In this section, we consider random processes such as those discussed previously (either with a discrete or continuous state space), but we no longer make them run on a single interval. Instead, we would like our processes to unfold on a rooted edge-weighted tree, to model a trait value evolving along

a tree. We state the main hypotheses considered in the literature, and the associated simulation and likelihood computation algorithms.

Classical hypotheses

We intend to provide a generic description of trait evolution on a fixed tree. We will thus consider that, if X is a discrete trait, then $p(X)$ denotes its probability mass function, whereas if X is continuous, then $p(X)$ denotes its probability density.

Definition 16 *Let $T_\rho = (V, E)$ be a rooted tree with vertex set $V = \{\rho, v_1, v_2, ..., v_n\}$. We say that a set of random variables $(X_\rho, X_{v_1}, X_{v_2}, ..., X_{v_n})$ follows a* Markov model on a tree *if its joint distribution can be expressed as:*

$$p\left(X_\rho, X_{v_1}, X_{v_2}, ...X_{v_n}\right) = p\left(X_\rho\right) \prod_{v \in V \setminus \rho} p\left(X_v \mid X_{mother(v)}\right)$$

A Markov model on a tree thus extends the Markovian property to a tree structure, by assuming that the value of the process at a given vertex does only depend on the past through the value of its last ancestor. This is also an example of a *directed graphical model*, on a directed graph with a particular tree structure, where edges are considered to be directed from the root to the tips of the tree.

This independence assumption of processes running on distinct branches is a very strong one. Ancestral organisms living at the same time in the past are likely to influence each other through ecological interactions like competition, predation, mutualism... Chapter 4 will be devoted to relaxing this hypothesis in the context of continuous trait evolution (see also Bartoszek et al. (2016)). In the remainder of the introduction, we will consider that traits evolve independently in distinct lineages, for this is the standard hypothesis in the field.

When applied on dated trees where vertices $\rho, v_1, v_2, ..., v_n$ are associated to times $t_\rho, t_1, t_2, ..., t_n$, Markov models on trees have transitions that can be naturally parametrized using the time at vertex v and its mother. In fact, most Markov models on trees considered in the literature can be obtained through the following procedure:

1. define a Markov process or a diffusion process:
   (a) its initial law denoted ν is respectively a probability mass function or a probability density that must be assumed.
   (b) the transition function q of the process is derived, i.e. for any two states (i, j) and for any two dates $t_1 < t_2$, we know the quantity:

$$q(i,\ j)_{t_1,\ t_2} := p\left(X_{t_2} = j \mid X_{t_1} = i\right)$$

2. consider the Markov model on a dated tree where the joint distribution is expressed as:

$$p\left(X_\rho, X_{v_1}, X_{v_2}, ...X_{v_n}\right) = \nu(X_\rho) \prod_{v \in V \setminus \rho} q\left(X_{\text{mother}(v)},\ X_v\right)_{t_{\text{mother}(v)},\ t_v}$$

Simulations on a tree

The previous definition becomes very intuitive as soon as we describe its simulation, from the root to the tips.

Algorithm 2 *(Simulation of a Markov model on a tree)*

1. *The process starts at the root of the tree under a given law ν.*
   We simulate a value x_ρ following the law ν, which is the observed value of the state at the root.

2. *The process then runs independently in the k edges incident with the root.*
   For each child v_i of ρ, simulate a value x_{v_i} according to the transition law $q(x_\rho, \cdot)_{t_\rho,\, t_{v_i}}$.
3. *Cut the k edges, forget ρ, and repeat recursively in the new subtrees rooted at vertices $\{v_i\}_{1\leq i\leq k}$, until reaching the leaves.*

Figure 1.20 presents examples of trajectories for both a Markov process and a diffusion process unfolding on a fixed tree. The tree is traversed *depth-first* in *pre-order*, meaning that the value at each vertex is computed using the (previously computed) value of its mother.

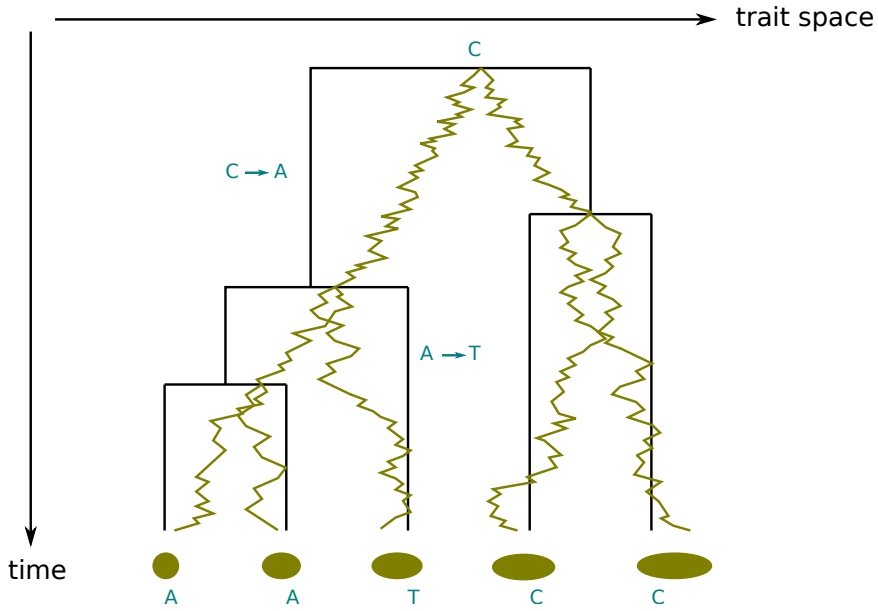


Figure 1.20 – Two stochastic processes running on a tree. In green, a diffusion process modeling the size of the organisms. In blue, a Markov process modeling one nucleotide.

Likelihood computation

Unfortunately, we never have access to the full trajectory on real world applications. We could get it on short timescales, considering experimental evolution studies with model organisms. But we are generally interested in longer timescales, and we thus only have access to the trait values at present. The probabilistic approach exposed before nevertheless allows us to compute the probability of the present-day observation at the tips of a tree. The random variables at the internal vertices are called *hidden* or *latent* variables, and we need ways to numerically integrate over their unknown values.

Consider a tree rooted at vertex ρ, with internal vertices $\{v_1, v_2, ..., v_k\}$ and tip vertices $\mathcal{L} = \{l_1, l_2, ..., l_n\}$. Consider a trait evolving in a trait space S as a Markov model on a tree. We seek to compute:

$$p\left((X_l)_{l\in\mathcal{L}} = (x_l)_{l\in\mathcal{L}}\right) = \int_{S^{k+1}} p\left((X_l)_{l\in\mathcal{L}} = (x_l)_{l\in\mathcal{L}}, X_{v_1} = s_1, X_{v_2} = s_2, ..., X_{v_k} = s_k, X_\rho = s_{k+1}\right) ds_1 ds_2 ... ds_{k+1}$$

The widely known *pruning algorithm* consists in computing this quantity efficiently by following the tree structure. If v is a vertex of T, we will denote $\mathcal{L}(v)$ the set of leaves subtended by v, i.e. $\mathcal{L}(v) := \mathcal{L} \cap \text{desc}(v)$. The algorithm consists in computing at each node v the quantity:

$$\eta_v(s_v) := p\left((X_l)_{l\in\mathcal{L}(v)} \mid X_v = s_v\right), \quad \forall s_v \in S$$

Until reaching the root of the tree, where we get the probability of tip values. It is described in the following algorithm:

Algorithm 3 *(Pruning algorithm: Probability of tip values under a Markov model on a tree)*

1. *compute for all tip vertices $l \in \mathcal{L}$:*

$$\forall s_l \in S, \quad \eta_l(s_l) := p\left(X_l = x_l \mid X_l = s_l\right) = \mathbb{1}_{s_l = x_l}$$

2. *after all children of a vertex v have been traversed, compute,*

$$\begin{aligned}
\eta_v(s_v) :&= p\left((X_l)_{l \in \mathcal{L}(v)} \mid X_v = s_v\right) \\
&= \prod_{u \in child(v)} \int_S p\left((X_l)_{l \in \mathcal{L}(u)} \mid X_u = s_u\right) p\left(X_u = s_u \mid X_v = s_v\right) ds_u \\
&= \prod_{u \in child(v)} \int_S \eta_u(s_u) p\left(X_u = s_u \mid X_v = s_v\right) ds_u
\end{aligned} \tag{1.3}$$

3. *stop at the root of the tree, where the probability of tip values is computed as:*

$$\begin{aligned}
P\left((X_l)_{l \in \mathcal{L}}\right) &= \int_S \nu(s_\rho)\ p\left((X_l)_{l \in \mathcal{L}} \mid X_\rho = s_\rho\right) ds_\rho \\
&= \int_S \nu(s_\rho)\ \eta_\rho(s_\rho) ds_\rho
\end{aligned} \tag{1.4}$$

Note that this algorithm traverses the tree *depth-first* in a *post-order*, because all children of a given vertex v must be observed before v. This algorithm is particularly useful when equation (1.3) can be simplified analytically, or computed efficiently.

This is in particular the case when S is finite. In that case, η_v can be represented as a line vector of size $|S|$ and the transitions $\left(p\left(X_u = s_u \mid X_v = s_v\right)\right)_{s_v,\ s_u}$ can be represented as a square matrix $Q_{v,u}$. The integral in equation (1.3) is a finite sum, that could be rewritten as the product of the vector line η_v with the square transition matrix $Q_{v,u}$. Moreover, the product over the children of v corresponds to the entry-wise product (denoted $\otimes$) of a sequence of line vectors. Putting everything together, equation (1.3) simplifies as:

$$\eta_v = \eta_{u_1} Q_{v,u_1} \ \otimes\ \eta_{u_2} Q_{v,u_2} \ \otimes\ \ldots \ \otimes\ \eta_{u_n} Q_{v,u_n}$$

Last, equation (1.4) simplifies as a product of a vector line η_ρ with a column vector ν giving the initial distribution.

Equations (1.3) and (1.4) can also be simplified when $S = \mathbb{R}$ and all transitions are Gaussian densities. In that case, it can be shown that η_v has also a Gaussian shape, with a mean and variance that can be computed using the mean and variance of (η_u), previously computed for any child u of v (Lartillot, 2013).

We presented in the above section the basic toolkit one needs to understand probabilistic models of trait evolution unfolding on a fixed tree. Together with probabilistic models of tree generation discussed in section 1.1.4, they are at the core of biodiversity evolution modeling. In the next section, we discuss tools available to confront these models to biological data, and produce inferences.

1.3 From probabilistic models to statistics

Statistics generally consist in analysing and interpreting data, using a mathematical formalism borrowed from probability theory.

In the field of macroevolution, we aim at inferring knowledge on the past history of organisms, using phenotypic observations as the main source of raw data. Interpreting this data requires to propose a model, stating the set of possible events and the probability that they occurred. Probabilistic models for generating phenotypic traits have been introduced in section 1.2, while probabilistic models for generating trees have been introduced in section 1.1.4.

We present now some generalities on statistics that are useful for macroevolution studies. We open up with an introduction on the maximum likelihood and Bayesian framework, before turning to numerical algorithms useful to both. We attempt to remain general and models will be truly detailed in the last section 1.4.

1.3.1 The maximum likelihood and Bayesian frameworks

Two competing ways to conceive the statistical approach coexist in the literature. The first one, known as *frequentist approach* considers that the investigator has no prior belief on his model's parameter values. Only the frequency of different observations can provide their estimation. In contrast, the second one, known as *Bayesian approach*, considers that the investigator has a prior belief, which can be adjusted using data observations. We describe both approaches straightaway.

Maximum likelihood estimation

We consider that we are given empirical data x, that is supposed to be an observation of a random variable X having law $p(X\ ;\ \theta)$, parametrized by $\theta \in \Theta$.

Definition 17 *The* likelihood *is the probability of the data, seen as a function of the parameters of the model, i.e.*

$$\forall \theta \in \Theta, \quad L(\theta) = p(X = x\ ;\ \theta)$$

The maximum likelihood principle *consists in estimating θ by:*

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{argmax}\ L(\theta)$$

This estimator is called maximum likelihood estimator *of θ.*

In a maximum likelihood framework, methodological developments generally aim at finding efficient algorithms to first compute, and second optimize, the likelihood function. After the optimization is performed, it might also be interesting to sample the law $p(X\ ;\ \hat{\theta}_{\mathrm{MLE}})$ and compare sample values to the data x, to check whether the model was *appropriate*. Optimization and sampling procedures will be soon discussed.

A popular way to compare distinct models in a likelihood framework consists in comparing their maximum likelihood. However, the goodness of fit of the model, as measured by the maximum likelihood, generally increases with the number of parameters. To tackle this issue, called *over-fitting* and avoid selecting always the most parameter-rich model, appropriate statistical tests might be designed.

Standard theory comes up with well-grounded likelihood ratio tests to compare nested models. In more complex model comparisons scenarios, the Akaike Information Criteria is often invoked. The AIC value of a model with k parameters is defined as $2k - 2\ln L(\hat{\theta})$. The model with the highest AIC value is then selected. Even without solid theoretical grounds, this approach can still be seen as a heuristic to correct for over-fitting.

Bayesian inference framework

The Bayesian framework consists in considering what we called *parameters* of the model as random variables. The model contains more hypotheses, as we now not only need the *conditional law* of X given θ, which is denoted $p(X \mid \theta)$, but we also need to assume a *prior distribution* of θ, which we denote $p(\theta)$.

Using Bayes formula enables us to compute the *posterior distribution* of θ as:

$$p(\theta \mid X = x) = \frac{p(X = x \mid \theta)\ p(\theta)}{\int_{\theta \in \Theta}\ p(X = x \mid \theta)\ p(\theta) d\theta}$$

In a Bayesian framework, methodological developments more generally aim at sampling the posterior distribution efficiently. The whole distribution is obtained, allowing one in particular to look at the maximum a posteriori $\hat{\theta}_{\text{MAP}}$ if a point estimate is really needed:

$$\hat{\theta}_{\text{MAP}} := \underset{\theta \in \Theta}{\text{argmax}}\ p(\theta \mid X = x)$$

In a Bayesian framework, model comparison is routinely achieved through the computation of Bayes factor. If we want an index of support for model 1 versus model 2, then we can compute:

$$K := \frac{p_1(X = x)}{p_2(X = x)} = \frac{\int_\theta\ p_1(X = x \mid \theta) p_1(\theta) d\theta}{\int_\theta\ p_2(X = x \mid \theta) p_2(\theta) d\theta}$$

Because the likelihood is integrated over the values of the parameters, this index is not sensible to the increase in the number of parameters.

1.3.2 Numerical algorithms for likelihood optimization

The parameter space Θ can be very large, and prevents a direct exhaustive search for the maximum. Instead, researchers need to rely on numerical algorithms for likelihood maximization. Many of these (the gradient descent, the Newton's method, the Nelder-Mead's method) are general minimization procedures than can be applied to any function (and may be applied to $-L$). Some others (the Expectation- Maximization related algorithms) are specifically design to maximize a likelihood function. We provide here a quick overview of these methods that are often used in practice.

Optimizing a differentiable function

The *gradient ascent* (gradient descent on $-L$) is a very naïve, but effective, algorithm that is a very good choice if L can be differentiated in Θ and if the second order derivative is impossible, or difficult, to compute. The algorithm iteratively finds values (θ_k) converging toward a local maximum by computing:

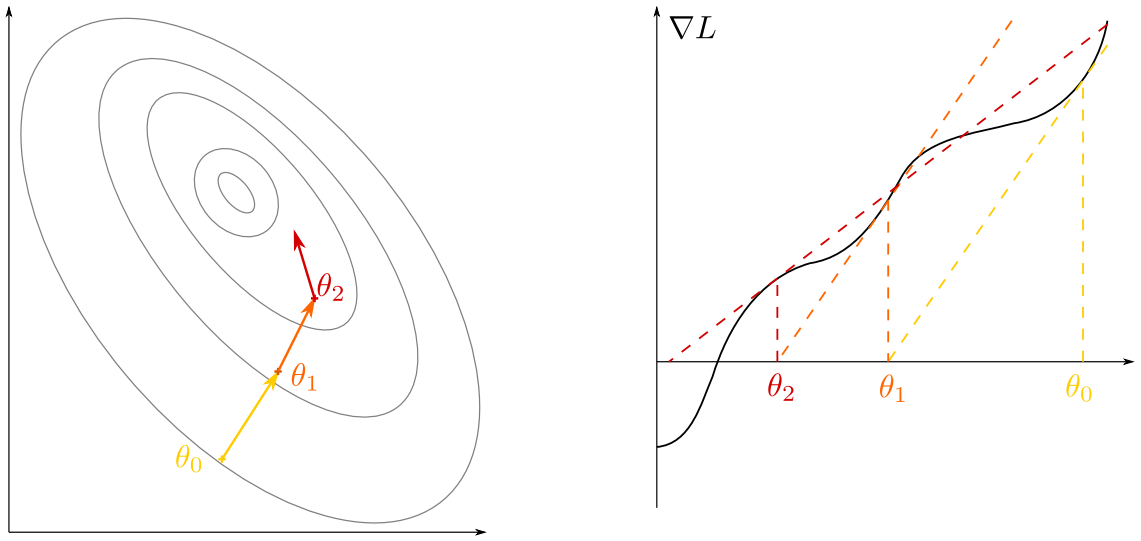
$$\theta_{k+1} := \theta_k + \alpha_k \nabla L(\theta_k)$$

The gradient ∇L evaluated in the previous point provides the direction of increase, and a step α_k is chosen in this direction, in order to increase L. This is illustrated in Figure 1.21a.

When the second order derivatives can be computed, Newton's method achieves better performances. The method consists in approaching the roots of the derivative by the zero of the tangent of the derivative. Again, values (θ_k) converging toward a local optimum are iteratively computed, using the Hessian matrix H_L:

$$\theta_{k+1} := \theta_k - H_L(\theta_k)^{-1} \nabla L(\theta_k)$$

The procedure is illustrated in Figure 1.21b.



(a) Gradient ascent.

(b) Newton's method.

Figure 1.21 – Three first iterations for (a) the gradient ascent and (b) Newton's method, two optimization procedures looking for a local maximum of L.

Optimizing a non-differentiable function

Other heuristic algorithms have been devised for cases where L cannot be differentiated. We quickly describe here the *Nelder-Mead's method*, for it is very popular, natively implemented in many programming languages, and incidentally used as a *black box* in chapter 3.

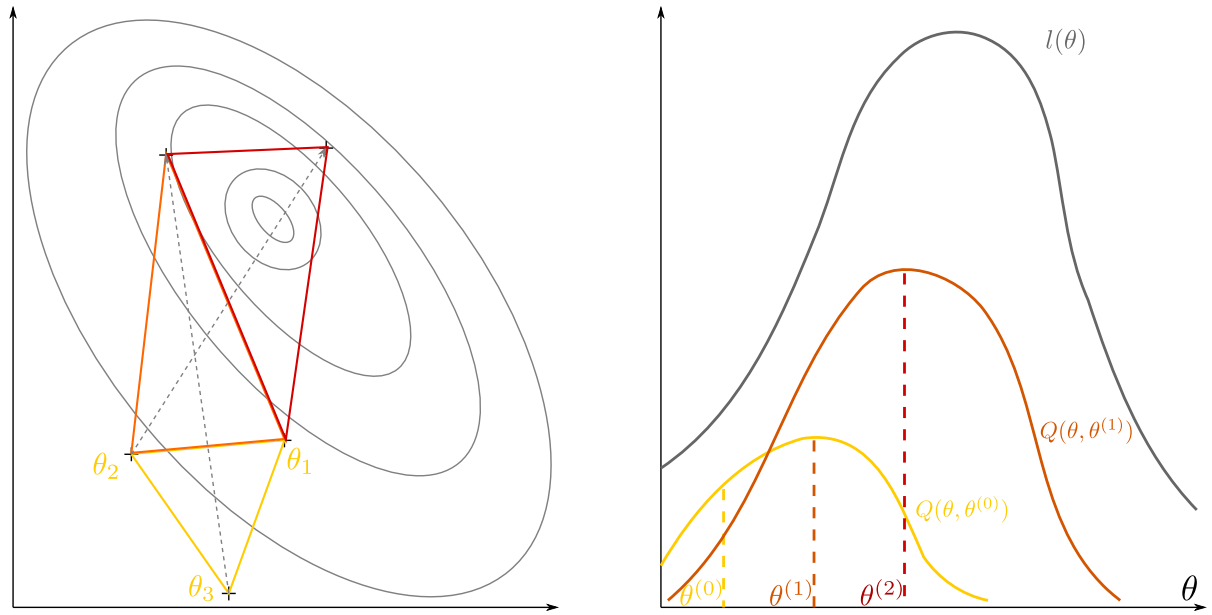
We suppose that $\Theta = \mathbb{R}^k$. Nelder-Mead's method, also called *downhill simplex method* consists in iteratively computing $k+1$ point coordinates $(\theta_1, \theta_2, ..., \theta_{k+1})$ that will move around according to heuristic rules. They are, in the end, expected to converge all to a local maximum of L. The rules are the following:

Algorithm 4 *(Nelder-Mead optimization procedure)*

1. *Reorder* $\theta_1, \theta_2, ..., \theta_{k+1}$ *to satisfy* $L(\theta_1) \geq L(\theta_2) \geq ... \geq L(\theta_{k+1})$.
   Compute the coordinates of θ_0*, the centroid of* $(\theta_1, \theta_2, ..., \theta_k)$.
2. *The reflection* $\theta^{(1)} := \theta_0 + (\theta_0 - \theta_{k+1})$ *might be a good option.*
   (a) *if* $L(\theta^{(1)}) \geq L(\theta_k)$*, then a point further in the same direction might work even better.*
   Compute $\theta^{(2)} := \theta_0 + 2(\theta_0 - \theta_{k+1})$.
      i. *if* $L(\theta^{(2)}) \geq L(\theta^{(1)})$*, replace* θ_{k+1} *with* $\theta^{(2)}$.
      ii. *otherwise, replace* θ_{k+1} *with* $\theta^{(1)}$.
   (b) *if* $L(\theta^{(1)}) \leq L(\theta_k)$*, then a point closer to* θ_0 *might work better.*
   Compute $\theta^{(3)} := \theta_0 + (\theta_0 - \theta_{k+1})/2$.
      i. *if* $L(\theta^{(3)}) \geq L(\theta_{k+1})$*, replace* θ_{k+1} *with* $\theta^{(3)}$.
      ii. *otherwise, look for a better landscape around* θ_1.
      Replace all points θ_i *except* θ_1 *with* $\theta_1 + (\theta_i - \theta_1)/2$.

   Go back to step 1.

This procedure is illustrated in Figure 1.22a.



(a) Nelder-Mead procedure.

(b) Expectation-maximization algorithm.

Figure 1.22 – Three first iterations for (a) the Nelder-Mead procedure and (b) the EM algorithm, two optimization procedures looking for a local maximum of the likelihood function.

The Expectation-Maximization algorithm

Last, we describe the *expectation-maximization (EM)* algorithm, a method specifically designed to optimize a likelihood function in problems with latent variables. Suppose that, additionally to a random variable X with observation x, there is a random variable Z which value cannot be observed (hence called latent variable). We aim at optimizing the likelihood, that can be written as:

$$\begin{aligned} l(\theta) &= \ln p(X \ ; \ \theta) \\ &= \ln \left(p(X \ ; \ \theta) \frac{p(Z \mid X \ ; \ \theta)}{p(Z \mid X \ ; \ \theta)} \right) \\ &= \ln p(X, Z \ ; \ \theta) - \ln p(Z \mid X \ ; \ \theta) \end{aligned}$$

Taking the expectation with respect to the conditional law of $Z \mid X = x \ ; \ \theta^{(0)}$, it rewrites as:

$$\sum_z p(Z = z \mid X \ ; \ \theta^{(0)}) \ln p(X \ ; \ \theta) = l(\theta) = Q(\theta, \theta^{(0)}) + H(\theta, \theta^{(0)})$$

Where expressions of Q and H are:

$$\begin{aligned} Q(\theta, \theta^{(0)}) &= \sum_z p(Z = z \mid X \ ; \ \theta^{(0)}) \ln p(X, Z = z \ ; \ \theta) \\ H(\theta, \theta^{(0)}) &= -\sum_z p(Z = z \mid X \ ; \ \theta^{(0)}) \ln p(Z = z \mid X \ ; \ \theta) \end{aligned}$$

Because Gibb's inequality ensures that, $\forall \theta, \ H(\theta^{(0)}, \theta^{(0)}) \leq H(\theta, \theta^{(0)})$, it results that, $\forall \theta$,

$$l(\theta) - l(\theta^{(0)}) \ \geq \ Q(\theta, \theta^{(0)}) - Q(\theta^{(0)}, \theta^{(0)})$$

Choosing a new value θ such that $Q(\theta, \theta^{(0)}) - Q(\theta^{(0)}, \theta^{(0)}) \geq 0$, thus leads to an increase in likelihood. The EM algorithm consists in specifically taking the argmax, over θ, of $Q(\theta, \theta^{(0)})$. It is summarized as follows.

Algorithm 5 *(EM algorithm)*

1. *initialize a first guess* $\theta^{(0)}$ *of the parameters.*
2. *while the* $\theta^{(i)}$ *have not converged, at step e perform:*
   E step Compute, $\forall z,\ p(Z = z \mid X\ ;\ \theta^{(e)})$.
   M step Optimize Q *over* θ*, i.e fix*

$$\theta^{(e+1)} := \underset{\theta}{argmax}\ Q(\theta, \theta^{(e)})$$

The EM algorithm is well suited when the M step can be performed quickly, for example when the maximum joint likelihood estimators (considering the joint law of X and Z) are obtained analytically.

All these optimization procedures nevertheless share the same inherent issue: they are sensible to local optima in the likelihood function. There are two classical ways to tackle this issue. The first one consists in trying multiple distinct initializations, to see if some of them can lead to a local maximum above others. The second way consists in adding some stochastic noise to the algorithm, in order to favour potential jumps in a part of the likelihood landscape leading to the global maximum.

1.3.3 Numerical algorithms for probability sampling

In many situations, modelers need to be able to numerically sample values of a random variable X having a probability law $\mathcal{L}_X$. We expose here three recurrent numerical algorithms used in practice in the field.

Inverse transform sampling

In some tractable enough situations, the cumulative distribution function F_X of a random variable X taking values in $\mathbb{R}$ might be analytically obtained and invertible, in closed form. We can then denote $\forall u \in (0,1),\ F_X^{-1}(u) := \inf\{x \in \mathbb{R} :\ F_X(x) \geq u\}$. When these conditions are met, if $U \sim \mathcal{U}(0,1)$, then $F_X^{-1}(U) \sim \mathcal{L}_X$. Indeed, we have:

$$\mathbb{P}(F_X^{-1}(U) \leq x) = \mathbb{P}(U \leq F_X(x)) = F_X(x)$$

One can thus draw a realisation of X by first drawing a realisation u of U, before computing $F_X^{-1}(u)$. This procedure is called *inverse transform sampling*. It is illustrated in Figure 1.23a.

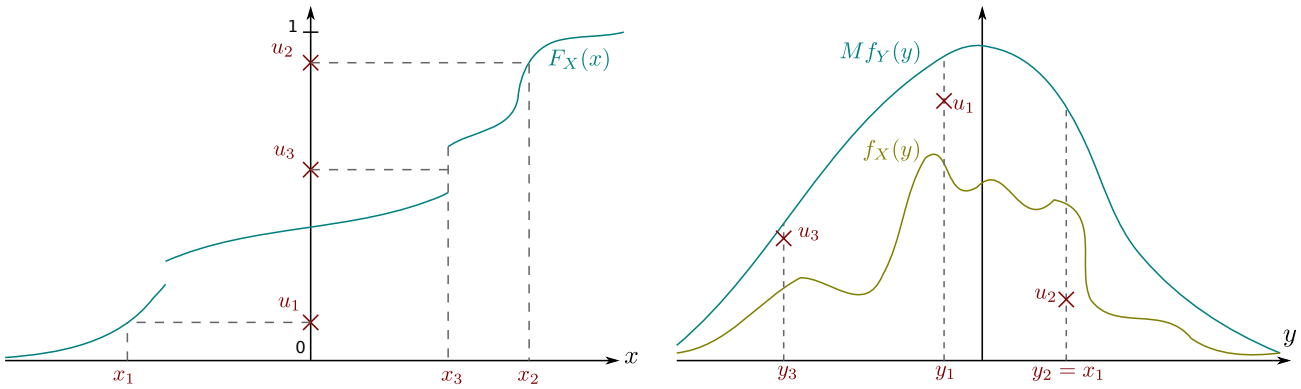
It is directly useful to sample, for example, a particular CPP law, as described in section 1.1.4. Provided we know the cumulative distribution function F_H of the node depths, one can simulate a CPP of height $t_{\max}$ by iteratively simulating uniform draws u_i on $(0,1)$, before computing the node heights $h_i = F_H^{-1}(u_i)$ until the first $h_i > t_{\max}$.

Rejection sampling

In other very common situations, we might not be able to directly sample the law $\mathcal{L}_X$, but we can show that X has a density f_X on a domain D, that we know how to compute. Suppose additionally that we can find a random variable Y with density f_Y defined on D, and a value $M \in \mathbb{R}$ such that $\forall x \in D,\ \ f_X(x) \leq M f_Y(x)$. Then, if $U \sim \mathcal{U}(0,1)$ is independent from Y, we can show that the law of X is the same as the law of Y conditionally on the event $U \leq f_X(Y)/(M f_Y(Y))$.

We then draw realisations of X by sequentially proposing realisations y_i, u_i of Y, U, and accepting the first y_i such that $u_i \leq f_X(y_i)/(M f_Y(y_i))$. This procedure is called *rejection sampling* and is illustrated in Figure 1.23b.

In most commonly encountered situations, the law of X can be directly expressed as the law of another variable Y conditioned on a specific event A, that may not depend on the density f_X. Observations of Y are drawn until the first time it verifies A.



(a) Inverse transform sampling.

(b) Rejection sampling.

Figure 1.23 – Two sampling procedures. In (a), the inverse of the cumulative distribution function at a random, uniformly drawn, point, is computed. In (b), a point y is drawn first according to the law of Y, and it is accepted as an observation of X with probability $f_X(y)/(Mf_Y(y))$.

Markov Chain Monte Carlo sampling

Last, we want to mention the very popular *Markov Chain Monte Carlo* techniques (MCMC), widely used to sample values in very diverse probability distributions. They consist in building a Markov chain $(X_i)_{i\geq 0}$ which equilibrium distribution is the target distribution ν of $\mathcal{L}_X$. The Markov chain can be initialized anywhere, the first steps are called *burn-in* and are erased. After a large number of steps, we assume that the chain has converged to its limiting distribution, and we sample the chain to sample from the target distribution as illustrated in Figure 1.24.

The two most popular ways to build a Markov chain with equilibrium distribution ν are the *Metropolis-Hastings* procedure and the *Gibbs sampling* procedure, that we detail a bit below.

The Metropolis-Hastings algorithm rely on the design of a movement proposal q. It is such that, $\forall x,\ q(x,\cdot)$ represents the distribution of the next proposed step, starting in state x. Moreover, q must be chosen such that there is a non-zero probability mass path between any two states, and such that, if $q(x,x')=0$, then $q(x',x)=0$. The algorithm is then the following:

Algorithm 6 *(Metropolis-Hastings MCMC)*

1. *Initialize a first state x_0.*
2. *At step i, the chain being in state x_i,*
   (a) *Propose a next state y_{i+1} by drawing a realisation in distribution $q(x_i,\cdot)$.*
   (b) *Compute the ratio:*

$$r(x_i, y_{i+1}) := \frac{\nu(y_{i+1})q(y_{i+1},x_i)}{\nu(x_i)q(x_i,y_{i+1})}$$

   i. *if $r \geq 1$, then $x_{i+1} := y_{i+1}$.*
   ii. *otherwise, draw a realisation u of $\mathcal{U}(0,1)$.*
   if $u \leq r$, then $x_{i+1} := y_{i+1}$.
   otherwise, $x_{i+1} := x_i$.

This procedure ensures that the Markov Chain is reversible (see definition 14), because we can verify that:

$$\forall x, x',\quad \nu(x)q(x,x')\min(1, r(x,x')) = \nu(x')q(x',x)\min(1, r(x',x))$$

It results that ν is the stationary distribution of the chain. The conditions that we imposed on q additionally ensure that it is ergodic, and that the distribution of the chain converges toward ν.

The art of building MCMC with the Metropolis-Hastings procedure consists in choosing, or *tuning* an appropriate movement proposal q, ensuring that the convergence empirically happens fast enough.

The second popular way to build a MCMC is called *Gibbs sampling* procedure, and is preferred when X can be decomposed into distinct random variables $X = \left(X^{(1)}, X^{(2)}, ..., X^{(k)}\right)$ and when the conditional distribution of each $X^{(i)}$ knowing all others can be sampled, i.e. we can express and sample $\forall i,\ p\left(X^{(i)} \mid \left(X^{(j)}\right)_{j\neq i}\right)$. The algorithm is detailed below:

Algorithm 7 *(MCMC by Gibbs sampling)*

1. *Initialize the chain in a first state* x_0.
2. *At step* n*, the chain being in state* $x_n = (x_n^{(1)}, x_n^{(2)}, ..., x_n^{(k)})$,
   (a) *Draw a realisation* i *uniformly on* $\{1, 2, ..., k\}$.
   (b) *Draw a realisation* $x_{n+1}^{(i)}$ *in the conditional law* $p\left(X^{(i)} \mid \left(X^{(j)}\right)_{j\neq i} = \left(x_n^{(j)}\right)_{j\neq i}\right)$.
   (c) *Fix* $x_{n+1} := \left(x_n^{(1)}, x_n^{(2)}, ..., x_n^{(i-1)}, x_{n+1}^{(i)}, x_n^{(i+1)}, ..., x_n^{(k)}\right)$.

The Markov Chain defined through this procedure is, again, reversible. Indeed, if x, y are such that $x^{(i)} = y^{(i)}$ and $\forall j \neq i, x^{(j)} = y^{(j)}$, then

$$\begin{aligned}
\nu_x p_{xy} &= \frac{1}{k}\nu_x p\left(X^{(i)} = y^{(i)} \mid \left(X^{(j)}\right)_{j\neq i} = (y^{(j)})_{j\neq i}\right) \propto \frac{1}{k}\nu_x\nu_y \\
\nu_y p_{yx} &= \frac{1}{k}\nu_y p\left(X^{(i)} = x^{(i)} \mid \left(X^{(j)}\right)_{j\neq i} = (x^{(j)})_{j\neq i}\right) \propto \frac{1}{k}\nu_x\nu_y \\
\Rightarrow \nu_x p_{xy} &= \nu_y p_{yx}
\end{aligned}$$

Otherwise, we also have $\nu_x p_{xy} = 0 = \nu_y p_{yx}$.

It results that ν is the stationary distribution of the chain. It will also be the asymptotic distribution of the chain, empirically considered to be attained after a large amount of steps.
In practice, most people replace the random uniform choice of component i by regularly cycling through the successive components.

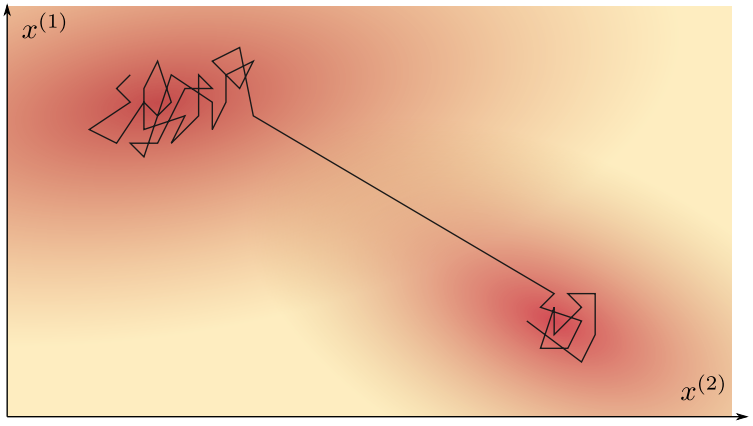


Figure 1.24 – Trajectory of a MCMC (in gray) *tuned* to sample from the likelihood landscape. Red and white zones represent respectively high and low likelihood values.

1.3.4 Monte-Carlo method for numerical integration

The sampling methods previously described are sometimes used to compute (numerical approximations of) integrals. In a Bayesian framework, one integral that might be worth computing is simply the probability of the observations:

$$p(X = x) = \int_{\theta \in \Theta} p(X = x \mid \theta)\ p(\theta) d\theta$$

In a maximum likelihood framework too, one often needs to compute integrals in order to have access to the likelihood. This happens in cases where the probability can be computed only if we know the values of some latent variables Z, distributed according to a distribution f on Ω_Z. We then have a problem very similar to the Bayesian one presented above:

$$p(X = x) = \int_{z \in \Omega_Z} p(X = x \mid Z = z) f(z) dz$$

In the rest of this section, we consider this last problem.

Monte-Carlo Principle

Monte-Carlo methods rely on the law of large numbers, allowing us to estimate expectations through empirical means. More precisely, if the series of (Z_i) are independent and identically distributed variables, distributed in the same law as Z, and if the (z_i) are observations of these, then:

$$\frac{1}{N} \sum_{i=1}^{N} p(X = x \mid Z_i = z_i) \quad \underset{N \to \infty}{\longrightarrow} \quad \mathbb{E}_Z \left(p(X = x \mid Z) \right) = p(X = x)$$

The basic Monte-Carlo estimation thus consists in drawing a large number of realisations (z_i) in the law of Z. These are used to compute the empirical mean in the left-hand term of the previous equation. However, depending on the problem at hand, the convergence might be very slow, in which case having a good estimate of $p(X = x)$ might require a lot of simulations and numerical power.

Importance Sampling

The Importance Sampling is a refinement of Monte-Carlo methods that has been proposed as a mean to reduce the number of iterations necessary to get a precise estimate. The idea is to choose another distribution g on Ω_Z, and to rewrite the integral as:

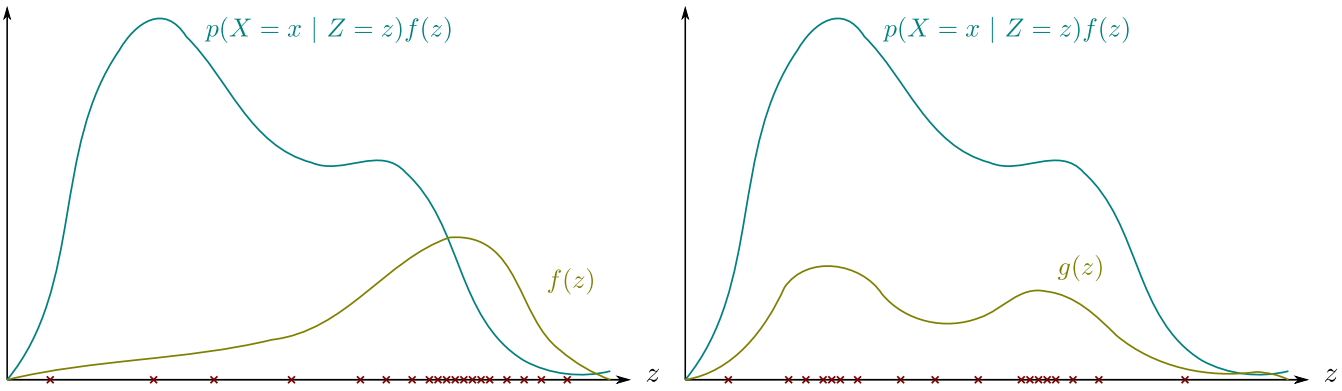
$$p(X = x) = \int_{\Omega_Z} \frac{p(X = x \mid Z = z)\ f(z)}{g(z)}\ g(z) dz$$

If the (z_i) are now observations drawn in distribution g, then the empirical mean one needs to compute is now:

$$\frac{1}{N} \sum_{i=1}^{N} \frac{p(X = x \mid Z_i = z_i)\ f(z_i)}{g(z_i)}$$

This estimate has a variance that can be lower than the estimate of a *basic* Monte-Carlo approach. The idea is to choose g to mimic $p(X = x \mid Z) f(Z)$, i.e. to put preferentially weight in areas of Ω_Z such that $p(X = x \mid Z) f(Z)$ is non-negligible (see Fig. 1.25b).

In practice, this boils down to *tuning* an appropriate probability distribution g on Ω_Z such that the empirical variance of the estimate is lowered.



(a) Monte-Carlo integration.

(b) Importance Sampling integration.

Figure 1.25 – Two stochastic methods for integrating the area under the blue curve. In (a), samples are drawn according to density f. In (b), samples are drawn according to a density g, tuned to sample more frequently areas with the most weight. Red crosses on the axis represent sampled points.

Integration with a MCMC sampler

The MCMC samplers described in the previous section can be used as a way to compute numerical approximations of integrals too. In contrast to the two previous approaches, successive samples of a MCMC are not independent, thus preventing the direct application of the law of large numbers.

Fortunately, a very analogous result can be stated for MCMC samples, known under the name of *ergodic theorem*. If (Z_i) is a Markov chain that has been tuned to have limiting distribution with density f, then we have as before:

$$\frac{1}{N}\sum_{i=1}^{N} p(X = x \mid Z = Z_i) \underset{N\to\infty}{\longrightarrow} \int_{\Omega_Z} p(X = x \mid Z = z)f(z)dz$$

In practice, people usually discard the beginning of the chain (called burn-in) to reduce the sensibility to the initial condition. They further sample only a fraction of the remaining chain, to reduce the auto-correlation.

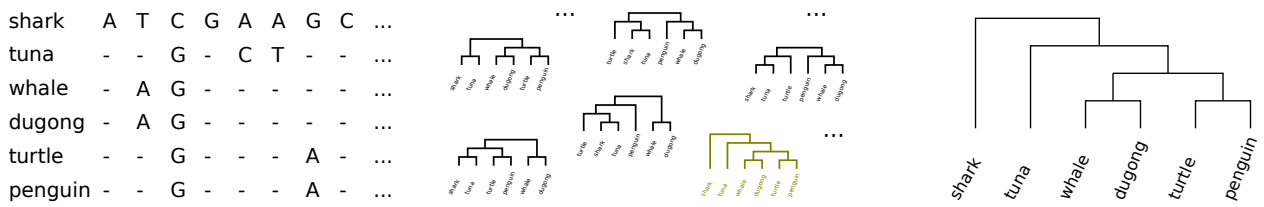
In summary, all tools briefly described in this section allow one to fit models described in sections 1.1.4 and 1.2 to empirical data. In the last section, we present applications of our first three sections to empirical data and discuss some inferences that can be made about the past history of organisms.

1.4 Applications to biological data

We finally present examples of empirical applications of both diversification and trait evolution models. These applications are divided in two parts: when the phylogeny linking organisms is considered as unknown on the one hand, and when the phylogeny is known on the second hand.

1.4.1 Tree reconstruction

In this section, we consider that we are given n extant organisms, for which we would like to estimate a phylogeny. The most important data that people have at hand are the alignment of a DNA fragment, of length, say m nucleotides, that is present, with some differences, in our n organisms. We would like to make use of these differences to reconstruct somehow *reasonable* ancestral relationships. The literature proposes three main interpretations of the term *reasonable*, that we sketch here.



(a) Data. (b) Tree space exploration. (c) Result.

Figure 1.26 – General tree reconstruction procedure. In (a) we consider that we are given a perfect alignment of m nucleotides. Each position is present in all species, and they are supposed to derive from the same nucleotide in a common ancestor. In (b) the tree space is explored, and a criterion of *reasonability* is assessed on each tree. In (c) the reconstruction either outputs one tree, or a distribution of *reasonable* trees.

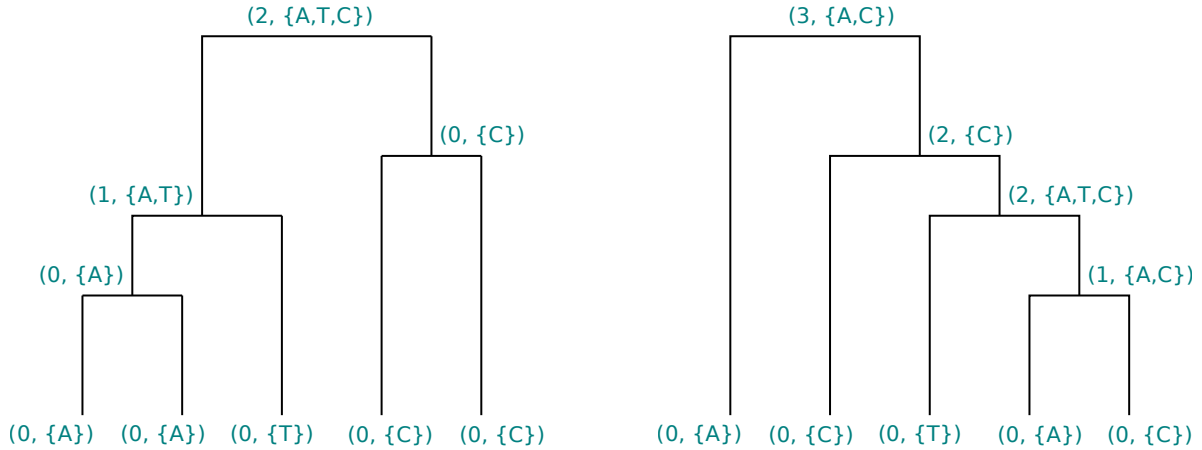
Maximum parsimony

The first historical attempt to answer this question has made use of the *parsimony* principle. According to this view, the most reasonable scenario of phylogeny is the one that minimizes the total number of substitutions needed to explain the present day data.

Given a fixed binary tree and the nucleotide state at one position, at the tips of the tree, an elegant algorithm (due to Fitch (1971), but see also Sankoff (1975) for an extension) allows us to compute the minimal number of substitutions that happened along the tree. It relies on a depth-first traversal in post-order, i.e. it traverses all child vertices before their parent. At each vertex v, it computes:

$s_v :=$ minimal number of substitutions needed to explain states at tips descending from v
$S_v :=$ nucleotide states at vertex v compatible with s_v substitutions.

Here is the principle, illustrated in Figure 1.27 and detailed in algorithm 8:



(a) On a first tree. (b) On a second tree.

Figure 1.27 – Fitch algorithm for one site, on two distinct trees. Note that the tree represented in (a) is more parsimonious than in (b), at least with respect to this site's data.

Algorithm 8 *(Fitch algorithm)*
Recursively compute, at each vertex v of the tree, until reaching the root:

1. *if v is a tip, then $s_v := 0$ and $S_v :=$ observed nucleotide state at tip v.*

2. *otherwise, let* v_1 *and* v_2 *be the two children of* v.
   Compute (s_{v_1}, S_{v_1}) *and* (s_{v_2}, S_{v_2}).
   (a) *if* $S_{v_1} \cap S_{v_2} \neq \emptyset$ *then*
   $s_v := s_{v_1} + s_{v_2}$
   $S_v := S_{v_1} \cap S_{v_2}$
   (b) *otherwise,*
   $s_v := s_{v_1} + s_{v_2} + 1$
   $S_v := S_{v_1} \cup S_{v_2}$

For each position i of the alignment, the minimal number of substitutions needed to explain it is obtained at the root of T and called $s_T^{(i)}$. The total *parsimony score* associated to the tree is then $\sum_{i=1}^{m} s_T^{(i)}$. The maximum parsimony tree reconstruction is defined as:

$$\hat{T}_{\text{pars}} := \underset{T}{\text{argmin}} \sum_{i=1}^{m} s_T^{(i)} \tag{1.5}$$

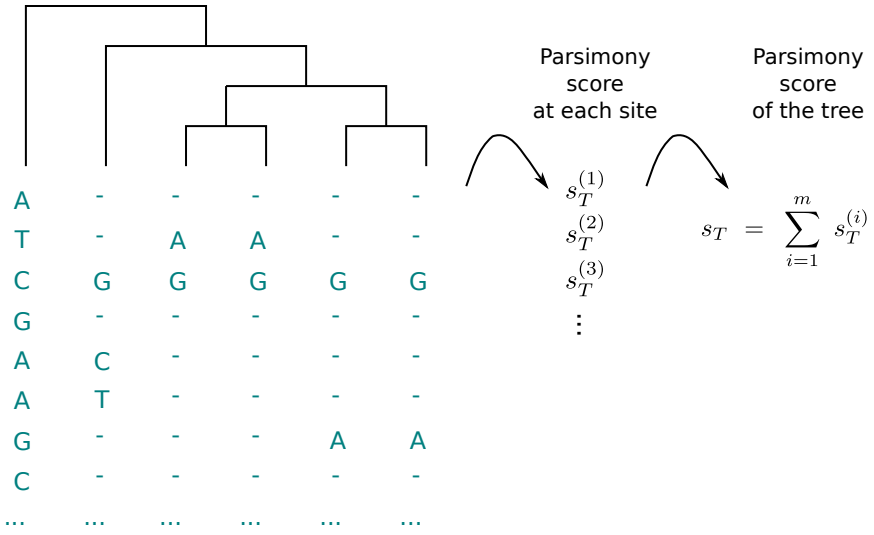


Figure 1.28 – Tree assessment by parsimony. For each tree, the parsimony score must be assessed, in order to look for the maximum parsimony tree.

Finding the trees minimizing this parsimony score still requires to explore all possible trees. Because exhaustive search is too computationally intensive, we have to rely on heuristics to explore the space of all trees through successive modifications of a first tree. As this would open a way too long digression, we refer the interested readers to the paragraphs on Nearest Neighbor Interchange, Subtree Pruning and Regrafting, or Tree Bissection and Reconnection, in molecular evolution textbooks such as Perrière and Brochier-Armanet (2010).

Maximum likelihood

After Felsenstein (1978) showed that parsimony reconstructions could, under some conditions on the generating model of nucleotide change, not converge to the true tree, alternative tree reconstruction methods based on the maximum likelihood principle have been proposed.

These methods rely on a given model of nucleotide evolution, such as the ones discussed around Figure 1.17. Two categories of methods can then be highlighted.

The first category considers that we could make use of the whole data. In order to do so we compute, at each position i of the alignment, the probability $p_T^{(i)}$ of the observed tip states on tree T. This is done using the previously described *pruning algorithm* 3. The log likelihood of the whole alignment is then obtained as $\sum_{i=1}^{m} \ln p_T^{(i)}$. In the maximum likelihood framework, the tree is seen as a parameter, which can be estimated through its maximum likelihood estimator. It thus remains to search for the tree maximizing the likelihood:

$$\hat{T}_{\text{mle}} := \underset{T}{\text{argmax}} \sum_{i=1}^{m} \ln p_T^{(i)} \tag{1.6}$$

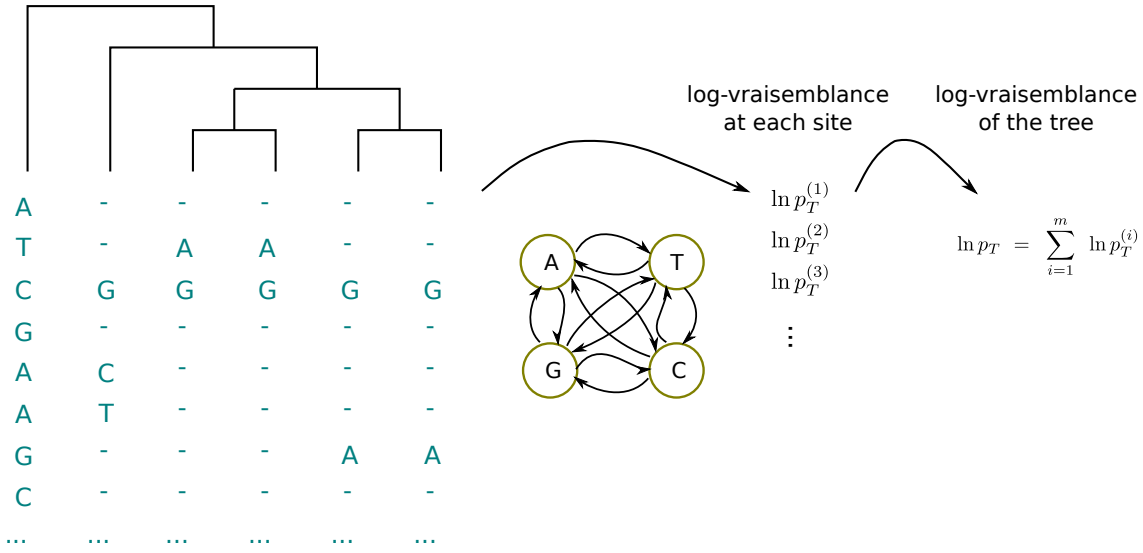


Figure 1.29 – Tree assessment by likelihood. The likelihood of each tree must be assessed, in order to look for the maximum likelihood tree. A model of molecular evolution, illustrated here with the diagram giving transitions between nucleotides, need to be assumed.

Coming right after the paragraph on tree reconstruction by parsimony, there is an evident similarity between the two approaches. The pruning algorithm is somehow analogous to the Fitch algorithm. Both use the tip data and require a depth traversal of the tree to compute a score at the root. The last step consists in exploring the space of trees to optimize a score (either parsimony or likelihood).

However, in the maximum likelihood framework, we aim to explore the space of all dated trees, whereas we were exploring non-dated trees before. In practice, this is done through two nested explorations. Non-dated trees are explored thanks to heuristics. For each non-dated trees, the branch-lengths are then estimated by their maximum likelihood estimator, using a likelihood optimization procedure like Newton's method or EM (section 1.3.2).

The second category of maximum likelihood reconstruction approaches considers that the previously proposed procedure is too computationally intensive. In order to lower the time needed to estimate the tree, the data is simplified in a first phase and treated in a second phase.

Only the first phase makes use of the model assumption and of the maximum likelihood principle. It consists in estimating the number of substitutions that happened between any two taxa. The resulting matrix is a distance matrix, which can be used as data for the tree reconstruction.

The second phase takes as input a pairwise distance matrix (d_{ij}). It then looks for a binary rooted X-tree T which would induce a distance d_T on X *close to* the pairwise distances. Distinct interpretations of the term *close* have been proposed, but we provide here only one to fix our mind:

$$\hat{T}_{\text{dist}} := \underset{T}{\text{argmin}} \sum_{i<j} (d_{ij} - d_T(i,j))^2 \tag{1.7}$$

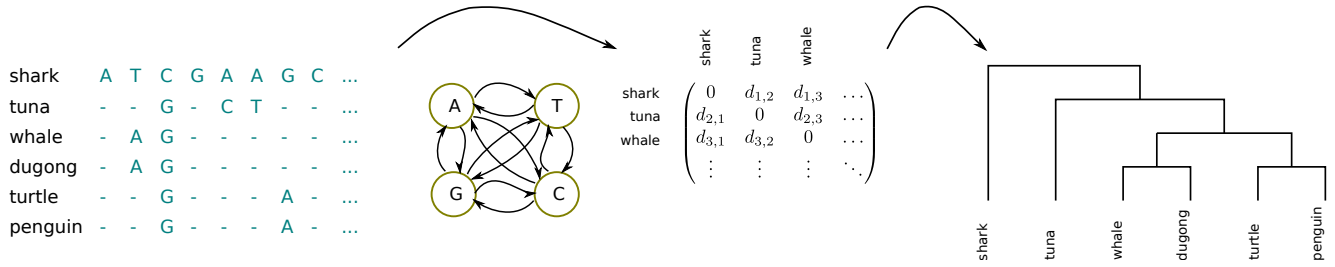


Figure 1.30 – Tree reconstruction by *distance method*. A model of molecular evolution is used to infer the matrix of pairwise distances between any two taxa. It remains to look for a tree inducing distances between taxa that would be close enough to this matrix.

In practice, as for the previous maximum likelihood problem (1.6), two nested explorations can be performed. Non-dated trees are proposed thanks to heuristics, and for each non-dated tree, the branch lengths are proposed so as to induce a tree distance close enough to the pairwise distances. Other algorithms have been proposed to approximate the solution to problem (1.7) by adding taxa one by one. More details about these approximations can be found in Perrière and Brochier-Armanet (2010).

Bayesian framework

The increase of computational power available, together with the increase in complexity of models of nucleotide change, led to the rise of tree reconstructions in a Bayesian framework. So-called Bayesian tree reconstructions rely on a different paradigm. Whereas the tree was considered to be a parameter in the previous maximum likelihood approach, it will now be considered to be a random variable distributed according to an *a priori* law.

The Bayesian framework thus draws a link between probabilistic models for tree generation presented in section 1.1.4 and models of nucleotide change on a fixed tree presented in section 1.2. The first allows us to assume the prior tree distribution $p(T)$ while the second provides the conditional distribution for the data X knowing T : $p(X \mid T)$. Together, they allow us to derive the target posterior distribution knowing the data x:

$$p(T \mid X = x) = \frac{p(X = x \mid T)p(T)}{\int_t p(X = x \mid T = t)p(T = t)dt}$$

This posterior tree distribution is numerically sampled using MCMC approaches (as described in section 1.3.3). This approach is thus particularly computationally intensive, as it requires to evaluate a very large number of times the numerator of the previous equation.

However, the result of a Bayesian tree reconstruction is more easily interpretable, as it gives direct access to the tree distribution. In practice, a fixed size collection of trees (t_i) such that $p(T = t_i \mid X = x)$ is high enough is retained, providing an idea of the uncertainty of the tree estimation procedure.

In contrast, previously described approaches provided only a point estimate of the underlying tree. More work was needed to get an idea of the uncertainty of these procedures. One commonly used approach consists in subsampling the data, i.e. reconstructing the tree for distinct subsets of taxa, to assess the robustness of the reconstruction. More details are provided, again, in the textbook by Perrière and Brochier-Armanet (2010).

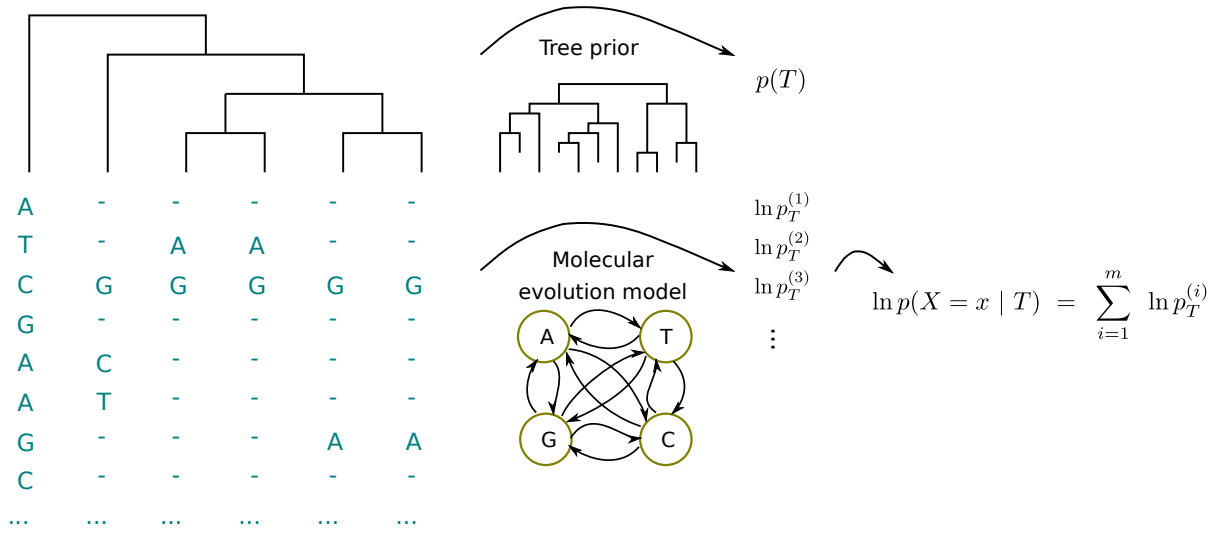


Figure 1.31 – Tree reconstruction in a Bayesian framework. A model of molecular evolution and a model of tree must be assumed. The posterior probability is computed using both. Note that the framework can incorporate more than 2 model *layers*, with for example a prior on trees, a model of molecular rate evolution knowing the tree, and the molecular evolution knowing the tree and the rate.

The Bayesian and maximum likelihood frameworks that we sketched and opposed here are in fact sometimes difficult to distinguish so easily. The prior probability distribution and the conditional distribution may rely on other parameters, that one might want to infer. For example, the prior probability distribution on trees might be given by a birth-death process with birth rate b and death rate d (see section 1.1.4). The conditional probability distribution of the tip nucleotide state knowing a fixed tree might be given by a Kimura model with transition over transversion ratio α (see section 1.2.1). These three parameters can in turn be estimated in a maximum likelihood or Bayesian fashion. In a completely Bayesian fashion, one needs to assume a prior distribution on these parameters too, and eventually a prior distribution on the parameters of the prior distribution of the first generation of parameters, and so on until, hopefully, assuming parameter-free distributions at some point.

All methods discussed above allow us to get estimates of the phylogeny of organisms. These reconstructed phylogenies are considered as data by another, partly disjoint, community of researchers in macroevolution, interested in studying the diversification, or the phenotypic evolution, of these organisms. We present here an overview of some questions and models for diversification study first, then for phenotypic evolution study, and last for the joint study of diversification and trait evolution.

1.4.2 Diversification studies

Diversification studies are interested in the relative tempo at which distinct species appear and go extinct. We attempt here to present a condensed overview of the models that have been fit to real data. For more on this topic, we refer to reviews by Stadler (2013) and Morlon (2014).

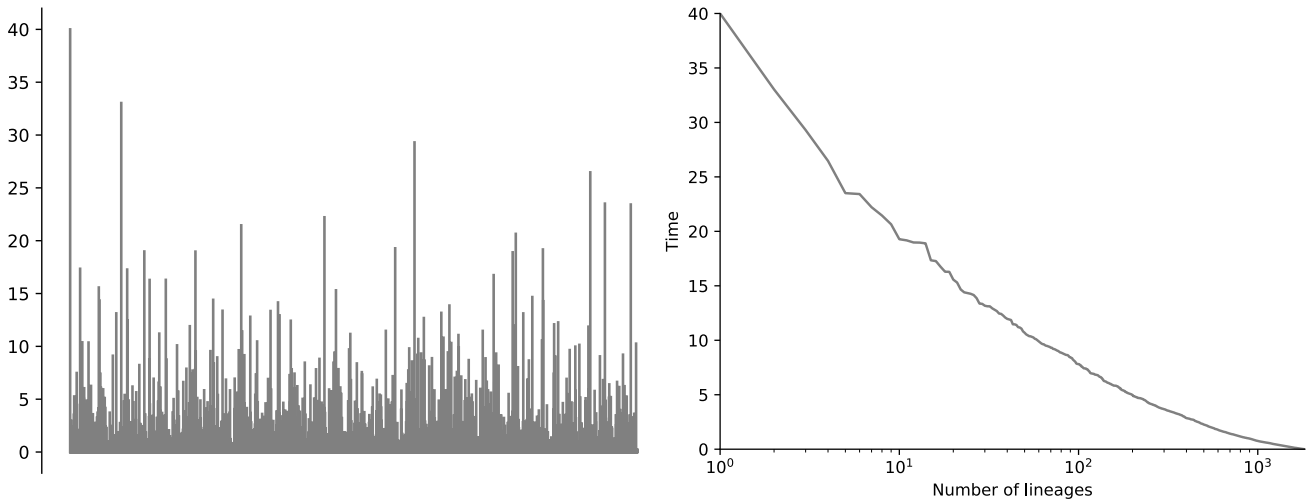
Information in branching times and imbalance

Most commonly fitted models in this research area are lineage-based branching processes. We described in section 1.1.4 the most simple ones, namely pure constant-rate birth process, and birth-death processes with constant birth and death rates. These models have been fitted and compared to empirical

data early in the history of diversification studies (Nee et al., 1994). Two interesting results emerged:

1. They do not fit well the observed patterns of tree imbalance. More precisely, they lead to trees more balanced than empirical ones.

2. They produce a distribution of branching times different from the one found on empirical phylogenies. In particular, a non-zero constant death rate leads to a phenomenon called *pull of the present* (see Fig. 1.32). The logarithm of the number of lineages in the reconstructed phylogeny increases linearly in the past, and accelerates near the present. In contrast, empirical phylogenies tend to be less *tippy*, i.e. their nodes are preferentially distributed near the root of the tree.



(a) Birth-death tree.

(b) Lineage Through Time (LTT) plot.

Figure 1.32 – Comb representation of a reconstructed tree simulated under a birth-death process with birth rate $b = 1$ and death rate $d = 0.8$ (a) and its associated lineage through time (LTT) plot (b). The logarithm of the number of lineages in the reconstructed tree first increases linearly, before accelerating near the present.

Rate heterogeneities among lineages

The imbalance of empirical phylogenies, as compared to the (constant birth and death rate) null model, has been quickly explained by heterogeneities in rates in distinct parts of the tree. If the place in the tree where the change of rates happens is considered to be known, the problem consists in estimating four parameters, say (b_1, d_1) in some part of the tree and (b_2, d_2) in a subtree. The likelihood formula presented in section 1.1.4 can be readily adapted.

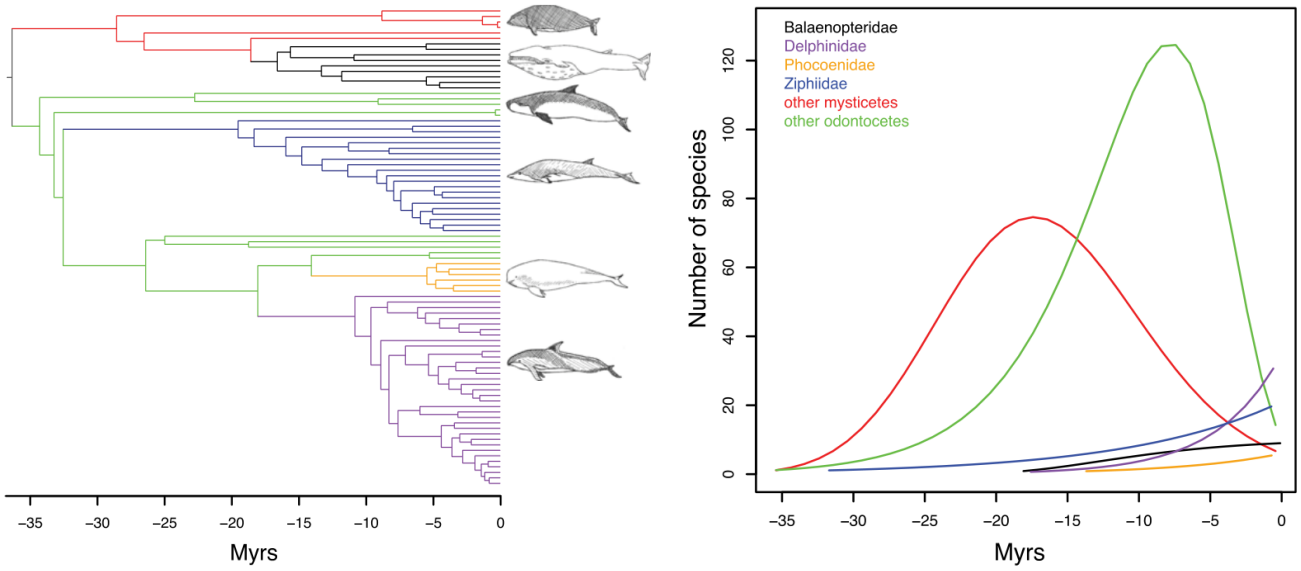
However, the problem is much more complex if we do not fix a priori the points at which rates shift. The method should then be able to retrieve the most likely rate values, together with the shift points. One approach that has been proposed to tackle this issue consists in sequentially proposing potential shifts on the phylogeny, and computing a penalized likelihood to assess the support for the newly proposed shift. The procedure is repeated until no shift can increase the penalized likelihood anymore. This approach has been first proposed by Alfaro et al. (2009), and has allowed them to spot 6 main accelerations, called *radiations* on the Gnathostome tree, with for example notable ones at the root of the clade comprising coral fishes, or at the root of eutherian mammals. In a Bayesian framework, the popular and recently much debated *BAMM* package has been proposed to tackle the exact same issue (Rabosky, 2014).

Rate heterogeneity through time

Other extensions have been proposed in the literature to bring the distribution of branching times closer in agreement with empirical patterns. In order to repress the pull of the present phenomenon and display more profound nodes, birth-death processes with non-constant rates through time have been considered. The exact likelihood formula of a tree can still be derived for time-varying rates, but it may be intensive to assess numerically due to the need to compute integrals.

Stadler (2011) proposed that rates might be piece-wise constant, with some shifts in time that could be inferred from the data. Applying the method to the mammalian phylogeny, they could not find a significant shift at the Cretaceous-Tertiary boundary (65 Mya). The first increase in the mammalian diversification rate seemed to occur instead around 30 Mya.

Morlon et al. (2011a) proposed that rates might vary as linear or exponential functions of time. Combining this with the *a priori* placement of shifts on a tree, they were able to take into account scenarios of *waxing-waning* of diversity through time. They fitted the model to the cetacean phylogeny, and inferred from the phylogeny of extant cetaceans only a scenario in agreement with the known fossil record (see Fig. 1.33). Overall, models including a decreasing birth rate near the present fit much better empirical phylogenies.



(a) Cetacean phylogeny.

(b) Number of species through time.

Figure 1.33 – Fit of a diversification model with heterogeneity of rate both among clade (distinct colors on the phylogeny) and through time, in Morlon et al. (2011a).

Other authors have proposed that the rates could depend on some environmental characteristics. Diversity-dependent models have been proposed, where the birth rate would decrease with the number of coexisting species. Rabosky and Lovette (2008) fitted such a model to north American wood warblers, suggesting that their radiation has been diversity-dependent. Rates have been also proposed to vary with abiotic factors, such as temperature for example. Explored by Condamine et al. (2013), the fit of this model on the cetacean tree suggests a positive dependence between speciation rate and temperature.

Individual-based models

Last, a few individual-based models have been developed to study diversification, following the rise of the Neutral Theory of Biodiversity (NTB) (Hubbell, 2001). By explicitly modeling the fate of individuals, they allow researchers to incorporate more realistic processes, for instance taking into account

available geographic space (Pigot et al., 2010). However, these models have generally proved much less handy to fit, thus preventing their spread in the community.

Because they explicitly take into account the individual level, these models need to make strong hypotheses on the way different species can emerge from the individual level, and on the way the phylogeny can emerge from the knowledge of the genealogy of individuals. Hypotheses considered in the literature will be discussed more in-depth in chapter 2, before presenting two alternatives.

These models can be compared to real data. Patterns of branching times and patterns of imbalance that they can produce in simulations are compared to empirical ones. Davies et al. (2011) showed that the popular species definitions considered in NTB, like the speciation by *point mutation*, where each mutation happening on an individual leads to a new species, does produce phylogenies exhibiting the right level of imbalance. However, those phylogenies are way too *tippy* as compared to real ones, i.e. the branching times occur massively near the present.

This has led to the emergence of another model of speciation, called *incipient speciation*, aimed at reconciling the patterns of branching times with empirical phylogenies. In this model, mutations arising at the individual scale lead to new *incipent species*, which will become true species only after a fixed or random time span (Rosindell et al., 2010).

We propose in this thesis another model of speciation. It is based on completely distinct hypotheses at the individual level, and yet, also produces patterns of imbalance and branching times observed in empirical data. Using one of the two species definitions that naturally emerge in chapter 2, we present in chapter 3 a comparison of the shape of simulated phylogenies with the shape of empirical ones. We also provide a method to fit our model to empirical phylogenies.

1.4.3 Phenotypic evolution studies

Phylogenetic signal of distinct traits

In the field of trait evolution, the Brownian motion has played the role of a null model analogous to the constant-rate birth-death model in diversification. Originally introduced in Felsenstein (1973) to use continuous traits in maximum likelihood tree reconstruction procedures, it has quickly become a standard model to apprehend trait evolution on a fixed tree, as well as a null model against which to compare more sophisticated ones.

In particular, one important characteristic of models of continuous evolution concerns their ability to produce the right level of *phylogenetic signal* in traits. Traits are said to exhibit phylogenetic signal if, loosely speaking, their present-day correlation reflects the phylogenetic tree structure. The OU process, proposed as an alternative to the BM by Hansen (1997) and already described in section 1.2.2, tends to lower the phylogenetic signal in the past as compared to the BM. Support for one or another model can be assessed in model selection studies (Blomberg, 2017; Harmon et al., 2010). Figure 1.34 presents an example of such a model comparison on traits related to body size in distinct groups. No clear, general, picture emerged from these studies. In particular, the precise knowledge of the taxa and the type of trait does not seem to predict well if the BM or OU will show the best fit. However, this type of model comparison is still performed on a case-by-case basis, leading to distinct biological interpretations. Traits best fitted by an OU are generally interpreted as being selectively stabilized around an optimal value assumed to be quite stable, whereas traits best fitted by a BM are interpreted as more labile, or as selectively stabilized around an ever-moving optimum.

Multivariate trait evolution

Most studies on trait evolution are interested in the covariation of distinct traits evolving jointly along the same tree. All of these could be continuous, in which case the vector of traits is supposed to evolve as a multivariate diffusion process. Or it could be a mix of discrete and continuous traits, in which

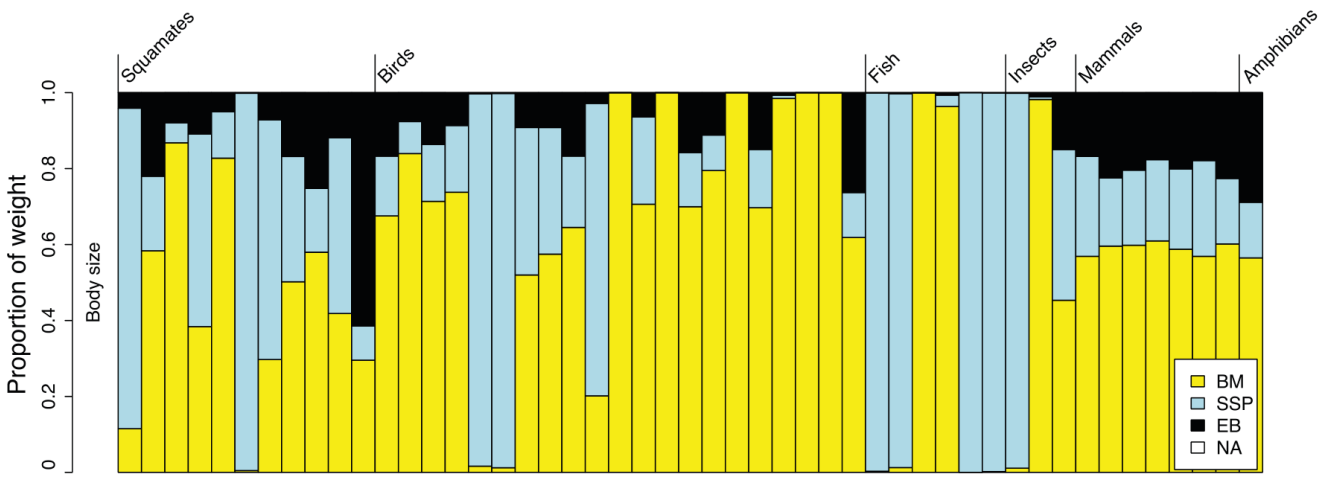


Figure 1.34 – Relative support for three models of trait evolution (from Harmon et al. (2010)). SSP stands for Single Stationary Peak and corresponds to the Ornstein-Uhlenbeck model. BM, OU and EB have been presented in section 1.2.2.

case discrete traits evolve according to a Markov process unfolding on the tree, and the evolution of the continuous traits knowing the discrete trait value is placed on top of the discrete trait trajectory.

All these traits at present are more strongly correlated among closely related species either (i) because they evolve on the same tree or (ii) because they further evolved in a correlated fashion throughout the clade history. The relative fit of models describing only (i) v.s. both (i) and (ii) can be assessed to infer the most likely scenario. Using this type of approach, Thomas et al. (2006) for example showed that the evolution of a precocial development (seen as a discrete trait) in birds leads to a higher rate of sexual size dimorphism.

This framework has also been used in more complex modeling studies. Lartillot and Poujol (2011) modeled for example the joint evolution of the substitution rate and body mass, or the substitution rate and the longevity, as a multivariate diffusion process. They further superimposed the evolution of nucleotides knowing the substitution rates. The resulting process, fitted in a Bayesian framework, allows the authors to demonstrate a negative correlation between substitution rates and body mass, as well as between substitution rates and longevity, on the mammal tree. This correlation can be appreciated visually on Figure 1.35, showing the inferred substitution rate along the mammal tree.

Exploring the role of biotic or abiotic factors

Analogously to what was discussed for models of diversification, rates of phenotypic evolution have been proposed to depend on abiotic factors, such as temperature, or biotic ones, such as the number of species in a clade. Clavel and Morlon (2017) compared models of BM where the diffusion coefficient σ is either a linear or exponential function of past mean temperature. The fit of the model to the observed body mass of birds and mammals supports a higher rate of body mass evolution during cold climatic periods. Weir and Mursleen (2013) proposed that the rate of evolution could be a function of the number of lineages in the reconstructed tree, as a proxy to assess diversity-dependence. They fitted the model to data on bill shape in the auks, and suggest that there is indeed a signal for a slowdown of trait evolution as diversity increases in the clade.

Punctuated trait evolution

All models described in section 1.2.2 suggest that the evolution of a continuous trait is continuous through time. This hypothesis has been challenged a lot in evolutionary biology, leading to a long standing

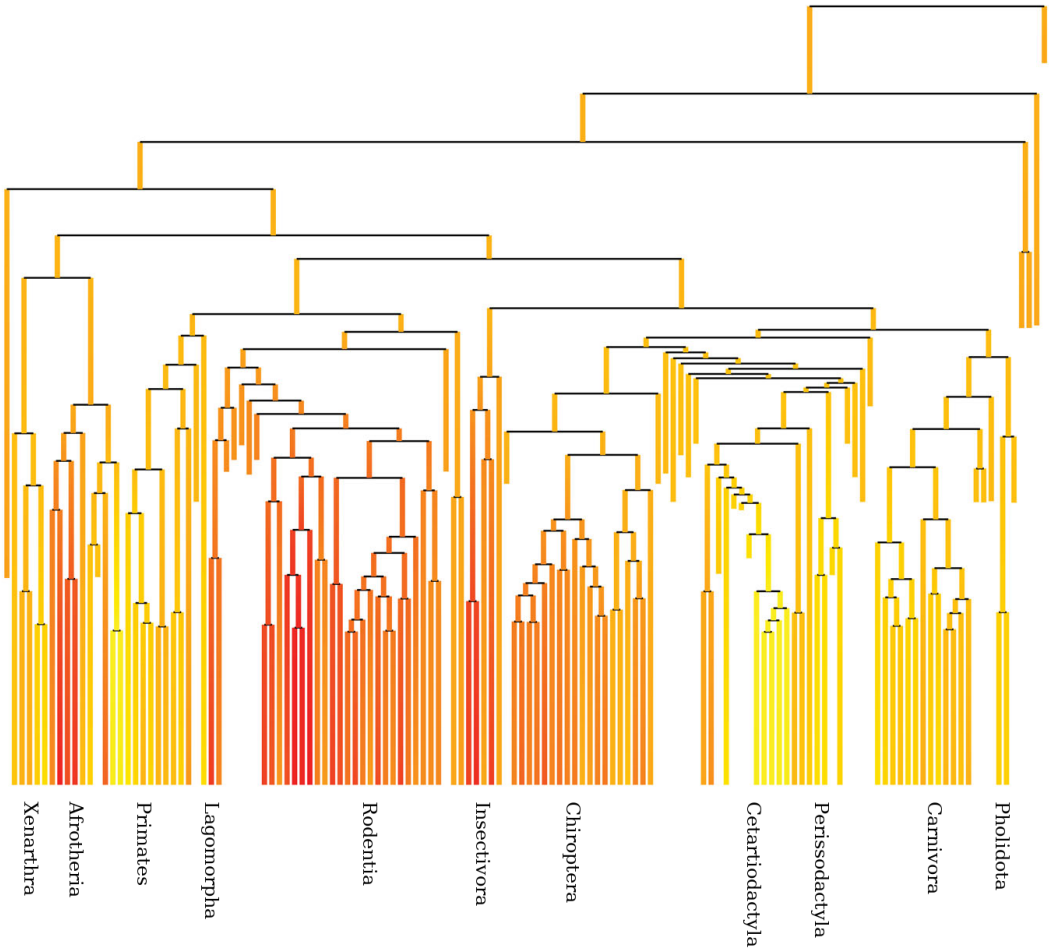


Figure 1.35 – High (red) and low (yellow) substitution rates reconstructed on the mammal tree (from Lartillot et al. (2016)). Even visually, the high rate values seem to correlate well with clades having the lowest body mass.

debate between supporters of *gradualism* and supporters of *punctualism*. This latter hypothesis proposes that traits evolve through rapid shifts, separated by long periods of stasis (Pennell et al., 2014). Two very different implementations of the same idea have been proposed in recent modeling work.

The first one proposes that traits evolve according to an OU unfolding along the tree, with rare shifts in the optimal value θ. These shifts could be then inferred in a maximum likelihood (Beaulieu et al., 2012; Bastide et al., 2016) or Bayesian (Khabbazian et al., 2016) framework. Figure 1.36 illustrates the inference of shift positions on the Chelonian tree using their body mass data (Bastide et al., 2016).

The second one proposes that traits evolve according to Levy processes. These are processes incorporating both the possibility for a continuous evolution, and for shifts occurring directly in the trait value (Bokma, 2002; Landis et al., 2013). Applying them on empirical morphological data, Uyeda et al. (2011) and Landis and Schraiber (2017) suggest that a lot of traits would actually be best fitted by such models of *punctuated equilibriums*.

All these models of trait evolution are also used to infer the ancestral states of either lineages (edges), or only common ancestors (vertices), on the phylogeny. The quantity of interest is then the probability distribution of the trait at internal vertices, knowing the tip trait values. The output is generally summarized as a colouring of the phylogeny, as is shown in Figures 1.35 and 1.36.

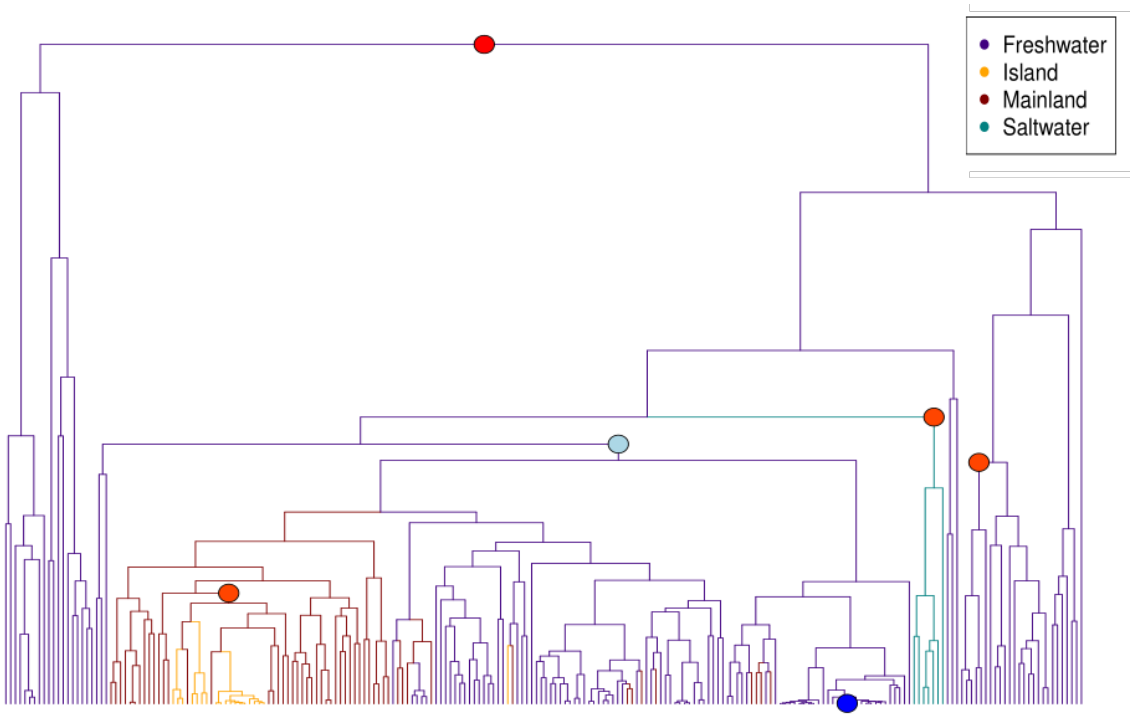


Figure 1.36 – Best fit scenario of an OU model with shifts in the *optimum parameter*, to model the body mass on the chelonian phylogeny. Red or blue color on the shifts represent respectively high or low optimal body mass. Colors on the branches represent the habitat reconstruction. Shifts for high body mass occur in particular for saltwater turtles and tortoises living on islands (from Bastide et al. (2016)).

1.4.4 Trait-dependent diversification studies

The first State-dependent Speciation and Extinction (SSE) model was introduced in Maddison et al. (2007), who proposed the Binary State Speciation and Extinction (BiSSE) model as a way to estimate the influence of a given trait on the speciation-extinction dynamics of a clade.

Lineages can be either of type 0 (e.g. 'able to fly') in which case they give birth to new lineages at rate b_0 and go extinct at rate d_0, or they can be of type 1 (e.g. 'unable to fly'), in which case they speciate at rate b_1 and go extinct at rate d_1. Additionally, lineages of type 0 can acquire type 1 at rate q_{01}, and reciprocally lineages of type 1 can transform into type-0-lineages at rate q_{10}. Results of a simulation thus look like the coloured reconstructed tree in Figure 1.37.

The model can be fitted to empirical data, comprising a fixed phylogeny with tip trait data. The maximum likelihood estimators of the 6 parameters can be inferred, in order to assess the effect of the trait on the speciation-extinction dynamics. This type of model has been used for example by Rolland et al. (2014), who show that migratory birds have a higher diversification rate than sedentary ones. In another study, Price et al. (2012) considered three diet types (herbivores, carnivores and omnivores) that could impact the diversification rate. They show that among mammals, herbivores diversify fastest, followed by carnivores and omnivores. It also allowed them to estimate the transition rates between the three diets.

SSE models have exploded in the recent literature, extending the idea to a variety of distinct traits, among which for example:

QuaSSE 'Qua' standing for quantitative traits (FitzJohn, 2010).
The model assumes that a quantitative trait x (e.g. the body mass) evolves through time as a diffusion process. A dependence of the speciation and extinction rates on this trait is further assumed

in two functions: $b(x)$ and $d(x)$. The fit of the model to empirical phylogenies with data on the tip body mass provides estimates of the strength of this dependence.

GeoSSE 'Geo' standing for geography of species (Goldberg et al., 2011).
Each lineage is characterized by a trait describing its spatial distribution over two patches A and B. The trait can be either A, B, or AB. Transitions include range expansion (e.g. from A to AB) and range contraction (e.g. from AB to A, but also from A to extinction). Finally, speciation events occur according to site-specific parameters. All these parameters, as well as ancestral spatial distributions, can be obtained by fitting the model to an empirical phylogeny with known spatial distribution of tip species.

HiSSE 'Hi' standing for hidden traits (Beaulieu and O'meara, 2016).
This model was developed after Rabosky and Goldberg (2015) pointed out a risk of model inadequacy when applying the BiSSE model to empirical data. The BiSSE model could indeed detect a spurious relation between the studied trait and diversification rates when the diversification rates were in fact dependent on another, non-studied trait evolving along the same phylogeny. To tackle this issue, Beaulieu and O'meara (2016) proposed a model were two uncorrelated traits evolve, one being observed at the tips, the second one being hidden. The speciation and extinction rates depend on these two traits, and the fit to empirical data provides estimates of this dependence.

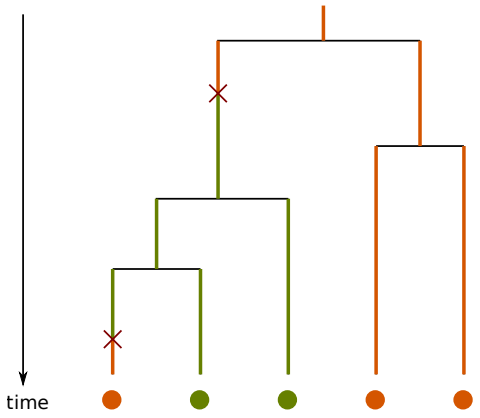


Figure 1.37 – Reconstructed tree under the BiSSE model. Green and orange represent the two states of the trait.

We end up our background introduction on macroevolution modeling with these numerous examples of applications. In the following chapters, we will turn to the presentation of our specific work, spreading across the study of individual-based modeling (chapters 2 and 3), diversification modeling (chapter 3), continuous phenotype evolution modeling (chapter 4) and molecular evolution modeling (chapter 5). On each subject, we will try to question commonly accepted hypotheses in the field, and propose unconventional ones which may help put macroevolution work into perspective. In the last chapter 6 we discuss the various connections which could be further considered between the distinct chapters of this thesis.

Chapter 2

The species definition from the modeler's point of view

Species are central entities in evolutionary biology, and yet no consensus on the definition of the term has been reached. On the other hand, in a simple modeling world, the picture should be much different and above all clearer. In this chapter, we discuss the many species definitions that have been proposed in an individual-based framework, where individuals reproduce clonally and diversity arises through point mutations. This framework is very popular, for it has been put forward by the Neutral Theory of Biodiversity (Hubbell, 2001). Originally aimed at studying macroecological patterns such as species abundance distribution, it has subsequently found its way into studies of diversification.

By modeling explicitly the fate of individuals and species, the modeler has a direct access to the genealogy of individuals and the phylogeny of species. Motivated by the complex articulation between genealogies and phylogenies under all current species definitions, we propose two new species definitions aimed at more easily coupling the two trees.

These two species definitions naturally emerge from the consideration of one key unconventional hypothesis for the field, namely, the monophyly of species.

This chapter corresponds to the submitted manuscript (Manceau and Lambert, 2016). The first section presents the article information. Material from the second to the last sections corresponds to the core of the article.
This chapter is also closely linked to appendix chapter A, presenting proofs of our results.

Contents of the chapter

2.1 Article information

Title

The species problem from the modeler's point of view.

Authors

Marc Manceau[1,2] and Amaury Lambert[1,3]

[1] Center for Interdisciplinary Research in Biology (CIRB), Collège de France,
PSL Research University, CNRS UMR 7241, INSERM U1050, 75005 Paris, France

[2] Institut de Biologie de l'École Normale Supérieure, École Normale Supérieure,
PSL Research University, CNRS UMR 8197, INSERM U1024, 75005 Paris, France

[3] Laboratoire de Probabilités, Statistique et Modélisation (LPSM),
Sorbonne Université, CNRS UMR 8001, 75005 Paris, France

Keywords

Genealogy, Phylogeny, Microevolution, Macroevolution, Individual-based model, Species concept

Abstract

How to define a partition of individuals into species is a long-standing question called the *species problem* in systematics. Here, we focus on this problem in the thought experiment where individuals reproduce clonally and both the differentiation process and the population genealogies are explicitly known. We specify three desirable properties of species partitions: (A) Heterotypy between species, (B) Homotypy within species and (M) Monophyly of each species. We then ask: How and when is it possible to delineate species in a way satisfying these properties?

We point out that the three desirable properties cannot in general be satisfied simultaneously, but that any two of them can. We mathematically prove the existence of the finest partition satisfying (A) and (M) and the coarsest partition satisfying (B) and (M). For each of them, we propose a simple algorithm to build the associated phylogeny out of the genealogy.

The ways we propose to phrase the species problem shed new light on the interaction between the genealogical and phylogenetic scales in modeling work. The two definitions centered on the monophyly property can readily be used at a higher taxonomic level as well, e.g. to cluster species into monophyletic genera.

Aknowledgments

The authors are very grateful to F. Débarre, R.S. Etienne, M. Steel, S. Türpitz and A. Hoppe for their comments on this paper, and to D. Baum for helpful literature advice. The authors thank the Center for Interdisciplinary Research in Biology (CIRB, Collège de France) for funding, as well as the École Normale Supérieure for MM PhD funding. We declare no conflict of interest.

2.2 Introduction

Models in macro-evolution have traditionally been centered on species. The so-called *lineage-based models* of diversification form a wide class of models considering species as key evolutionary units, thought of as particles that can give birth to other particles (i.e., speciate) during a given lifetime (i.e., before extinction) (see Stadler, 2013; Pyron and Burbrink, 2013; Morlon, 2014 for reviews). In contrast,

evolutionary processes amenable to direct empirical measurement (differentiation, reproduction, selection) are usually described at the level of individuals or populations.

The Neutral Theory of Biodiversity (NTB) (Hubbell, 2001) opened a new way of thinking about species in macro-evolution. The birth, death, differentiation and speciation processes are described at the level of individuals, under the assumption of selective neutrality. In the last two decades, a popular way of studying macro-evolution has followed, consisting in performing computer-intensive simulations of individual-based stochastic processes of species diversification (Jabot and Chave, 2009; Aguilée et al., 2011; Rosindell et al., 2015; Gascuel et al., 2015; Missa et al., 2016). These models rely on three major steps: (i) The genealogy of individuals is produced under a stochastic scenario of population dynamics; (ii) A process of phenotypic differentiation superimposed on the genealogy generates a partition of individuals into phenotypic groups; (iii) A species definition is postulated and is used to cluster individuals into different species, in relation to both the genealogy and phenotypic groups. These three steps allow modelers to track the evolutionary history of species, where extinction and speciation events emerge from the genealogical history of individual organisms.

The scenario of population dynamics (i) has for example been modeled with the Wright-Fisher or the Moran model from population genetics or with density-dependent branching processes (Durrett, 2008). We will not focus on that step, but we will nonetheless consider throughout the paper that individuals reproduce clonally. The genealogical relationships within a sample $\mathcal{X}$ of present-day individuals will thus be a known rooted tree denoted T. Each tip of T is labelled by an element of $\mathcal{X}$, each internal vertex corresponds to some ancestor of elements of $\mathcal{X}$, and an edge between vertices represents a parent-child relationship. After running through step (ii), all individuals in $\mathcal{X}$ can be grouped into clusters of individuals on the basis of their phenotype. We call the associated $\mathcal{X}$-partition the *phenotypic partition*, denoted $\mathscr{P}$.

In this paper, we review how steps (ii) and (iii) have been handled in the literature, before asking the following theoretical question:

> *How* and *when* is it possible to delineate species, in a way satisfying *biologically meaningful properties*, in an ideal situation where the phenotypic partition $\mathscr{P}$ is specified and the entire genealogy T is known?

Because we work under the simplifying assumption of clonally reproducing organisms, this question is formally identical to the problem of defining and delineating *genera* when the *species phylogeny* is known, or any similar question formulated at a higher-order level (Aldous et al., 2008, 2011).

In the biological literature, the problem of agreeing on what should be considered as the most relevant concept of species is a long-standing question called the *species problem*. The species problem is both a conceptual question (defining the species concept) and a practical problem (classifying individuals into species) (Bock, 2004). Several of the most notable evolutionary biologists (e.g. Darwin, Dobzhansky, Mayr, Simpson, Hennig...) took a stand on the species problem often leading to vigorous debates (see Mayden, 1997; De Queiroz, 2007 for overviews of historical disagreements).

In the very simple setting that we are looking at, two classes of species concepts appear relevant in the quest for *biologically meaningful properties* of species. The 'typological species concepts' (Regan, 1925; Sneath, 1976) correspond to the clustering of individuals on the sole basis of their observed phenotype. This can be translated into two desirable properties: (A) any two individuals in distinct clusters differ for at least one characteristic, (B) individuals belonging to the same cluster all share the same characteristics. The later foundation and spread of cladistics by Hennig (Hennig, 1965) marked a radical change of paradigm in the systematic classification, which quickly resulted in the proposition of so-called 'phylogenetic species concepts' (De Queiroz and Donoghue, 1988; Avise and Ball, 1990; Baum, 2009). These definitions, which brought to the forefront the notion of common ancestry, provide us with a third desirable property: (M) species are monophyletic groups of individuals, i.e. any two individuals in one cluster are more closely related to each other than either is to any individual in another cluster.

Ideally, defining putative species in our framework amounts to finding clusters of individuals satisfying these three desirable properties that we will denote throughout the paper:

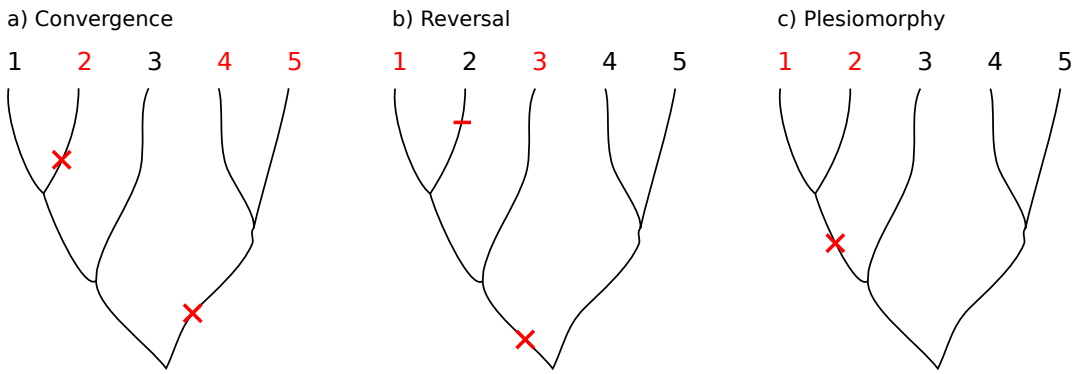


Figure 2.1 – Different scenarios leading to non-monophyletic phenotypic partitions, for the 'red' phenotype. a) The same phenotype arises twice independently in two branches. b) A new phenotype arises, and disappears later. c) The group of individuals showing the ancestral phenotype is not monophyletic.

(A) Heterotypy between species

(B) Homotypy within species

(M) Monophyly

We quickly observe that by definition, the phenotypic partition $\mathscr{P}$ satisfies the two desirable properties: (A) and (B). Unfortunately, phenotypically similar individuals may not be more genealogically related to one another than to any different individual. In other words $\mathscr{P}$ does not in general satisfy (M). Convergence and reversal events, by making a trait either appear several times independently in different parts of the tree or by making a trait disappear in subtrees, are classically invoked to explain the non-monophyly of $\mathscr{P}$. However, let us stress that even with traits evolving without convergence or reversal, individuals characterized by an ancestral phenotype may define a non-monophyletic subset of $\mathcal{X}$, a phenomenon called 'ancestral type retention' or 'plesiomorphy' (see Figure 2.1).

The paper is organized as follows. In Section 2.3, we review the main species definitions used in the context of individual-based modeling of macro-evolution. In Section 2.4, we introduce the formalism required to study species partitions, and we make some preliminary observations on the three desirable properties. In Section 2.5, we prove that it mathematically makes sense to define the finest species partition satisfying (A) and (M) and the coarsest species partition satisfying (B) and (M). We call these respectively the *loose* and the *lacy* species partitions. Finally, we discuss the relevance of these definitions from both empirical and theoretical points of view.

2.3 Five species definitions in individual-based models

We provide here an overview of five modes of speciation that have been proposed so far in the context of individual-based models of diversification (see Kopp, 2010 for a review and Figure 2.2 for illustration). Among these five modes, only the second one is intended to model specifically the geographical isolation of two subpopulations. The four other modes focus on modeling sympatric speciation by means of gradual accumulation of mutations.

Speciation by point mutation. This mode of speciation was proposed in the original framework of the NTB (Hubbell, 2001; Jabot and Chave, 2009). Differentiation occurs as the product of neutral mutations modeled by a point process on the genealogy. Each mutation confers a new type to the lineage carrying it (infinite-allele model) and to its descent before any new mutation arises downstream.

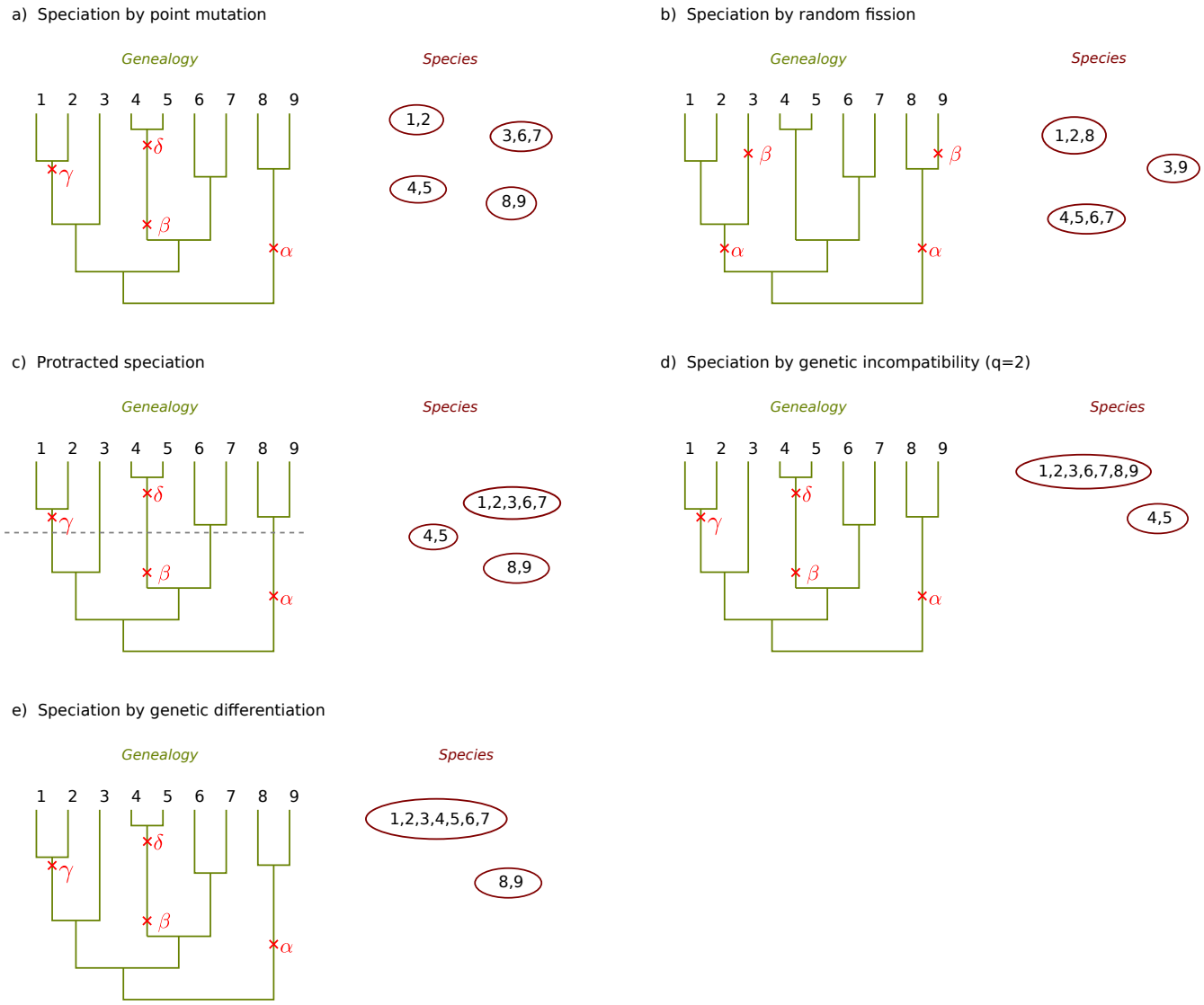


Figure 2.2 – The five modes of speciation proposed in individual-based models of macro-evolution. In each panel, the genealogy (tree) of individuals (integer labels) is given on the left, along with mutations (crosses) that confer new types (Greek letters) to individuals. The corresponding species partition is represented on the right of each panel (subsets of labels circled).

Species are then defined as groups of individuals carrying the same type. The phenotypic partition and the species partition thus coincide by definition.

Speciation by random fission, or peripheral isolates. These two closely related models have also been proposed first in the framework of the NTB (Hubbell, 2001, 2003), but see also Lambert and Ma (2015). In these models, independently of the genealogy, each phenotypic class of individuals, interpreted as a geographic deme, may split at random times into two new demes. In Figure 2.2b, this is illustrated as mutations hitting simultaneously several lineages in the same phenotypic class, which endows them with the same new phenotype. The two models differ only with regard to the size of the newly formed deme, whether the split is even (random fission) or uneven (peripheral isolate).

Protracted speciation. This model intends to reflect the general idea that speciation is not instantaneous (Rosindell et al., 2010; Lambert et al., 2015; Etienne et al., 2014). The differentiation process is usually assumed to be differentiation by point mutation under the infinite-allele model. A new phenotypic class is called an incipient species but becomes a so-called good species only after a fixed or random time duration. In other words, two individuals belong to different species if they carry different phenotypes and if they diverged far enough in the past. In Figure 2.2c, the species arisen from the mutation labelled γ and δ are still incipient at present time. More complex models of protracted speciation feature several stages that incipient species have to go through before becoming good species.

Speciation by genetic incompatibility. This generalization of the point mutation mode of speciation (Melián et al., 2012) is inspired by the model of Bateson-Dobzhansky-Muller incompatibilities (Orr, 1995). Again, a first step consists in endowing the genealogy of individuals with neutral mutations. Then two individuals are said compatible if there are fewer than q mutations on the genealogical path linking them. Finally, species are the connected components of the graph associated to the compatibility relationship between individuals. For $q \neq 1$, there can be incompatible pairs of individuals in the same species, as can be seen in Figure 2.2d with individuals labelled 1 and 9 for example. Speciation by point mutation corresponds to the particular case $q = 1$.

Speciation by genetic differentiation. This model of speciation also assumes that phenotypic differentiation is driven by point mutations on the genealogy. Species are then defined as the smallest monophyletic groups of individuals such that any pair of individuals carrying the same phenotype are always in the same species (Manceau et al., 2015). We will show later that this definition, hereafter called *loose species definition*, always makes sense once given a phenotypic partition and a genealogy.

As can be seen in Figure 2.2, the first four models out of the five described in the previous section yield partitions of individuals into species that are in general non-monophyletic with respect to the underlying genealogy. This is problematic when it comes to measuring the phylogenetic relationship between species, reflecting their shared evolutionary history (Velasco, 2008). In particular based on the true genealogy, there are multiple, arbitrary ways of defining the divergence time between two non-monophyletic species. We illustrate three of them in Figure 2.3.

The first two possibilities consist in relying on a time of divergence between individuals of the newly derived species and individuals of the ancestral, mother, species. One could imagine taking the shortest (scenario a), as well as the longest (scenario b), time of coalescence between these two groups of individuals. Among these two, scenario a) is the preferred option among population geneticists, for it provides a phylogeny consistent with the genealogy of alleles. A second possibility, and the one which is in ordinary usage in NTB-based studies (Jabot and Chave, 2009), is to consider the date of appearance of the derived character as the time of divergence between the two species. We now argue that none of these possibilities is really appropriate from an evolutionary point of view.

Let us show for more clarity what these three possibilities lead to, in a scenario with three species (See Fig. 2.4). While scenarios a) and b) provide an unnecessary multifurcating tree, scenario c) even leads to a phylogeny topology distinct from what we would expect from the topology of the genealogy.

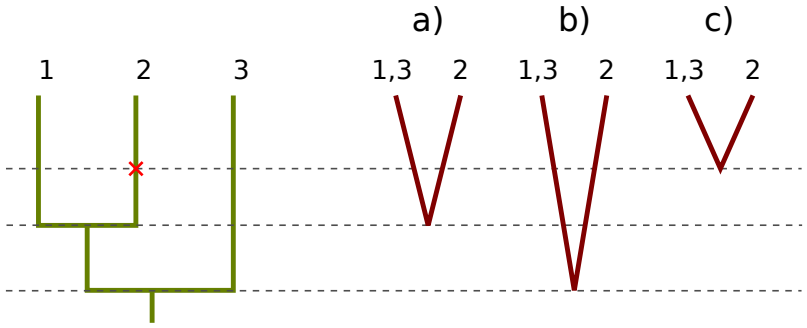


Figure 2.3 – Building the phylogeny out of the genealogy. Left panel: A fixed genealogy with one mutational event giving rise to a derived character responsible for partitioning $\{1, 2, 3\}$ into the two distinct phenotypic groups $\{1, 3\}$ and $\{2\}$, assumed to be different species. Right panel: Phylogenies associated with three possible choices of divergence times, from left to right: a) Shortest or b) Longest coalescence time between individuals of different species; c) Date of origin of the derived character.

Any representation of a phylogeny for non-monophyletic species would suffer such inconsistencies with the genealogy. To the contrary, the phylogeny of monophyletic species can be obtained from the genealogy by considering as phylogenetic nodes all genealogical nodes that are most recent common ancestors between present-day species. As part of the current effort to bridging the gap between micro- and macro-evolution (Graham and Fine, 2008; Rosindell et al., 2011; Pennell and Harmon, 2013), it appears crucial to interlock the genealogical and phylogenetic scales. This is the rationale behind considering monophyly as a desirable property of species constructions.

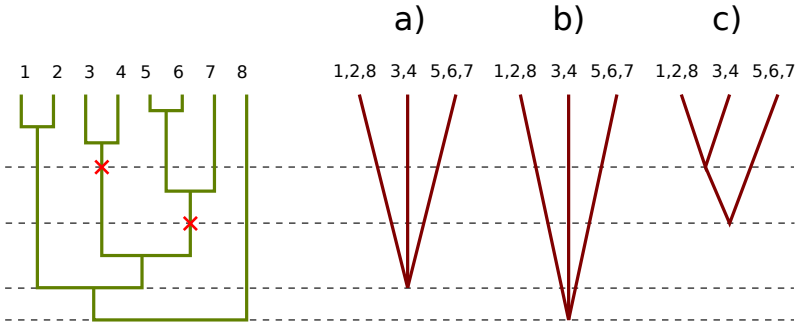


Figure 2.4 – Building the phylogeny out of the genealogy with three species. Left panel: A fixed genealogy with mutational events giving rise to three species. Right panel: Phylogenies associated with three possible choices of divergence times, from left to right: a) Shortest or b) Longest coalescence time between individuals of the ancestral and the derived species; c) Date of origin of the derived character.

In Sections 2.4 and 2.5, we will consider as given the genealogical tree T of the set of all present-day organisms $\mathcal{X}$, and the partition $\mathscr{P}$ of $\mathcal{X}$ into phenotypic groups. The tree T may have been generated under any model of population dynamics and the partition $\mathscr{P}$ may have been produced by any process of differentiation unfolding through time. With these data at hand, we formalize the three desirable properties of the species partition mentioned in the introduction and then study different ways to fulfill them.

2.4 Three desirable properties of species definitions

For each internal node of T, by a slight abuse of terminology, we call *clade* the subset of $\mathcal{X}$ comprising exactly all tips descending from this node. We denote by $\mathscr{H}$ *the collection of all clades of* T. Note that as a subset of $\mathcal{X}$, $\mathcal{X}$ itself is an element of $\mathscr{H}$, and that for every $x \in \mathcal{X}$, the singleton $\{x\}$ is an element of $\mathscr{H}$. Moreover, any two clades C and D elements of $\mathscr{H}$, are always either nested or mutually exclusive, meaning that $C \cap D$ can only be equal to C, D or $\varnothing$. Mathematically, a collection of nonempty subsets of $\mathcal{X}$ satisfying these properties is called a *hierarchy*, and it can be shown that to any hierarchy corresponds a unique rooted tree with tips labelled by $\mathcal{X}$. Therefore, we will equivalently speak of T or of its hierarchy $\mathscr{H}$. For a nice discussion around the notion of hierarchy and neighboring concepts, see Steel (2014).

One should keep in mind that $\mathscr{H}$ and $\mathscr{P}$ are both collections of subsets of $\mathcal{X}$, but that $\mathscr{H}$ is not a partition. With this formalism, we define the species problem as : Given $\mathscr{H}$ and $\mathscr{P}$ find a partition $\mathscr{S}$ of $\mathcal{X}$, called the *species partition*, whose elements are called *species clusters* or simply *species*, satisfying one or more of the following three desirable properties:

(A) Heterotypy between species. Individuals in different species are phenotypically different, i.e. for each phenotypic cluster $P \in \mathscr{P}$ and for each species cluster $S \in \mathscr{S}$, either $P \subseteq S$ or $P \cap S = \varnothing$;

(B) Homotypy within species. Individuals in the same species are phenotypically identical, i.e. for each phenotypic cluster $P \in \mathscr{P}$ and for each species cluster $S \in \mathscr{S}$, either $S \subseteq P$ or $P \cap S = \varnothing$.

(M) Monophyly. Each species is a clade of T, i.e. $\mathscr{S} \subseteq \mathscr{H}$;

As mentioned in the introduction, if $\mathscr{S}$ satisfies both (A) and (B), then it is immediate from the preceding definitions that $\mathscr{S} = \mathscr{P}$. If in addition $\mathscr{S}$ satisfies (M) then $\mathscr{P} = \mathscr{S} \subseteq \mathscr{H}$. We record this as a first observation.

Observation 1 *Unless we are given $\mathscr{P}$ and $\mathscr{H}$ such that $\mathscr{P} \subseteq \mathscr{H}$ (that is, each phenotypic cluster is a clade in the first place), no species partition satisfies simultaneously (A), (B) and (M).*

Then our next question is: 'Is there a species partition $\mathscr{S}$ for which two of them hold?' For X, Y equal to A, B or M, we will write (XY) the property (X AND Y).

We already saw that $\mathscr{S} = \mathscr{P}$ satisfies (AB). Now let us go for species partitions $\mathscr{S}$ satisfying (M). To fulfill (A), each $S \in \mathscr{S}$ must contain all the phenotypic clusters it intersects. So in particular the partition $\mathscr{S}_1 := \{\mathcal{X}\}$ fulfills (AM). This trivial solution corresponds to assigning all the individuals of $\mathcal{X}$ to one single species. Symmetrically, to fulfill (B), each $S \in \mathscr{S}$ must be contained in all the phenotypic groups it intersects. So in particular the partition $\mathscr{S}_0$ made of all singletons fulfills (BM). This trivial solution corresponds to assigning each individual of $\mathcal{X}$ to a different species. This is recorded in the following observation.

Observation 2 *For any $\mathscr{P}$ and $\mathscr{H}$ and for any two desirable properties among (A), (B) and (M), there is at least one species partition $\mathscr{S}$ satisfying both properties.*

The species partitions $\mathscr{S}_1$ and $\mathscr{S}_0$ are obviously not biologically relevant. In particular, we would like to find species partitions that are *finer* than assigning all individuals to one single species, and *coarser* than assigning each individual to a different species.

We use the standard notions of finer and coarser partitions of a set (Bóna, 2011). Let $\mathscr{S}$ and $\mathscr{S}'$ be two partitions of the set $\mathcal{X}$. We say that $\mathscr{S}$ is finer than $\mathscr{S}'$, and we write $\mathscr{S} \leq \mathscr{S}'$ if for each $S \in \mathscr{S}$ and each $S' \in \mathscr{S}'$, either $S \subseteq S'$ or $S \cap S' = \varnothing$. If $\mathscr{S} \leq \mathscr{S}'$, we say equivalently that $\mathscr{S}$ is finer than $\mathscr{S}'$ or that $\mathscr{S}'$ is coarser than $\mathscr{S}$.

Remark that two species partitions $\mathscr{S}$ and $\mathscr{S}'$ cannot always be compared, in the sense that they can satisfy neither $\mathscr{S} \leq \mathscr{S}'$ nor $\mathscr{S}' \leq \mathscr{S}$. The relation $\leq$ is thus not a linear order on all the partitions of $\mathcal{X}$, but is known to be a partial order (see Appendix A.1 for details).

Now observe that properties (A) and (B) can precisely be stated in terms of inequalities associated with the partial order $\leq$ as follows.

Observation 3 *Consider a given phenotypic partition $\mathscr{P}$ and a species partition $\mathscr{S}$.*
$\mathscr{S}$ satisfies (A) if and only if $\mathscr{P} \leq \mathscr{S}$, and $\mathscr{S}$ satisfies (B) if and only if $\mathscr{S} \leq \mathscr{P}$.
As a consequence, if $\mathscr{S}_A$ satisfies (A) and $\mathscr{S}_B$ satisfies (B), then

$$\mathscr{S}_B \leq \mathscr{P} \leq \mathscr{S}_A$$

This leads us to investigate the possibility of defining 'the finest partition satisfying (AM)' as well as 'the coarsest partition satisfying (BM)'.

2.5 The lacy and loose species definitions

In general, there is no guarantee that the coarsest or finest partition of a given collection of partitions does belong to this collection. In particular, there is no guarantee that the coarsest (resp. finest) partition satisfying (BM) (resp. (AM)), does itself satisfy (BM) (resp. (AM)). However, we state the following result that ensures the existence of the finest partition satisfying (AM) and the coarsest partition satisfying (BM).

Theorem 1 *Given $\mathscr{P}$ and $\mathscr{H}$, there exists a unique finest partition of $\mathcal{X}$ satisfying (AM), and a unique coarsest partition of $\mathcal{X}$ satisfying (BM).*

This result is proved in Appendix A.2. It allows us to highlight and name two new different species definitions:
The *loose species definition* is the finest partition satisfying (AM).
The *lacy species definition* is the coarsest partition satisfying (BM).

For any species partition $\mathscr{S}$ satisfying (M), there is a unique phylogenetic tree $T_{\mathscr{S}}$ which represents the evolutionary relationships between the species in $\mathscr{S}$ consistently with the genealogy T. For every species S, since S is monophyletic there is a unique internal node $u(S)$ of T such that S is exactly constituted of the labels of the tips subtended by $u(S)$. Then $T_{\mathscr{S}}$ is obtained from T by merging, for every species S, the subtree descending from $u(S)$ into a single edge. This is expressed in terms of hierarchies in the following observation.

Observation 4 *Consider a given genealogical hierarchy $\mathscr{H}$ and species partition $\mathscr{S}$ satisfying (M). The hierarchy $\mathscr{H}_{\mathscr{S}}$ corresponding to the phylogenetic tree $T_{\mathscr{S}}$ can be defined by*

$$\mathscr{H}_{\mathscr{S}} := \{H \in \mathscr{H} : \ \exists S \in \mathscr{S}, \ S \subseteq H\}.$$

So for both the loose and the lacy species partition, there is a phylogeny consistent with the genealogy. Figure 2.5 shows both the lacy phylogeny and the loose phylogeny associated with a simple genealogy and a simple phenotypic partition.

For any genealogy T and phenotypic partition $\mathscr{P}$, we now describe a procedure to get the phylogeny corresponding either to the lacy or to the loose definition, without requiring the knowledge of species partitions. Interestingly, building $\mathscr{H}_{\mathscr{S}}$ this way offers a quick way to get $\mathscr{S}$ under the lacy and loose definitions, because species are the smallest sets of labels in $\mathscr{H}_{\mathscr{S}}$. The different steps of the algorithm are explained hereafter, illustrated in Figure 2.6 and formalized in Appendix A.3.

First, we classify all interior nodes of the genealogy as *convergent node* or *divergent node*. An interior node is *convergent* if there are at least two tips, one in each of its two descending subtrees, carrying the same phenotype. Otherwise the node is said to be *divergent*. Note that convergent nodes may be ancestors of divergent nodes when the phenotypic partition is not monophyletic. Second, we build a phylogeny by deciding which interior nodes are *phylogenetic nodes*, that is, appear in the phylogeny.

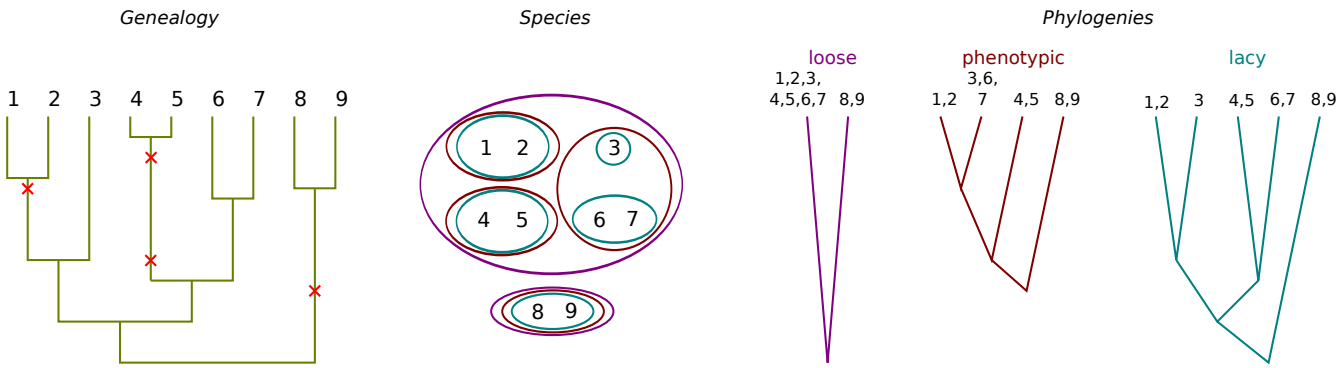


Figure 2.5 – Species partitions associated to each of three definitions. Left panel: The fixed genealogy with point mutations (infinite-allele model) leading to the phenotypic partition $\mathscr{P} = \{\{1,2\},\{3,6,7\},\{4,5\},\{8,9\}\}$. Middle panel: Inclusion relations between the three species partitions, as discussed in Observation 3, the loose partition is coarser than the phenotypic partition, which is coarser than the lacy partition. Right panel: Phylogenies corresponding to the three species partitions, from left to right: loose, phenotypic (under the arbitrary convention that divergence times are taken as mutation times), lacy.

Observation 5 *The loose phylogeny is obtained by declaring non-phylogenetic (i) all convergent nodes and (ii) all divergent nodes descending from a convergent node. Other nodes are declared phylogenetic. The lacy phylogeny is obtained by declaring phylogenetic (i) all divergent nodes and (ii) all convergent nodes ancestral to divergent nodes. Other nodes are declared non-phylogenetic.*

The last observation holds due to the following reasons:

By definition, the two clades C and C' subtended by a convergent node satisfy $C \cap P \neq \varnothing$ and $C' \cap P \neq \varnothing$ for some phenotypic cluster $P \in \mathscr{P}$. As a consequence, these two clades have to be included in the same species cluster in a species partition satisfying the heterotypy property (A), that is, a convergent node cannot appear in a phylogeny satisfying (A). Conversely, any phylogeny whose nodes are included in the set of divergent nodes of the genealogy satisfies (A). The finest partition satisfying (A) corresponds to the phylogeny containing the largest number of divergent nodes, and only divergent nodes, as in the construction of the loose phylogeny proposed in the observation.

Symmetrically, for the two clades C, C' subtended by a divergent node, we have that $C \cap P \neq \varnothing$ implies $C' \cap P = \varnothing$ for any phenotypic cluster $P \in \mathscr{P}$. As a consequence, these two clades have to belong to two different species clusters in a species partition satisfying the homotypy property (B), that is, any divergent node has to appear in a phylogeny satisfying (B). Conversely, any phylogeny whose nodes contain all divergent nodes of the genealogy satisfies (B). The coarsest partition satisfying (B) corresponds to the phylogeny containing the smallest number of convergent nodes, but all divergent nodes, as in the construction of the lacy phylogeny proposed in the observation.

2.6 Discussion

The present study builds on recent attempts to describe evolutionary trees on various scales simultaneously. The multi-species coalescent is one of the most influential of these attempts (Maddison, 1997). In this model, a species tree is first specified and given the species tree, the gene genealogies are drawn from a censored coalescent (i.e., ancestral lineages can coalesce only if they lie in the same ancestral species). In particular, this approach has been used to assess the relevance of the reciprocal monophyly criterion to recognize species (Hudson and Coyne, 2002; Mehta et al., 2016). Note that this framework introduces a top-down coupling between the macro-evolutionary scale and the micro-evolutionary scale, thus

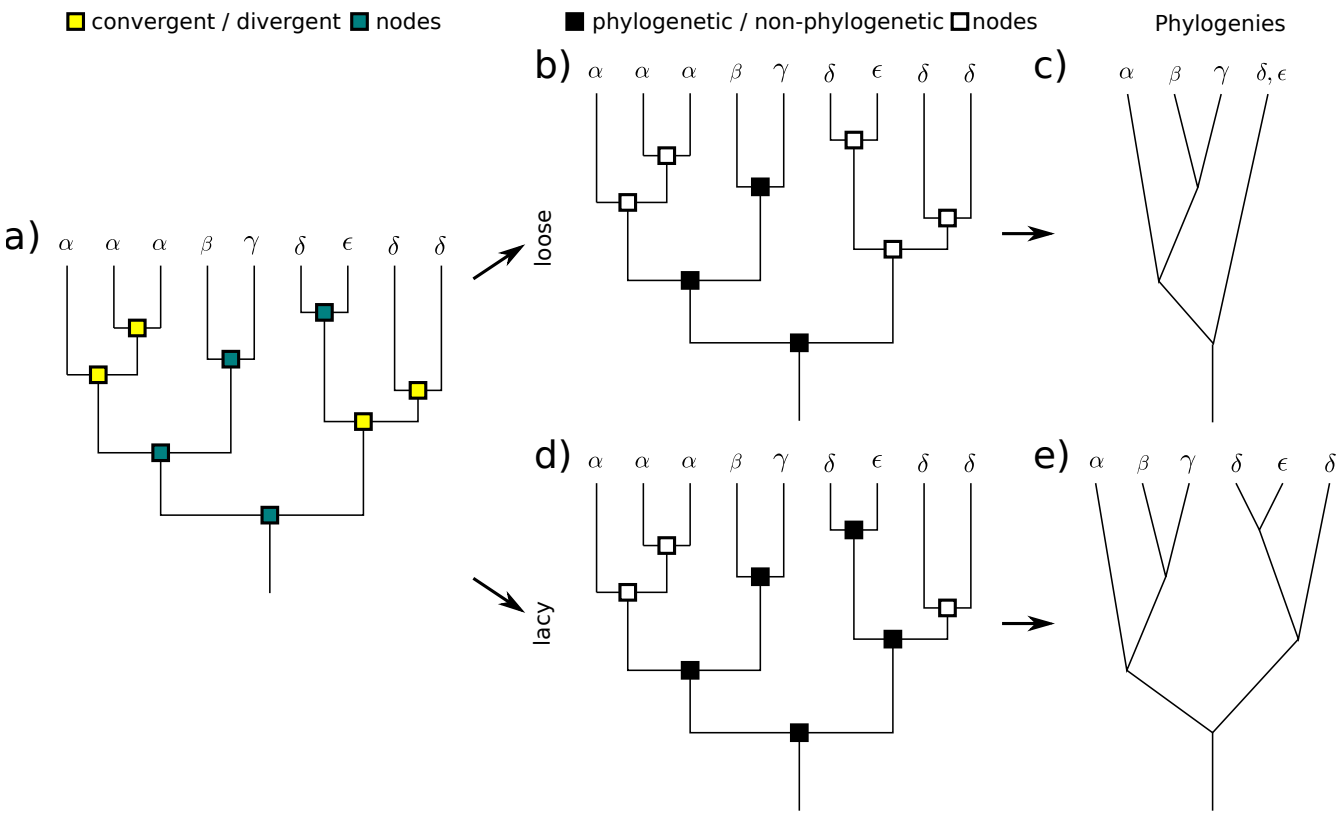


Figure 2.6 – Construction of the phylogeny under the lacy and loose species definitions. Greek letters correspond to phenotypes. a) The genealogy with interior nodes classified as convergent (light yellow) or divergent (dark blue). bd) The genealogy with interior nodes classified as non-phylogenetic (white) or phylogenetic (black). ce) The corresponding phylogeny. b) Loose: Light nodes and dark nodes descending from light nodes are colored white. d) Lacy: Dark nodes and light nodes ancestors of a dark node are colored black.

relying on an external species definition. In contrast, the bottom-up approach that we adopted consists in assuming that macro-evolutionary patterns are shaped by micro-evolutionary processes. Let us point out two studies similar in spirit to ours, which make several proposals to lump together lower-order taxa (e.g. species) in order to build trees on higher-order taxa (e.g. genera) (Aldous et al., 2008, 2011). The authors ground their definitions on the knowledge of a phylogeny with point mutations, the main differences being that each branching event distinguishes a mother and a daughter lineage, and that monophyly is not a desirable property in their work.

On the contrary, we put to the forefront the monophyly property (M), together with (A) heterotypy between species and (B) homotypy within species. We explicitly defined and compared three species definitions, each of these satisfying a different set of properties: Phenotypic (AB), Loose (AM) and Lacy (BM).

Additionally, we stress that the most popular way of defining species in individual-based modeling studies of macro-evolution (i.e., speciation by point mutation) in general leads to non-monophyletic species. In contrast, the loose species definition, previously used in the context of macro-evolution (Manceau et al., 2015), systematically yields monophyletic species. Here we extended this study and compared it to a third species definition also satisfying (M), the lacy species definition. Finally, we provided a standardized procedure to build the lacy and loose species partitions given a genealogy and a phenotypic partition.

In practice, the task of systematists is the inference of ancestral relationships between individual organisms

from molecular sequence and phenotype data and the characterization of species from those data. Classifying diversity is notoriously difficult for many reasons, including the difficulty of choosing the appropriate level of description, the ubiquitous presence of convergent evolution and reversal events, and the difficulty to agree on a unique species concept (Mayden, 1997; De Queiroz, 2007; Baum, 2009).

On the other hand, the task of modelers, assuming a fully known individual-based evolutionary history, appears at first sight trivial. They face, however, the same difficulty in defining a proper species concept. Even within the very simple framework that we considered, three distinct species definitions came out. They all fit the general species definition of *'separately evolving metapopulation lineages'* (De Queiroz, 2007), while satisfying distinct desirable properties. We argue that comparing species definitions based on the properties they fulfill in simple models might help shed light on the species problem.

Let us draw parallels between our theoretical considerations and the habits of taxonomists. Practically, the sole phenotypic information is usually sufficient do decide whether a taxon above the species level is monophyletic. And indeed in systematics, monophyly has long been a criterion for defining taxa above the species level, even before the rise of molecular methods. On the contrary, phenotypic information alone is certainly not sufficient to diagnose monophyletic species. The use of molecular markers has brought the question of intra-species monophyly (M) to the forefront. Today, it is standard to use multiple sequence alignments to automatically delineate putative species: from a single-locus phylogeny (Fujisawa and Barraclough, 2013), from multiple gene trees (Yang and Rannala, 2010), or from the raw alignment (Puillandre et al., 2012). Species descriptions based on these methods are thus more likely to concern monophyletic groups of individuals than earlier. This recent requirement for species monophyly puts taxonomists in front of new dilemmas. Loose or lacy? Crudely speaking, one could say that 'splitter' taxonomists more often lean for the lacy definition, while 'lumper' taxonomists are more willing to use the loose definition. More precisely, the diagnosis of species as phenotypically homogeneous groups of individuals that can be separated solely on a molecular basis into what is known as cryptic species (Bickford et al., 2007) corresponds to the lacy definition. On the contrary, taxonomists preferring to ensure that species are diagnosable and monophyletic units, two properties stressed as 'priority taxon naming criteria' (Vences et al., 2013) use the loose definition.

Note that advanced theoretical work has been undertaken in a context of sexually-reproducing organisms (Dress et al., 2010; Kwok, 2011; Alexander, 2013; Alexander et al., 2015). While we based our study on the knowledge of only one genealogy, even the genealogical history of supposedly 'asexual' real-world organisms such as bacteria shows evidence for horizontal gene transfer events (Puigbò et al., 2013). The genealogical history of organisms should in general be represented as a non-tree network, or as a collection of gene genealogies, making far more complex the question of grouping individuals into taxa (Hudson and Coyne, 2002; Samadi and Barberousse, 2006). Their framework is closer to biological reality, but much less connected to most modeling studies in macro-evolution.

Individual-based modeling is a promising avenue for understanding macro-evolution from first principles, as it may allow evolutionary biologists to describe explicitly the stochastic demography of whole metacommunities and the ecological interactions between different types of individuals in each community. We believe that these processes may have left enough signal in both the shape of evolutionary trees and the patterns of contemporary biodiversity, so as to be unraveled by statistical inference. Understanding how species, the elementary units of macro-evolution, are formed and deformed by these processes remains a major challenge, to which the present work hopefully contributes.

Chapter 3

Toward an individual-based modeling of diversification

The study of diversification over long timescales is commonly addressed using lineage-based models, considering species as the basal particles of a birth-death process (see section 1.1.4). However, there has been a growing interest in recent years for individual-based models, for they hold a promise to take into account all observable processes at a human scale and integrate them over long timescales to deduce the long term behaviour of the system.

In this chapter, we consider a (not so unconventional now) individual-based framework inspired by the influential Neutral Theory of Biodiversity (Hubbell, 2001). However, our model relies on two unconventional hypotheses for the field. First, the metapopulation of individuals follows a birth-death process, while the standard hypothesis, called *zero-sum assumption*, consists in assuming a constant-size metapopulation. Second, we rely on the *loose* species definition presented in previous chapter 2, which had never been envisaged before.

We show that this set of hypotheses is adequate to reproduce coarse-grained patterns observed in empirical phylogenies, thus suggesting the importance of out-of-equilibrium scenarios (here, constantly increasing metapopulation) to fit macroevolutionary diversification patterns. We further propose an inference procedure to recover the maximum likelihood parameters of the model from the observation of a phylogeny.

This chapter corresponds to a published article (Manceau et al., 2015). A first section introduces the paper information. The second to last sections correspond to the published manuscript.
This chapter is closely linked to appendix chapter B, presenting details of formulas and methods used in the main text.

Contents of the chapter

3.1 Article information

Title

Phylogenies support out-of-equilibrium models of biodiversity.

Authors

Marc Manceau[1,2], Amaury Lambert[2,3], Hélène Morlon[1,2]

[1] École Normale Supérieure, Institut de Biologie, CNRS UMR 8197, Paris, France

[2] Collège de France, Center for Interdisciplinary Research in Biology, CNRS UMR 7241, Paris, France

[3] UPMC Univ Paris 06, Laboratoire de Probabilités et Modèles Aléatoires, CNRS UMR 7599, Paris, France

MM, AL & HM designed research, MM, AL & HM performed research, MM & AL contributed analytical tools, MM analyzed the data, MM, AL & HM wrote the paper.

Keywords

Neutral Biodiversity Theory, Macroevolution, Speciation by Genetic Differentiation, Birth-death models, Metacommunity dynamics, Diversification, Speciation, Extinction.

Abstract

There is a long tradition in ecology of studying models of biodiversity at equilibrium. These models, including the influential Neutral Theory of Biodiversity, have been successful at predicting major macroecological patterns, such as species abundance distributions. But they have failed to predict macroevolutionary patterns, such as those captured in phylogenetic trees. Here, we develop a model of biodiversity in which all individuals have identical demographic rates, metacommunity size is allowed to vary stochastically according to population dynamics, and speciation arises naturally from the accumulation of point mutations. We show that this model generates phylogenies matching those observed in nature if the metacommunity is out of equilibrium. We develop a likelihood inference framework that allows fitting our model to empirical phylogenies, and apply this framework to various Mammalian families. Our results corroborate the hypothesis that biodiversity dynamics are out of equilibrium.

Full citation

Marc Manceau, Amaury Lambert, and Hélène Morlon. *Phylogenies support out-of-equilibrium models of biodiversity.* Ecology Letters, 18(4):347–356, 2015.

Aknowledgments

We thank Alexandre Pigot for sharing his data. We also thank Franck Jabot as well as former and current members of the BioDiv team at the ENS for discussions. HM acknowledges the CNRS and grants ECOEVOBIO-CHEX2011 from the French National Research Agency (ANR) and PANDA from the European Research Council (ERC). AL thanks the Center for Interdisciplinary Research in Biology (Collège de France) for funding.

3.2 Introduction

Ever since MacArthur and Wilson (1967) proposed their equilibrium theory of island biogeography, equilibrium models have played a major role in ecology. Of particular influence has been the Neutral Theory of Biodiversity (NTB) (Hubbell, 2001), which simplicity has allowed to analytically derive major macroecological patterns at equilibrium, including the species abundance distribution (Etienne and Alonso, 2005), the species area relationship (O'Dwyer and Green, 2010), and the distance-decay relationship (Chave and Leigh, 2002; O'Dwyer and Green, 2010). The NTB has been relatively successful at predicting realistic macroecological patterns, making this model a central model in ecology (but see e.g. McGill et al. 2006 for a debate on the empirical support of NTB). The theory, however, has been much less successful at predicting realistic macroevolutionary patterns, in particular phylogenetic tree shapes (Davies et al., 2011). At a time when ecologists are increasingly interested in the role of history on present day patterns of biodiversity (Webb et al., 2002; Wiens et al., 2010), in understanding phylogenetic patterns of diversity (Graham and Fine, 2008; Morlon et al., 2011b), and in preserving evolutionary history (Nee and May, 1997; Lambert and Steel, 2013), designing a model of biodiversity predicting realistic phylogenetic trees is critically needed.

There are macroevolutionary models that predict realistic phylogenies (see Morlon 2014 for a recent review). However, most of these models are based on so-called *birth-death models of cladogenesis*, which were historically designed to estimate rates of speciation and extinction in groups where fossil data is scarce (Nee et al., 1992). Since they were first introduced, lineage-based models have been further developed to account for diversity-dependent effects, as well as heterogeneities in diversification rates across time and species groups (Rabosky and Lovette, 2008; Alfaro et al., 2009; Morlon et al., 2010; Etienne et al., 2012; Morlon et al., 2011a; Stadler, 2011; Lambert and Stadler, 2013; Rabosky, 2014). The simplest models, which assume time-constant speciation and extinction rates, produce trees that are more balanced and *tippy* than empirical trees (Blum and François, 2006; Mooers et al., 2007). Realistic balance can be obtained by allowing diversification rates to vary across lineages, while realistic branching times (sensu Morlon 2014) can be obtained by allowing diversification rates to vary through time.

Birth-death models of cladogenesis have tremendous applications for understanding biodiversity patterns (Morlon, 2014). They however have serious limitations. In particular, they consider the *birth* (speciation) and *death* (extinction) events of lineages, or species, ignoring the numbers of individuals constituting these species. By not incorporating population dynamics, they implicitly assume that speciation and extinction events are independent from species' population sizes. However, several lines of evidence suggest that species' abundances and the extant of their geographic range influence probabilities of speciation and extinction (Rosenzweig, 1995). Larger areas likely offer greater opportunities for geographical isolation due to a higher incidence of dispersal barriers, greater habitat heterogeneity, and the limits to gene flow (Pigot et al., 2010). Large ranges also provide a buffer against stochastic or environmentally driven fluctuations in size that may lead to extinction (McKinney, 1997). In addition, it would seem more natural to model extinction by a process in which all individuals die, rather than a process independent of population sizes.

There exists very few evolutionary models that explicitly incorporate population or range sizes and yield predictions for phylogenetic trees (Hubbell, 2001; McPeek, 2008; Pigot et al., 2010). Contrary to traditional lineage-based birth-death models for which likelihood expressions allow parameter inference and model comparison, the phylogenetic trees arising from these models have mainly been investigated with simulations. Parameter inference approaches have been developed only for Hubbell's neutral theory of biodiversity, using approximate Bayesian computation and data on local species abundance and phylogenetic relatedness (Jabot and Chave, 2009). Inference methods for McPeek's model of ecological differentiation (McPeek, 2008) and Pigot's model of geographic speciation (Pigot et al., 2010), which produce trees with realistic branching times, have yet to be developed.

In this paper, we develop a new individual-based neutral model inspired from Hubbell's neutral model. One of the big contributions of Hubbell's model has been to provide a *unified* theory of biodiversity

accounting for both the short time-scale processes of individuals' birth and death, and the long time-scale processes of speciation and extinction. As a result, the theory generates predictions for both macroecological patterns, such as species abundance distributions, and macroevolutionary patterns, such as phylogenies. Our model keeps this same *unifying* particularity, thus also generating both types of patterns. Hubbell's original NTB model relies on three main assumptions. The first one, known as the hypothesis of neutrality, is that individuals behave similarly whichever species they belong to. In the continuity of this hypothesis, we stick to the assumption that individual demographic rates are independent of species identity. The second assumption, known as the zero-sum assumption, is that the metacommunity size is constant. Each death event is assumed to occur simultaneously with a birth event, as in the Moran process of population genetics. In our model, we instead allow metacommunity size to vary according to the stochastic birth and death of individuals. Metacommunity size is not bounded; for example, if the birth and death rates remain constant through time, the metacommunity grows exponentially. Unlike what happens in a metacommunity of constant size where diversity necessarily reaches an equilibrium limit, diversity may not be bounded in our model. Given that previous analyses found little support for equilibrium diversity models in terms of phylogenetic branching times (Morlon et al., 2010), we hypothesized that relaxing the zero-sum assumption could lead to more realistic branching times. The third assumption of NTB is linked to the speciation process. Here, we design a mode of speciation based on gradual genetic differentiation that presents several advantages compared to previously considered speciation modes. We analyze phylogenies arising from this model, provide related likelihood formulas, and apply the model to Mammalian trees. Finally, we discuss the implication of the results for our understanding of biodiversity dynamics.

3.3 The model of Speciation by Genetic Differentiation (SGD)

We consider a model of biodiversity incorporating population dynamics, mutations, and speciation events, thereafter referred to as the model of Speciation by Genetic Differentiation (SGD). This model and the resulting phylogenies are illustrated in Figure 3.1 and summarized in Box 1. Population dynamics are given by a stochastic birth-death process in which individuals give birth and die with rates $b(t)$ and $d(t)$ that are identical across individuals and can potentially vary with time t (see Figure 3.1A). Genetic mutations arise at per-individual rate $\nu(t)$. We derive all our analytical results in the most general case, with the three rates varying through time. In our empirical applications of the model, however, we consider the case of constant rates, denoted b, d, and ν.

Similarly to the infinite-allele model in population genetics, each mutation gives rise to an entirely new genetic type. These mutations are assumed to be neutral, meaning that they do not affect the demography of individuals. We define species as being the smallest monophyletic groups of extant individuals such that any two individuals of same genetic type always belong to the same group. Hence, speciation occurs when two sister populations no longer contain individuals of same genetic type. This typically happens as follows: A first birth event in an ancestral individual generates two descents. At least one individual in either descent undergoes a mutation. Genetic drift makes the two descents fully differentiated (e.g., if there is one mutation, this mutation invades the population by drift) leading to speciation. In the context of sexually reproducing species, mutations can be seen as barriers to hybridization, either prezygotic (e.g. in the form of mechanical, behavioural, or habitat isolation), or post-zygotic (hybrid inviability or sterility). Species are then the smallest monophyletic groups of individuals such that any two individuals that are interfertile always belong to the same group (see Figure 3.1 and Box 1). Finally, SGD naturally includes extinction events, which occur when all individuals of a species die without leaving any descendant. The constant-rate SGD model is thus entirely characterized by only three parameters (b, d, and ν).

As long as the ancestral type is alive in two descents from one ancestral individual, these descents form a single species. Hence, the speciation rate is negatively correlated with the time it takes for the ancestral type to disappear in at least one of the two descents. Therefore, one expects to see different behaviors of the model as a function of a trade-off between population growth rate ($b-d$) and the mutation

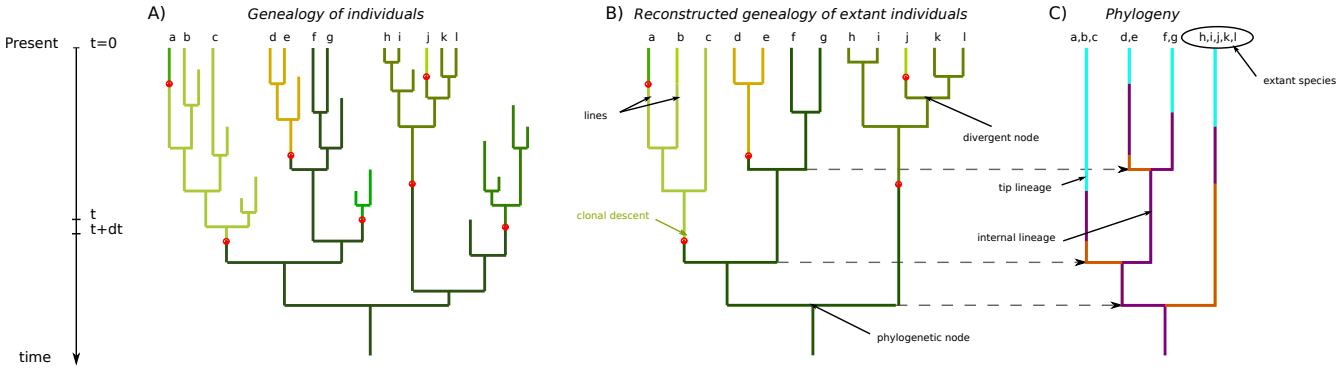


Figure 3.1 – Phylogeny arising from the model of speciation by genetic differentiation (see Box 1 for details). A) Genealogy arising from the stochastic birth and death of individuals. Red dots denote mutations. Each mutation gives rise to a new genetic type (represented with a new color) characterizing the mutated individual and all its descent until the next mutation. B) Resulting reconstructed genealogy of extant individuals, obtained by removing all dead lines from the genealogy. C) Resulting phylogeny, obtained as explained in Box 1. Our derivations and simulations involve defining lineages of different types. Purple lineages are type 0 lineages (extant genetic types that are not frozen), orange lineages are type 1 lineages (extant genetic types) and the blue lineages are *frozen* lineages (lineages that cannot experience further speciation or extinction events). Letters at the tips of the genealogy and phylogeny represent individuals. The set of all extant individuals form the present-day metacommunity.

rate (ν) (see Figure B.1 for illustration). As mutations are drawn following a Poisson process of parameter ν on the reconstructed genealogy, it adds a death rate of parameter ν for clonal lines. A clonal population therefore follows a birth-death process with parameters b and $d+\nu$. When ν is larger than $b-d$, clonal populations have a negative net growth rate ($b-d-\nu$) and thus cannot survive indefinitely (Champagnat and Lambert, 2013). When ν is smaller than $b-d$, however, clones can coexist indefinitely. Hence, there should be a phase transition with long-lived lineages when $b-d > \nu$ and fast lineage turnover when $b-d < \nu$.

The mode of speciation resulting from SGD is more biologically realistic than the point mutation mode of speciation under which most of NTB's previous analytical results have been derived. Contrary to the point mutation model in which speciation happens instantaneously, mutations in the SGD model give rise to new genetic types rather than new species. Speciation thus takes time to complete, as a result of a gradual accumulation of genetic differentiation. In this respect, our model holds some analogies with the protracted mode of speciation introduced by Rosindell and Etienne (Rosindell et al., 2010; Etienne and Rosindell, 2011; Etienne et al., 2014; Lambert et al., 2015). In the protracted speciation model, speciation is modeled as a gradual rather than instantaneous process, such that a population of a new type gives rise to a new species only after a fixed or random time span. In our model, the time span is not an input of the model, but rather arises naturally from the accumulation of mutations. In addition, our model generates monophyletic species which are not clonal (that is, there is genetic diversity within species).

Our primary interest lies in the shape of phylogenies arising from the SGD model. These phylogenies are obtained by a three-step process: first population dynamics generate a stochastic genealogy of individuals (Figure 3.1A); second, mutations arise on the genealogy according to a Poisson process (Figure 3.1A & 3.1B); finally, the phylogeny is a subtree of the reconstructed genealogy obtained according to our species definition, as detailed in Figure 3.1 and Box 1.

3.4 Theoretical results

3.4.1 Key formulas

We aim to analyze phylogenies arising from the SGD model and to develop tools for fitting the model to empirical phylogenies. We measure time from the present to the past, such that $t = 0$ denotes the present and t increases into the past (Figure 3.1). As we show below, we can simulate phylogenies under SGD efficiently (i.e. without simulating the whole individual-based process) and compute associated likelihood formulas using the analytical expressions of two key probabilities. The phylogeny strongly relies on the reconstructed genealogy of individuals, and the two probabilities relate to events happening on this genealogy. The probability of observing a branching event in the reconstructed genealogy between t an $t + dt$ is denoted $g(t)dt$. It corresponds to the probability that an ancestral individual gives birth to two individuals whose descents do not go extinct before the present. The probability that an ancestral individual living at time t has at least one descendant individual at present carrying its genetic type (i.e. there is a descendant line without mutation), conditioned on the survival of at least one descendant, is denoted $m(t)$.

Using results from Kendall (1948), we show (see Supplementary Material in section B) that $g(t)$ and $m(t)$ are given by:

$$g(t) = \frac{b(t)e^{\int_0^t b(z)-d(z)dz}}{1 + \int_0^t b(s)e^{\int_0^s b(z)-d(z)dz}ds} \tag{3.1}$$

$$m(t) = \frac{e^{\int_0^t b(z)-d(z)-\nu(z)dz}}{1 + \int_0^t b(s)e^{\int_0^s b(z)-d(z)-\nu(z)dz}ds} \frac{1 + \int_0^t b(s)e^{\int_0^s b(z)-d(z)dz}}{e^{\int_0^t b(z)-d(z)dz}} \tag{3.2}$$

These probabilities relate to events happening on the genealogies of individuals and therefore do not depend on population sizes. Intuitively, $m(t)$ is an inverse measure of genetic drift and depends on a trade-off between population growth and mutation events.

3.4.2 Simulating phylogenies arising from the model

We show (Supplementary Material) that phylogenies under SGD can be generated by a multi-type branching process with the three following types (see Figure 3.1C).

i) a lineage of type 0 is an extant genetic type. It corresponds to a line from the underlying genealogy that has at least one descendant of same genetic type at present.

ii) a lineage of type 1 is an extinct genetic type. It corresponds to a line from the underlying genealogy that has no descendant of same genetic type at present.

iii) a lineage of type 0 *freezes* when it cannot experience further splitting or extinction events up to the present. This occurs if there exists at least two individuals of same genetic type, one in each of the two descents from the incident node in the underlying genealogy. In this case, the whole descent of this node is collapsed into a single species.

We derive the rates of the following events, at any given time t (Supplementary Material):
A lineage of type 1 becomes of type 0:

$$\rho_{1\to 0}(t) = \frac{\nu(t)m(t)}{1 - m(t)}$$

A lineage of type 1 branches and gives rise to two lineages of type 1 :

$$\rho_{1\to +1}(t) = g(t)(1 - m(t))$$

A lineage of type 0 branches and gives rise to one lineage of type 0 and one lineage of type 1:

$$\rho_{0\to+1}(t) = 2g(t)(1-m(t))$$

A lineage of type 0 freezes, giving rise to a tip lineage in the phylogeny:

$$\rho_{0\to\varnothing}(t) = g(t)m(t)$$

To simulate a phylogeny for a total time duration T, we start with a single lineage which type is 0 with probability $m(T)$ and 1 with probability $1-m(T)$. We then simulate the above events with the corresponding rates, until time $t = 0$ is reached. This provides a very efficient way of simulating phylogenies arising from SGD.

The detailed protocol for these simulations is provided in Supplementary Material.

3.4.3 Computing the likelihood of phylogenies arising from the model

We assume that a clade has evolved according to the SGD model. We allow for the possibility that some extant species are missing from the phylogeny of this clade by assuming that each extant species was sampled with probability f. We derive differential equations governing the probability of observing any ultrametric tree given our model. We obtain a set of coupled differential equations involving the two key functions g and m, that can be computed analytically in the case $f = 1$, and integrated numerically in the case $f < 1$. This allows us to follow a natural *peeling algorithm* (Felsenstein, 1981), which consists in computing recursively the likelihood of a tree, by decomposing it into subtrees until finding tip lineages. Likelihoods satisfy ordinary differential equations that we solve using numerical integration. The differential equations, analytical solutions and details of the algorithm are given in Supplementary Material.

3.4.4 Estimating the parameters of the model

Given a phylogeny, the parameters of the model can be estimated by maximum likelihood. To test the ability of the approach to recover the true parameters, we simulated phylogenies under a wide range of parameter values, and applied our maximum likelihood inference algorithm. We found that the approach performs well to recover the net growth rate $b-d$ and the mutation rate ν (Figure 3.2). Estimates of b alone are biased, due to the fact that this parameter has a weak influence on the likelihood surface.

Codes for the simulations, likelihood computations and parameter estimations are available in Python from the authors upon request. They are also implemented in the R package RPANDA (Morlon et al., 2015).

3.5 Empirical results

3.5.1 Phylogenies arising from the model have realistic balance and branching times

To test whether our model produces realistic trees, we analyzed the branching times and balance of both simulated and empirical trees. We use the γ statistic (Pybus and Harvey, 2000) to measure branching times. This statistic reflects the relative position of nodes in a phylogeny: stemmy phylogenies (i.e. phylogenies with many nodes close to the root) are characterized by negative γ values, while tippy ones are characterized by positive γ values. We use the β statistic (Blum and François, 2006) to measure phylogenetic balance, computed by maximum likelihood using the R package *apTreeshape*.

We begin by evaluating how each of the three parameters of the time-constant SGD model influences phylogenetic trees. To do this, we vary each parameter while constraining the others (Figure 3.3).

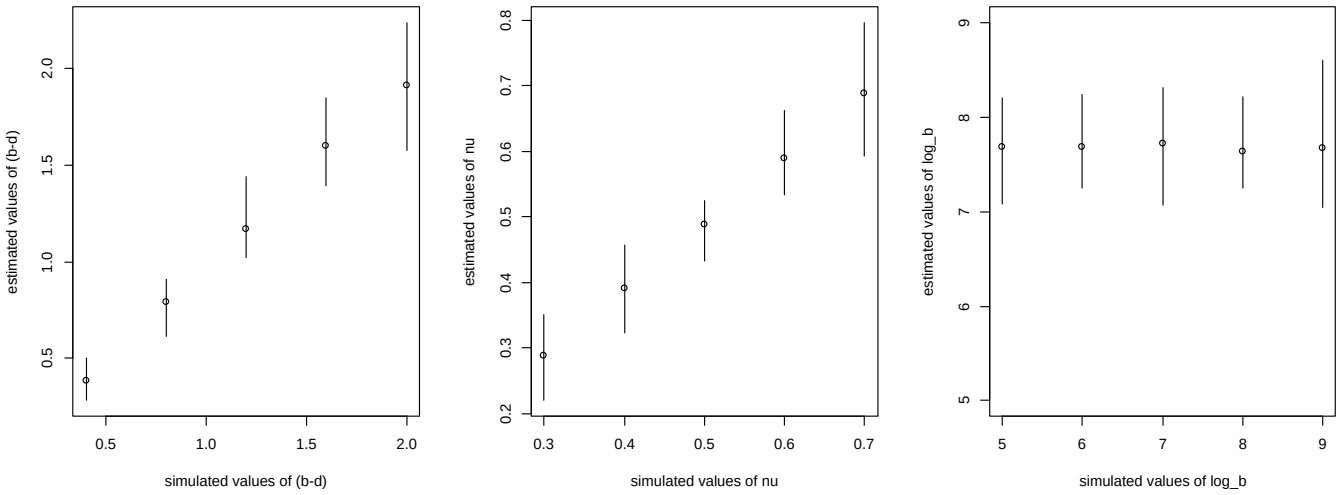


Figure 3.2 – Growth rates and mutation rates can be robustly inferred from molecular phylogenies, but not birth rates. The figure shows maximum likelihood parameter estimates for phylogenies simulated under different parameter sets. The true, simulated parameters are indicated on the x axis, while inferences are indicated on the y axis (expressed in number of events per time unit). Points and error bars indicate the median and 95% quantile range of the maximum likelihood parameter estimates. Left panel: estimates of $b - d$ are unbiased ($b = 10^6$ and $\nu = 0.5$). Middle panel: estimates of ν are unbiased ($b = 10^6$ and $b - d = 0.8$). Right panel: estimates of b are biased ($b - d = 0.8$ and $\nu = 0.5$). Units are expressed in Myr^{-1}.

These analyses confirm that tree shape is principally constrained by a balance between the population growth rate $b-d$ and the mutation rate ν (Figure B.1). The higher the mutation rate ν at $b-d$ constant, the higher γ and β, meaning trees tend to be tippy and balanced. On the contrary, higher $b-d$ values at ν constant lead to lower γ and β, that is, stemmy and unbalanced trees. The parameter b alone has little if any impact on both β and γ, thus explaining why there is little signal in phylogenies to infer this parameter.

We compare the γ and β values of simulated phylogenies to the γ and β values of the 84 empirical binary trees from McPeek's repository with more than 10 species (McPeek, 2008). We considered only trees with more than 10 species because the variance in β values increases very rapidly for trees of small size, and thus estimates of β can be inaccurate for small trees (Blum and François, 2006). We find that the SGD model can generate trees with a wide range of β and γ values, including those of empirical trees (Figure 3.3). The model produces trees with levels of imbalance and branching times similar to those observed in nature, but only when the growth rate is sufficiently large (of the same order of magnitude as ν), meaning in out-of-equilibrium dynamics.

3.5.2 Fit to Mammalian phylogenies

We infer the parameters of our model for fourteen Mammalian phylogenies used by Pigot et al. (2012). For each fit, the value of f is fixed, computed by dividing the number of species in the tree by the total number of known species in the clade. Figure 3.4 shows the likelihood surface corresponding to four of these phylogenies, which are typical of likelihood surfaces obtained for the data. These likelihood surfaces confirm that there is no ambiguity in finding the maximum likelihood parameters, with a single, well defined peak, and no local optima. We do not report estimates of b, which showed an order of magnitude of difference across clades, confirming that phylogenetic data is not useful for estimating this parameter.

Interestingly, we find parameter estimates for $b-d$ and ν that are rather consistent across Mam-

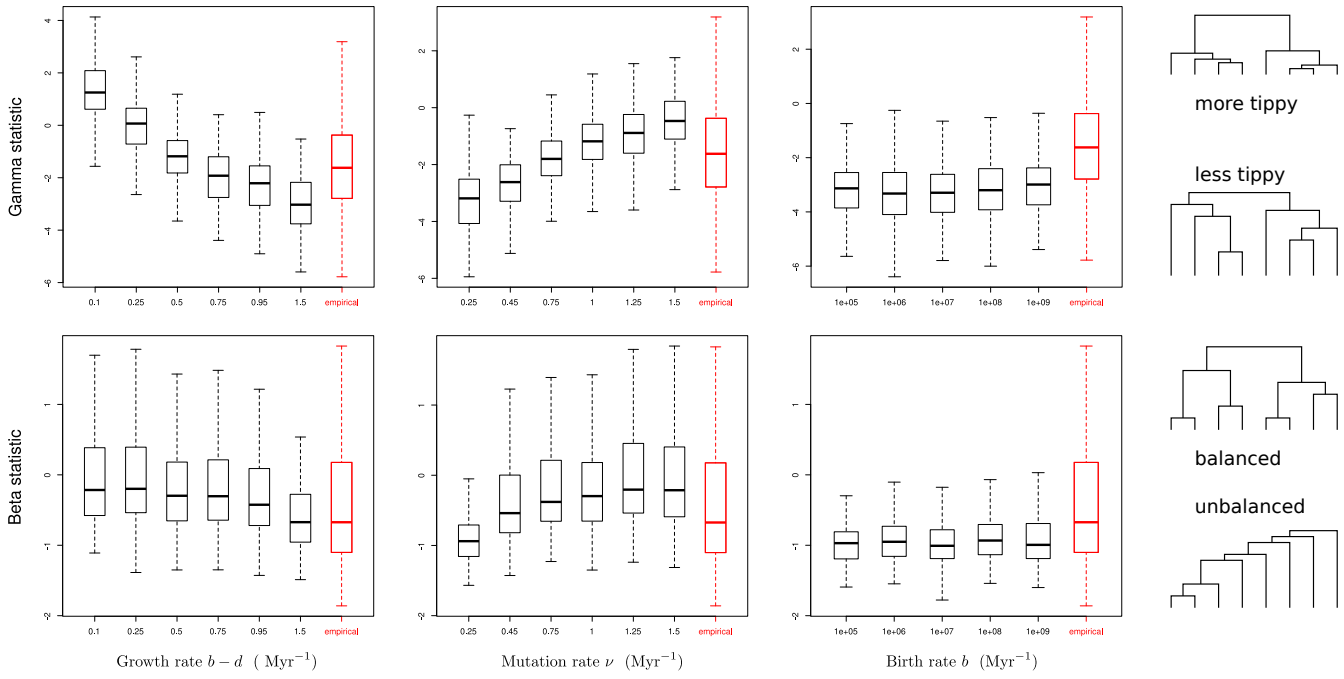


Figure 3.3 – Branching times and balance under the model of speciation by genetic differentiation. First column : high growth rates $b - d$ at constant birth and mutation rates lead to phylogenies that are stemmy and unbalanced ($b = 10^6$, $\nu = 1$). Second column : high mutation rate ν at constant birth and growth rates ($b = 10^6$, $b - d = 0.5$) lead to phylogenies that are tippy and balanced. Third column : the birth rate b has little effect on phylogenies at constant growth and mutation rates ($b-d = 1$, $\nu = 0.5$). Units are expressed in Myr^{-1}. Each box-plot summarizes results for 200 simulated phylogenies. Empirical box-plot corresponds to the 84 binary phylogenies in the McPeek repository comprising more than 10 species.

malian groups (Table 3.1). The mutation parameter is constrained to a narrow range (from 0.16 to 0.39 Myr^{-1}) for 11 out of 14 phylogenies; only three outliers (Calomys, Microtus, and Macaca) have higher values (up to 1.72 for Calomys). The range is slightly broader for the growth rate (from 0.05 to 0.53 Myr^{-1}), with only one outlier (1.84 for Microtus). In agreement with results presented above, we find that the estimated growth rate is of the same order as the mutation rate (slightly higher for exactly half of the phylogenies, and slightly lower for the other half). Hence, the growth rate is far from being null, suggesting that the metacommunity is not at equilibrium.

3.6 Discussion

We developed a neutral, out-of-equilibrium model of biodiversity that produces realistic phylogenetic trees. We developed a fast simulation algorithm for this model, as well as a method of inference that allows fitting the model to empirical data efficiently. We illustrated our method using fourteen Mammalian phylogenies. Our results corroborate the hypothesis that phylogenies are better explained by out-of-equilibrium models of biodiversity.

Our model can be seen as an extension to the neutral theory of biodiversity. The first main difference lies in the speciation process. Similarly to the point mutation model, speciation arises as a result of mutation events. However, contrary to the original point mutation model, a single mutation is typically not enough to induce speciation. The SGD process leads to the split of an ancestral species' population into two daughter species. In this respect, it can be seen as providing a mutational basis to the

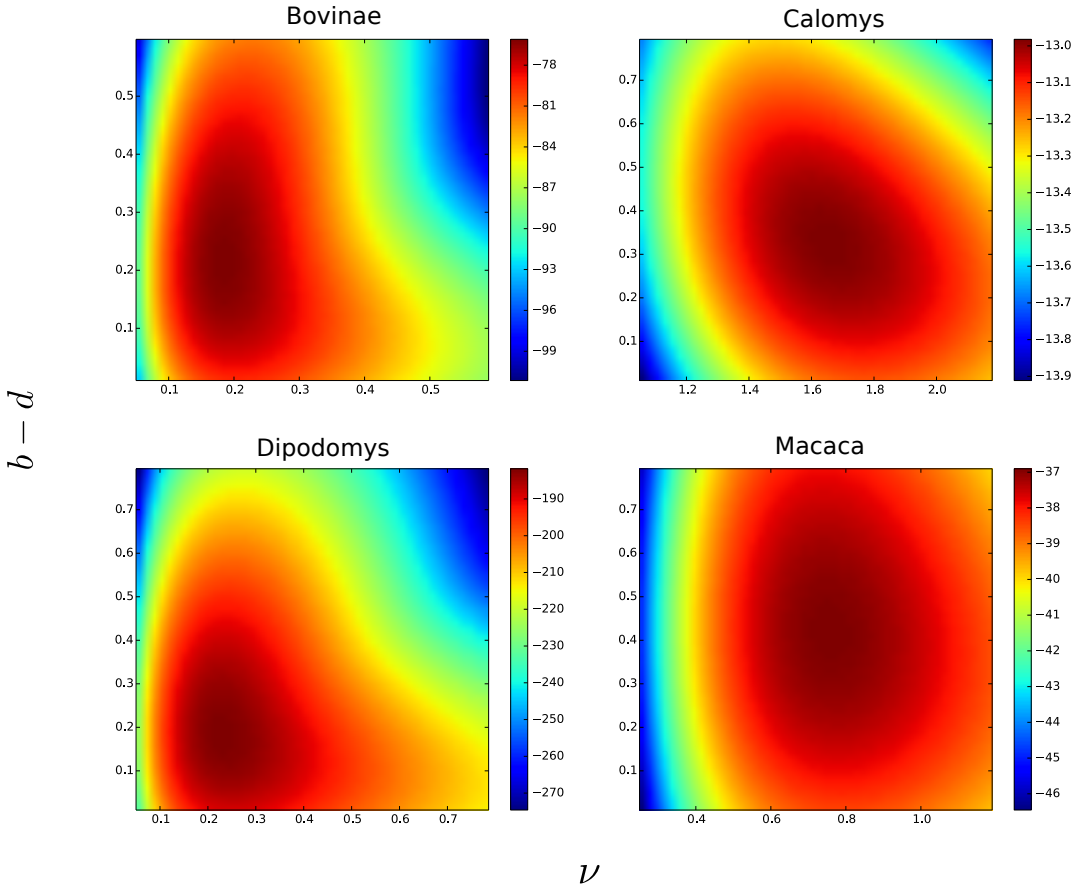


Figure 3.4 – Growth rates and mutation rates under SGD estimated for four Mammalian phylogenies. Colors correspond to likelihood values. The likelihood landscapes have a single peak, demonstrating the ability to infer the parameters of SGD from phylogenies.

random fission mode of speciation. The SGD model also offers a good microscopic basis to the hypothesis of *protracted speciation* by which there is a time lag between the initiation of population divergence and the time when gene flow completely stops and distinct species are recognized (Coyne and Orr, 2004; Rosindell et al., 2010; Etienne and Rosindell, 2011). Indeed, divergence starts in SGD with a mutation event, but a new species is formed and recognized only if (and after) enough mutations have accumulated.

The second main difference between our model and the classical neutral theory of biodiversity is that we relax the constant metacommunity size hypothesis. We find that when the growth rate $b - d$ decreases, meaning that the metacommunity is close to equilibrium, phylogenies become unrealistic in terms of branching times. This confirms results from the classical equilibrium NTB model showing that realistic branching times are hardly ever obtained (Davies et al., 2011).

Here, non-equilibrium refers to a growing metacommunity size, as opposed to a constant (equilibrium) metacommunity size. Data on metacommunity sizes over evolutionary time scales is generally missing. However, a growing metacommunity size seems more realistic than a constant one at time scales spanning the history of entire clades. Indeed, the number of individuals in radiating clades has to have increased lastingly during diversification, during expansion phases corresponding (for example) to the colonization of new territories. Still, the exponential growth model considered here is simplistic with regards to the complex history of clades. More complex scenarios with time-inhomogeneous rates could be analyzed with our framework. We are hopeful that such developments could allow us to detect broad trends in the way metacommunity size varies over long time scales.

Clade	f	Clade age (Myr)	N	β	γ	$b-d$	ν	$b-d-\nu$
Bovinae	1	19,58	25	-1,28	-1,2	0.19	0.16	0.03
Calomys	0.85	2,99	13	-1,77	0,58	0.45	1.72	-1.27
Caprinae	0.89	9,94	38	-1,06	-1,38	0.40	0.39	0.01
Dasyuridae	0.92	29,5	72	-0,52	-5,22	0.20	0.23	-0.03
Dipodomys	0.95	20,66	65	-0,33	-2,41	0.05	0.30	-0.25
Duikers	0.83	17,16	58	-1,03	-0,79	0.38	0.30	0.08
Viverrinae	0.88	14,92	25	-1,55	0,64	0.36	0.31	0.05
Hylobatidae	1	8,81	14	-1,57	-1,47	0.53	0.32	0.21
Alouatta	0.91	16,46	14	-0,7	-0,42	0.20	0.39	-0.19
Macaca	0.95	5,8	22	0,9	-1,45	0.42	0.76	-0.34
Microtus	0.69	4,08	154	-0,37	-5,33	1.84	1.05	0.79
Mustelidae	0.85	22,53	59	-1,28	-1,65	0.38	0.22	0.16
Ochotona	0.92	13,69	39	-0,5	-0,89	0.25	0.35	-0.10
Talpa	0.77	26,85	35	-0,99	-1,58	0.16	0.21	-0.05

Table 3.1 – Parameters of the SGD model inferred for various clades of mammals. f : sampling fraction; clade age: crown age; N : number of extant species in the clade (not all species are sampled); the inferred parameters $b-d$ and ν are expressed in Myr^{-1}.

Our results refer to equilibrium in terms of metacommunity size, not diversity. Still, it is more likely that diversity reaches equilibrium when metacommunity size reaches equilibrium than when metacommunity size is expanding. This confirms earlier results stemming from lineage-based models that have found a better support for non equilibrium models compared to stationary ones (Hey, 1992; Morlon et al., 2010), and suggest that non equilibrium models should be preferred over equilibrium models to predict the loss of evolutionary history due to species loss (Nee and May, 1997) as well as the way phylogenetic diversity scales spatially (Morlon et al., 2011b).

Applying our model to Mammalian phylogenies, we found rather consistent parameter estimates across groups. As expected, demographic parameter estimates ($b-d$) were more heterogeneous than mutational parameter estimates (ν). In our results, the two clades showing the highest values of inferred ν were Microtus and Calomys. These clades also had high values of $b-d$ compared to other clades, such as Bovinae, Dipodomys and Talpa. Microtus and Calomys are small rodent species, indeed known to have a high reproduction rate (Golley et al., 1975), and small generation times potentially leading to high mutation rates. We do not see any other obvious history traits that could explain differences across groups in terms of growth and mutation rates. It would be interesting to compare our growth estimates to effective population size curves obtained from genetic data, although such data for entire clades are not yet available. We could also look at these growth estimates in light of the age and current global population sizes of clades. Estimates for $b-d$ may seem low (in the order of one event per Myr) in comparison with the usual instantaneous growth rate of population dynamics. At the time scales considered here, the growth rate $b-d$ reflects the long-term growth of the entire metacommunity, that is, an average trend rather than the fast oscillating dynamics of populations. Similarly, estimates for the mutation rate ν may seem low in comparison with the usual genomic mutation rate, and high in comparison to estimates of the point-mutation speciation rate in NTB (Condit et al., 2002). However, mutation rate in SGD model refers to mutations that have an effect on speciation. They represent only a small fraction of the mutations arising on a DNA sequence, leading to values much lower than genomic mutation rate. As a high number of these mutations do not directly lead to speciation, SGD's mutation rate is indeed expected to be larger than NTB's mutation rate. Finally, the mutation rate we infer is orders of magnitude lower than reasonable birth rates, which shows that the mutations playing a role in speciation in SGD arise in a much slower timescale than the one of population dynamics. Typical values of b (e.g., 10^5 Myr^{-1}) yield a ratio of

birth to mutation in the order of a population size, which is in line with the traditional assumption in population genetics.

Our study provides an alternative to previous interpretations of patterns observed in empirical trees, in terms of both branching times and balance. Negative γ values have traditionally been interpreted as the effect of adaptive diversification (Phillimore and Price, 2008; Rabosky and Lovette, 2008), biogeographical processes (Pigot et al., 2010), or protracted speciation (Rosindell et al., 2010; Etienne and Rosindell, 2011) (see Moen and Morlon 2014b for a review). Here, we show that a non-adaptive, non-spatial model can explain the branching times of real trees. In our model, stemmy phylogenies arise both from the chosen mode of speciation, which naturally accounts for protractedness, and as a result of an expanding metacommunity. Phylogenetic imbalance suggests that some groups of organisms are more species rich than others. This variation in species richness across taxonomic groups has traditionally been interpreted as evidence that non-neutral, ecological differences among lineages drive differences in speciation and extinction rates (Alfaro et al., 2009). In agreement with previous studies (Jabot and Chave, 2009; Pigot et al., 2010; Davies et al., 2011), our analyses demonstrate that the levels of phylogenetic imbalance observed in nature can arise from purely neutral processes. In the NTB with point mutation, phylogenetic imbalance arises as a by-product of stochastic differences in population sizes (per-lineage speciation rate is a linear function of abundance). Similarly, in Pigot's biogeographic model (2010), which is also neutral, phylogenetic imbalance arises from stochastically driven differences in range sizes (species with wider ranges are more likely to experience vicariance events). In our model, however, the link between speciation rates and abundance is not as straightforward. On the one hand, abundant species *see* more mutations, which could promote speciation. On the other hand, the ancestral type survives longer in rapidly expanding populations, such that speciation may become more difficult. Another explanation for imbalance in phylogenetic trees is differences in diversification linked to a heritable trait (Heard, 1996). In general this has been interpreted as different abilities for species with different ecological characteristics to speciate and/or go extinct. Here the model is neutral, meaning individuals across species all have the same birth, death, and mutation rates. However, the process of speciation generates differences across species in terms of the interconnection of individuals through potential hybridization. Some species are big hubs (those where the ancestral type has not disappeared) that do not easily speciate, while others (those where the ancestral type has disappeared) speciate more easily. This hidden trait is heritable, generating imbalance without invoking ecological differences between species.

While our study provides an alternative to previous interpretations of patterns observed in empirical trees, assessing the goodness of fit of our models compared to other models is not yet possible. Such comparisons cannot currently be performed, as we are lacking a robust approach for fitting models such as the adaptive (McPeek, 2008) and the biogeographic (Pigot et al., 2010) models to phylogenetic trees. However, we could use the framework presented here to compare the fit of models with constant birth and death rates, leading to an exponentially growing metacommunity, to that of models with time-varying growth rates. We could consider models with population-level density-dependence that could lead to clade-wide diversity-dependence (Phillimore and Price, 2008; Rabosky and Lovette, 2008; Etienne et al., 2012). This would provide a (non-adaptive) diversity-dependent model certainly worth exploring. We could also consider models in which the metacommunity net growth rate switches from positive (expanding metacommunity) to negative (shrinking metacommunity) along history, which could result in periods of diversity expansion followed by diversity decline (Morlon et al., 2011a; Quental and Marshall, 2013; Morlon, 2014). It would also be particularly interesting to consider a spatial version of the model accounting for dispersal limitation (MacArthur and Wilson, 1967; Etienne and Alonso, 2005; Jabot and Chave, 2009; Rosindell and Phillimore, 2011), which would allow us to fit the model to a much broader array of datasets at the community-scale.

An interesting aspect of our model is that it not only produces predictions for macroevolutionary patterns (phylogenies), but also for macroecological patterns (species abundance distributions). We have not yet fully explored the shape of species abundance distributions arising from SGD, but preliminary results suggest that the model can produce shapes covering the classical log-series and log-normal shapes

depending on the choice of the parameters.

3.7 Conclusion

Our study is one of the first attempts at proving analytical solutions for phylogenies arising from an individual-based model. Further work in this direction will be clearly needed for a better integration of macroevolution into macroecology and community ecology. Importantly, we showed that considering out-of-equilibrium models will be crucial to this integration. In macroevolution, out-of-equilibrium models are the norm, but they had not been previously linked to non-equilibrium metacommunity sizes. Our framework provides perspectives for better understanding how diversity dynamics relate to metacommunity dynamics. In macroecology and community ecology, our results call for a major shift from our current focus on steady state predictions to a focus on transient dynamics.

Box 1: From the genealogy of individuals to a species-level phylogeny in the model of speciation by genetic differentiation.

The process of speciation by genetic differentiation (SGD) and the resulting phylogeny are illustrated in Figure 3.1. Here we clarify the terms used throughout and how phylogenies are obtained from the underlying genealogies of individuals.

Genealogy of individuals The *genealogy* is the tree representing ancestor-descendant relationships for all individuals arising from the individual-based birth-death process. A *line* is a path on the genealogy joining an ancestral individual at the base to its successive descendants. The *reconstructed genealogy* is the genealogy in which extinct lines have been removed. The *descent* of an individual is the genealogical subtree of all individuals descending from this ancestral individual (on the genealogy or the reconstructed genealogy). The *metacommunity* is the set of all individuals alive at a given time.

Mutations A mutation event on the genealogy is an event that changes the *genetic type* of the mutant individual and all its descent. The *clonal descent* of an individual is the set of all individuals of same genetic type descending from this ancestral individual. In Figures 3.1A) and 3.1B), each clonal descent of a mutant is represented with a different color. We need to consider *divergent nodes* on the reconstructed genealogy, defined as nodes such that any two pairs of individuals picked at random, one in each of the two extant descents from the node, are of different genetic types.

Species A *species* is the smallest monophyletic group of extant individuals such that any two individuals of same genetic type always belong to the same group. This ensures consistency between genealogical and phylogenetic relationships, in agreement with the genealogical concordance species concept (Avise and Ball, 1990).

Phylogenies The reconstructed phylogeny (simply called *phylogeny* throughout) is the tree representing the evolutionary relationships between extant species. A *phylogenetic node* is a node that appears on the reconstructed phylogeny. Given our monophyletic definition of species, we can take phylogenetic nodes as subsets of genealogical nodes. Phylogenetic nodes are obtained recursively as follows. The oldest node in the reconstructed genealogy is phylogenetic if it is divergent (otherwise the phylogeny is made of a single extant species). Each other genealogical node is phylogenetic if its parent node is phylogenetic and if it is divergent. *Branching times* are the times when there is a phylogenetic node. A *lineage* is a path on the phylogeny (in contrast to a *line*, which is a path on the genealogy). A *tip lineage* is a lineage joining an extant species to its most recent ancestral node. It can correspond to one or several genealogical lines. All other lineages are *internal lineages* and correspond to a single line in the genealogy.

Lineage types A *type 0 lineage* is a lineage starting from an ancestral individual which clonal descent survived to the present. A *type 1 lineage* is a lineage starting from an ancestral individual which clonal descent did not survive to the present. A *frozen lineage* is a lineage that cannot experience further speciation or extinction events. This happens when the ancestral individual at the base of the lineage gives birth to two individuals with both clonal descents surviving up to the present.

Chapter 4

Integrating species interactions into models of phenotypic evolution

The diversification of species studied in chapters 2 and 3 is what ultimately generates present-day patterns of species richness; but another important aspect of biodiversity is phenotypic disparity. We now turn to the study of patterns of species phenotypic differentiation through time, which we will link to diversification studies only in the last chapter 6.

Continuous phenotypes, such as body mass, leg length, brain size, can be compared and studied among phylogenetically related organisms using so-called *Phylogenetic Comparative Methods* (PCMs). These methods commonly rely on diffusion processes unfolding along a known phylogeny, as presented in section 1.2.2. They allow one for example to address questions on the correlated evolution of distinct traits.

Importantly, most of these PCMs assume that traits evolve independently on two sister branches following a branching event. As a result, the method retains the phylogeny as the only factor explaining the proximity between tip phenotypes. In this chapter, we will relax this strong assumption and propose instead an unusual hypothesis where interactions between species living at the same time can also impact the phenotype evolution. We present an inference procedure that allows us to fit models including interactions among lineages. In particular, we helped Drury et al. (2016) to carry out an empirical application of such a model, which enabled to assess the importance of between lineages competition in driving some traits' evolution.

This chapter corresponds to our published work on models of continuous phenotype evolution (Manceau et al., 2017). The first section presents the article information. The second to last sections are the core of the article.
This chapter is linked to appendix chapter C, presenting details of the methods, together with a presentation of codes used for numerical computation. We additionally provide an example of empirical application by reproducing the article of Drury et al. (2016) in appendix E.

Contents of the chapter

4.1 Article information

Title

A Unifying Comparative Phylogenetic Framework Including Traits Coevolving Across Interacting Lineages.

Authors

Marc Manceau [1,3,4], Amaury Lambert[2,3], Hélène Morlon[4]

[1] Muséum National d'Histoire Naturelle, 75005 Paris, France;

[2] Laboratoire Probabilités et Modèles Aléatoires, UPMC - University of Paris 06, 75005 Paris, France;

[3] Center for Interdisciplinary Research in Biology, Collège de France, CNRS UMR 7241, 75005 Paris, France;

[4] Institut de Biologie de l'École Normale Supérieure, CNRS UMR 8197, 75005 Paris, France.

Keywords

Comparative phylogenetics, trait evolution, coevolution, interspecific interactions, character displacement, linear stochastic differential equations

Abstract

Models of phenotypic evolution fit to phylogenetic comparative data are widely used to make inferences regarding the tempo and mode of trait evolution. A wide range of models is already available for this type of analysis, and the field is still under active development. One of the most needed developments concerns models that better account for the effect of within- and between-clade interspecific interactions on trait evolution, that can result from processes as diverse as competition, predation, parasitism, or mutualism. Here, we begin by developing a very general comparative phylogenetic framework for (multi)-trait evolution that can be applied to both ultrametric and non-ultrametric trees. This framework not only encapsulates many previous models of continuous univariate and multivariate phenotypic evolution, but also paves the way for the consideration of a much broader series of models in which lineages co-evolve, meaning that trait changes in one lineage are influenced by the value of traits in other, interacting lineages. Next, we provide a standard way for deriving the probabilistic distribution of traits at tip branches under our framework. We show that a multivariate normal distribution remains the expected distribution for a broad class of models accounting for interspecific interactions. Our derivations allow us to fit various models efficiently, and in particular greatly reduce the computation time needed to fit the recently proposed phenotype matching model. Finally, we illustrate the utility of our framework by developing a toy model for mutualistic coevolution. Our framework should foster a new era in the study of coevolution from comparative data.

Full citation

Marc Manceau, Amaury Lambert, and Hélène Morlon. *A unifying comparative phylogenetic framework including traits coevolving across interacting lineages.* Systematic Biology, 66(4):551-568, 2016.

Funding

MM acknowledges his PhD funding from the École Normale Supérieure; AL acknowledges funding from the Center for Interdisciplinary Research in Biology (Collège de France); HM acknowledges funding from ERC, grant ERC-CoG PANDA.

Acknowledgements

We are very grateful to Jonathan Drury, Florence Débarre and Luke Harmon, as well as members of our research teams for helpful discussions and comments on previous versions of the manuscript. We also aknowledge three reviewers for their helpful comments that helped us improve this paper.

4.2 Introduction

Evolutionary biologists have long been interested in the long-term evolution of phenotypic traits (Simpson, 1944). In 1973, Felsenstein introduced one of the first models of phenotypic evolution, with the initial goal to account for shared ancestry when testing for statistical correlation between pairs of traits in extant species. In this founding paper, Felsenstein proposed that a one-dimensional quantitative trait evolving on a tree could be modeled as a Brownian process that splits into two independent Brownian processes at branching times. This model mimics a trait that would evolve as a mere effect of stochastic drift; it is now often used as a null model, but also to estimate the relative lability (or rate of evolution) of various traits in a given group of organisms or of a given trait across different groups of organisms (Thomas et al., 2006; Harmon et al., 2010).

Since these early developments, evolutionary biologists have designed a series of models to better understand the evolutionary processes that shape phenotypic evolution (see Pennell and Harmon, 2013 for a review). The Ornstein-Uhlenbeck (OU) process has been proposed to model evolution under stabilizing selection, i.e. with a selective pressure pushing trait values toward a given optimum (Hansen, 1997; Hansen and Martins, 1996; Butler and King, 2004). The ACDC model has been proposed to account for accelerating (AC) or decelerating (DC) rates of phenotypic evolution through time (Blomberg et al., 2003). The latter scenario, where the evolutionary rate is high early in the history of a clade and subsequently declines toward the present, well known as the early burst (EB) model, has often been used to test support for adaptive radiation theory (Harmon et al., 2010; Moen and Morlon, 2014a). These univariate models representing the evolution of a single trait have been extended to multivariate models representing the simultaneous evolution of multiple traits, thus allowing to directly test hypotheses about the coevolution between several phenotypic traits (Hansen et al., 2008; Bartoszek et al., 2012; Jhwueng and Maroulas, 2014). Other extensions have been developed to account for variations in model parameters across clades (O'Meara et al., 2006; Revell and Collar, 2009; Eastman et al., 2011; Butler and King, 2004; Beaulieu et al., 2012). Finally, some of these models have been developed in the context of phylogenies including fossil data (i.e. non ultrametric trees, see Ruta et al., 2006; Slater, 2015) in addition to phylogenies with only extant taxa (i.e. ultrametric trees). Most of these models have been implemented in open-access packages (Martins, 2004; Harmon et al., 2008; Butler and King, 2004; Thomas and Freckleton, 2012; Clavel et al., 2015; Morlon et al., 2015), allowing their application to a broad variety of questions and datasets (see, e.g. Labra et al., 2009; Mahler et al., 2010; Dale et al., 2015; Quintero et al., 2015; Slater, 2015).

Despite these developments, most currently available models ignore the effect of interspecific interactions on trait evolution. Given the importance of species interactions in classical evolutionary theories, such as Simpson's adaptive radiation (Simpson, 1944), Ehlrich & Raven's escape and radiate (Ehrlich and Raven, 1964) and Van Valen's Red Queen (Van Valen, 1973) theories, building models that better account for such interactions is fundamental. In a first attempt to take into account the role of competition for niche space on character evolution, a diversity-dependent (DD) model has been introduced, where the rate of phenotypic evolution declines as the number of lineages in the clade increases (Mahler et al., 2010; Weir and Mursleen, 2013). While this model represents an important first step, it still assumes that trait changes in one lineage are independent from the value of traits in other, interacting lineages, therefore ignoring the widespread idea of trait- (or ecologically-) driven interspecific interactions. More recently, the phenotype matching (PM) model relaxed these hypotheses and more explicitly accounted for interspecific interactions by modeling either the attraction or the repulsion of traits from a clade-wise average trait value. In the first case, referred to as matching mutualism, species traits tend to converge to similar values,

whereas in the second case, referred to as matching competition, species traits tend to diverge (Nuismer and Harmon, 2014; Drury et al., 2016).

The comparative phylogenetic approach developed by Drury et al. (2016) allows fitting a version of the model introduced by Nuismer and Harmon (2014) where the evolution of trait values in one lineage is influenced by the trait values of other lineages. This approach focused on the evolution of traits within one clade. While within-clade interactions can be particularly relevant for some types of interactions (e.g. in the case of competitively driven character displacement, Brown and Wilson, 1956), the effect of other types of antagonistic or mutualistic interactions on trait evolution is often most relevant between distantly related species. For example, host-parasite interactions are thought to drive a coevolutionary race between traits involved in host defence and parasite ability to infect (e.g. the production rate of a toxic compound v.s. the inhibitory concentration of parasite growth). Similarly, prey-predator interactions may lead to the coevolution of prey traits involved in camouflage, repulsion, or escape strategies, together with predator traits involved in the ability to detect and capture its prey (e.g. escape speed v.s. hunting speed) (Ehrlich and Raven, 1964; Dawkins and Krebs, 1979). Mutualistic plant-pollinator interactions also are thought to drive the coevolution between plant traits involved in pollen accessibility or flower attractiveness to their pollinator (secondary metabolites, floral traits), and pollinator traits involved in the ability to detect suitable plants and to exploit plant rewards (Fenster et al., 2004; Weiblen, 2004; Sletvold et al., 2016). While these types of biotic interactions likely play a key role in trait evolution and have been crucial in the development of coevolutionary theories (Ehrlich and Raven, 1964; Van Valen, 1973), there currently exists no framework for fitting models of phenotypic evolution incorporating the effect of clade-clade interactions.

The current paper expands the work of Bartoszek et al. (2012) who presented a unified framework for studying coevolving traits in independently evolving lineages by providing a unified framework for coevolving traits in coevolving lineages. Throughout the paper, the term coevolution refers to the evolution of traits on fixed phylogenies, i.e. without any effect on the speciation-extinction dynamics. In short, we aim to provide the phylogenetic comparative tools allowing to study co-evolutionary scenarios such as the one depicted on Figure 4.1, where a plant trait and a pollinator trait coevolve as a result of mutualistic interactions. We focus on quantitative rather than discrete traits (Pagel, 1994) and on gradual rather than punctuated evolution (Bokma, 2002, 2008; Landis et al., 2013; Bartoszek, 2014). Our framework, based on linear stochastic differential equations (SDE), encompasses all models of continuous (multi-)trait evolution mentioned above, and allows the treatment of a broad set of co-evolutionary models. We show that the tip trait distribution under all these models is Gaussian, and we highlight a general procedure that allows one to compute its expectation and covariance structure. We also provide analytical developments that speed up the computation of the likelihood in comparison with the general procedure. This leads for example to a faster algorithm for the likelihood computation of the PM model. The goal of the paper is two-fold: first, by providing general solutions to the distribution of traits at tip branches under our unified framework, we hope to help users find their way in a dense and potentially overwhelming literature; second, by showing how the framework can be used to treat a broad class of within-clade and clade-clade coevolutionary scenarios, we hope to foster the development of models to test long-standing hypotheses on the role of competition, predation, parasitism and mutualism in evolution.

We begin by presenting our framework and showing how previous models as well as novel clade-clade coevolutionary models fit within this framework; next, we provide general solutions for the distribution of tip trait values under this framework; then, we illustrate how the framework can be used to formalise and study a toy model of clade-clade coevolution.

4.3 A general framework for phenotypic evolution

We introduce a general formalism to study (multi-)trait (co)evolution when the interaction between distinct phenotypic traits, distinct lineages within a clade, and/or distinct lineages among several clades potentially affects how phenotypic traits evolve.

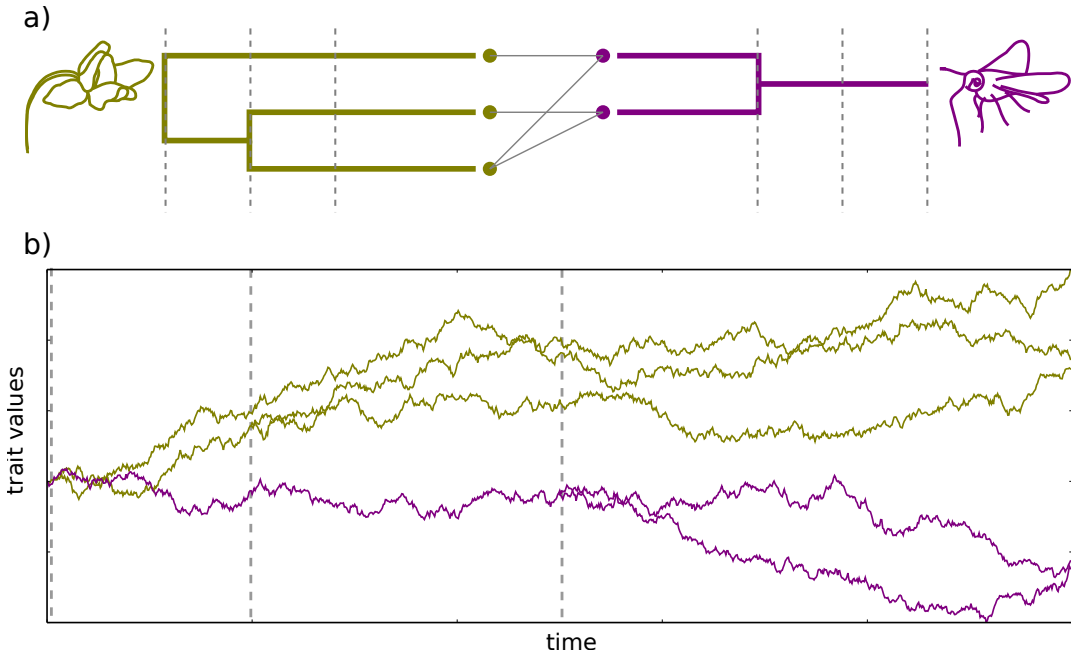


Figure 4.1 – Phenotypic evolution in coevolving clades. We aim at developing a framework for analyzing how traits evolve as a response of potentially complex ecological interactions. Here, the mean proboscis length of pollinators and the mean floral tube length of pollinated flowers coevolve as a result of clade-clade mutualistic interactions.

4.3.1 Trait evolution through time

We begin by considering the evolution of n traits at a given time t; these can for example represent n distinct traits evolving on a single lineage or a single trait evolving in n lineages. We denote by X_t the column vector of the n trait values at time t. Throughout the paper we assume that the evolution of the traits is driven by a linear stochastic differential equation of the form:

$$\begin{cases} dX_t &= (a(t) - AX_t)dt + \Gamma(t)dW_t \\ X_0 &= x_0 \end{cases} \tag{4.1}$$

where a is a vector of $\mathbb{R}^n$ whose coefficients can vary with time, A is a constant square matrix of size $\mathbb{R}^n \times \mathbb{R}^n$, Γ is a square matrix of the same size whose coefficients can vary with time, and W_t is a n-Brownian motion (i.e. a vector composed of n independent standard Brownian motions). Schematic examples are presented in Figure 4.2.

The formulation above implies that we consider interaction effects that operate gradually in time rather than as punctuated events. Intuitively, the deterministic part $(a(t) - AX_t)dt$ reflects the direct effects of trait values on the evolution of these traits, including the effect of a trait value in one lineage on both its own evolution (as in the OU process) and the evolution of traits in the other lineages (as in the PM process). The stochastic part $\Gamma(t)dW_t$ reflects drift and the environmental noise influencing trait evolution. It has been proposed that correlations within the covariance matrix Γ represent non-causal correlations, for example linked to joint evolutionary responses to shared environmental conditions, while correlations within the interaction matrix A represent causal effects (Reitan et al., 2012; Liow et al., 2015). For simplicity, we avoid the 'causal/non-causal' dichotomy here; we stick to the term 'correlations' and consider only models making the simplifying assumption that Γ is diagonal. The framework is however equally adapted to deal with correlations incorporated through Γ.

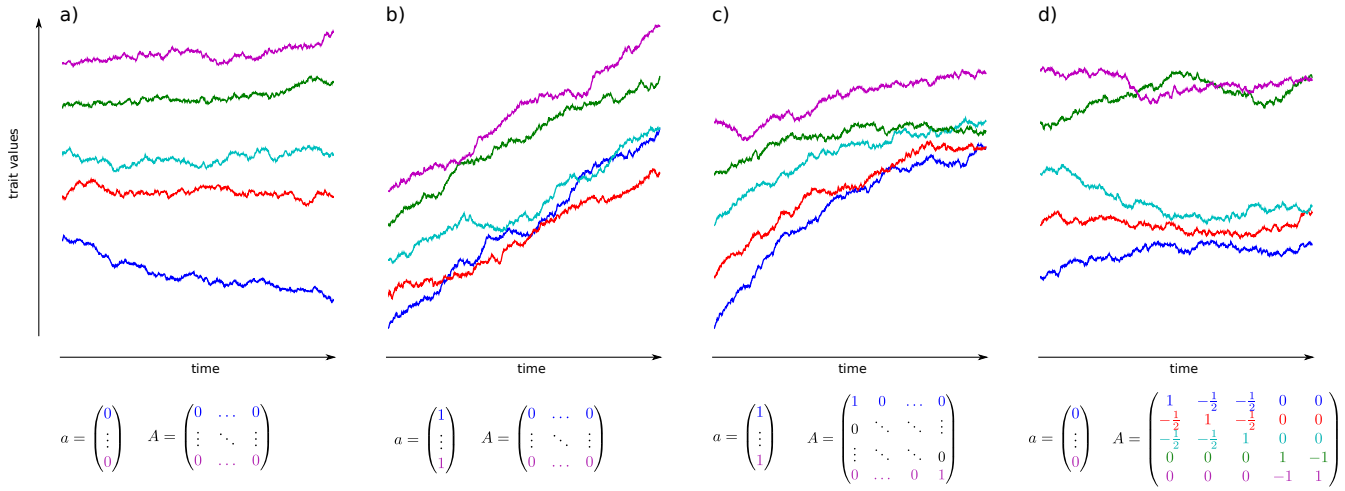


Figure 4.2 – Schematic examples of trait evolution under various models covered by our framework. a) BM model b) BM model with drift c) OU model. In a), b) and c), traits evolve independently from one another. d) illustrates a new class of model that can be handled in our framework, where a given trait value can influence the evolution of other traits: the two top traits evolve independently from the three bottom traits, but within each of these independently evolving groups of traits, trait values are attracted to the mean of the interacting traits, as in the PM model. The a vector and the A matrix corresponding to each model are represented, with trivial parameter values (e.g. the strength of attraction and optimal value of the OU model are set to 1). The parameters in each row dictate the evolution of the trait to which the specific row is associated. Each column represents the effect of the corresponding trait on each of the evolving traits. Under all these models, Γ is the identity matrix.

4.3.2 Notation for trees and traits

We now consider a single or several clades, each of them represented by a single fixed, binary, time-calibrated phylogenetic tree. Our framework could easily be modified to treat non-binary trees including polytomies. Trees are not necessarily ultrametric, meaning that they may include non-contemporary tips (fossils). When considering multiple clades, all associated trees share the same absolute time calibration. Time t runs from the root of the oldest tree ($t = \tau_0 = 0$) to the most recent tip of all trees ($t = T$ is the present if at least one of the phylogenies includes extant species). The K successive branching and extinction times when considering the various trees altogether are denoted by $(\tau_1, \tau_2, \dots, \tau_K)$ and the time-intervals between two such events are called epochs, following Butler and King (2004). We denote by n_t the total number of lineages that arose before (and at) time t.

In the case of trait evolution within a single clade, we assign numbers (from 1 to n_t) to lineages by order of origination. At each branching event τ, one daughter lineage inherits the number assigned to the ancestral lineage while the other one is assigned n_τ.

We model the evolution of d one-dimensional quantitative traits. We denote by $X_t^{(i,j)}$ the value of trait j ($1 \leq j \leq d$) on branch i at time t and X_t the column vector containing the values of all traits on all lineages at time t, ordered as follows : $X_t = {}^{tr}(X_t^{(1,1)}, X_t^{(1,2)}, ...X_t^{(1,d)}, X_t^{(2,1)}..., X_t^{(n_t,d)})$, where ^{tr} stands for the transposition.

In the case of trait evolution in c distinct (co)evolving clades, we begin by arbitrarily ordering the clades from 1 to c; then, we assign numbers to lineages following the formalism introduced above, first numbering lineages from clade 1, then clade 2, and so on. As above, we denote by X_t the column vector containing the values of all traits on all lineages at time t, which now is a concatenation of the c column vectors corresponding to each clade.

4.3.3 Trait evolution on trees

Given one (or several) phylogenetic tree(s), a model of phenotypic evolution is entirely defined by initial conditions X_0 on the trait values at the root(s) and a set of rules dictating how the vector of traits X_t is updated (i) at branching times, (ii) through each epoch (i.e. between two branching or extinction times), and (iii) after a death time.

In line with most models of phenotypic evolution, we consider anagenetic character evolution, meaning that traits do not change at cladogenesis. Hence, at a given branching time τ, each of the daughter lineages inherits the trait value of their mother lineage. In practice, in the case of evolution within a single clade, the new vector X_τ is obtained by concatenating the d trait values of the branching lineage at time τ at the end of $X_{\tau-}$ (where $\tau-$ is the time just preceding the branching event). In the case of evolution in several clades, the new vector X_τ is obtained by inserting the d trait values of the branching lineage at time τ at the appropriate location in $X_{\tau-}$ (i.e., at the end of the part of $X_{\tau-}$ corresponding to the clade in which the branching event is occuring).

On each given epoch (τ_i, τ_{i+1}) $(i \in \{0, 1, ..., K-1\})$, we assume that the evolution of the d traits on the n lineages is driven by a linear stochastic differential equation of the form introduced earlier in Equation (4.1) :

$$\begin{cases} dX_t &= (a_i(t) - A_i X_t)dt + \Gamma_i(t)dW_t \\ X(\tau_i) &= X_{\tau_i} \end{cases} \tag{4.2}$$

where a, A and Γ are now indexed by i, the label of the focal period. The content, as well as the size of a, A and Γ can hence vary with the period. Here, a_i is a vector of $\mathbb{R}^{nd}$ whose coefficients can vary with time, A_i is a constant square matrix of size $\mathbb{R}^{nd} \times \mathbb{R}^{nd}$, Γ_i is a square matrix of the same size whose coefficients can vary with time, and W_t is a nd-Brownian motion.

Finally, when a lineage goes extinct at a given time τ, its d trait values no longer evolve (i.e. they are frozen at the extinction time), and they no longer have any influence on the evolution of the traits of other lineages until reaching the end of the process at time $t = T$. In practice, this means that the vector X_τ is simply equal to $X_{\tau-}$, and that the d lines and columns in a_i, A_i and Γ_i corresponding to the now extinct lineage are all set to zero.

We will show later that this general formulation encapsulates many classical models of phenotypic evolution, ensures analytical tractability, and further allows the incorporation of a broad set of interspecific coevolutionary scenarios.

Given the above, initial conditions on X_0, and the collection of (a_i), (A_i) and (Γ_i) associated to each epoch fully define a process of trait evolution on one or several trees. This formalism is illustrated in Figure 4.3 for a single trait evolving on a single small tree.

All models written under the formalism that we propose can easily be simulated numerically. First, the whole trajectory of the process can be simulated using a numerical scheme for SDE such as the Euler-Maruyama scheme (Gardiner et al., 1985) through each epoch, and augmenting the vector of traits at branching times with traits corresponding to the branching lineage (see Online Appendix C.4.1 in Dryad doi:10.5061/dryad.52636 and Fig. 4.7). Second, we show in the next section how to compute numerically the tip distribution. Tip values can then directly be drawn in a fast way from the tip distribution.

4.3.4 Application: existing and novel models of trait evolution

We first show that the general formulation above encapsulates many classical models of phenotypic evolution, before showing how it further allows considering a much broader set of models, including models of within and between clades coevolution.

Models of phenotypic evolution have traditionally been characterized by a stochastic differential equation specifying how a given trait evolves along a single lineage. Applying Equation (4.2) to trait k on epoch i yields:

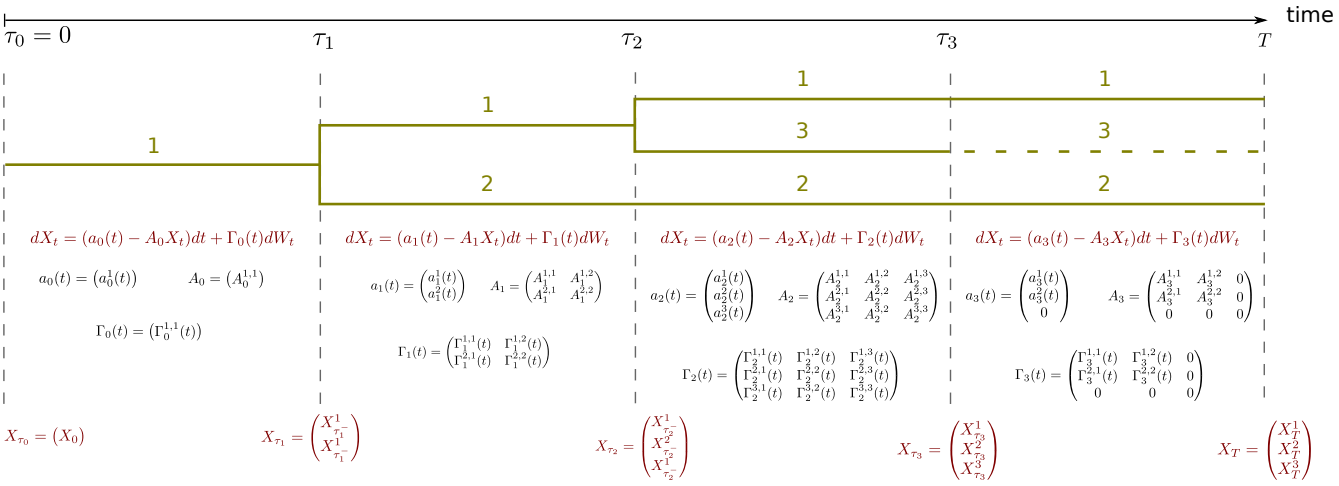


Figure 4.3 – Formalism used throughout the paper to model the evolution of one trait on a non-ultrametric tree. Epochs are separated with vertical dashed lines.

$$dX_t^{(k)} = \left(a_i^{(k)}(t) - \sum_{l=1}^{n_t d} A_i^{(k,l)} X_t^{(l)} \right) dt + \sum_{l=1}^{n_t d} \Gamma_i^{(k,l)}(t) dW_t^{(l)} \tag{4.3}$$

where the two sums are taken over all traits and all lineages. The term $\sum_{l=1}^{n_t d} A_i^{(k,l)} X_t^{(l)}$ is the term that specifies how the value of trait k and all other traits in all other lineages influence the evolution of trait k. Given a well-known differential equation specifying how a given trait evolves along a single lineage for a previously proposed model of phenotypic evolution (second column in Table 4.1), deriving the corresponding expressions for a, A and Γ using Equation (4.3) is straightforward. Table 4.1 summarizes these expressions for existing univariate models running on ultrametric trees.

The first three models (BM, ACDC and DD) are models in which trait evolution along a lineage is influenced neither by the trait value of this lineage nor the trait value of any other lineage. The corresponding A matrices are null matrices, as would be the case for any model with the latter property. The fourth model (OU) is a model in which trait evolution along a lineage is influenced by its own trait value, but not the trait values of other lineages. The corresponding A matrix is diagonal, as would be the case for any model with this property. Finally, the last model (PM) is a model in which trait evolution along a lineage is influenced by its own trait value and the trait values of other lineages, such that A has non-negative off-diagonal values. A remarkable property of A under this model is that all its off-diagonal values are identical. This is explained by the fact that the PM model is a neutral model, in the sense that the effect $A^{(k,l)}$ of lineage l on lineage k is the same for all lineages $k \neq l$. All other models in which the off-diagonal elements of A are identical would have this same property, known in probability theory as exchangeability.

Several variations around these models can still be embedded in our general framework: i) Models in which the rate of phenotypic evolution depends on a variable $Y(t)$ that itself varies through time (see, e.g. global temperature $T(t)$ in Clavel and Morlon, 2017) can be formalised similarly to ACDC, with time t replaced by $Y(t)$. ii) Models accounting for the biogeographic background in which species coevolved (e.g. all the "+GEO" models in Drury et al., 2016) can be incorporated in our framework through the design of the A matrix when ancestral geographic ranges are known or reconstructed (see details in Online Appendix C.3.3). iii) Considering subclades in which trait evolution follows distinct models or similar models with distinct parameter values (as in Butler and King 2004) is also straightforward. One just needs to specify distinct parameters in a, A and Γ on the lines and columns corresponding to lineages in the distinctive subclade. iv) Multivariate trait evolution models, in which several distinct traits evolve in a correlated

Key	Model name Evolution along lineage k	a	A	Γ
BM	Brownian motion, random genetic drift $dX_t^{(k)} = \sigma dW_t^{(k)}$	0	0	σI
ACDC	Accelerating or decelerating rate, early burst $dX_t^{(k)} = \sigma_0 e^{rt} dW_t^{(k)}$	0	0	$\sigma_0 e^{rt} I$
DD	Diversity-Dependent $dX_t^{(k)} = \sigma_0 e^{r n_t} dW_t^{(k)}$	0	0	$\sigma_0 e^{r n_t} I$
OU	Ornstein-Uhlenbeck, stabilizing selection $dX_t^{(k)} = \psi(\theta - X_t^{(k)})dt + \sigma dW_t^{(k)}$	$\psi\theta V$	ψI	σI
PM	Phenotype Matching $dX_t^{(k)} = \psi(\theta - X_t^{(k)})dt$ $+S\left(\frac{1}{n_t}\sum_{l=1}^{n_t} X_t^{(l)} - X_t^{(k)}\right)dt + \sigma dW_t^{(k)}$	$\psi\theta V$	$(\psi + S)I - \frac{S}{n_t}U$	σI

Table 4.1 – Examples of classical models of trait evolution fitting our framework.

The unity vector (vector full of 1) is denoted by V, I refers to the identity matrix (diagonal matrix with diagonal values equal to 1), and U refers to the unity matrix (matrix full of 1). Their size is the same as the size of the vector of traits X_t considered. Parameters are σ: rate of neutral phenotypic evolution; ψ: strength of stabilizing selection; θ: optimal phenotype; S: strength of between-lineage competition driving individual phenotypes away from clade-wise average phenotype; σ_0 : rate of phenotypic evolution at the root of the tree; r : parameter controling the exponential rise or decay of the rate of phenotypic evolution with time (ACDC) or with the number of lineages (DD). Considering non-ultrametric trees including fossils amounts to replacing vector V and matrices I and U by their homologs V_{alive}, I_{alive} and U_{alive}, where the subscript specifies that the vector and matrices have 0 on lines and columns corresponding to lineages that are extinct in the given epoch.

manner (Hansen et al., 2008; Bartoszek et al., 2012) are easily written in our framework, as shown with some examples in Online Appendix C.2.2. In multivariate models with lineages evolving independently from one another (e.g. multivariate combinations of BM, ACDC, DD and OU models), A and Γ are block diagonal matrices, with blocks of size the number of traits, each of them describing correlated multivariate evolution along a particular lineage. In this case, trait-trait correlations introduced through the A matrix correspond, as in Bartoszek et al. (2012), to the case when a given trait on a lineage is attracted to (or repulsed from) a linear combination of other traits in this lineage. v) Finally, accounting for observation errors when available only requires to adjust the variance-covariance matrix as described in Hansen and Bartoszek (2012).

By considering previous models under this light, it becomes very clear that the set of models that have been considered so far represents a very small fraction of all the models that could potentially be considered. In particular the A matrix, which dictates how the value of a given trait influences the evolution of other traits – either different traits in the same lineage, or the same trait in other lineages, or yet different traits in other lineages – has so far been very constrained. It has been considered to be zero (BM, ACDC, DD), diagonal (OU), block diagonal (multivariate), and only recently with non-zero off-diagonal values (PM). Relaxing these constraints means that a much broader array of models incorporating the effect of interspecific interactions on phenotypic evolution can be considered. In particular, lineages do not need to be interchangeable. Evolution in complex networks of interactions can be considered by designing a priori the A matrix according to the known network. The effect of clade-clade interactions can be modeled by filling the A matrix with non-zero entries $A^{(k,l)}$ with k and l corresponding to lineages from different clades. For example, under a scenario of two clades coevolving with no effect of within-clade interactions, this leads to a A matrix with two off-diagonal blocks.

We can thus imagine a variety of coevolutionary scenarios, the only major constraint being that

the effect of a trait value on the evolution of other traits is assumed to be linear (Equations (4.2) & (4.3)). Given a scenario, we can write the corresponding evolution of each trait on a given lineage through each epoch (Equation (4.3)), and deduce the collection of (a_i), (A_i) and (Γ_i) defining the evolutionary process (Equation (4.2)). Below, we first show how to derive the probabilistic distribution of traits at tip branches for any model that can be written under this framework before illustrating the approach with a particular model of clade-clade interaction.

4.4 Distribution of tip trait values

4.4.1 The distribution of traits is Gaussian

Deriving the probabilistic distribution of traits at tip branches is key to our ability to fit phenotypic models to comparative data using maximum likelihood or Bayesian approaches. It also provides a very efficient way to simulate tip values for specific models, by drawing from the expected tip distribution.

When X_0 has a Gaussian distribution (including the particular case when X_0 is constant) the linear equations considered in the framework ensure that X_t remains a Gaussian vector at each time t (see details in Appendix C.1.1). The trait vector X_t, of size $n_t d$, is thus uniquely defined by its expectation vector m_t and covariance matrix Σ_t, and has the following density:

$$\forall x \in \mathbb{R}^{n_t d}, \quad f(x) = \frac{1}{\sqrt{(2\pi)^{n_t d}\det(\Sigma_t)}} e^{-\frac{1}{2}tr(x-m_t)\Sigma_t^{-1}(x-m_t)}$$

In particular, the distribution of tip trait values at present time T is Gaussian with expectation vector m_T and covariance matrix Σ_T. We can compute m_T and Σ_T iteratively: starting with initial conditions m_0 and Σ_0 for $X_{\tau_0} = X_0$, we compute, until reaching the present:

i) $m_{\tau_{i+1}^-}$ and $\Sigma_{\tau_{i+1}^-}$ at the end of each epoch i

ii) $m_{\tau_{i+1}}$ and $\Sigma_{\tau_{i+1}}$ at the branching time τ_{i+1}

4.4.2 Evolution of the distribution through each epoch

Knowing the expectation vector and covariance matrix $(m_{\tau_i}, \Sigma_{\tau_i})$ at the beginning of epoch i, we show (Appendix C.1.2) that $m_{\tau_{i+1}^-}$ and $\Sigma_{\tau_{i+1}^-}$ at the end of epoch i are given by the following analytical expressions:

$$m_{\tau_{i+1}^-} = e^{(\tau_i - \tau_{i+1})A_i} m_{\tau_i} + \int_{\tau_i}^{\tau_{i+1}} e^{(s-\tau_{i+1})A_i} a_i(s) ds \tag{4.4a}$$

$$\Sigma_{\tau_{i+1}^-} = \left(e^{(\tau_i-\tau_{i+1})A_i}\right) \Sigma_{\tau_i}{}^{tr}\left(e^{(\tau_i-\tau_{i+1})A_i}\right) + \int_{\tau_i}^{\tau_{i+1}} \left(e^{(s-\tau_{i+1})A_i}\Gamma_i(s)\right)^{tr}\left(e^{(s-\tau_{i+1})A_i}\Gamma_i(s)\right) ds \tag{4.4b}$$

Alternatively, we can write the evolution of m and Σ on epoch i as a set of ordinary differential equations (ODE), and integrate these ODEs numerically, with initial conditions given by (m_{τ_i} and Σ_{τ_i}). On each epoch i, each component k of the expectation vector evolves according to equation (4.5a) and each component (k,l) of the covariance matrix evolves according to equation (4.5b) (see derivation in Appendix C.1.3):

$$\frac{d}{dt} m_t^{(k)} = a_i^{(k)}(t) - \sum_{m=1}^{n_t d} A_i^{(k,m)} m_t^{(m)} \tag{4.5a}$$

$$\frac{d}{dt} \Sigma_t^{(k,l)} = -\sum_{m=1}^{n_t d} A_i^{(k,m)} \Sigma_t^{(m,l)} + A_i^{(l,m)} \Sigma_t^{(m,k)} - \Gamma_i^{(l,m)}(t) \Gamma_i^{(k,m)}(t) \tag{4.5b}$$

Equations (4.4a, 4.4b) and the ODE system described by Equations (4.5a, 4.5b) are mathematically equivalent. The first formulation is more computationally efficient when the integrals can be simplified analytically. For example when A is symmetric Equations (4.4a, 4.4b) can be simplified (Appendix C.3.1) and computed very efficiently. This first formulation also reveals that if Σ_{τ_i} is positive definite, then $\Sigma_{\tau_{i+1}^-}$ remains positive definite (and thus invertible) even when Γ_i is not. Inverting Σ is important for computation of the Gaussian distribution. The second formulation provides a more intuitive interpretation of the components that influence the evolution of trait distribution, and is easily implementable for any model.

4.4.3 Evolution of the distribution at branching times

Knowing the expectation vector and covariance matrix $(m_{\tau_{i+1}^-}, \Sigma_{\tau_{i+1}^-})$ at the end of epoch i, which precedes the branching of a given lineage j, we build $m_{\tau_{i+1}}$ and $\Sigma_{\tau_{i+1}}$ at the branching event, as illustrated in Figure 4.4.

Recall that in the case of evolution within a single clade, $X_{\tau_{i+1}}$ is obtained by concatenating the d trait values of lineage j at time τ_{i+1}^- at the end of $X_{\tau_{i+1}^-}$. The d new components in $X_{\tau_{i+1}}$ are thus the exact copies of the trait values of lineage j, and have the same expectation and covariance matrix. Hence, the expectation vector $m_{\tau_{i+1}}$ is simply obtained by concatenating $m_{\tau_{i+1}^-}$ with the d components of $m_{\tau_{i+1}^-}$ corresponding to lineage j. The covariance matrix $\Sigma_{\tau_{i+1}}$ is obtained as follows: the covariance matrix corresponding to the previously existing lineages is unchanged, given by $\Sigma_{\tau_{i+1}^-}$; to this main block, we add below a copy of the d lines corresponding to the covariances between the d traits in lineage j and all the other traits, and we add to the right the same components arranged in d columns; finally, we fill the last missing block in the bottom right corner of $\Sigma_{\tau_{i+1}}$ with the block corresponding to the covariance matrix among the d traits in lineage j (i.e. the $d \times d$ diagonal block of $\Sigma_{\tau_{i+1}^-}$ starting from line $(j-1)d+1$).

In the case of evolution in multiple clades, $m_{\tau_{i+1}}$ and $\Sigma_{\tau_{i+1}}$ are constructed following a similar procedure, by augmenting $m_{\tau_{i+1}^-}$ and $\Sigma_{\tau_{i+1}^-}$ with copies of blocks corresponding to lineage j, inserted at the appropriate location. We illustrate this update step in Figure 4.4.

4.4.4 Tip trait distribution for particular models

Applying this general iterative procedure along a phylogenetic tree provides closed analytic tip distribution formulae for a wide variety of models. In Appendix C.2, we re-derive known tip distributions for models without lineage-lineage interaction, thus providing a unified review of mathematical results associated to these models. Tip distributions for classical univariate models (BM, ACDC, DD, OU) on ultrametric and non-ultrametric trees are summarized in Appendix Table C.1. We confirm, as has been shown before (Uyeda et al., 2015), that the OU and AC models have identical tip distributions on ultrametric trees. We also re-derive results that can be found in Bartoszek et al. (2012) providing tip distributions for multivariate models.

Analytical formulae of tip distributions for models with lineage-lineage interactions have not yet been proposed. Drury et al. (2016) developed the inference tools that allow fitting the PM model, using the ODE system given in Equations (4.5a, 4.5b) (thereafter referred to as 'ode' method). Here, we develop the inference tools based on analytical reduction of Equations (4.4a, 4.4b), (thereafter referred to as 'analytical' method, see Appendix C.3.2), and compare the efficiencies of the two methods. Specifically, we simulated 50 pure-birth Yule trees with a per-lineage speciation rate of 1 per time unit, conditioned to having a given number of tips at present, using the 'phytools' R package (Revell, 2012). We then computed the tip distribution corresponding to the PM model with parameters fixed at $(m_0, v_0, \theta, \psi, S, \sigma) = (0, 0, 1, 0.1, 1, 2)$ using both the analytical and the ode methods. The new analytical method is much more efficient than the previous ode method (Fig. 4.5). While we were previously limited to fitting the PM model to trees of less than 150 tips due to memory issues, the analytical methods allows fitting trees with up to 600 tips

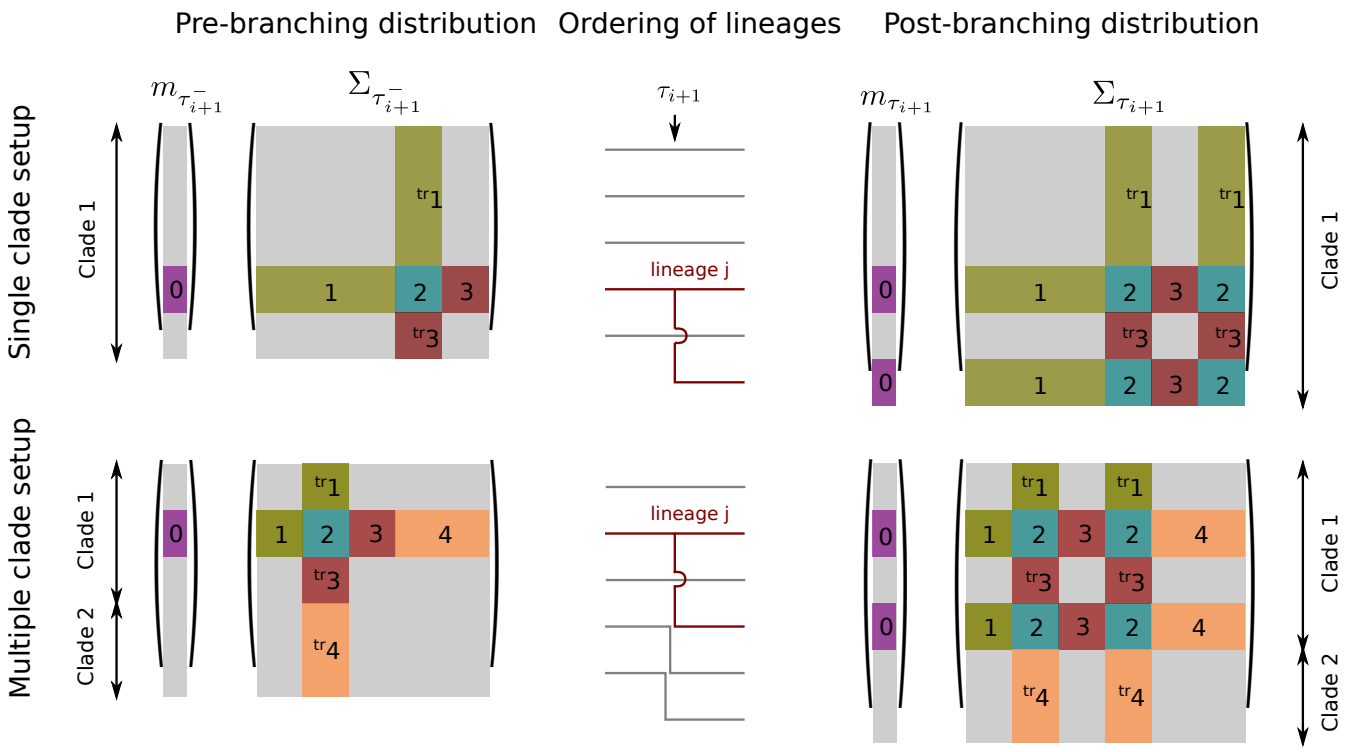


Figure 4.4 – Update step for the expectation vector and covariance matrix at branching times when there is one (top row) or two (bottom row) clades. The middle panel highlights the branching lineage j, as well as the ordering of lineages before and after the branching event. The vector $m_{\tau_{i+1}}$ and matrix $\Sigma_{\tau_{i+1}}$ (displayed on the right) are constructed by augmenting $m_{\tau_{i+1}^-}$ and $\Sigma_{\tau_{i+1}^-}$ (displayed on the left) with copies of blocks corresponding to lineage j (materialized by numbers).

on a desktop computer.

Drury et al. (2016) also proposed an extension of the PM model accounting for the biogeographic history of lineages. In the case when each lineage is present in at most one location, the 'analytical' method can be extended, providing fast likelihood computation (see Online Appendix C.3.3). When there are lineages occurring in more than one location at the same time, we need to resolve numerically the ODE system in order to compute the likelihood of tip traits. While this is more time-consuming than finding a good 'analytical' reduction, the new implementation is more efficient than the one we previously proposed (Drury et al., 2016).

4.5 Modeling trait evolution on coevolving clades

We illustrate how our framework can be used to study trait coevolution in scenarios of clade-clade interactions. We consider a simple model with two interacting clades (numbered 1 and 2), in which a given trait in clade 1 coevolves with another given trait in clade 2. Following the approach introduced above, we define X_t the vector of trait values containing first the trait values for clade 1, and then the trait values for clade 2, and we write a stochastic differential equation specifying how trait value evolves along a single lineage k. In the spirit of the phenotype matching model (Nuismer and Harmon, 2014), we propose here a formulation in which the trait of lineage k is attracted to (or repelled from, depending on the sign of S) the average trait value of the lineages it interacts with, plus (or minus) a shift:

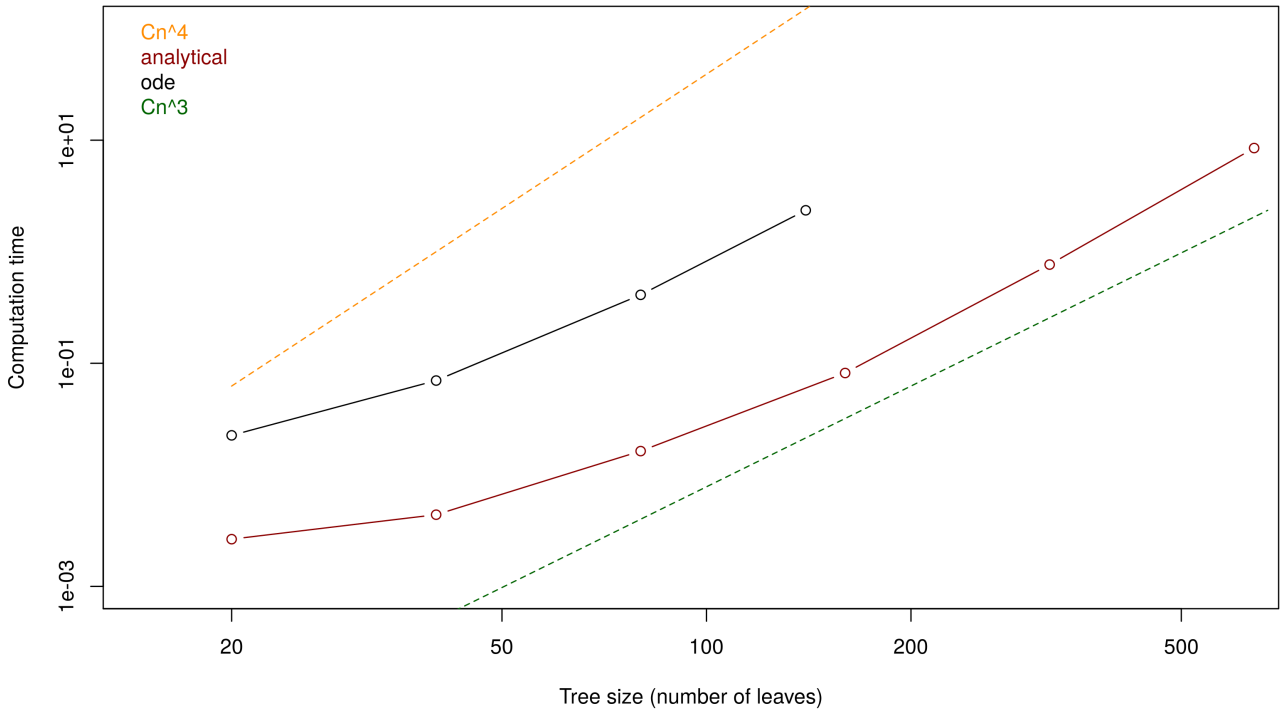


Figure 4.5 – Time needed to compute the distribution of tip data following the PM model with parameters $(m_0, v_0, \theta, \psi, S, \sigma) = (0, 0, 1, 0.1, 1, 2)$. Trees are simulated under a pure-birth model conditioned on having a given number of leaves. Bottom dots : 'analytical' implementation; Top dots: 'ode' implementation. The top dashed curve represents the slope of time increase as a power 4 of the number of leaves while the bottom dashed curve represents the slope of time increase as a power 3 of the number of leaves.

$$dX_t^{(k)} = S\left(\delta_k d_1 + (1-\delta_k)d_2 + \frac{1}{n_k}\sum_{l=1}^{n} p_{k,l}X_t^{(l)} - X_t^{(k)}\right)dt + \sigma dW_t^{(k)} \tag{4.6}$$

where S represents the attracting or repelling strength of species interactions on trait evolution, d_1 represents the shift for lineages from clade 1, d_2 the shift for lineages from clade 2, σ is the drift parameter, δ_k equals one if lineage k belongs to clade 1 and zero if it belongs to clade 2, $p_{k,l}$ equals one if lineages k and l interact and zero otherwise, $n_k = \sum_l p_{k,l}$ is the number of lineages interacting with lineage k, and n is the total number of lineages.

When S is positive, the trait value of lineage k is attracted to an optimal trait value given by the average trait value of the interacting species (plus a shift d_1 or d_2).

An example of such a scenario of clade-clade matching mutualism is the coevolution between the length of floral tubes and the length of butterfly proboscis in a plant-pollinator mutualistic network (illustrated in Fig. 4.6). In this example, we assume that the optimal length of a butterfly proboscis is the average length of the plant floral tubes it pollinates plus a shift d_2, while the optimal length of a plant floral tube is the average proboscis length of its butterfly pollinators plus a shift d_1. With $d_1 + d_2 = 0$, both traits can reach their optimal state, leading to a stable situation with butterfly proboscis a bit longer (if $d_1 > 0$) or shorter (if $d_1 < 0$) than plant floral tubes. With $d_1 + d_2 \neq 0$, traits cannot reach their

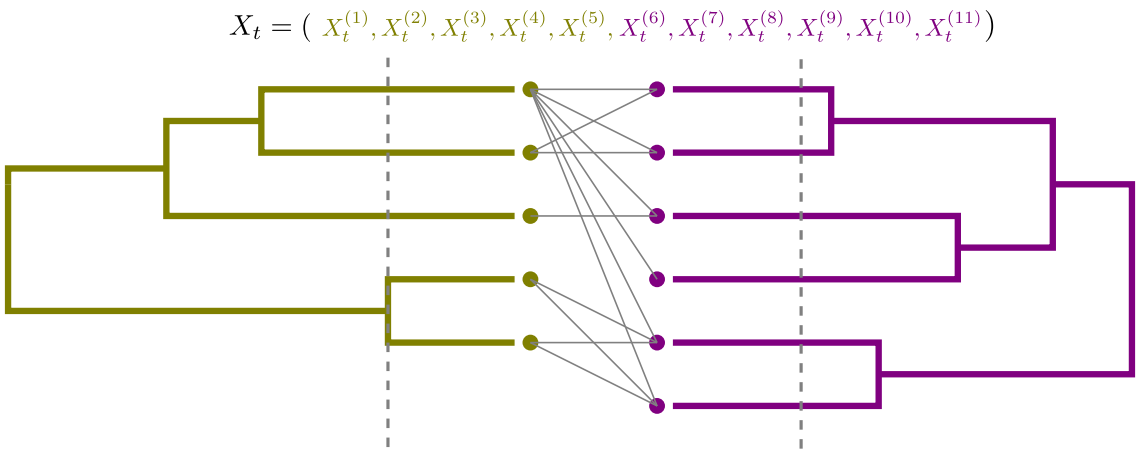


Figure 4.6 – Hypothetical clade-clade coevolutionary scenario. Vertical dashed lines delimitate the successive epochs. The vector X_t contains the trait values on the third (last) epoch, P_3 is the matrix of network interactions, and a_3, A_3 and Γ_3 together define trait evolution according to the clade-clade matching model defined in Equations (4.6) and (4.7).

optimal state, resulting in a runaway process where both traits tend to evolve toward an ever-moving optimum. For example, with positive d_1 and d_2, the butterflies proboscis tends to get longer to better access the nectar, while the floral tube also tends to get longer to force the butterfly's body to touch the stamen. The parameters S and σ control respectively the strength of the interaction effect and the rate of stochastic phenotypic change. The bigger S, the closer the traits will track the optimum; the bigger σ, the bigger the fluctuations around this optimum.

When S is negative, the traits are repelled from the average trait value of the interacting species (plus a shift d_1 or d_2). This may capture natural situations of clade-clade competition driving trait displacement. Finally, some antagonistic interactions between traits could require to introduce two parameters $S_1 > 0$ and $S_2 < 0$ to capture match-vs-escape scenarios. For example, parasites might tend to develop cues matching those of their hosts while hosts develop cues to escape their parasites in a co-evolutionary arms race.

From Equation (4.6) we deduce the corresponding a, A and Γ through each epoch:

$$\begin{aligned} a &= S(\Delta d_1 + (V - \Delta) d_2) \\ A_{k,l} &= S(\mathbb{1}_{k=l} - \frac{p_{k,l}}{n_k}) \\ \Gamma &= \sigma I \end{aligned} \tag{4.7}$$

where Δ is the vector of elements δ_k (see Fig. 4.6 for an illustration). Matrix A is in general not symmetric anymore, as all species k do not have the same number n_k of species that they interact with.

As shown by Equation (4.7), entirely defining a model of clade-clade coevolution requires introducing a constant network of interaction during each epoch (the P matrix with elements $p_{k,l}$). We can potentially re-define epochs to account for events of change in the interaction network in addition to speciation and extinction events, thus allowing interaction networks to evolve along branches. In practice, we typically (at best) have access to the current interaction network (Fig. 4.6), but not the ancestral networks. A solution to this would consist in treating the ancestral P matrices as parameters of the model, and searching the ancestral network(s) that maximize the fit to the data. Another approach would consist in reconstructing ancestral networks over each period according to rules regarding the inheritance of interactions at speciation times. Developing these approaches is outside the scope of the current study, but we have shown how to compute tip trait distributions once they are developed.

We illustrate the computation of tip trait distributions for a model in which the ancestral networks are known: a generalist model where all species from clade 1 interact with all species from clade 2. We consider a 'Generalist Matching Mutualism' model of trait evolution (thereafter referred to as GMM, and illustrated in Fig. 4.7a), which is captured by Equation (4.6) with S positive and $p_{k,l} = 1$ for any two lineages k and l from different clades and $p_{k,l} = 0$ for any two lineages k and l from the same clade. Given that the model fits within our framework, we know that the trait distribution at the tips is Gaussian, and we can compute the expectation vector and covariance matrix corresponding to the model using Equations (4.4a, 4.4b), which we can reduce for this specific model in order to speed up the computation (Appendix C.3.4).

The tip distribution is relatively fast to compute (e.g. in the order of 0.8 seconds with two 100-tip trees on a desktop computer), such that fitting the model by maximum likelihood or in a Bayesian framework should not be problematic for trees with a few hundred tips. However, we do not aim here to carry an in-depth study of this particular model, nor to fit it to empirical data. Rather, we use our ability to rapidly compute tip trait distribution to get a first glimpse of the model behaviour under distinct sets of parameter values.

In Figure 4.7 (b,c,d,e), we plotted the distribution of the average $\bar{X}^1$ of trait values in clade 1 and the average $\bar{X}^2$ of trait values in clade 2 for traits evolving under the GMM model with four parameter sets chosen to lead to four distinct qualitative behaviours. From Equation (4.6), we can easily show that under GMM $\bar{X}^1 + \bar{X}^2$ is a drifted Brownian motion with drift term $S(d_1 + d_2)$ and $\bar{X}^1 - \bar{X}^2$ is an OU process with optimum $(d_1 - d_2)/2$ and selection strength S. The shift parameters d_1 and d_2 thus directly determine the position of the optimum of the distribution. An 'equilibrium' scenario corresponds to $d_1 = -d_2$: the more likely values for the two average traits $\bar{X}^1$ and $\bar{X}^2$ are such that $\bar{X}^1 = \bar{X}^2 + d_1$ (see Fig. 4.7b). In contrast, when $d_1 \neq -d_2$, the two communities have optimal trait values that are non-compatible, and the traits will tend to increase if $d_1 + d_2 > 0$ and decrease if $d_1 + d_2 < 0$ (see Fig. 4.7c) in a 'runaway' process. In this case, the position of the peak in the tip distribution will also depend on the depth of the root, the trait values at the root, and the value of the parameter S. The parameter S plays an important role in the hump thickness: the bigger S, the more constrained $\bar{X}^1 - \bar{X}^2$ around $(d_1 - d_2)/2$ (see Fig. 4.7e). The parameter σ also plays a role in the thickness of the hump, but in the orthogonal direction : increasing σ flattens the distribution by allowing different $\bar{X}^1$ and $\bar{X}^2$ values while retaining the constraint on $\bar{X}^1 - \bar{X}^2$ (see Fig. 4.7d). In future work, it would be interesting to assess whether these results can be used to build a statistical test for distinguishing between the runaway and equilibrium scenarios.

Implementation

Our framework is implemented in the R package 'RPANDA' (Morlon et al., 2015), including functions to compute tip distributions, to simulate trait evolutionary trajectories using the Euler-Maruyama scheme (see Online Appendix C.4.1), and to simulate tip data by drawing from the expected tip distribution. We also implemented optimisation functions to infer model parameters by maximum likelihood, and to compare the fit of distinct models using information criteria. In the most general user-defined use of our framework, the input is one or several potentially non-ultrametric phylogenetic tree(s) and the collection of (a_i, A_i, Γ_i) matrices during each epoch that define a specific model. In this case, the tip distribution is computed using the most general 'ode' method that solves numerically the ODEs. In addition, we implemented all models mentioned in Table 4.1 as well as GMM with the fastest described algorithm to compute their tip distribution. In Online Appendix C.5, we provide a tutorial explaining the structure of our code and illustrating how to use it. We however recommend potential users to thoroughly test the statistical properties of the models they design (parameter estimation, type I and II error rates) using simulations before applying them to empirical data. These properties are model-dependent and thus should be assessed case by case. We are in the process of performing these tests for the GMM model.

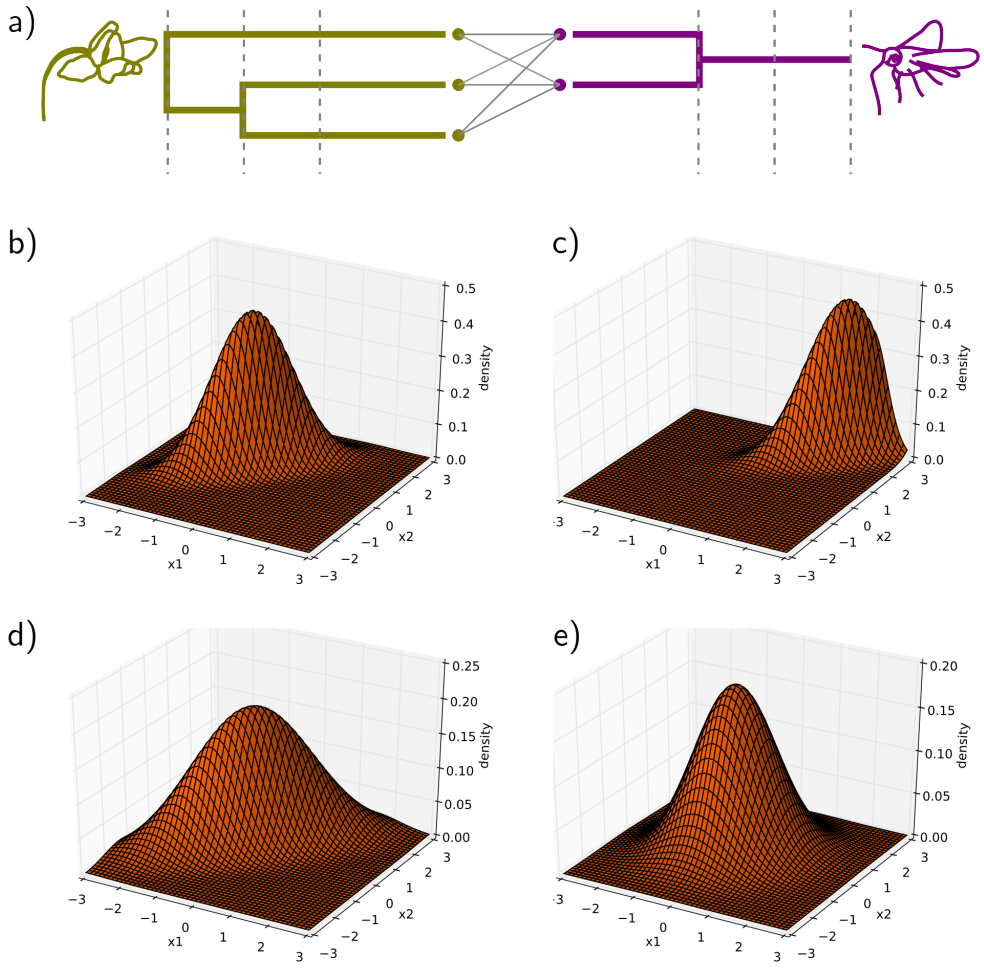


Figure 4.7 – Trait evolution under the Generalist Matching Mutualism (GMM) model. a) an illustrative generalist network of interactions between two clades. Vertical dashed lines delimitate the successive epochs. bcde) Expected tip distribution for the average trait value in each clade, with parameter values $(S, d_1, d_2, \sigma) =$ b) $(2, -1, 1, 1)$, c) $(2, 0, 2, 1)$. d) $(2, -1, 1, 1.5)$, e) $(0.2, -1, 1, 1)$.

4.6 Discussion

We developed a modeling framework for traits coevolving in coevolving lineages and clades. We highlighted that under a wide variety of models where the evolution of a given trait on a given lineage is linearly related to its own value and the value of other traits on the same lineage, of the same trait on other lineages, and/or of other traits on other lineages, the expected tip trait distribution is Gaussian. We showed how to compute this tip distribution in general, as well as for specific models, including classical models of phenotypic evolution and new models of clade-clade coevolution.

Many classical models of phenotypic evolution, such as univariate and multivariate BM, OU, ACDC and DD fit within our framework. They correspond to the situation where the evolution of traits on a given lineage is independent of trait values on other lineages. For these models, we already know that the tip trait distribution is Gaussian. However, finding the relevant computation of the expectation vectors and covariance matrices associated with each model in the dense literature of comparative phylogenetics can be overwhelming for neophytes. Our Appendix C.2 unifies these computations under a common formalism,

providing both the expressions for the various existing models and their mathematical underpinning. This is done in the context of trees that are not necessarily ultrametric, meaning that all models can be applied to phylogenies including fossils. We hope that this Appendix can serve as a useful review for navigating phylogenetic approaches for understanding trait evolution.

The fact that the distribution of traits remains Gaussian when traits from different lineages coevolve is a convenient result, because it means that computing the tip distribution only requires computing the expectation vector and covariance matrix associated with the different models. For example, we used this result in Drury et al. (2016) to compute tip trait distributions for the phenotype matching model (Nuismer and Harmon, 2014) and fit it to comparative data by maximum likelihood. Here we vastly extend the set of potential coevolutionary models for which tip trait distributions can be computed and provide two general approaches for computing the expectation vector and covariance matrix. One of these two approaches (the 'ode' approach) consists in numerically integrating a set of ODEs. This is the approach that was used in Drury et al. (2016). The other approach (the 'analytical' approach) involves computing integrals and is more efficient when these integrals can be analytically reduced, which depends on the form of the model. Applying the 'analytical' approach to the PM model, we greatly improved its computational efficiency.

We provide a framework for computing tip trait distributions for a wide class of models accounting for within-clade and clade-clade interactions. We hope that this flexibility will foster the development and study of various models adapted to the specificities of particular scientific questions and biological systems. We did not study at length a particular coevolutionary model in this paper, but the PM model was thoroughly studied elsewhere (Drury et al., 2016). The Generalist Matching Mutualism model that we introduce here can be seen as a clade-clade analogue to the PM model (Nuismer and Harmon, 2014). Both models are 'generalist' in the sense that all lineages are assumed to interact (within-clade in the case of PM and between clades in the case of GMM). This assumption can be relaxed by incorporating additional information. In our biogeographic models for example, lineages can only interact if they are sympatric (Drury et al., 2016). More generally, any information or hypothesis concerning the network of interactions between lineages can be accounted for into the A matrices.

There are two main limitations to the modeling framework presented here. The first one is that trait evolution is always assumed to respond linearly to trait values in other lineages. Thus, non-linear effects such as a stronger selection for divergence when phenotypes are similar cannot be accounted for. Nuismer and Harmon (2014) originally developed an individual-based model accounting for such non-linear effects, and then linearized the effects to derive the model of phenotypic evolution emerging at the lineage level (their Equation S38). This linear expression inspired the development of the present framework, and assures that the tip trait distribution is Gaussian. It would however be particularly interesting to model non-linear effects. The second limitation is the issue of model and parameter identifiability, in particular in the absence of fossils. A Gaussian distribution in $\mathbb{R}^{nd}$ can potentially allow identifying several models and parameters, but there are distinct combinations for which a similar (or even identical) distribution is expected. For example, we already know that parameters of the OU model are non identifiable on phylogenies with only extant species (Ho and Ané, 2014) and that OU and AC have identical tip distributions on ultrametric trees (Uyeda et al., 2015). Thus, while we wrote our framework in all generality, with a, A and Γ encompassing as many parameters as desired, and parameters that potentially vary between epochs, it is clear that simplifying assumptions need to be made in order to reduce this parameter space. Identifiability cannot always be checked analytically, as in the case of the OU and AC models. In addition, there can be differences between theoretical and *de facto* identifiability, with models that are identifiable in theory but are difficult to identify in practice. For example, we can show analytically that ψ and S from the PM model are theoretically identifiable, but in practice in most cases only $\psi + S$ can be estimated with precision. Also, *de facto* identifiability depends on the data available, such as the size and shape of a particular phylogeny, and whether it includes fossils or not (Slater et al., 2012). Furthermore, models taking into account interactions among lineages will have to assess the influence of extinct lineages in the past. This has been studied in Drury et al. (2016) for the PM model, by simulating trait evolution on trees including dead branches, before fitting the model on the reconstructed tree only. Our recommendation is

to check identifiability on a case-by-case basis, by fitting the set of models under consideration to trait datasets simulated directly on the specific empirical phylogenies in hands. We provide the tools for rapidly simulating tip values under various models by sampling expected distributions.

One of the most challenging and exciting developments that we see ahead is to move from generalist models to models that account for specific interaction networks. We show in this paper how to compute tip trait distributions for such models, assuming that the ancestral networks are known. While some fossil species interaction networks have been compiled (Dunne et al., 2008), such data is typically not available. Thus, if we are to really understand if and how species interactions affected long-term phenotypic evolution, we need to start developing models for reconstructing ancestral networks, analogous to the use of ancestral biogeographic models (see Ronquist and Sanmartín 2011 for a review) to incorporate biogeography into models of phenotypic evolution (see e.g. DD+GEO or MC+GEO, in Drury et al., 2016). Interestingly, our modeling framework could provide an approach to do so, informed by species phylogenies, the interaction network of present day species, and current species phenotypes. Indeed, rather than assuming that the ancestral networks are known, we could treat them as additional parameters to optimize upon, and find the ancestral networks that maximize the likelihood of the current data. These approaches have not been experimented yet and are not part of the available code in RPANDA. Whether there will be enough information in the data to distinguish the probability of alternative ancestral networks remains to be tested, but the observed phylogenetic signal in empirical networks of interactions is encouraging (Ives and Godfray, 2006; Rafferty and Ives, 2013; Hadfield et al., 2014; Hayward and Horton, 2014; Martín González et al., 2015). Our ability to distinguish the probability of alternative ancestral networks will be increased by proposing various scenarios regarding the inheritance of interactions at speciation times, such as scenarios in which daughter species interact with many or few of the species that interacted with their mother lineage. These upcoming developments can draw upon the existing literature on the cophylogeny problem (Conow et al., 2010), and will certainly have an important role to play in the on-going effort of understanding the evolution of species interaction networks (Loeuille and Loreau, 2005; Martinez, 2006; Nuismer et al., 2013).

Our framework for modeling continuous trait evolution on phylogenetic trees includes most previously proposed models and can be used to develop a series of new models of within-clade and clade-clade coevolution. We hope that this will motivate new theoretical and empirical applications aimed at unravelling how species interactions evolve and influence phenotypic evolution over macro-evolutionary time-scales.

Chapter 5

The relaxed molecular clock hypothesis with episodes of fast divergence

This chapter focuses on a very old debate in evolution, opposing two different views of phenotypic evolution through time. The first one, called *phyletic gradualism*, assumes that traits evolve through small, gradual, mutations arising throughout the lifetime of a species. It is commonly contrasted to the second view, named *punctuated equilibrium*, proposing that changes happen in a less continuous manner. Periods of stasis would be interspersed by large effect transformations, preferentially occuring at speciation events.

While these two hypotheses have been modeled in the context of continuous trait evolution, much less has been done to integrate them in models of molecular evolution. In this chapter, we propose a way to model both a smooth, basal, molecular evolution, together with sudden episodes of *spikes* of mutation arising at speciation events. Our approach can be seen as a *relaxed clock model*, where rate heterogeneities are replaced by punctuational events correlated to the diversification process. It is intended to provide first a way to jointly date phylogenies and study patterns of spikes across the tree and across loci. Second, it could provide a way to assess the importance of *gradualism* vs. *punctualism* in driving sequence divergence.

This chapter presents work in progress. The inference protocol is still in a preliminary development stage, and has not been applied on empirical data. This future work will be carried out and hopefully published with the help of Julie Marin.

Contents of the chapter

5.1 Introduction

The debate between *gradualism* and *punctualism* has a long history in evolutionary biology. *Gradualism*, which consists in considering that trait evolution occurs through the constant accumulation of small, gradual, changes through time, was the prevailing idea before the seventies. Yet another view has emerged, considering that trait evolution over deep times may be better described through rare, punctual, large effect changes. First proposed by paleontologists Eldredge and Gould (1972), the theory of punctuated equilibrium has been widely studied and discussed for morphological traits evolution.

While gradualism has been the dominant idea in the very beginning of continuous trait evolution modeling (Felsenstein, 1973; Hansen, 1997; Butler and King, 2004), both gradualism and punctualism are now well considered in modern comparative tools focusing on continuous traits (Bokma, 2002, 2008; Pennell et al., 2014). On one hand, gradualism is the underlying idea justifying the use of diffusion processes to model continuous trait evolution. On the other hand, punctualism might be included through the use of processes including jumps, known as Levy processes (Landis et al., 2013; Landis and Schraiber, 2017), or through discrete shifts in the parameters of the diffusion processes themselves (Bastide et al., 2016; Khabbazian et al., 2016).

In contrast, punctualism has been much less considered in models of molecular evolution. The modern literature commonly considers that sequences evolve through the accumulation of isolated substitutions arising as a Poisson process through time. The rate of the substitution process, which was first supposed to be constant (an hypothesis called *strict clock hypothesis*, Zuckerkandl and Pauling (1962)), is now considered to vary through time (see the review on *relaxed molecular clocks* by Lepage et al. (2007)). But all this family of models rely on a an underlying gradualist view, considering only gradual anagenetic changes, i.e. isolated mutations happening along lineages of the tree.

A first attempt to introduce the idea in the field was carried out by Webster et al. (2003) and Pagel et al. (2006). Starting from the observation that, in trees reconstructed by maximum parsimony, there is often a correlation between the number of substitutions inferred from the tip to the root and the number of nodes on this path (a phenomenon known as the *node-density artefact*, Fitch and Beintema (1990)), Webster et al. (2003) and Pagel et al. (2006) hypothesized that this correlation was due to frequent cladogenetic mutation events. They designed a statistical test aimed at establishing whether this correlation was indeed due tu such punctual events or to an artifact in the phylogenetic reconstruction. However their studies suffered from methodological artifacts and internal inconsistencies (Brower, 2004; Witt and Brumfield, 2004), and punctualism in the evolution of molecular sequences faded from memory.

Yet, ecological theory predicts that some situations could lead to episodes of fast accumulation of mutations that would be perceived as punctual at the macroevolutionary scale. Some episodes of speciation might for example include founder effects, with species originating from a very small population size. A strong genetic drift might then lead to the fast fixation of otherwise weakly selectively stabilized substitutions (Bromham, 2009). Strong divergent selection is also expected to happen between two sister species at speciation, when they specialize in a distinct habitat, in a distinct resource use, or in a different conspecific recognition mechanism (Peichel and Marques, 2017). When strong enough, we can expect this divergent selection to be observed at the gene level too.

In this project, we propose to model these episodes of fast accumulation of mutations as punctual events called *spikes*. These spikes are jointly modeled with the diversification process, and are supposed to happen at speciation events. They are superimposed on a background evolution through gradual substitutions at constant rate. As a result, the spike process mediates a tight link between the substitution and the diversification process. The observation of present-day sequences has thus the potential to better inform the diversification model. We aim at inferring joint knowledge on the diversification model and on the position of spikes, both over the DNA sequence and along the phylogeny of organisms. The determination of gene sequences impacted by the speciation process would bring further insights on the genomics of speciation (Seehausen et al., 2014).

In a first section, we detail the model, together with its simulation and likelihood computation

procedures. We then design a statistical inference protocol aimed at inferring parameters of the model, using molecular data on a known dated tree. We present preliminary tests of the inference protocol on simulated data. We finally discuss the links between this model and other relaxed molecular clocks, and present the future directions that we see ahead for this project.

5.2 Model

5.2.1 Joint law of trees and spikes

We intend to model the evolution of a sequence of m homologous nucleotides among n phylogenetically related organisms.

We consider that the n organisms are related through a phylogeny which is generated by a birth-death process with a constant birth rate $b > 0$ and constant death rate $d \geq 0$, originating at time 0 and stopped at present time T. Furthermore, at each birth event, a spike occurs with probability $\nu \in (0, 1)$ at the very beginning of each of the two sister lineages. This process thus generates a *spiked tree*, which we will represent in the following as a rooted dated tree with dots superimposed where spikes occurred.

When $d \neq 0$, the generated tree is not necessarily ultrametric, meaning that some lineages die before reaching present time T. However, we are only interested in the phylogeny of present-day organisms. This one is called the *reconstructed spiked phylogeny*, and is obtained by erasing all extinct lineages in the complete phylogeny, as illustrated in Figure 5.1. Note that, because of hidden speciation events with now extinct lineages, spikes may occur along branches on the reconstructed spiked phylogeny.

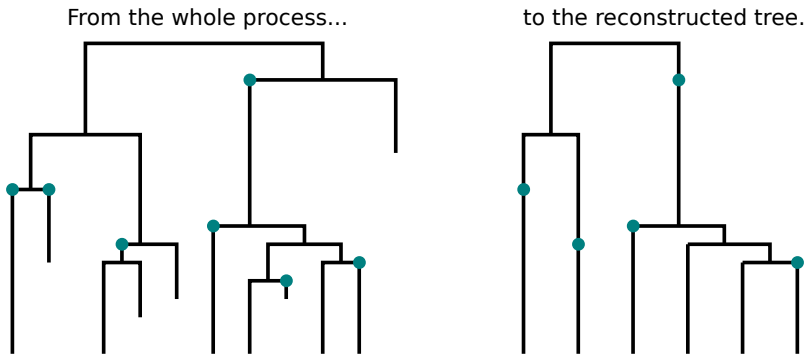


Figure 5.1 – Spikes (blue dots) happen along branches of the reconstructed phylogeny. They correspond to spikes associated to birth events of an extinct lineage.

5.2.2 Law of spikes on a fixed tree

Let $\mathscr{T}$ denote the reconstructed phylogeny, and $\mathscr{S}$ denote the spike positions on the tree. Because we have no mean to ever infer the precise position of the spikes along a branch, we consider that $\mathscr{S}$ denotes only the number of spikes on each branch of a fixed tree. We aim at describing the law of $\mathscr{S}$ on a fixed, known, realisation $\mathcal{T}$ of $\mathscr{T}$ following a birth-death process with known birth rate b and death rate d.

The time is oriented from the tips ($t = 0$) to the root of the tree. We call $u(t)$ the extinction probability before present of the descent of an individual living at time t. This probability is given by Kendall (1948):

$$u(t) = \frac{1 + \frac{d}{b-d} e^{(b-d)t}}{1 + \frac{b}{b-d} e^{(b-d)t}}$$

Along any branch of the reconstructed tree, we have:

$\mathbb{P}$(having a spike on $[t, t+dt]$)
$= \mathbb{P}$(having a birth event on $[t, t+dt]$, death of a lineage, survival of the other | survival of one lineage)
$= 2b\nu u(t)dt$

We can thus simulate spikes along branches of $\mathcal{T}$ as a Poisson Process with rate $2b\nu u(t)$. On a branch originating at time t_0 and ending at time t_1 (with $t_1 < t_0$ being closer to the tips), the number of spikes is Poisson distributed with parameter :

$$\begin{aligned}
\alpha = \int_{t_1}^{t_0} 2b\nu u(s)ds &= 2b\nu \int_{t_1}^{t_0} \frac{1-e^{(b-d)s}}{1-\frac{b}{d}e^{(b-d)s}}\, ds \\
&= 2b\nu \int_{t_1}^{t_0} \frac{1-\frac{b}{d}e^{(b-d)s}+\frac{b}{d}e^{(b-d)s}-e^{(b-d)s}}{1-\frac{b}{d}e^{(b-d)s}}\, ds \\
&= 2b\nu \int_{t_1}^{t_0} 1+\frac{e^{(b-d)s}(\frac{b}{d}-1)}{1-\frac{b}{d}e^{(b-d)s}}\, ds \\
&= 2b\nu \left[s - \frac{1}{b}\ln\left(1-\frac{b}{d}e^{(b-d)s}\right)\right]_{t_1}^{t_0} \\
&= 2b\nu(t_0-t_1) - 2\nu \ln \frac{1-\frac{b}{d}e^{(b-d)t_0}}{1-\frac{b}{d}e^{(b-d)t_1}}
\end{aligned}$$

This leads us to the following simulation procedure:

Algorithm 9 *(spike simulation on a fixed tree with known (b,d) rates)*
The algorithm is initialized at the root (t_0 is the height of the tree), and all branches are recursively explored.

i) Let t_0, t_1 be respectively the origination and ending times of the branch ($t_1 < t_0$).
Compute $\alpha = \int_{t_1}^{t_0} 2b\nu u(s)ds$.

ii) Draw random values :
(a) a realisation n_B under the law $\mathcal{B}(\nu)$
(b) a realisation n_P under the law $\mathcal{P}(\alpha)$
Fix n_B spikes at the very beginning of the branch (time t_0), and n_P spikes along the branch.

iii) Go back to step i) on the two daughter branches.

Futhermore, for any realisation $\mathcal{S}$ of $\mathscr{S}$, and considering again that we are given the birth and death rates of the branching process, we are able to compute $\mathbb{P}(\mathscr{S} = \mathcal{S} \mid \mathscr{T} = \mathcal{T})$ using the following procedure :

Algorithm 10 *(likelihood of a spike realisation on a fixed tree with known (b,d) rates)*
The algorithm is initialized by fixing $p = 1$.
We start at the root and recursively explore the branches of the tree.

i) Let t_0, t_1 be respectively the origination and ending times of the current branch ($t_1 < t_0$ is closer to the tips) and let n_S be the number of spikes on the branch.
Compute $\alpha = \int_{t_1}^{t_0} 2b\nu u(s)ds$.

ii) Use it to compute the probability to see n_S spikes on the branch: $\mathbb{P}(N_S = n_S)$. It is the convolution of a Bernoulli random variable with parameter ν and a Poisson random variable of parameter α, i.e.

$$\mathbb{P}(N_S = 0) = (1-\nu)e^{-\alpha}$$
$$\forall n_S \geq 1,\ \ \mathbb{P}(N_S = n_S) = \nu e^{-\alpha}\frac{\alpha^{n_S-1}}{(n_S-1)!} + (1-\nu)e^{-\alpha}\frac{\alpha^{n_S}}{n_S!}$$

iii) Update $p \leftarrow p * \mathbb{P}(N_S = n_S)$.

iv) Go back to step i) on the two daughter branches.

We now need to describe the second ingredient of our model, i.e. the evolution of molecular sequences on a reconstructed spiked tree.

5.2.3 Molecular evolution on a reconstructed spiked tree

We now model the evolution of a sequence of m nucleotides on a known reconstructed spiked tree $\mathcal{T}, \mathcal{S}$. On any node w of $\mathcal{T}$, We call $X_w^{(k)}$ the state of nucleotide k in the sequence displayed at node w. We will make use of the following notation :

$$\begin{aligned} \rho &:= \text{ the root of the tree.} \\ \mathcal{F}(w) &:= \text{ all leaves descending from node } w. \\ \mathcal{N}(\mathcal{T}) &:= \text{ all nodes of } \mathcal{T}. \end{aligned}$$

The alignment of sequences at the leaves will obviously play an important role, for this is our raw data. We will denote it :

$$\mathscr{A} := \left(\left(X_f^{(k)} \right)_{1 \leq k \leq m} \right)_{f \in \mathcal{F}(\rho)}$$

Conditionally to $\mathcal{T}, \mathcal{S}$, all nucleotides of the sequence evolve independently and identically (we will say shortly *are iid* in the following). We may thus call for ease X_w instead of $X_w^{(k)}$ when the index does not matter. We consider that each nucleotide evolves according to a Markov process with discrete state space $\{A, T, C, G\}$. We moreover assume that it is reversible, with invariant law $\pi = (\pi_A, \pi_T, \pi_C, \pi_G)$, characterized by the intensity matrix :

$$Q = (q_{ij}) = \begin{pmatrix} -(a\pi_T + b\pi_C + c\pi_G) & a\pi_T & b\pi_C & c\pi_G \\ a\pi_A & -(a\pi_A + d\pi_C + e\pi_G) & d\pi_C & e\pi_G \\ b\pi_A & d\pi_T & -(b\pi_A + d\pi_T + f\pi_G) & f\pi_G \\ c\pi_A & e\pi_T & f\pi_C & -(c\pi_A + e\pi_T + f\pi_C) \end{pmatrix}$$

This is known as the *Generalized Time Reversible* model in the literature (GTR, Lanave et al. (1984)). Many special cases of the model have been described in the literature and recalled in section 1.2.1. We suppose that we fixed one, allowing us to integrate analytically the Kolmogorov equation to derive the transition probabilities between any two states displayed by nodes w and w_1 separated by time t.

$$P(t) = (\mathbb{P}(X_{w_1} = j | X_w = i))_{i,j} = e^{tQ}$$

When a spike is encountered, the transition probabilities P_S for a nucleotide state just before and just after the spike are such that:

$$\begin{aligned} \forall i, j, \quad (P_S)_{i,i} &= (1 - \kappa) \\ (P_S)_{i,j} &= \kappa \frac{q_{i,j}}{-q_{i,i}} \end{aligned}$$

We will need the transition probabilities on a branch with duration t and n_S spikes, which we will denote $P(n_S, t)$. For now, we suppose that P_S and $P(t)$ commute, which happens if and only if Q and P_S commute. This in turn happens iff $q_{ii} = q_{jj}, \ \forall i, j$. We can then write :

$$P(n_S, t) = (P_S)^{n_S} P(t)$$

Models JC69 (Jukes and Cantor, 1969) and K80 (Kimura, 1980) satisfy this assumption. In order to use others, we would need to work a bit more and integrate over the positions of spikes along a branch.

Because all nucleotides are iid conditionally on $\mathcal{T}, \mathcal{S}$, we describe the simulation procedure for one given position:

Algorithm 11 *(simulation of a present-day alignment conditional on $\mathcal{T}, \mathcal{S}$)*

i) Initialize the simulation at the root : $X_\rho \sim \pi$.

ii) Let w_1 be a daughter of node w, displaying nucleotide state X_w, after a branch of duration t on which n_S spikes appeared.
The law of X_{w_1} is given by the line corresponding to state X_w in matrix $P(n_S, t)$.

iii) Recursively simulate nucleotide states in descending nodes, until reaching present.

Moreover, our description of molecular evolution allows us to compute the likelihood of sequences at the leaves, conditional on $\mathcal{T}, \mathcal{S}$. Here again, we only focus on one nucleotide position. This is done using a popular *pruning algorithm*, already described in a more general way as algorithm 3 in section 1.2. Recall that we need to define, at any node w, the following conditional likelihood:

$$L_w = \left(\mathbb{P}((X_f)_{f \in \mathcal{F}(w)} \mid X_w = i)\right)_{i \in \{A,T,G,C\}}$$

The adaptation to our specific problem is detailed below.

Algorithm 12 *(likelihood of one column of the tip alignment, knowing $\mathcal{T}, \mathcal{S}$)*

i) Initialize, for any leaf f :

$$L_f = (\mathbb{1}_{X_f=A}, \mathbb{1}_{X_f=T}, \mathbb{1}_{X_f=G}, \mathbb{1}_{X_f=C})$$

ii) Then recursively compute L_w at any node w.
If w is the mother of nodes w_1 and w_2, with respectively n_{S1} et n_{S2} spikes on branches leading to w_1 and w_2, and with branch lengths t_1 and t_2, then

$$L_w = (P(n_{S1}, t_1) L_{w_1}) \cdot (P(n_{S2}, t_2) L_{w_2})$$

iii) Once the root ρ is reached, the likelihood of the tip alignment can be expressed as :

$$L = \pi L_\rho$$

Note that, when the Markov process is reversible, this likelihood computation does not depend on the position of the root on the tree.

In the following section, we will present tools to perform statistical inference under our model. We will in particular make use of the simulation, and likelihood computation, algorithms of:

— $\mathscr{S}$ on a fixed reconstructed tree $\mathcal{T}$,

— $\mathscr{A}$ on a fixed reconstructed spiked tree $\mathcal{T}, \mathcal{S}$.

5.3 Statistical inference in a Bayesian framework

In this section we aim at proposing an inference framework for the previously described model of molecular evolution with spikes.

We first tried to make inferences based on the maximum likelihood. We thus designed Monte-Carlo algorithms to compute the likelihood of an alignment, integrating over the space of all spike positions. However, because we did not come with a practical way to optimize this likelihood yet, these derivations are kept only in the appendix section D.2.

We also followed a second lead, which proved to be more successful, consisting in considering our model in a Bayesian framework. We simplify a bit the notation previously introduced, first giving a name to the density of spikes on a tree:

$$f(\mathcal{S}) := \mathbb{P}(\mathscr{S} = \mathcal{S} \mid \mathscr{T} = \mathcal{T})$$

And second, calling all other densities l. The name of the quantity considered giving us a hint of which density is considered, e.g. :

$$\begin{aligned} l(\mathcal{A}|\mathcal{S}) &:= \mathbb{P}(\mathscr{A} = \mathcal{A} \mid \mathscr{S} = \mathcal{S}, \mathscr{T} = \mathcal{T}) \\ l(\mathcal{A}) &:= \mathbb{P}(\mathscr{A} = \mathcal{A} \mid \mathscr{T} = \mathcal{T}) \end{aligned}$$

We now use these notations to sketch a Markov Chain Monte Carlo (MCMC) procedure aimed at inferring the joint posterior distribution of parameters and spike positions.

5.3.1 Method principle

Suppose we have fixed the tree law and the tree realisation, i.e. we know values of the diversification rates (b, d), the tree topology and the times at which branching events happened. Furthermore, suppose parameters ν, κ, α are not fixed anymore, but are instead random variables with uniform law on $(0, 1)$. We aim at determining the *posterior distribution* :

$$\begin{aligned} l(\mathcal{S}, \nu, \kappa, \alpha \mid \mathcal{A}) &= \frac{l(\mathcal{A} \mid \mathcal{S}, \kappa, \alpha) f(\mathcal{S} \mid \nu) l(\nu, \kappa, \alpha)}{l(\mathcal{A})} \\ &\propto l(\mathcal{A} \mid \mathcal{S}, \kappa, \alpha) f(\mathcal{S} \mid \nu) \end{aligned}$$

This type of question is classically resolved using a MCMC algorithm to sample the desired distribution. Because we already presented how to compute $f(\mathcal{S} \mid \nu)$ (cf. algorithm 10) and how to compute $l(\mathcal{A} \mid \mathcal{S}, \kappa, \alpha)$ (cf. algorithm 12), we need only describe two additional components: (i) the initialization of the chain, which provides the first values of the parameters, as well as the spike positions, and (ii) the movement proposal, which provides the transitions between two steps of the chain.

5.3.2 Initialization of the chain

Without any knowledge on the parameter values, we initialize them using a uniform law on $(0, 1)$. However, the first spike positions could be chosen more carefully, by first analyzing the alignment. We wish to use present-day data to annotate branches on which there is *more differences than expected*, in order to tune an interesting law g.

We call D_{w_1,w_2} the random variable counting the number of nucleotides that are in a different state in two extremities w_1 and w_2 of a branch. More precisely, we wish to compare :

i) the expected number of differences on the branch : $\mathbb{E}(D_{w_1,w_2})$.

ii) the expected number of differences conditional on present-day data $\mathbb{E}(D_{w_1,w_2} \mid \mathscr{A} = \mathcal{A})$.

We provide the procedure allowing us to compute these quantities for each branch in the appendix section D.1. The first one is derived analytically from the knowledge of the substitution model. The second one requires a slight modification of the pruning algorithm 12, incorporating an additional depth traversal of the tree.

We measure the difference between the two quantities, informing us on the departure from what we would expect without any spike :

$$x := \frac{\mathbb{E}(D_{w_1,w_2} \mid \mathscr{A} = \mathcal{A}) - \mathbb{E}(D_{w_1,w_2})}{\mathrm{Var}(D_{w_1,w_2})}$$

When x is *large enough*, we would like to place a spike with higher probability along the branch. We chose to put a spike on a branch with a probability $b(x)$, where b is the logistic function :

$$b(x) := \frac{1}{1 + ae^{-x}}$$

We further chose a such that, as soon as x is large enough (i.e. $x > \gamma$), there is a probability $b(x) > \nu$ to place a spike. This gives us :

$$b(\gamma) = \nu \iff a = e^{\gamma}\frac{1-\nu}{\nu}$$

Because f places more than one spike on a branch with non-zero probability, we would like g to have the same behaviour. We will consider that g is the convolution of a Bernoulli with parameter $b(x)$ and a Poisson with parameter α.

5.3.3 Movement proposal

We now describe the movement proposal, that we write $q((\mathcal{S}, \nu, \kappa, \alpha)\ ,\ \cdot)$, aimed at drawing a new state $(\mathcal{S}', \nu', \kappa', \alpha')$ at each step of the chain.

First, we consider the same transitions for the three parameters ν', κ', α', which will be drawn in a Gaussian distribution centered respectively on ν, κ, α, with variance σ^2, and conditioned on staying in the interval $(0, 1)$.

The transition from the spike configuration $\mathcal{S}$ to another configuration $\mathcal{S}'$ is built on the same principle than the g distribution previously discussed. We compute, on each branch of the tree, the quantity:

$$x := \frac{\mathbb{E}(D_{w_1,w_2} \mid \mathcal{A}, \mathcal{S}) - \mathbb{E}(D_{w_1,w_2} \mid \mathcal{S})}{\mathrm{Var}(D_{w_1,w_2} \mid \mathcal{S})}$$

We then either:

i) add a spike on the branch with probability $b(x)$,

ii) or remove a spike (if there is one) with probability $b(-x)$.

This ends the description of the distribution proposal q, which is central in the following MCMC algorithm. Note that q depends on two parameters that can be chosen by hand so as to achieve a faster convergence: σ adjusts the size of the steps for the new parameter set, while γ adjusts the propensity to add new spikes (see the description of function b above).

Algorithm 13 *(Metropolis-Hastings MCMC to sample $l(\mathscr{S}, \nu, \kappa, \alpha \mid \mathscr{A})$)*

i) Initialize the chain by drawing:

$$\begin{aligned}(\nu_0, \kappa_0, \alpha_0) &\sim \mathcal{U}(0,1) \otimes \mathcal{U}(0,1) \otimes \mathcal{U}(0,1)\\ \mathcal{S}_0 &\sim g\end{aligned}$$

ii) Repeat the following procedure a large number *of times :*

(a) At step t, draw :

$$(\mathcal{S}'_{t+1}, \nu'_{t+1}, \kappa'_{t+1}, \alpha'_{t+1}) \sim q((\mathcal{S}_t, \nu_t, \kappa_t, \alpha_t)\ ,\ \cdot)$$

(b) Compute :

$$\beta = \min\left(1,\ \frac{l(\mathcal{A} \mid \alpha'_{t+1}, \kappa'_{t+1}, \mathcal{S}'_{t+1}) f(\mathcal{S}'_{t+1} \mid \nu'_{t+1}) q((\nu'_{t+1}, \kappa'_{t+1}, \alpha'_{t+1}, \mathcal{S}'_{t+1}), (\nu_t, \kappa_t, \alpha_t, \mathcal{S}_t))}{l(\mathcal{A} \mid \alpha_t, \kappa_t, \mathcal{S}_t) f(\mathcal{S}_t \mid \nu_t) q((\nu_t, \kappa_t, \alpha_t, \mathcal{S}_t), (\nu'_{t+1}, \kappa'_{t+1}, \alpha_{t+1}, \mathcal{S}'_{t+1}))}\right)$$

(c) With probability β, fix $(\nu_{t+1}, \kappa_{t+1}, \alpha_{t+1}, \mathcal{S}_{t+1}) = (\nu'_{t+1}, \kappa'_{t+1}, \alpha'_{t+1}, \mathcal{S}'_{t+1})$.

(d) With probability $1 - \beta$, keep $(\nu_{t+1}, \kappa_{t+1}, \alpha_{t+1}, \mathcal{S}_{t+1}) = (\nu_t, \kappa_t, \alpha_t, \mathcal{S}_t)$.

The chain $(\nu_t, \kappa_t, \alpha_t, \mathcal{S}_t)$ tends to its stationary distribution when t tends to ∞. This stationary distribution is precisely our target : $l(\mathscr{S}, \nu, \kappa, \alpha \mid \mathscr{A})$.
We delete the beginning of the chain (called burn-in*) and use the remainder as an estimate of the target distribution.*

5.3.4 Inferences on simulations

We fix parameter values of $(b, d, \nu, \kappa, \alpha)$ by hand. We first simulate a reconstructed tree under a birth-death process with parameters b, d. Then, using algorithms 9 and 11, we simulate a spike scenario on a fixed reconstructed tree, on which we further simulate sequence evolution. This provides us with a known tree, known diversification rates (b, d) as well as simulated sequences at the leaves. Everything else is forgotten.

We then apply algorithm 13, and recover a sample of the posterior distribution, i.e. the distribution of $\alpha, \kappa, \nu, \mathcal{S}$ knowing the alignment at the tips of the tree.

In order to get a visual output, we summarize the marginal distribution of spike configurations. On each branch, we compute the average number of spikes over the whole distribution. The inferred distribution can be compared to the parameter values that were chosen, and to the initially simulated spike scenario. An inference run on one specific simulated data set is summarized as on Figure 5.2.

This inference procedure is still in its infancy, and more work would be needed before applying it on real data. We detail below the future developments that we see ahead for this project.

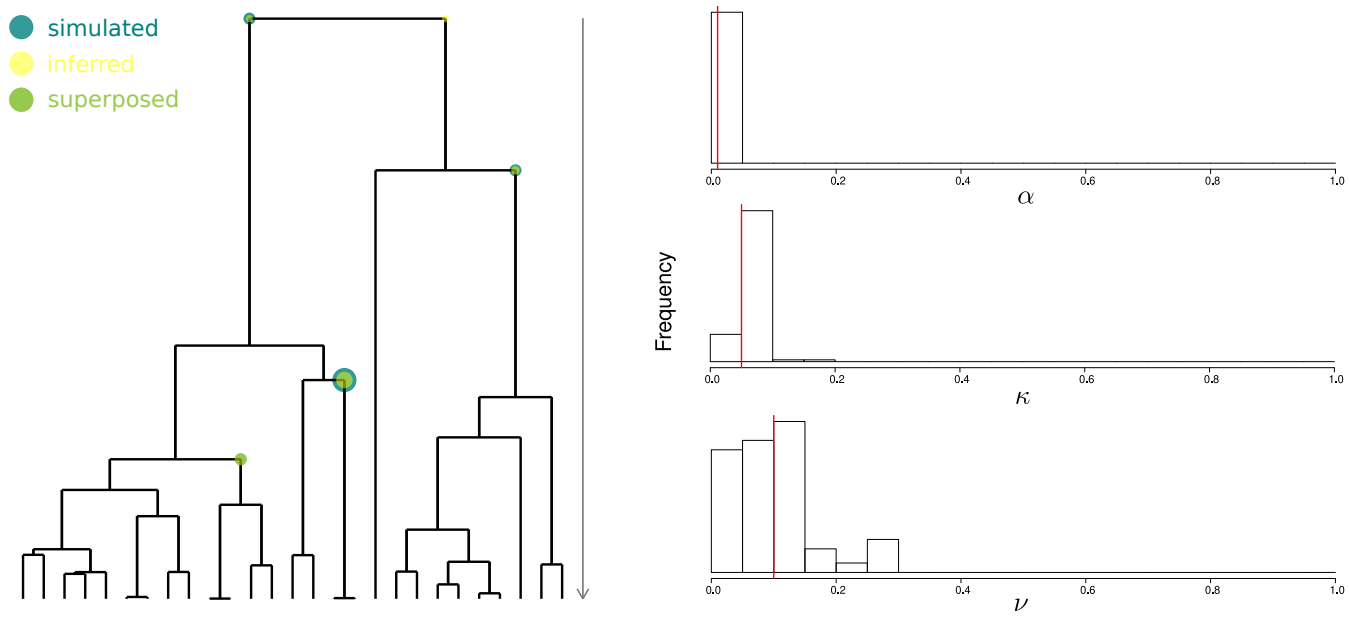
5.4 Future developments of the project

We specify below the next directions that we would like to explore in this ongoing study.

5.4.1 Improvement of the inference method

The very first modifications that we need to undertake concern the inference method that we just presented above. Although it converges to the target distribution, and recovers the simulated scenario, we are not satisfied with its current efficiency. The *mixing* of the MCMC is rather poor, meaning that the chain remains for a long time in exactly the same state. As a result, the convergence toward the target distribution is rather slow and a quite high number of steps is needed to perform meaningful inferences. The first axis of future developments will thus be to work on the proposal distribution to improve the algorithm efficiency.

The research of a better efficiency would not be so essential if it were only to infer the spike positions and the parameter values on a fixed tree with known diversification rates and branch lengths as we presented above. However, we would very much like the inference method to jointly infer the diversification parameters and the branch lengths of the tree as well. Without this development, we would



(a) Spike summary.

(b) Parameter distribution.

Figure 5.2 – An example of posterior distribution of spikes and parameters, on simulated data. In (a) simulated spikes on each branch are in blue, and the inferred average number of spikes on each branch is in transparent-yellow. Both are placed at the beginning of the branch. The radius of the disk is proportional to the number of spikes. When both superpose, it leads to a green disk. In (b) parameters used in the simulation are represented with red lines. The histogram corresponds to the posterior distribution.

need to assume that a totally disjoint DNA alignment has been used to reconstruct a dated tree, before studying the spike patterns of another gene alignment. However, in case the spike pattern of the first and second alignments are linked, the departure from a strict clock would already be taken into account in the tree dating, thus biasing our study. Moreover, jointly inferring the branch lengths would make this study conform to the standards in the field of *relaxed tree dating*, and comparisons with other relaxed molecular clocks would be facilitated.

5.4.2 Comparison to other relaxed molecular clocks

Assuming that we will be able to improve the inference procedure and implement the joint inference of branch lengths, spikes positions, and parameter values, the next step consists in comparing our outputs to the outputs of other relaxed molecular clocks.

A *molecular clock* is a model of molecular evolution specifically designed to infer the branch lengths of phylogenies. The history of molecular clocks goes back to the work of Zuckerkandl and Pauling (1962) who observed that pairwise differences between extant sequences were approximately compatible with the branch lengths of an ultrametric tree. They proposed a model that is now referred to as the *strict molecular clock*, which supposes that sequences evolve at a fixed rate along the branches of an ultrametric tree. Two decades later, Felsenstein (1981) introduces its method to date phylogenies by maximum likelihood, which relies on a model that we could refer to as a *fully relaxed molecular clock*, because he assumes that sequences evolve according to a fixed rate, on a non-ultrametric tree (i.e. with no constraint on any branch length). In between these two extremes, *relaxed molecular clocks* have been flourishing in recent years, for they propose to accomodate the departure from the strict molecular clock in a much less parameter rich way than the fully relaxed clock. The review by Lepage et al. (2007) provides a thorough, if not so recent, review of these models. They all rely on a superposition of model layers, which we briefly sketch below.

The first model layer consists in assuming a tree generating model. The law of the tree branch lengths (T in Fig. 5.3) could be a birth-death process (as we supposed in our study), but any other distribution might work as well. Second, a model of substitution rate variation (R in Fig. 5.3) along the tree is assumed. These substitution rates have been proposed to be drawn independently on each branch in a given distribution (a modeling choice referred to as a *non-autocorrelated clock*). Alternatively, the substitution rate can evolve as a Markov process along the tree (a modeling choice referred to as an *autocorrelated clock*). The third and last layer is a model of molecular evolution like the ones presented in section 1.2.1, unfolding along a tree displaying the substitution rate, and generating a sequence alignment (A in Fig. 5.3).

Figure 5.3 – A relaxed molecular clock relying on three layers. Letter T stands for Tree, R stands for Rate of substitution, and A stands for Alignment.

Bayesian inference relying on MCMC is well suited for this category of models, because the posterior probability is proportional to the product of probabilities that our three layers allow us to compute (see Fig. 5.3).

Following such a description, it is clear that our study fits well within the framework of relaxed molecular clocks. The two unconventional hypotheses considered here being:

i) instead of a rate variation R along branches, our model accomodates the departure from the strict clock by placing punctual spikes of mutations.

ii) these spikes are highly correlated with the branching events.

It will thus be very interesting to compare the outputs of models assuming rate change along the tree, with the outputs of this model with spikes. The branch-lengths are expected to be very close, but the main difference will probably be in terms of the biological interpretations that researchers will be inclined to express in front of a colored tree (with distinct rates along branches) vs a spiked tree. Finally, distinct relaxed clocks can be compared in a Bayesian model selection framework, following the methodology of Lepage et al. (2007) who proposed to compute the Bayes factor (i.e. the probability of the alignment, integrated over all other random variables) under each model. In particular, this could allow us to assess the relative fit of the spike model versus other relaxed clocks, on specific datasets.

5.4.3 Application to empirical data

The last step forward that we envision for this study will be to make inferences on empirical data. We have not yet decided the precise dataset, and it may depend on our ability to conduct inferences on large sized datasets or not.

The application to empirical data will probably require another technical improvement of the method. Indeed, we did not yet take into account the substitution rate heterogeneities across nucleotides in the sequence, while the phenomenon seem ubiquitous and is taken into account in virtually all tree dating studies. The mot popular way to tackle the issue consists in considering that each nucleotide's mean rate is drawn in a fixed distribution. Usually, it is a Gamma distribution, discretized for computational reasons into four distinct rate classes. The probability of the alignment is further obtained by integrating out the rate of each nucleotide's position.

The application to empirical data will be the opportunity to study the heterogeneities in the molecular evolution process across loci at a larger scale too. It would be interesting to look for distinct patterns of spikes positions across distinct genes, especially for genes having a known function (e.g. genes involved in resource use, or in conspecific recognition) that could have an impact or be impacted by the speciation process. Empirical studies at the intra-species level suggest indeed that recurrent mutations at a few pleiotropic loci coordinate the divergence of populations (Peichel and Marques, 2017).

Last, it would be interesting to question the link between the diversification and the spike process that we hypothesized. For now, and it is the originality of the approach, spikes happen only at branching times. But nothing ensures that it really fits the data better than a naive model were spikes would be drawn as a Poisson process along the tree. These two alternatives could be tested on empirical data.

5.5 Conclusion

As this is still ongoing work, we provide only a temporary conclusion. We proposed a relaxed clock model incorporating two unconventional hypotheses: (i) departure from the strict molecular clock is accomodated through the presence of punctual events called *spikes*, representing a large number of synchronous substitutions along the sequence, and (ii) these spikes are tightly linked to the tree generating model, for they arise only at speciation events.

We propose an inference procedure in a Bayesian framework, which allows us to recover the spike positions and the parameters of the model, on a fixed dated tree. We will hopefully extend this procedure to the joint inference of the tree branch lengths in future work.

The originality of this approach is mainly a matter of biological interpretation. Whereas all other molecular clocks build on the idea of *anagenetic gradual change*, we propose to reconsider the concept of *cladogenetic punctuated evolution* in the field of molecular evolution. We highlight that, whereas autocorrelated clocks can be explained by their link with other life-history traits evolving through time (like body mass or longevity, see Lartillot and Poujol (2011)), non-autocorrelated clocks do not rely on any precise biological process. Moreover, recent work by Lartillot et al. (2016) found support for both an autocorrelated part and a non-autocorrelated part in a molecular clock fitted on empirical data. We suggest that replacing the non-autocorrelated part in the clock by a spike process could ease the interpretation.

Chapter 6

Conclusion

We presented our work in diverse sub-areas of macroevolution modeling in previous chapters. We focused on individual-based modeling of diversification (chapters 2 and 3), continuous phenotype evolution (chapter 4) and molecular evolution (chapter 5).

In this concluding chapter, we first present a synthesis of our results. We then discuss how these distinct sub-areas of macroevolution modeling can be articulated together in order to get a more accurate picture of life evolution over long timescales. We finally discuss the many modeling choices and challenges that researchers in macroevolution face.

Contents of the chapter

6.1 Synthesis

We presented four distinct research chapters in this thesis, each challenging some commonly used hypotheses and each trying to propose statistical methods that could help researchers choose the right ones using empirical data.

These four chapters address this question in three closely related, yet distinct, sub-domains of macroevolution: chapter 2 and chapter 3 will particularly appeal to researchers interested in diversification and individual-based modeling, while chapters 4 and 5 focus on trait evolution (either continuous traits or discrete, genomic, ones). These two topics are explored by an only partially overlapping research community, and yet they both consist in modeling processes happening over similar timescales.

In chapter 2 we proposed two new species definitions that naturally emerge from the consideration of biologically reasonable desirable properties. These two definitions can readily be used in modeling work relying on the same very popular, yet crude, framework, i.e. taking into account clonally reproducing particles subject to point mutations. They challenge the most commonly used hypotheses of speciation at the individual level, and aim at facilitating the integration of two distinct scales together, e.g. clustering individuals into species, or species into genera. These two species definitions have been further studied and extended recently by Hoppe et al. (2017).

In chapter 3 we considered one of these two definitions (the *loose* one), and implemented it in an individual-based model of diversification. This model challenges two common hypotheses in the literature based on the Neutral Theory of Biodiversity: (i) the speciation process, and (ii) the underlying metapopulation dynamics of individuals, which is chosen to be a birth-death process instead of a constant-size metapopulation. We showed that this set of hypotheses is sufficient to produce phylogenies with shapes close to empirical ones. The likelihood of a tree can further be computed and optimized to make inferences on the population dynamics parameters.

In chapter 4 we turned to models of continuous trait evolution, and tried to incorporate more ecology into those. While most models of trait evolution suppose that traits evolve independently in distinct lineages, we describe a way to take into account ecological interactions between species, that could drive the trait evolution through time. We designed a method to compute the likelihood of trait data on a phylogeny, under these models of trait evolution with interspecific interactions. This can allow one to assess, e.g. the importance of interspecific competition in driving some traits evolution (Drury et al., 2016).

Last, we sketched a new project in chapter 5 aimed at contrasting two hypotheses on trait evolution. The first one, generally considered in the literature on *relaxed clock models*, considers that the rate of molecular evolution evolves on the tree, and that molecular substitutions happen regularly along branches according to this rate. The second one considers that substitutions would happen regularly along branches, but that additionally *spikes* of mutations would take place at some particular speciation events. These spikes of mutations would correspond to a number of mutations happening very quickly at the macroevolution scale, which might have played a role in the speciation process. Intuitively, contrasting both hypotheses boils down to asking whether the departure from the strict clock is or is not correlated with the number of speciation events separating distinct lineages.

This might now be the right time to introduce table 6.1 summarizing the various hypotheses considered throughout the distinct chapters.

Because each chapter already comes with its own discussion and perspectives, we dedicate the following section 6.2 of this general discussion on examining some ideas to articulate them together. We then turn, in section 6.3, to general considerations on modeling work in macroevolution, and seize the opportunity to provide our thoughts on what we believe could be the most promising research directions in macroevolution.

Chapter	Unconventional hypothesis	Motivation
2	Presents two new species definitions for individual-based models, called *lacy* and *loose* species definitions.	Tying the individual and species level, while being able to base the phylogeny of species on the genealogy of individuals.
3	Implements the *loose* species definition, together with a metapopulation dynamics following a birth-death process.	Reproducing known macroevolutionary patterns from simple micro-evolutionary processes. Fitting the birth and death rate of the metapopulation from the phylogenies.
4	Incorporates the effect of interspecific interactions into models of continuous phenotypic evolution.	Assessing the effect of ecological interactions on trait evolution over macroevolutionary timescale.
5	Proposes a relaxation of the strict molecular clock, with *spikes* of mutations happening at speciation times.	Proposing a novel framework for modeling the interaction between molecular evolution and diversification. Testing the relative support for punctualism vs. gradualism in molecular data. Looking for genes supporting patterns of spikes at speciation.

Table 6.1 – A summary of *unconventional hypotheses* considered in the distinct chapters of this thesis, together with the rationale for considering them.

6.2 Drawing links between chapters

We provide first some hints on further linking chapters 2 and 3, to contrast distinct species definitions based on empirical data. We then give our thoughts on how to tie chapters (2,3) and chapters (4,5), i.e. on how to take into account the individual level within models of trait evolution. Finally, we discuss the connections that may be possible to draw between chapters 4 and 5, i.e. between models of phenotypic trait and genomic trait evolution. As artificial as this exercise might seem, because these distinct models have clearly not been designed in the first place to be combined together, it nevertheless allows us to give general thoughts on what we believe could be future directions for people interested in bridging gaps between (i) individual-based, (ii) diversification, and (iii) trait evolution, models.

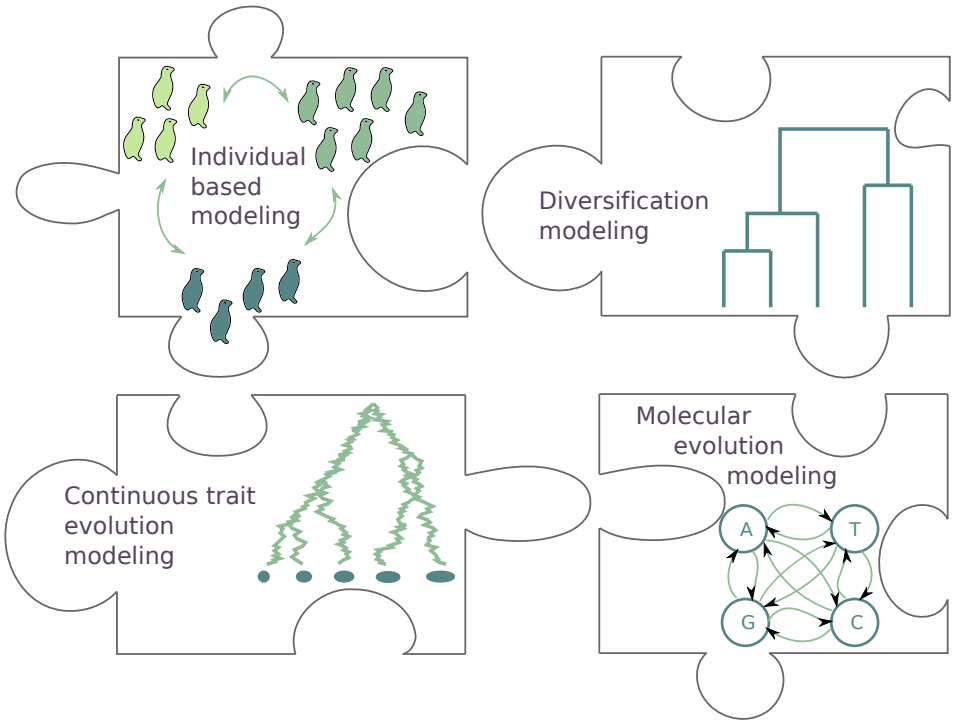


Figure 6.1 – Assembling a modeling *jigsaw puzzle* with the four thematics developed in this thesis.

6.2.1 Linking the individual-based modeling of species and diversification studies

In chapter 3, we proposed the first individual-based model implementing one of the two species definitions presented in chapter 2, and concluded that this model was overall able to reproduce patterns of imbalance and branching times observed in empirical phylogenies.

Because the model differs from what has been proposed in the literature with respect to two important hypotheses, two related questions arise:

i) Could we confront the lacy and loose species definitions with empirical data ?

ii) What is the importance of the metapopulation dynamics scenario to fit empirical data?

Answering these questions would require building other models, incorporating the missing hypotheses (see Table 6.2), before studying their relative fit to empirical phylogenies. We do not have the ambition to rigorously carry out such a study, but we nevertheless aim at providing some clues on this issue in what follows.

Metapopulation dynamics	Loose species	Lacy species	Phenotypic species
Birth-death process	✓	✗	✗
Constant size	✗	✗	✓

Table 6.2 – Missing sets of hypotheses in the literature.

Confronting species concepts with empirical data

Both the lacy and loose species can be determined based on the knowledge of a genealogy with point mutations, using the algorithm provided in chapter 2.

Suppose the metapopulation is constant, with very large size N, and the genealogy of individuals is given by a Wright-Fisher model. Then the genealogy of $n \ll N$ individuals, sampled irrespective of the species they belong to, and rescaled in time by a factor N, is a Kingman coalescent (see section 1.1.4 in introductory chapter 1).

Suppose additionally that phenotypic differentiation is superimposed on this genealogy, using a Poisson process with parameter ν and an infinite allele model, as is most often assumed in individual-based models. The resulting Kingman tree with point mutations can be transformed into the lacy and loose phylogenies, and the shapes of these phylogenies can be compared. We extract in particular, for each tree, two summary statistics that were already used in chapter 3:

γ an index summarizing information of branch lengths.
The bigger γ, the *tippier* the tree, i.e. nodes occur preferentially near the tips.

β an index summarizing information on tree balance.
The bigger β the more balanced the tree. See the precise description in section 1.1.4.

Results are shown in Figure 6.2.

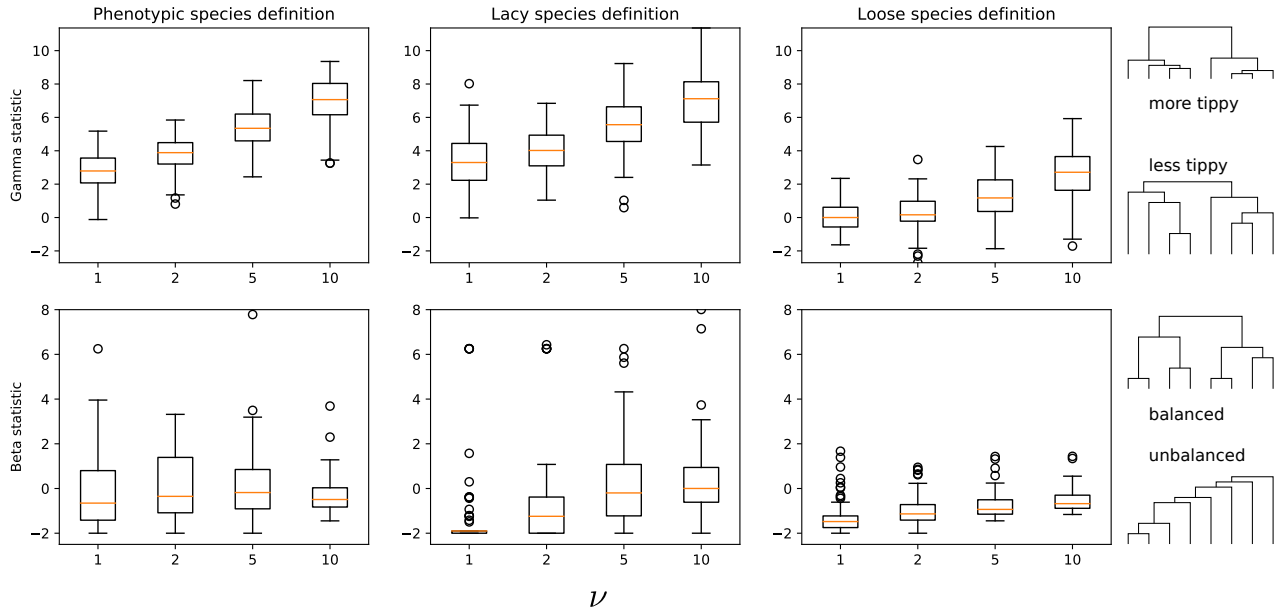


Figure 6.2 – Distribution of γ and β indices over simulated trees. Kingman trees with 200 leaves are simulated. Point mutations are superimposed as a Poisson process with parameter ν. The phenotypic, lacy and loose phylogenies are obtained according to the procedure exposed in chapter 2.

The first observation is that these simulated trees have a shape that differs from the one typically found in empirical data. The distribution of the γ statistic on empirical trees is indeed centered around -1. A closer look at the γ boxplot reveals a similar pattern of γ statistics for the phenotypic and lacy species definition applied on these Kingman genealogies. Only the loose species definition, which only retains the deepest nodes of the genealogy up to a given level, has the potential to lower significantly the γ statistic on a genealogy initially very tippy.

In contrast, the comparison of β values obtained with these three species definitions provides much less guidance to discriminate the three definitions. Phylogenies can roughly exhibit the same range of imbalance than empirical ones. The distributions seem only narrower in the loose phylogenies as compared to the other ones.

Differentiating the effects of species concepts and population dynamics

The loose species definition is one of the two unconventional hypotheses that were considered in chapter 3. The second one, concerning the metapopulation dynamics, has also an influence on the shape of the resulting phylogenies. In chapter 3, we emphasized the ability of the method to take into account any shape of the birth and death rates with time : $b(t)$ and $d(t)$. We now take advantage of this to demonstrate the influence of the metapopulation dynamics on the tree shape summary statistics.

We chose four distinct metapopulation dynamic scenarios. Under the four scenarios, the mutation rate ν and the birth rate b remain constant, respectively equal to 1 event per Myr and to 10^6 events per Myr and the tree height is 5 Myr. We varied the growth rate $b - d$ through time as follows:

scenario1 constant growth-rate.

$$\forall t \in (0, 5), \quad b - d(t) = 0.5$$

scenario2 period of positive growth-rate followed by short period of negative growth-rate:

$$\begin{aligned} \forall t \in (0, 4), \quad & b - d(t) = 0.7 \\ \forall t \in (4, 5), \quad & b - d(t) = -0.3 \end{aligned}$$

scenario3 short period of negative growth-rate followed by period of positive growth-rate:

$$\begin{aligned} \forall t \in (0, 1), \quad & b - d(t) = -0.3 \\ \forall t \in (1, 5), \quad & b - d(t) = 0.7 \end{aligned}$$

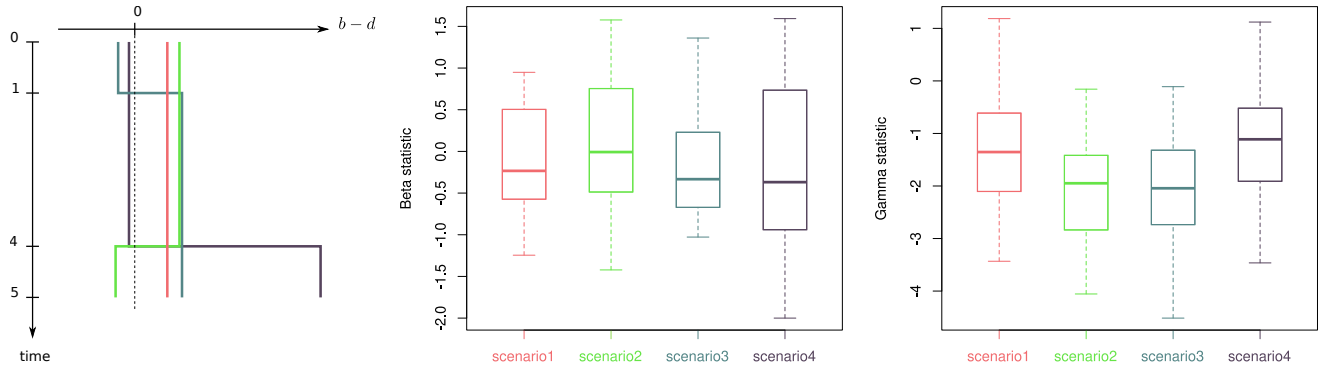
scenario4 period of negative growth-rate followed by short period of positive growth-rate:

$$\begin{aligned} \forall t \in (0, 4), \quad & b - d(t) = -0.1 \\ \forall t \in (4, 5), \quad & b - d(t) = 2.9 \end{aligned}$$

Figure 6.3 shows the results of 50 simulations under each scenario. The β statistic does not seem much different in the four distributions. However, metapopulation dynamics have an impact on the γ statistic. More or less tippy trees can be obtained by varying the timing of parameter change. The tippier trees are obtained under scenario 4, where an exponential increase of metapopulation size occurs near present-time.

6.2.2 Linking trait evolution and individual-based species modeling

The literature already provides nice leading examples intenting to bridge the gap between individual-based modeling and phenotypic evolution modeling. We review here some of these and attempt to transpose these ideas to our work.



(a) Parameters. (b) β statistic. (c) γ statistic.

Figure 6.3 – Tree shape statistics under four distinct metapopulation dynamics. In (a) the $b - d$ value through time is provided for the four scenarios. The middle dotted line represents zero growth rate. In (b) and (c) the β and γ distributions of 50 simulations are plotted.

From a single mean species phenotype to a set of individual phenotypes

Most phylogenetic comparative methods (PCMs) were originally derived from microevolutionary hypotheses (Lande, 1985; Hansen and Martins, 1996). Even today, microevolutionary hypotheses continue to be the motivation for the development of new PCMs. Nuismer and Harmon (2014) justified e.g. the *Matching Competition* model, which inspired both the framework that we presented in chapter 4 and the empirical application presented in appendix E, with microevolutionary principles. In spite of finding their roots in quantitative genetics, modern tools derived in the field now exclusively aim at modeling the evolution of the mean lineage phenotype through deep time, without any reference to the individual-scale parameters. Some propositions have nevertheless been made to take into account information on individual scale.

The first, and most commonly used, method to take into account across-individual variability in PCMs consists in measuring the variance of the trait within each species. This variance is considered as a measurement error, and is supposed to correspond to the variance of the diffusion process at the leaves (Ives et al., 2007). The advantage of this approach is that it can be readily implemented in virtually all PCMs, and in particular in the type of model that we studied in chapter 4. The main drawback is that it is not process-based, and it therefore does not take into account properly the individual-scale. The signal in intra-specific variability could be better captured by fully specifying the relatedness between individuals within species, and the within-species evolution of the trait.

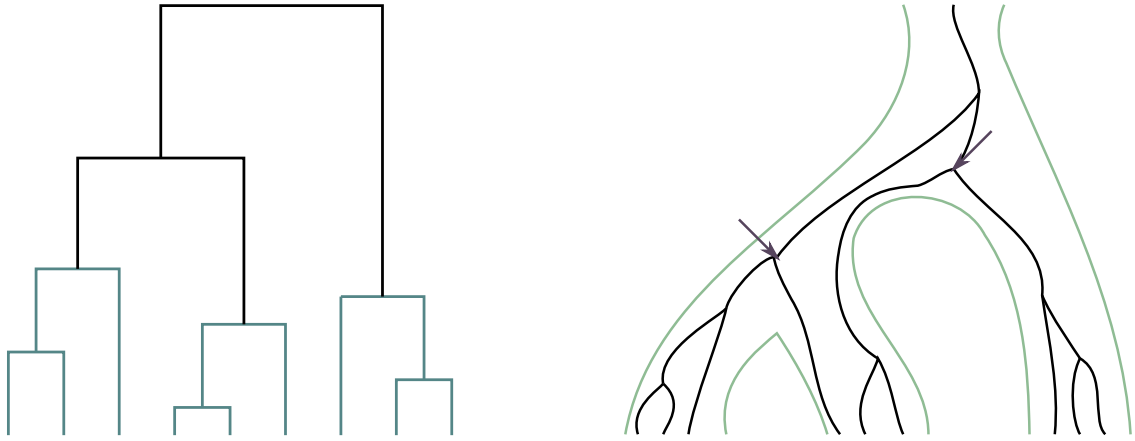
Other authors have proposed to go back to the quantitative genetic framework roots, modeling the evolution of a phenotype within a population through a small number of generations, and study how it scales up to the macroevolutionary scale for parameters chosen in a realistic range (Estes and Arnold, 2007). The model that best fitted empirical observations of divergence at all scales was a model of stabilizing selection around a fitness optimum moving within fixed limits. Hohenlohe and Arnold (2008) proposed an inference framework that includes both individual-scale measurements and mean measurements across species. Last, we refer to an article by Hadfield and Nakagawa (2010), who summarized the many connections existing between models and parameters stemming from population genetics and PCMs.

Clustering species based on individuals genotypes

The idea of taking into account the individual scale in models of trait evolution has been much more studied when the traits are molecular. This part of the literature is mostly interested in proposing methods to cluster individuals into distinct putative species based on their genotype. Distinct models have

been proposed, but they all share the same underlying idea, which consists in looking for a gap between small intra-species divergence and large inter-species divergence.

The most process-based models assume a prior on the intra-species gene tree that is different from the prior of the inter-species tree. They further assume a transition from the inter- to the intra-species level. For example, Fujisawa and Barraclough (2013) proposed a model called *GMYC* for *Generalized Mixed Yule Coalescent* which assumes that the intra-species gene tree is a Kingman tree, and that the inter-species tree is a birth-death process (see Fig. 6.4a). Zhang et al. (2013) proposed another closely related model called *PTP* for *Poisson Tree Process*, which assumes that intra- and inter-species trees follow a birth-death process with distinct rates. In both models, molecular evolution then unfolds along the tree, resulting in the observed alignment comprising many sequences for each species.



(a) Generalized Mixed Yule Coalescent (b) Multispecies coalescent

Figure 6.4 – In (a), the blue parts correspond to the intra-species coalescences, followed by the black, inter-species coalescences. In (b), the species tree is a guide that allows the coalescence between individuals. Arrows highlight examples of Incomplete Lineage Sorting. The left one leads to non-monophyletic species, while the right one is responsible for the incongruence between the gene tree and the species tree topologies.

Bayesian methods have also been proposed to tackle the issue of species delimitation sketched above, under more complex underlying models (Yang and Rannala, 2010). They adopt the *multispecies coalescent* framework introduced by Maddison (1997), which considers the gene trees to be embedded in the phylogeny. Each gene has its own tree, and gene lineages can coalesce backward in time only if they stand in the same phylogenetic lineage, as illustrated in Figure 6.4b. This framework can take into account multiple independent genes, as well as multiple individual sequences for each gene. It is the most flexible framework, for it can take into account the phenomenon of *Incomplete Lineage Sorting*, allowing multiple gene trees to have distinct topologies. It has been further extended to take into account more complex scenarios including secondary contacts after a first period of isolation, in order to study the genomic structure of pairs of extant populations (Roux et al., 2016).

Last, other species clustering methods have been proposed to look for a molecular signature of intra vs. inter species divergence, without specifying the underlying model of species divergence. The *ABGD* approach, for *Automatic Barcode Gap Discovery*, considers that intra-species genealogies follow a Kingman coalescent, and infers from the data the genetic pairwise difference threshold corresponding to the inter-species limit (Puillandre et al., 2012).

All these species clustering techniques are based on the observation of genetic data alone. As such, they have been criticized for unravelling population structure more than a biologically meaningful species status, thus leading to an inflation of species descriptions (Sukumaran and Knowles, 2017). This criticism is especially true for work based on the multispecies coalescent, which can unravel more subtle population

structure without requiring strong signal such as reciprocal monophyly. However, the Bayesian framework is also more flexible, and can be extended to integrate information from another source of data, such as continuous phenotypic traits (Solís-Lemus et al., 2014).

Using an individual-based process of diversification as tree prior

The previously described multispecies coalescent framework has been most importantly used as an individual-based prior for phylogenetic tree reconstruction (Rannala and Yang, 2003). Molecular evolution then unfolds along the resulting gene trees. Inferring the species tree under such a model thus necessitates a powerful numerical machinery to integrate over all gene tree topologies compatible with each species tree (Maddison and Knowles, 2006; Jones, 2017).

This modeling work can give us insights on how to link our chapters 3 and 5. The *SGD* model could replace the lineage-based birth-death prior in chapter 5. This would be a way to link both in the same tree dating analysis. Taking into account the individual-scale would make the method able to use the signal from within-species molecular heterogeneity. We could thus expect a better accuracy in the estimation of the underlying spiked tree, the molecular evolution parameters (α, κ, ν), and the (individual-scale) growth and mutation rates b, d, ν. Note that these two ν parameters, despite their similar name in the two models, do not refer to the same process at all for it would not make sense to require a spike of mutations to define new species.

This proposition could as well be extended to find applications in the field of phylogeography, exactly in the same way as the multispecies coalescent (Rannala and Yang, 2003; Knowles, 2004). More precisely, distinct scenarios of lineage-specific growth rate could be taken into account and inferred from the observation of present-day genetic data. However, this framework would be much less effective for this task, for it would not make use of the signal stemming from incongruences between gene trees.

6.2.3 Linking continuous trait evolution and genomic change

Last, we present some ideas on the joint study of phenotypes and molecular sequences evolution. Two main axes can be found in the literature:

i) considering them as independent data, that could provide us with a way to use both morphological and molecular data to reconstruct trees. This would allow us in particular to use contemporary observations (with morphological and molecular data) and fossil observations (with usually only morphological data) to perform phylogenetic tree reconstruction by *total evidence approach*.

ii) considering a joint correlated process, allowing us to address questions on the patterns of covariation that can be inferred from empirical data.

We examine below these two axes, and discuss how they could be transposed to our work.

Total evidence tree reconstruction

Considering both phenotypic traits and molecular sequences as independent data in the same tree reconstruction analysis should not in itself bring much difficulties. After all, tree reconstruction using either continuous trait data or molecular sequence data has been discussed since the seminal papers of Felsenstein (1973, 1981), more than 50 years ago. But if we are to really use this to reconstruct phylogenetic trees with both contemporary and fossil data, many complications still mess the picture up.

One first issue concerns the choice of a tree prior incorporating both extant and fossil taxa. Distinct propositions have been made in the literature (Pyron, 2011; Ronquist et al., 2012), the most process-based being the Fossilized Birth Death (FBD) process in Zhang et al. (2015). Similarly to a birth-death process, it assumes a rate of speciation and a rate of extinction. Moreover, extinct lineages can be sampled at another rate, which can vary through time. It mainly differs from previous propositions in that it allows fossils to be placed both at tips of the tree, and along branches leading to extant species.

To earn the *total-evidence* name, one then needs to assume a model of phenotypic evolution unfolding along the tree. Papers cited above were using discrete phenotypic data observed in fossils and extant species, but extending their studies to continuous characters would be straightforward, provided we have access to measurements that can be thought to evolve independently.

Last, a model of molecular evolution unfolding along a tree also needs to be considered, in order to make use of the observed sequence data in contemporary species. State-of-the-art relaxed molecular clocks have been considered, such as the ones discussed in chapter 5. They can take into account heterogeneities of rates along the tree, by considering that the substitution rate evolves as a continuous process along the tree. They also take into account rate heterogeneity across sites, by assigning a rate class to each site.

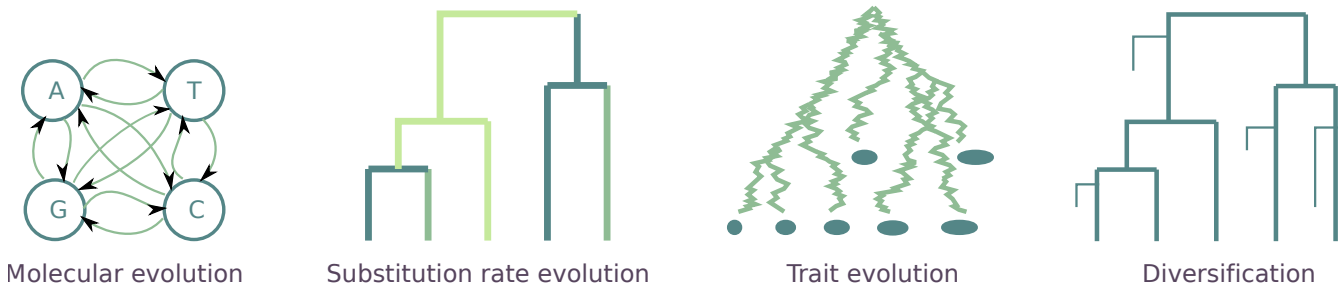


Figure 6.5 – The distinct modeling layers needed in a total-evidence approach to tree reconstruction. Letters stand for: A, alignment, R, rate of substitution, C, character, T, tree.

These total-evidence dating approaches represent the most promising, yet complex, methods to reconstruct and date phylogenies, place fossils, and estimate all other model parameters at the same time. The beauty of the model expressed in a Bayesian framework, with distinct layers superposed coherently, masks the difficulty to perform inferences. These rely on MCMC computations, which need to be carefully designed to keep a reasonable computation time.

Trait dependent parameters of molecular evolution

We already briefly sketched studies considering the joint correlated evolution of continuous traits with genomic ones, in the introduction section 1.4. Because they allow bridging the gap between molecular evolution and all other traits evolution over long timescales, we mention them again in this conclusion.

Quite similarly to what has been described for the total-evidence dating approach, Lartillot and Poujol (2011) described a model with superposed layers. First, their tree is fixed and known. Second, continuous traits and molecular evolution parameters (such as, e.g. substitution rate and ratio of synonymous to non-synonymous substitutions) evolve as a multivariate diffusion process along the tree. Third, nucleotides evolve along the tree according to the parameters of the substitution process.

The simple expression of the model masks here again the difficulty to fit it on empirical data. The authors carefully designed a MCMC inference procedure, that allowed them in particular to demonstrate a negative correlation between the rate of substitution and both body size and longevity among mammals.

More recently, Lartillot and Delsuc (2012) extended the previously described work and added a prior on tree branch lengths. The fit of the model to empirical data thus allows them to jointly date the tree, estimate the correlation between substitution parameters and continuous traits, and reconstruct these continuous traits in deep time.

Transposing these ideas to our work

The models that we studied in chapters 4 and 5 could both be integrated within a layered, more complex, model aimed at jointly reconstructing continuous traits and divergence times.

First, as already emphasized in chapter 5, the model of molecular evolution with spikes of mutations at branching times is no more than a new proposition of relaxed molecular clock, with the slight particularity that it is more tightly linked to the tree prior. As such, and provided our developments will work in the near future, it could be readily implemented as any other relaxed molecular clock in total-evidence studies. A model of continuous phenotypic evolution could even be more tightly linked to the spike model if one further assumes that shifts in some trait, or in some trait optimum, preferentially occur concurrently with spikes.

Second, the model of continuous trait evolution with interactions among lineages could also, in theory, find its place in a total-evidence approach linking it to other traits. However, we would argue that it is not really meant to this task, as it is already quite complex, and its likelihood is not fast to compute. The only situation in which it could prove useful would be when present-day interactions are well-known to have an important impact on trait values, and one can hypothesize a simple heritability rule of interactions in the past. For two interacting clades comprising a lot of specialists, the evolution of a continuous trait under the *clade-clade matching model* defined in chapter 4 could even bring useful additional information to tree reconstruction and tree dating studies. It could further be coupled with a model specifically designed to study the diversification of two interacting clades (Poisot and Stouffer, 2016).

6.3 On the many roads to modeling the long-term evolution of biodiversity

Our work constitutes one example among a vast literature on biodiversity modeling. In this final section, we end up with general considerations on this field, discussing the many types of models that can be found in the literature, and the many challenges that will stand in the near future.

6.3.1 Trade-off between biologically reasonable models and toy models

One of the main characteristics of biology, as a scientific field, is to be the science of complexity and special cases. This characteristic of biology is clearly inherited and exacerbated in evolutionary biology over long timescales, where a huge number of both observations and processes might be relevant to study.

Observations A virtually infinite number of features can be observed and analysed in contemporary species. The morphology, physiology, behaviour, molecular sequence, are as many characters that have evolved through time and that constitute clues of the past history of organisms. As stimulating as this profusion of data could be, it also brings along a real challenge for modelers. These are indeed hyper-diverse and non-independent data that require the development of new data analysis techniques.

Processes The observations that we get at present-time are the result of many processes, taking place over very different time and geographical scales. At the geological level, continents are drifting, making mountains rise and disappear, seas open and close, or still islands emerge and sink. These large-scale events are also at the root of changes in geochemical and climatic cycles affecting life on Earth. At the individual level, reproduction, dispersal and death are key events, impacted by ecological interactions such as competition, predation, parasitism or mutualism. In between these two scales, species are formed, which later hybridize with others, or go extinct.

For these reasons, the study of life history over long timescales has to rely on simplifying hypotheses, that need to be carefully chosen to make the research tractable while still able to refine our understanding of the past. Yet, this general statement leaves room for a wide spectrum of modeling work.

At the *toy model* extremity stand models relying on quite crude assumptions, but which have the ability to improve our mental representation of a process otherwise difficult to apprehend. The Wright-Fisher model is one of these: it would be really ambitious, if not impossible, to find an example of a clonally reproducing biological population of constant size. Yet, under this tractable scenario, the Kingman coalescent emerges, providing us with a clear picture of genealogical relationships.

At the more *biologically realistic* or *predictive* extremity of the spectrum, the same model can be refined to take into account a variable population size, asynchronous generations, sexual reproduction with recombination, and so on. These numerous refinements may burden our mental representation of the process. However, the fit of the model then allows us to infer realistic knowledge on the past evolution of organisms. Working on any part of this spectrum is mainly a matter of each researcher's affinity.

6.3.2 Deterministic or stochastic modeling

Virtually all the literature on models of evolution that we cited in this thesis lean on probabilistic modeling (as opposed to deterministic modeling). At least two reasons can be put forward to explain this.

The first one is highly practical. Considering a probabilistic framework indeed allows the modeler to design statistical tools directed towards performing inferences on real data. It allows taking into account both the uncertainty in the observed data, and in the assumed underlying processes.
In contrast, a deterministic model could also allow us to compare empirical patterns with the predicted ones. A fit of the model can be performed, by considering a distance measure between observed and predicted patterns that can be minimized with respect to the parameters of the model. However, the toolbox available to the modeler is much narrower, with no possibility to assess the uncertainty over the estimates, and far less theory on model comparison.

The second reason that can be invoked to explain the preference observed for probabilistic over deterministic modeling in evolutionary biology is more phenomenological. Evolution is generally thought of as being a stochastic process, and many evolutionary biologists keep Gould's metaphor in head (Gould, 1989):

> The pageant of evolution is a staggeringly improbable series of events, sensible enough in retrospect and subject to rigorous explanation, but utterly unpredictable and quite unrepeatable. Wind back the tape of life to the early days of the Burgess Shale; let it play again from an identical starting point, and the chance becomes vanishingly small that anything like human intelligence would grace the replay.

Not that it would really be an *intrinsically* stochastic process, governed by probabilistic laws. In fact, we could even argue that all microscopic processes involved in the evolution of organisms are deterministic at some point, considering an infinite amount of initial data. The stochasticity would more plausibly emerge from the chaotic nature of the system, and the impossibility to get access to a sufficient amount of initial data, at all scales. As a result, a probabilistic representation of biological evolution would prove more efficient and productive to make sense out of the data.

6.3.3 The quantity of data at hand

Overall all recent technological advances have been directed toward the acquisition of an ever increasing amount of data, across the full spectrum of biodiversity observations:

Automated environmental data acquisition Large-scale environmental surveys rely more and more on remote data acquisition using pictures, radar, or sound recordings from planes, drones, satellites.

Whole genome sequencing Molecular sequencing technologies have progressed at an outstanding pace in the last 20 years. Small-size complete genomes were already sequenced at the end of the 90' (Haemophilus influenzae, Saccharomyces cerevisiae, Caenorhabditis elegans). In 2003, the sequencing of the human genome was completed, after a 15 years global research effort. The so-called 1000

genomes project, launched in 2008, was already completed in 2012 (Consortium et al., 2012). It now references genetic variation across more than thousands of individuals. Whole genome sequencing has spread to many more species in recent years, including e.g. some primates, birds, plants...

Metagenomics We observe a trend towards large-scale ecological studies based on environmental genomics. The Tara project, aiming at sequencing the whole marine biodiversity in samples of water columns spread across the world and across time, is a typical example of such type of very ambitious surveys (Karsenti et al., 2011).

Morphometrics Recent effort has been put forward in many Natural History Museums around the world to digitize their collections. Examples include projects to reference the plumage colours and scan the beaks of all birds (Thomas et al., 2016), or to scan 20000 different vertebrate species (Cross, 2017).

And many more massive data The observations of many new microscopic features of organisms, e.g. transcriptomics (the record of all transcribed genes across distinct cellular types or across time) or epigenetics (the record of methylation and acetylation patterns across the genome) will continue to transform the field.

Keeping pace with the extraordinary inflation of available data brings to the forefront methodological challenges to handle the so-called *curse of dimensionality*. This expression refers to the fact that, as more and more precise but correlated data accumulate, the dimension of the parameter space needed to model them with standard tools increases too fast to really hope inferring them.

An example of such a phenomenon is offered by the study of body scans in a phylogenetic context. Standard multivariate methods were modeling the evolution of a set of n traits by a Gaussian diffusion process in dimension n, requiring a set of the order of n^2 parameters only to describe their correlations (Bartoszek et al., 2012). While it seems feasible to study 2,3,4 potentially correlated traits, it is much less adapted to the study of morphometric datasets including typically more landmark coordinate measurements than species. To alleviate this problem, methods of dimensionality reduction need to be applied (Tolkoff et al., 2017; Clavel et al., 2018). The same type of difficulty is found in molecular evolution studies looking for the coevolution between distinct sites along the genome (Galtier and Dutheil, 2007; Behdenna et al., 2016).

6.3.4 Where we stand

All the modeling work in macroevolution that we presented in this thesis was carried out in a probabilistic framework. As far as possible, we kept in mind the fact that the study of macroevolution is directed towards reconstructing the past history of life on Earth, and we provided statistical inference procedures in our chapters 3, 4 and 5. Yet, on a modeling spectrum going from toy models to predictive ones, we would place differently two groups of models.

Chapters 2 and 3 would find their place closer to the toy model extremity. Taken together they indeed mostly provide two thoughts: i) the speciation process is a key assumption that has a direct impact on the shape of the phylogeny; and ii) modifying the metapopulation dynamics of individual-based models has the power to bring them closer to agreement with empirical phylogenies. The inference procedure that we proposed can be seen more as a proof-of-concept, showing that it is feasible to build individual-based models of diversification which lead to a tractable likelihood computation. However as it stands, the inference is quite limited and we would not expect anyway to have the power to distinguish more complex metapopulation dynamics scenario by the sole observation of the resulting phylogeny. Moving this model to the more predictive extremity of the spectrum would necessitate to take into account various metapopulation dynamics scenarios, various speciation scenarios, and try to distinguish them by computing the likelihood using more observations. Computing the joint probability of a phylogeny and the species abundance distribution would be a good start.

On the other hand, chapters 4 and 5 would find their place closer to the predictive extremity of the spectrum. Both try to accommodate existing frameworks to take into account a specific biological

process. They are specifically designed to perform inferences, which could be of direct interest to other researchers.

All models that we considered are part of the most *process-based* body of the modeling literature, i.e. they precisely describe, with a mathematical formalism, a process already articulated in biology. Looking at the literature, the future for this type of modeling work seems to be in *total evidence approaches*, holding the promise to get the most of many types of data in the same study. The exact same trend seems to occur in the field of phylogenetic reconstruction (Zhang et al., 2015) and phylogeography (Richards et al., 2007; Papadopoulou and Knowles, 2016) with recent calls to take into account fossil data, morphological data, spatio-geographic data, to refine the signal carried by genetic data, at many loci and including many individuals per species. These integrated models will require to build specific datasets in order to address very precise questions. The most tractable models could be fitted in a Bayesian framework taking into account the whole set of observations (as described throughout this thesis). The less tractable ones might be fitted through intensive simulations under a broad range of parameter values and subsequent comparisons of well chosen simulated summary statistics to the observed ones (an approach called *Approximate Bayesian Computation*).

In the meantime, the computation time needed to perform statistical inferences under these models will hinder their application to the most recent (and large) datasets, thus leaving room for the development of novel, non-process-based, approaches. This part of model development will be typically interested in the optimization or in the adaptation of clustering techniques for new types of data.

Appendix A

Paper Appendix : The species problem from the modeler's point of view

We provide here the Appendix to the submitted paper (Manceau and Lambert, 2016) presented in chapter 2.

Contents of the chapter

Some of the results stated in Sections A.1 and A.2 are classical results in combinatorics for partially ordered sets (see Bóna, 2011, chapter 16). For the sake of self-containment and because all readers may not be familiar with these notions, we nevertheless expose them here.

A.1 'Finer than', a partial order relation on $\mathcal{X}$-partitions

Definition 18 *Let $\mathscr{S}_1$ and $\mathscr{S}_2$ be two $\mathcal{X}$-partitions. We say that $\mathscr{S}_1$ is finer than $\mathscr{S}_2$, and we write $\mathscr{S}_1 \leq \mathscr{S}_2$ if $\forall S_1 \in \mathscr{S}_1, \forall S_2 \in \mathscr{S}_2,\ S_1 \cap S_2 \in \{\varnothing, S_1\}$.*

We detail here the three criteria that make the 'finer than' relation a partial order on the set of $\mathcal{X}$-partitions.

Proof 1 *One must check the reflexivity, antisymmetry and transitivity properties.*

- *Reflexivity. Take any $\mathcal{X}$-partition $\mathscr{S}$. Then for all $S_1, S_2 \in \mathscr{S}$ we either have $S_1 \cap S_2 = S_1$ if $S_1 = S_2$, or $S_1 \cap S_2 = \varnothing$ otherwise. It follows that $\mathscr{S} \leq \mathscr{S}$.*
- *Antisymmetry. Take two $\mathcal{X}$-partitions denoted $\mathscr{S}_1$ and $\mathscr{S}_2$, verifying $\mathscr{S}_1 \leq \mathscr{S}_2$ and $\mathscr{S}_2 \leq \mathscr{S}_1$. Then for all $(S_1, S_2) \in \mathscr{S}_1 \times \mathscr{S}_2,\ S_1 \cap S_2 \in \{\varnothing, S_1\}$ and $S_1 \cap S_2 \in \{\varnothing, S_2\}$.*
  If $S_1 \cap S_2 \neq \varnothing$, it follows that $S_1 = S_2$, and finally $\mathscr{S}_1 = \mathscr{S}_2$.
- *Transitivity. Take now three $\mathcal{X}$-partitions denoted $\mathscr{S}_1, \mathscr{S}_2, \mathscr{S}_3$, verifying $\mathscr{S}_1 \leq \mathscr{S}_2$ and $\mathscr{S}_2 \leq \mathscr{S}_3$. Let $S_1 \in \mathscr{S}_1$ and $S_3 \in \mathscr{S}_3$ and assume that $S_1 \cap S_3 \neq \varnothing$. Then there is $x \in S_1 \cap S_3$ and we let S_2 be the unique element of $\mathscr{S}_2$ such that $x \in S_2$. Thus $S_1 \cap S_2 \neq \varnothing$ and $S_2 \cap S_3 \neq \varnothing$, which implies by assumption that $S_2 \cap S_1 = S_1$ and $S_2 \cap S_3 = S_2$. So we see that $S_1 \subseteq S_2 \subseteq S_3$, so that $S_1 \cap S_3 = S_1$.*

□

A.2 Proof of Theorem 1

Here we will consider sets of partitions verifying one or two desirable properties. Hence the following definitions

$$\begin{aligned}
\Sigma_A &:= \{\mathcal{X}\text{-partitions satisfying (A)}\} \\
\Sigma_B &:= \{\mathcal{X}\text{-partitions satisfying (B)}\} \\
\Sigma_M &:= \{\mathcal{X}\text{-partitions satisfying (M)}\} \\
\Sigma_{AM} &:= \{\mathcal{X}\text{-partitions satisfying (AM)}\} = \Sigma_A \cap \Sigma_M \\
\Sigma_{BM} &:= \{\mathcal{X}\text{-partitions satisfying (BM)}\} = \Sigma_B \cap \Sigma_M \\
\Sigma_{AB} &:= \{\mathcal{X}\text{-partitions satisfying (AB)}\} = \Sigma_A \cap \Sigma_B
\end{aligned}$$

We will see that the collection of $\mathcal{X}$-partitions Σ_M plays a singular role in Theorem 1. This is due to the characterization of Σ_M by the fact that there is a hierarchy $\mathscr{H}$ (here the hierarchy associated with the genealogy T) such that

$$\mathscr{S} \in \Sigma_M \Longleftrightarrow \mathscr{S} \subseteq \mathscr{H}.$$

Also recall that the collections of $\mathcal{X}$-partitions Σ_A and Σ_B can be defined as follows

$$\begin{aligned}
\mathscr{S} \in \Sigma_A &\Longleftrightarrow \forall P \in \mathscr{P},\ \forall S \in \mathscr{S},\ P \cap S \in \{\varnothing, P\} \\
\mathscr{S} \in \Sigma_B &\Longleftrightarrow \forall P \in \mathscr{P},\ \forall S \in \mathscr{S},\ P \cap S \in \{\varnothing, S\}.
\end{aligned}$$

In this section, we aim at giving a proof of Theorem 1 which can now be restated as follows

$$\exists \mathscr{S}_{\text{loose}} \in \Sigma_{AM},\ \text{such that } \forall \mathscr{S} \in \Sigma_{AM},\ \mathscr{S}_{\text{loose}} \leq \mathscr{S}$$
$$\exists \mathscr{S}_{\text{lacy}} \in \Sigma_{BM},\ \text{such that } \forall \mathscr{S} \in \Sigma_{BM},\ \mathscr{S} \leq \mathscr{S}_{\text{lacy}}$$

The proof is divided into two parts. First, given a set of partitions Σ (resp. $\Sigma \subseteq \Sigma_M$), we prove the existence of the finest (resp. coarsest) partition finer (resp. coarser) than any element of Σ, which we call $\inf \Sigma$ (resp. $\sup \Sigma$). Second, we show that $\inf \Sigma_{AM} \in \Sigma_{AM}$ and $\sup \Sigma_{BM} \in \Sigma_{BM}$, hence yielding the definitions $\mathscr{S}_{\text{loose}} := \inf \Sigma_{AM}$ and $\mathscr{S}_{\text{lacy}} := \sup \Sigma_{BM}$.

A.2.1 Defining the supremum and the infimum of a set of $\mathcal{X}$-partitions

Definition 19 *For any non-empty collection Σ of $\mathcal{X}$-partitions, we define the two relations $\underline{\mathcal{R}}_\Sigma$ and $\overline{\mathcal{R}}_\Sigma$ on $\mathcal{X}$ by*

$$\forall (x,y) \in \mathcal{X}^2,\ \ x\ \underline{\mathcal{R}}_\Sigma\ y \iff \forall \mathscr{S} \in \Sigma,\ \exists S \in \mathscr{S},\ x \in S \text{ and } y \in S$$
$$x\ \overline{\mathcal{R}}_\Sigma\ y \iff \exists \mathscr{S} \in \Sigma,\ \exists S \in \mathscr{S},\ x \in S \text{ and } y \in S.$$

Lemma 1 *For any non-empty collection Σ of $\mathcal{X}$-partitions, $\underline{\mathcal{R}}_\Sigma$ is an equivalence relation. For any non-empty collection Σ of $\mathcal{X}$-partitions such that $\Sigma \subseteq \Sigma_M$, $\overline{\mathcal{R}}_\Sigma$ is an equivalence relation.*

Proof 2 *The* reflexivity *and* symmetry *of the two relations are easily seen. Now let us prove their transitivity. Let Σ be a non-empty collection of $\mathcal{X}$-partitions, and $(x,y,z) \in \mathcal{X}^3$ such that $x\ \underline{\mathcal{R}}_\Sigma\ y$ and $y\ \underline{\mathcal{R}}_\Sigma\ z$. Let $\mathscr{S} \in \Sigma$. By definition,*

$$\exists S_1 \in \mathscr{S},\ x \in S_1 \text{ and } y \in S_1$$
$$\exists S_2 \in \mathscr{S},\ y \in S_2 \text{ and } z \in S_2$$

It follows that $y \in S_1 \cap S_2$, and because $\mathscr{S}$ is a partition, $S_1 = S_2$. Finally, with $S := S_1 = S_2$, there exists $S \in \mathscr{S}$ such that $x \in S$ and $z \in S$, so that $x\ \underline{\mathcal{R}}_\Sigma\ z$ and we can conclude that $\underline{\mathcal{R}}_\Sigma$ is transitive.

Now let $\Sigma \subseteq \Sigma_M$ be a non-empty collection of $\mathcal{X}$-partitions and $(x,y,z) \in \mathcal{X}^3$ such that $x\ \overline{\mathcal{R}}_\Sigma\ y$ and $y\ \overline{\mathcal{R}}_\Sigma\ z$. By definition,

$$\exists \mathscr{S}_1 \in \Sigma,\ \exists S_1 \in \mathscr{S}_1,\ x \in S_1 \text{ and } y \in S_1$$
$$\exists \mathscr{S}_2 \in \Sigma,\ \exists S_2 \in \mathscr{S}_2,\ y \in S_2 \text{ and } z \in S_2$$

Because $\Sigma \subseteq \Sigma_M$, $\mathscr{S}_1 \subseteq \mathscr{H}$ and $\mathscr{S}_2 \subseteq \mathscr{H}$, so that $S_1 \in \mathscr{H}$ and $S_2 \in \mathscr{H}$. From the definition of hierarchy, we get $S_1 \cap S_2 \in \{\varnothing, S_1, S_2\}$. Since $y \in S_1 \cap S_2$, we have $S_1 \cap S_2 \neq \varnothing$.
Suppose that $S_1 \cap S_2 = S_2$. It follows that $\exists \mathscr{S}_1 \in \Sigma,\ \exists S_1 \in \mathscr{S}_1,\ x \in S_1$ and $z \in S_1$.
Suppose that $S_1 \cap S_2 = S_1$. It follows that $\exists \mathscr{S}_2 \in \Sigma,\ \exists S_2 \in \mathscr{S}_2,\ x \in S_2$ and $z \in S_2$. So $x\ \overline{\mathcal{R}}_\Sigma\ z$ and we can conclude that $\overline{\mathcal{R}}_\Sigma$ is transitive. □

Definition 20 *For any non-empty collection Σ of $\mathcal{X}$-partitions, we call* $\inf \Sigma$ *the $\mathcal{X}$-partition induced by the equivalence relation $\underline{\mathcal{R}}_\Sigma$. For any non-empty collection Σ of $\mathcal{X}$-partitions such that $\Sigma \subseteq \Sigma_M$, we call* $\sup \Sigma$ *the $\mathcal{X}$-partition induced by the equivalence relation $\overline{\mathcal{R}}_\Sigma$.*

Readers familiar with lattice theory will note that these definitions match the usual 'meet' and 'join' operators used for lattices, and in particular the lattice of partitions of a set, ordered by refinement. For the other readers, recall first that any equivalence relation on a set $\mathcal{X}$ induces an $\mathcal{X}$-partition obtained by placing all elements in relation in one cluster. Further, the following lemma justifies the notation inf and sup.

Lemma 2 *Let Σ be any non-empty collection of $\mathcal{X}$-partitions. Then for any $\mathscr{S} \in \Sigma$, $\inf \Sigma \leq \mathscr{S}$. Let Σ be any non-empty collection of $\mathcal{X}$-partitions such that $\Sigma \subseteq \Sigma_M$. Then for any $\mathscr{S} \in \Sigma$, $\mathscr{S} \leq \sup \Sigma$.*

Proof 3 *Let Σ be any non-empty collection of $\mathcal{X}$-partitions and $S \in \inf \Sigma$. Let also $\mathscr{S} \in \Sigma$ and $S' \in \mathscr{S}$. We need to prove that $S \cap S' \in \{\varnothing, S\}$. Assume that $S \cap S' \neq \varnothing$ and $S \cap S' \neq S$. Then there is $x \in S \cap S'$ and $y \in S$ such that $y \notin S'$. Because $x, y \in S$, by definition of $\inf \Sigma$, we have $x \underline{\mathcal{R}}_\Sigma y$ and by definition of $\underline{\mathcal{R}}_\Sigma$, $\exists S'' \in \mathscr{S},\ x, y \in S''$. So S' and S'' are both elements of $\mathscr{S}$ containing x, which implies that $S' = S''$ and contradicts $y \notin S'$.*

Now let Σ be any non-empty collection of $\mathcal{X}$-partitions such that $\Sigma \subseteq \Sigma_M$ and $S \in \sup \Sigma$. Let also $\mathscr{S} \in \Sigma$ and $S' \in \mathscr{S}$. We need to prove that $S \cap S' \in \{\varnothing, S'\}$. Assume that $S \cap S' \neq \varnothing$ and $S \cap S' \neq S'$. Then there is $x \in S \cap S'$ and $y \in S'$ such that $y \notin S$. Because $x \in S$ and $y \notin S$, by definition of $\sup \Sigma$, x and y are not in relation $\overline{\mathcal{R}}_\Sigma$ and by definition of $\overline{\mathcal{R}}_\Sigma$, either $x \notin S'$ or $y \notin S'$ and we get a contradiction. □

Note that, in general, we can have $\inf \Sigma \notin \Sigma$ and $\sup \Sigma \notin \Sigma$. Here are two examples to provide the reader with some intuition.

Example 1. Take

$$\begin{aligned}
&\mathcal{X} = \{1, 2, 3, 4\} \\
&\mathscr{S} = \{\{1\}, \{2\}, \{3, 4\}\} \\
&\mathscr{S}' = \{\{1, 2\}, \{3\}, \{4\}\} \\
&\Sigma = \{\mathscr{S}, \mathscr{S}'\}
\end{aligned}$$

In this case, we get $\inf \Sigma = \{\{1\}, \{2\}, \{3\}, \{4\}\}$, which does not belong to Σ. Moreover, if we define the hierarchy $\mathscr{H} := \{\{1, 2, 3, 4\}, \{1, 2\}, \{3, 4\}, \{1\}, \{2\}, \{3\}, \{4\}\}$, we have $\Sigma \subseteq \Sigma_M$, which allows us to consider $\sup \Sigma = \{\{1, 2\}, \{3, 4\}\}$ which again does not belong to Σ.

Example 2. Take

$$\begin{aligned}
&\mathcal{X} = \{1, 2, 3, 4\} \\
&\mathscr{S} = \{\{1, 3, 4\}, \{2\}\} \\
&\mathscr{S}' = \{\{1, 2\}, \{3, 4\}\} \\
&\Sigma = \{\mathscr{S}, \mathscr{S}'\}
\end{aligned}$$

In this case, we get $\inf \Sigma = \{\{1\}, \{2\}, \{3, 4\}\}$, which does not belong to Σ. Moreover, there is no $\mathcal{X}$-hierarchy $\mathscr{H}$ such that $\mathscr{S}, \mathscr{S}' \in \mathscr{H}$. Then we can see that the relation $\overline{\mathcal{R}}_\Sigma$ is not an equivalence relation on $\mathcal{X}$, because $1\ \overline{\mathcal{R}}_\Sigma\ 2$ and $1\ \overline{\mathcal{R}}_\Sigma\ 3$, but we do not have $2\ \overline{\mathcal{R}}_\Sigma\ 3$. Thus, $\sup \Sigma$ is not defined.

A.2.2 Proving that $\inf \Sigma_{AM} \in \Sigma_{AM}$ and $\sup \Sigma_{BM} \in \Sigma_{BM}$

In order to prove that $\inf \Sigma_{AM} \in \Sigma_{AM}$ and $\sup \Sigma_{BM} \in \Sigma_{BM}$, we will rely on properties of $\inf \Sigma$ and $\sup \Sigma$ presented in the following lemma.

Lemma 3 *For any non-empty collection Σ of $\mathcal{X}$-partitions, for any $S \in \inf \Sigma$, S can be written in the form of the following non-empty intersection*

$$S = \bigcap_{\mathscr{S} \in \Sigma : S \subseteq S^* \in \mathscr{S}} S^* \tag{A.1}$$

For any non-empty collection Σ of $\mathcal{X}$-partitions such that $\Sigma \subseteq \Sigma_M$, for any $S \in \sup \Sigma$, S can be written in the form of the following non-empty union

$$S = \bigcup_{\mathscr{S} \in \Sigma : S \supseteq S^* \in \mathscr{S}} S^* \tag{A.2}$$

In addition,

$$\exists \mathscr{S} \in \Sigma,\ \exists S^* \in \mathscr{S},\ S^* = S. \tag{A.3}$$

Proof 4 *We begin with proving (A.1). Let Σ be any non-empty collection of $\mathcal{X}$-partitions and consider $S \in \inf \Sigma$. Now set*

$$S' := \bigcap_{\mathscr{S} \in \Sigma : S \subseteq S^* \in \mathscr{S}} S^*$$

and let us prove that $S = S'$. According to Lemma 2 that for any $\mathscr{S} \in \Sigma$, $\inf \Sigma \leq \mathscr{S}$ so $\exists! S^* \in \mathscr{S}$ such that $S \subseteq S^*$. This proves that the intersection in the definition of S' is not empty. Now by definition of S' we have $S \subseteq S'$, which also implies $S' \neq \varnothing$. We need to show now that $S' \subseteq S$. Let x be any element of S' and y be any element of S. Then for any $\mathscr{S} \in \Sigma$, there is (a unique) $S^* \in \mathscr{S}$ such that $S \subseteq S^*$ and by definition of S', we have $x \in S^*$. But since $S \subseteq S^*$ we also have $y \in S^*$. This shows that for any $\mathscr{S} \in \Sigma$ there is $S^* \in \mathscr{S}$ such that $x \in S^*$ and $y \in S^*$. This can be expressed equivalently as $x \underline{\mathcal{R}}_\Sigma y$, so that x and y are in the same element of $\inf \Sigma$, that is $x \in S$.

Now let us prove (A.2). Let Σ be any non-empty collection of $\mathcal{X}$-partitions such that $\Sigma \subseteq \Sigma_M$ and let $S \in \sup \Sigma$. Set

$$S' := \bigcup_{\mathscr{S} \in \Sigma : S \supseteq S^* \in \mathscr{S}} S^*$$

and let us prove that $S = S'$. According to Lemma 2 that for all $\mathscr{S} \in \Sigma$, $\mathscr{S} \leq \sup \Sigma$, so $\exists S^* \in \mathscr{S}$ such that $S^* \subseteq S$. In particular, the intersection in the definition of S' is not empty and $S' \neq \varnothing$. Now by definition of S' we have $S' \subseteq S$. We need to show now that $S \subseteq S'$. Let x be any element of S and y be any element of S'. Since $S' \subseteq S$, $y \in S$ so that x and y are in the same element of $\sup \Sigma$, which can be expressed equivalently as $x \overline{\mathcal{R}}_\Sigma y$. Now by definition of $\overline{\mathcal{R}}_\Sigma$, there is $\mathscr{S} \in \Sigma$ and $S^* \in \mathscr{S}$ such that $x, y \in S^*$. Now since $S^* \cap S \neq \varnothing$, we have $S^* \subseteq S$, which shows by definition of S' that $x \in S'$.

It remains to show (A.3)

$$\exists \mathscr{S} \in \Sigma,\ \exists S^* \in \mathscr{S},\ S^* = S.$$

Let us prove by induction on $n \geq 1$ that for any $F \subseteq S$ of cardinality n, there is $\mathscr{S} \in \Sigma$ and $S^* \in \mathscr{S}$ such that $F \subseteq S^* \subseteq S$. The result will follow by taking $F = S$. For $n = 1$, the property holds due to to (A.2). Let $n \geq 1$ strictly smaller than the cardinality of S and assume that the property holds for all integers smaller than or equal to n. Let F be any subset of S of cardinality $n+1$ and write $F = F_1 \cup \{x\}$, where $x \notin F_1$. Since F_1 is of cardinality n there is $\mathscr{S}_1 \in \Sigma$ and $S_1 \in \mathscr{S}_1$ such that $F_1 \subseteq S_1 \subseteq S$. Let $y \in F_1$. There is also $\mathscr{S}_2 \in \Sigma$ and $S_2 \in \mathscr{S}_2$ such that $\{x, y\} \subseteq S_2 \subseteq S$. Now because $\mathscr{S}_1, \mathscr{S}_2 \in \Sigma \subseteq \Sigma_M$, we have $\mathscr{S}_1, \mathscr{S}_2 \subseteq \mathscr{H}$, so that $S_1 \in \mathscr{H}$ and $S_2 \in \mathscr{H}$. From the definition of hierarchy, we get $S_1 \cap S_2 \in \{\varnothing, S_1, S_2\}$. Since $y \in S_1 \cap S_2$, we have $S_1 \cap S_2 \neq \varnothing$, so one of the two, denoted S^* contains the other one. In particular, $F_1 \subseteq S^* \subseteq S$ and $\{x, y\} \subseteq S^* \subseteq S$, which shows that $F = F_1 \cup \{x\} \subseteq S^* \subseteq S$ and terminates the proof. □

We can now go back to the proof of Theorem 1.

Proof 5 *(i) $\inf \Sigma_{AM} \in \Sigma_A$: Consider $S \in \inf \Sigma_{AM}$ and $P \in \mathscr{P}$. From Lemma 3, we get*

$$S \cap P = P \cap \left(\bigcap_{\mathscr{S} \in \Sigma_{AM} : S \subseteq S^* \in \mathscr{S}} S^* \right) = \bigcap_{\mathscr{S} \in \Sigma_{AM} : S \subseteq S^* \in \mathscr{S}} (P \cap S^*)$$

Now for each $S^* \in \mathscr{S} \in \Sigma_{AM}$, $P \cap S^* \in \{\varnothing, P\}$, *thus leading to* $S \cap P \in \{\varnothing, P\}$, *that is* $\inf \Sigma_{AM} \in \Sigma_A$.

(ii) $\inf \Sigma_{AM} \in \Sigma_M$*: Consider* $S \in \inf \Sigma_{AM}$. *From Lemma 3, we get*

$$S = \bigcap_{\mathscr{S} \in \Sigma_{AM} : S \subseteq S^* \in \mathscr{S}} S^*$$

Now for each $S^* \in \mathscr{S} \in \Sigma_{AM}$, $S^* \in \mathscr{H}$. *Moreover, the hierarchy* $\mathscr{H}$ *is closed under finite, non-disjoint intersections, thus leading to* $S \in \mathscr{H}$, *that is* $\inf \Sigma_{AM} \in \Sigma_M$.

(iii) $\sup \Sigma_{BM} \in \Sigma_B$*: Consider* $S \in \sup \Sigma_{BM}$ *and recall from Lemma 3 that there is* $\mathscr{S} \in \Sigma_{BM}$ *and* $S^* \in \mathscr{S}$ *such that* $S = S^*$. *Now for any* $P \in \mathscr{P}$,

$$S \cap P = S^* \cap P \in \{\varnothing, S^*\} = \{\varnothing, S\},$$

so that $\sup \Sigma_{BM} \in \Sigma_B$.

(iv) $\sup \Sigma_{BM} \in \Sigma_M$*: Consider* $S \in \sup \Sigma_{BM}$ *and* $S^* = S$ *as previously. Since* $S^* \in \mathscr{H}$, $S \in \mathscr{H}$, *so that* $\sup \Sigma_{BM} \in \Sigma_M$.

This shows that $\inf \Sigma_{AM} \in \Sigma_{AM}$ *and* $\sup \Sigma_{BM} \in \Sigma_{BM}$, *which completes the proof of Theorem 1.*□

A.3 Construction of the lacy and loose phylogenies

This section aims at formalizing mathematically the construction of the lacy and loose phylogenies presented in the main text.

Recall that an interior node is convergent if there are two tips, one in each of its two descending subtrees, carrying the same phenotype, otherwise this node is said to be divergent. We will say that the two clades subtended by a convergent (resp. divergent) node are convergent (resp. divergent). We define $\mathscr{H}_d$ as the collection of divergent clades, that is

$$\mathscr{H}_d = \{h \in \mathscr{H} : \exists h', h'' \in \mathscr{H},\ h' = h \cup h'',\ \forall P \in \mathscr{P},\ h \cap P = \varnothing \text{ or } h'' \cap P = \varnothing\} \cup \mathcal{X}$$

We similarly consider phylogenetic and non-phylogenetic clades for either the loose or the lacy definition. We call $\mathscr{H}_{\text{loose}}$ and $\mathscr{H}_{\text{lacy}}$ the collection of phylogenetic clades for the loose and lacy definitions respectively. The procedure described in the main text amounts to defining

$$\mathscr{H}_{\text{loose}} = \mathscr{H} \backslash \{h \in \mathscr{H} : \exists h_c \in \mathscr{H} \backslash \mathscr{H}_d,\ h \subseteq h_c\}$$
$$\mathscr{H}_{\text{lacy}} = \{h \in \mathscr{H} : \exists h' \in \mathscr{H},\ h \cup h' \in \mathscr{H},\ h \cap h' \in \{\emptyset, h'\},\ \exists h_d \in \mathscr{H}_d,\ h_d \subseteq h'\}$$

Appendix B

Paper Appendix : Phylogenies support out-of-equilibrium models of biodiversity

We provide here the Appendix to our paper (Manceau et al., 2015) presented in chapter 3.

Contents of the chapter

B.1 Effects of parameter values on the shape of the phylogeny

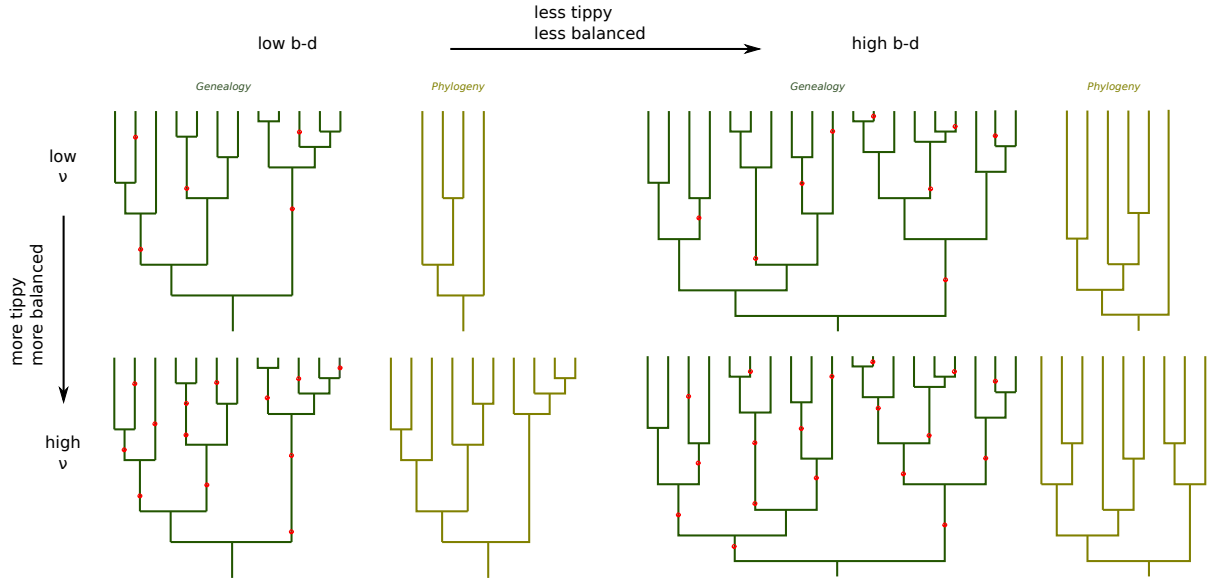


Figure B.1 – Effect of the parameter values $b - d$ and ν on the shape of phylogenies.

Decreasing the growth rate at constant mutation rate (from the right column to the left column in Figure B.1) has the same effect as increasing the mutation rate at constant growth rate (from the top to the bottom row). We detail here the later effect. By our definition of speciation, all nodes from the phylogenies are nodes from the genealogy.

Note first that deep nodes from the genealogy tend to be phylogenetic.

Increasing the mutation rate at constant growth rate increases the number of mutations on the genealogy, and thus the number of phylogenetic nodes. As more nodes from the genealogy are conserved on the phylogeny, nodes that are close to the tips are increasingly conserved and phylogenies become more tippy.

The genealogies are generated by a constant birth-death process and their expected balance is thus $\beta = 0$ by definition of β. When there are few mutations, they tend to fall on long lines in the genealogy according to the Poisson process, such that nodes from stemmy subtrees (with short lines) tend to not be phylogenetic. This creates imbalance in the resulting phylogeny. When there are more mutations, short lines are also hit by mutations, and phylogenies become more balanced.

B.2 Derivation of g(t) and m(t)

B.2.1 Survival probability of a population up to a time t

We denote $u_{b,d}(t)$ the extinction probability before time t of a population originally composed of one individual, following a birth-death process with inhomogeneous birth rate $b(t)$ and death rate $d(t)$. This probability is derived in Kendall (1948) :

$$u_{b,d}(t) = \frac{1 + \int_0^t b(s) e^{\int_0^s b(z)-d(z)dz} ds - e^{\int_0^t b(z)-d(z)dz}}{1 + \int_0^t b(s) e^{\int_0^s b(z)-d(z)dz} ds} \tag{B.1}$$

B.2.2 Branching rate $g(t)$ on the reconstructed genealogy

We consider the genealogy of individuals (Figure 3.1A) given by the linear birth-death model. We introduce the following notation to describe what happens at a birth time :

$M_t = \{$At least one descendant from an ancestral individual giving birth at time t is still alive at present.$\}$
$L_t = \{$The left descent from an ancestral individual giving birth at time t is still alive at present.$\}$
$R_t = \{$The right descent from an ancestral individual giving birth at time t is still alive at present.$\}$

A branching event on the reconstructed genealogy (Fig 3.1B) corresponds to a birth event that leads to two descents that are both alive at present. The instantaneous rate of such events at time t (conditioned on non extinction) is given by :

$$\begin{aligned} g(t)dt &= P(\text{ birth} \in dt,\ L_t,\ R_t|\ M_t) \\ &= b(t)\frac{P(L_t, R_t)}{P(M_t)}dt \\ &= b(t)\frac{(1-u_{b,d}(t))^2}{1-u_{b,d}(t)}dt \\ &= b(t)(1-u_{b,d}(t))dt \\ &= \frac{b(t)e^{\int_0^t b(z)-d(z)dz}}{1+\int_0^t b(s)e^{\int_0^s b(z)-d(z)dz}ds} \end{aligned} \tag{B.2}$$

B.2.3 Survival probability of a clonal population

We call clonal descent from an ancestral individual living at time t the whole descent from this individual in which no mutation occurred (see Figure 3.1A). We introduce the following notation :

$M_t^C = \{$The clonal descent from an ancestral individual living at time t is still alive at present$\}$
$m(t) = P(M_t^C \mid M_t)$

We get

$$m(t) = P(M_t^C \mid M_t) = \frac{P(M_t^C \cap M_t)}{P(M_t)} = \frac{P(M_t^C)}{P(M_t)}$$

Remember that the dynamics of the whole population is a birth-death process, with birth rate $b(t)$ and death rate $d(t)$, and the dynamics of the clonal population is a birth-death process, with birth rate $b(t)$ and death rate $d(t)+\nu(t)$. This gives us, $\forall t \in [0,T]$:

$$\begin{aligned} m(t) &= \frac{1-u_{b,d+\nu}(t)}{1-u_{b,d}(t)} \\ &= \frac{e^{\int_0^t b(z)-d(z)-\nu(z)dz}}{1+\int_0^t b(s)e^{\int_0^s b(z)-d(z)-\nu(z)dz}ds}\frac{1+\int_0^t b(s)e^{\int_0^s b(z)-d(z)dz}}{e^{\int_0^t b(z)-d(z)dz}} \end{aligned} \tag{B.3}$$

B.3 Forward-in-time phylogeny simulation

B.3.1 A three-type branching process

We need to define three types of lineages in order to simulate the process at the lineage-level (see Figures 3.1C and B.2 for an illustration) :

a "type 0" lineage is a line from the underlying genealogy that has at least one descendant of same genetic type at present.

a "type 1" lineage is a line from the underlying genealogy that has no descendant of same genetic type at present.

a type 0 lineage "freezes" at the first node (in forward time of the underlying phylogeny) that is not divergent. In other words, it freezes at the first node that has at least two descendants at present time, one in each of the two incident descents, having the same genetic type. In this case, the whole descent of this node is collapsed into a single species, and the lineage is "frozen", in the sense that no further splitting or extinction event can happen to this lineage up to the present.

The phylogenetic tree, considered as a time-inhomogeneous branching process with three types, is simulated forward-in-time. Figure B.2 illustrates the definition of types.

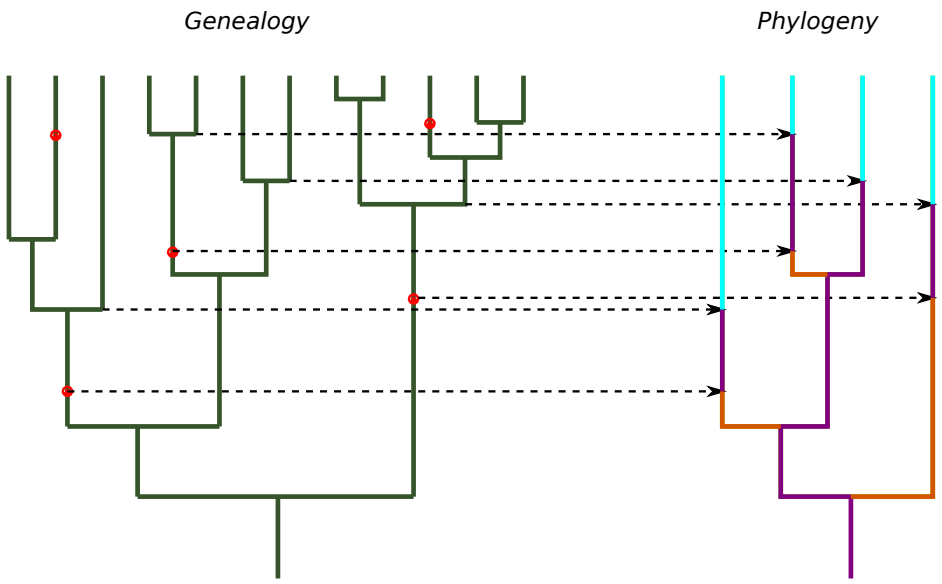


Figure B.2 – Left, genealogy is in green, and red dots are mutation events. Right, the corresponding phylogeny is in purple for type 0 lineages, orange for type 1 lineages and blue for frozen lineages.

B.3.2 Transition rates

Possible events

The following events can occur on a type 1 lineage between time t and $t + dt$:

— There is a mutation on $[t, t + dt]$ and the lineage is changed into a type 0 lineage.

— There is a mutation on $[t, t + dt]$ but the lineage remains of type 1.

— A branching occurs on $[t, t + dt]$ and gives rise to two type 1 lineages.

— Nothing happens on the genealogy between $[t, t + dt]$, nor in the phylogeny.

The following events can occur on a type 0 lineage between time t and $t + dt$:

— A branching occurs on $[t, t + dt]$ and the clonal type disappears in one population. It gives rise to one type 0 lineage, and one type 1 lineage.

— A branching occurs on $[t, t+dt]$ and the clonal type survives in both populations. The lineage is frozen.

— There is no birth on $[t, t+dt]$, the lineage remains of type 0.

We represent on Figure B.3 all events happening in our three-type process.

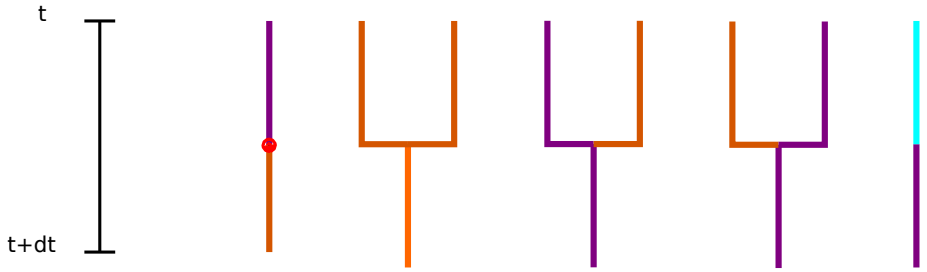


Figure B.3 – Type 0 lineages are in purple, type 1 lineages are in orange, and frozen lineages are in blue. The simulation goes "forward" from $t+dt$ to t, up to the present.

Derivation of the rates

Recall that M_t^C denotes the survival of a clonal descent from an ancestral individual living at time t up to time 0. Let $\overline{M_t^C}$ denote the extinction before present of the clonal descent from an ancestral individual at time t.

A lineage of type 1 becomes of type 0 between t and $t+dt$ when a mutation occurs in this time interval and the clonal descent from the ancestral individual carrying the mutation does not get extinct before present. This happens with rate :

$$\begin{aligned}\rho_{1\to 0}(t)dt &= P(\text{ mutation} \in dt,\ M_t^C |\ M_{t+dt},\ \overline{M_{t+dt}^C}\ ) \\ &= \frac{\nu(t)(1-u(t))m(t)}{(1-u(t))(1-m(t))}dt \\ &= \frac{\nu(t)m(t)}{(1-m(t))}dt\end{aligned}$$

A lineage of type 1 branches and gives rise to two lineages of type 1 when there is a birth event, survival of the two descents and extinction of the two clonal descents. This happens with rate :

$$\begin{aligned}\rho_{1\to +1}(t)dt &= P(\text{ birth} \in dt,\ L_t,\ R_t,\ \overline{L_t^C},\ \overline{R_t^C} |\ M_{t+dt}, \overline{M_{t+dt}^C}\ ) \\ &= b(t)\frac{(1-u(t))^2(1-m(t))^2}{(1-u(t))(1-m(t))}dt \\ &= g(t)(1-m(t))dt\end{aligned}$$

A lineage of type 0 branches and gives rise to one lineage of type 0 and one lineage of type 1 when there is a birth event, survival of the two descents, and extinction of the clonal descent in one of the two descents. This happens with rate :

$$\begin{aligned}\rho_{0\to +1}(t)dt &= P(\text{ birth} \in dt,\ L_t,\ R_t,\ (L_t^C, \overline{R_t^C})\ \cup\ (\overline{L_t^C}, R_t^C) |\ M_{t+dt}^C\ ) \\ &= b(t)\frac{2(1-u(t))^2 m(t)(1-m(t))}{(1-u(t))m(t)}dt \\ &= 2g(t)(1-m(t))dt\end{aligned}$$

A lineage of type 0 “freezes”, giving rise to a tip lineage in the phylogeny, when there is a birth event and survival of the two clonal descents :

$$\begin{aligned}\rho_{0\to\varnothing}(t)dt &= P(\text{ birth} \in dt,\ L_t^C,\ R_t^C |\ M_{t+dt}^C\ ) \\ &= b(t)\frac{(1-u(t))^2 m(t)^2}{(1-u(t))m(t)}dt \\ &= g(t)m(t)dt\end{aligned}$$

B.4 Likelihood of a tree

We aim to compute the likelihood of a tree arising from the SGD process from which each tip lineage has been sampled with probability f. Below, the "type of the tree" refers to the type of tree before the sampling procedure.

We define, for a given phylogenetic tree A, its likelihood under this model, up to time t, to be :

$$\begin{aligned}\mathcal{L}_{A,f}^i(t) = P(&\text{ a tree that started at one individual with stem age } t \text{ has shape } A \text{ and type } i \\ &| \text{ survival up to time } 0, \text{the model and the parameter set } (b,d,\nu))\end{aligned}$$

We will also need the following additional notation :

$$\begin{aligned}w_f^i(t) = P(&\text{ a tree that started at one individual with stem age } t \text{ has type } i \text{ but no species sampled} \\ &| \text{ survival up to time } t, \text{the model and the parameter set } (b,d,\nu))\end{aligned}$$

We study here how these probabilities change as t increases, and new subtrees appear. We have to take into account all events happening to the genealogy and leading to the observed phylogeny.

We slice the problem into four pieces :

— ODEs driving (w_f^0, w_f^1).

— Likelihood on a tip lineage : ODEs driving $(\mathcal{L}_{T,f}^0, \mathcal{L}_{T,f}^1)$, where T stands for "Tip".

— Likelihood on an internal lineage : ODE driving $(\mathcal{L}_{I,f}^0, \mathcal{L}_{I,f}^1)$, where I stands for "Internal".

— Likelihood at a node (branching time).

To ease notation, we will drop the dependence of all quantities upon t, including :

$$\nu = \nu(t)\ , \quad g = g(t)\ , \quad b = b(t)\ , \quad u = u(t)$$

We need to introduce some additional notation to describe different events. In the following, the type of a line (in the genealogy) or of a lineage (in the phylogeny) at a given time will be :

0 if its clonal descent has survived to the present.

1 if it has extant descent but no extant clonal descent at the present.

e if it has no extant descent at the present.

Additionally, a (phylogenetic) lineage (but not a genealogical line, because sampling concerns species and not individuals) can have two "sampling states" :

u unsampled (none of its descending tip species is sampled at present)

s sampled (at least one of its descending tip species is sampled at present)

We write the type of a line or lineage as a superscript, and the sampling state (for lineages only) as a subscript. We will also need the following event names :

$S :=$ { survival of the genealogical process up to time 0 }

$\emptyset :=$ { nothing happens in $[t, t+dt]$}

$M :=$ { a mutation event happens on $[t, t+dt]$}

$L^i_j :=$ { branching event in $[t, t+dt]$, the left tree has type i and sampling state j}

$R^i_j :=$ { branching event in $[t, t+dt]$, the right tree has type i and sampling state j}

$F_j :=$ { branching event in $[t, t+dt]$, both incident subtrees have type 0, frozen lineage has sampling state j }

Finally, we show on Figures B.4, B.5, B.6, and B.7 , the different events that could happen. Type 0 trees or lineages are in green, type 1 are in purple, extinct are in black. Dotted lines stand for trees with no species sampled, whereas solid lines stand for trees with at least one species sampled.

B.4.1 ODEs driving w^i_f

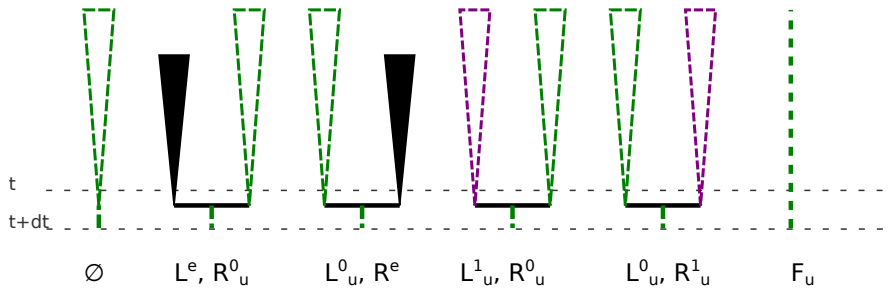


Figure B.4 – All events leading to a tree of type 0 with no species sampled at $t+dt$.

We know the initial condition for $w^0_f(0) = 1-f$, i.e., the probability of not sampling a given species. We will now derive ODEs driving w^0_f as t increases, with corresponding events shown in Figure B.4.

$$\begin{aligned} w^0_f(t+dt) &= w^0_f(t) P(\ \emptyset \mid S\ ) \\ &\quad + P(\ (R^0_u \cap L^e) \cup (R^e \cap L^0_u) \cup (R^1_u \cap L^0_u) \cup (R^0_u \cap L^1_u) \cup F_u \mid S\ ) \\ &= w^0_f(t) \left[1 - \nu dt - b\frac{1-u^2}{1-u} dt + 2bu dt + 2b(1-u) w^1_f dt\right] + b(1-u)m^2(1-f)dt \end{aligned}$$

This leads to the following differential equation driving w^0_f :

$$\begin{aligned} \frac{dw^0_f}{dt} &= w^0_f \left[-\nu - b(1+u) + 2bu + 2b(1-u)w^1_f\right] + b(1-u)m^2(1-f) \\ &= w^0_f \left[-\nu - g + 2g w^1_f\right] + g m^2 (1-f) \end{aligned} \tag{B.4}$$

Recall that the initial condition is $w^1_f(0) = 0$, because a short tip lineage has vanishing probability of carrying a mutation. We will now derive ODEs driving w^1_f as t increases, with corresponding events shown in Figure B.5 :

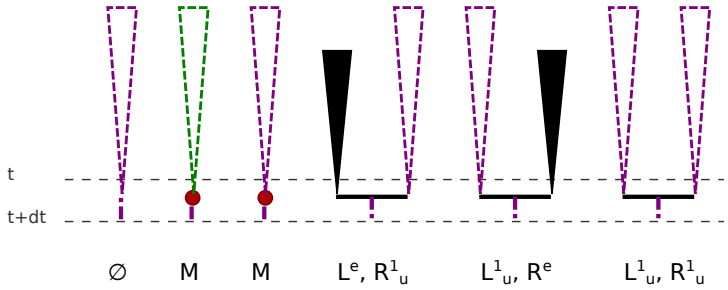


Figure B.5 – All events leading to a type 1 tree unsampled at $t + dt$.

$$\begin{aligned} w_f^1(t+dt) &= w_f^1(t)P(\ \emptyset \mid S\ ) + (w_f^1(t) + w_f^0(t))P(\ M \mid S\ ) \\ &+ P(\ (R_u^1 \cap L^e) \cup (R^e \cap L_u^1) \cup (R_u^1 \cap L_u^1) \mid S\ ) \\ &= w_f^1(t)\left[1 - \nu dt - b\frac{1-u^2}{1-u}dt + 2budt + b(1-u)w_f^1 dt + \nu dt\right] + w_f^0(t)\nu dt \end{aligned}$$

This leads to the following differential equation driving w_f^1 :

$$\begin{aligned} \frac{dw_f^1}{dt} &= w_f^1(t)\left[-b(1+u) + 2bu + b(1-u)w_f^1\right] + \nu w_f^0 \\ &= -gw_f^1(1-w_f^1) + \nu w_f^0 \end{aligned} \tag{B.5}$$

Note that for $f = 1$, the whole tree is sampled, and we verify that $\forall t \geq 0, \quad w_f^0(t) = w_f^1(t) = 0$.

B.4.2 Likelihood of a tip lineage

Likelihood of a type 0 tip lineage

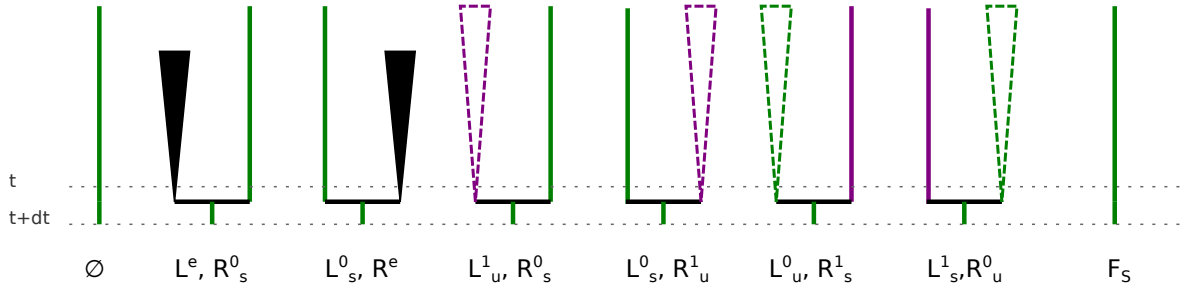


Figure B.6 – All events leading to a tip lineage of type 0 at $t + dt$.

The initial condition is $\mathcal{L}_{T,f}^0(0) = f$, i.e., the probability of sampling a given species. We will now derive ODEs driving $\mathcal{L}_{T,f}^0$ as t increases, with corresponding events shown in Figure B.6 :

$$\begin{aligned}
\mathcal{L}^0_{T,f}(t+dt) &= \mathcal{L}^0_{T,f}(t)P(\ \emptyset \mid S\ ) \\
&+ P(\ (R^0_s \cap L^e) \cup (R^e \cap L^0_s) \cup (R^0_s \cap L^1_u) \cup (R^1_u \cap L^0_s) \cup (R^0_u \cap L^1_s) \cup (R^1_s \cap L^0_u) \cup F_s \mid S\ ) \\
&= \mathcal{L}^0_{T,f}(t)\left[1 - \nu dt - b\frac{1-u^2}{1-u}dt + 2budt + 2b(1-u)w^1_f dt\right] \\
&+ \mathcal{L}^1_{T,f}(t)2b(1-u)w^0_f dt + b(1-u)m^2 f dt
\end{aligned}$$

This leads to the following differential equation driving $\mathcal{L}^0_{T,f}$:

$$\begin{aligned}
\frac{d\mathcal{L}^0_{T,f}}{dt} &= \mathcal{L}^0_{T,f}\left[-\nu - b(1+u) + 2bu + 2b(1-u)w^1_f\right] + \mathcal{L}^1_{T,f}2b(1-u)w^0_f + b(1-u)m^2 f \\
&= \mathcal{L}^0_{T,f}\left[-\nu - g + 2gw^1_f\right] + \mathcal{L}^1_{T,f}2gw^0_f + gm^2 f \qquad \text{(B.6)}
\end{aligned}$$

Solving the equation in the particular case $f = 1$, we derive the likelihood expression :

$$\mathcal{L}^0_{T,1}(t) = e^{-\int_0^t g(z)+\nu(z)dz} + \int_0^t g(s)m(s)^2 e^{-\int_s^t g(z)+\nu(z)dz} ds$$

Likelihood of a type 1 tip lineage

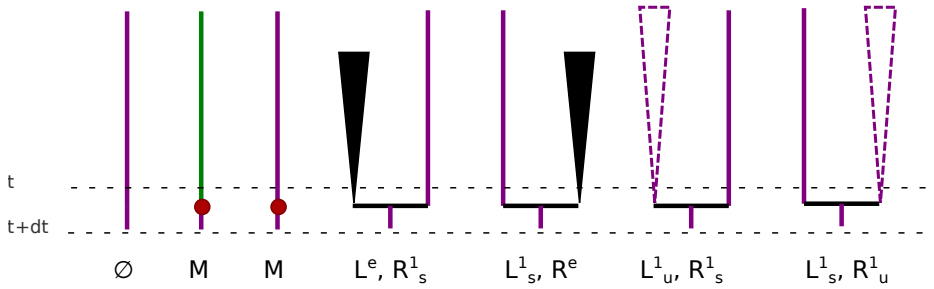


Figure B.7 – All events leading to a type 1 tip lineage at $t + dt$.

The initial condition is $\mathcal{L}^1_{T,f}(0) = 0$, because no mutation can occur exactly at time 0. We will now derive ODEs driving $\mathcal{L}^1_{T,f}$ as t increases, with corresponding events shown in Figure B.7 :

$$\begin{aligned}
\mathcal{L}^1_{T,f}(t+dt) &= \mathcal{L}^1_{T,f}(t)P(\ \emptyset \mid S\ ) + (\mathcal{L}^1_{T,f}(t) + \mathcal{L}^0_{T,f}(t))P(\ M \mid S\ ) \\
&+ P(\ (R^1_s \cap L^e) \cup (R^e \cap L^1_s) \cup (R^1_s \cap L^1_u) \cup (R^1_u \cap L^1_s) \mid S\ ) \\
&= \mathcal{L}^1_{T,f}(t)\left[1 - \nu dt - b\frac{1-u^2}{1-u}dt + 2budt + 2b(1-u)w^1_f dt + \nu dt\right] + \mathcal{L}^0_{T,f}(t)\nu dt
\end{aligned}$$

This leads to the following differential equation driving $\mathcal{L}^1_{T,f}$:

$$\begin{aligned}
\frac{d\mathcal{L}^1_{T,f}}{dt} &= \mathcal{L}^1_{T,f}(t)\left[-b(1+u) + 2bu + 2b(1-u)w^1_f\right] + \nu\mathcal{L}^0_{T,f} \\
&= g\mathcal{L}^1_{T,f}(2w^1_f - 1) + \nu\mathcal{L}^0_{T,f} \qquad \text{(B.7)}
\end{aligned}$$

Solving the equation in the particular case $f = 1$, we derive the likelihood expression :

$$\mathcal{L}^1_{T,1}(t) = e^{-\int_0^t g(z)dz}\left(1 - e^{-\int_0^t \nu(z)dz}\right) + \int_0^t g(s)m(s)^2 e^{-\int_s^t g(z)dz}\left(1 - e^{-\int_s^t \nu(z)dz}\right) ds$$

B.4.3 Likelihood on internal lineages

We call internal lineages all segments of the phylogenies between two nodes. Similarly as for tip lineages, we get :

$$\frac{d\mathcal{L}^0_{I,f}}{dt} = \mathcal{L}^0_{I,f}\left[-\nu - g + 2gw^1_f\right] + 2\mathcal{L}^1_{I,f}gw^0_f \tag{B.8}$$

And :

$$\frac{d\mathcal{L}^1_{I,f}}{dt} = g\mathcal{L}^1_{I,f}(2w^1_f - 1) + \nu\mathcal{L}^0_{I,f} \tag{B.9}$$

Solving the equation in the particular case $f = 1$, between two nodes with depths $t_1 \leq t_2$, we find the likelihood expressions :

$$\begin{aligned}
\mathcal{L}^0_{I,1}(t_2) &= \mathcal{L}^0_{I,1}(t_1)e^{-\int_{t_1}^{t_2}(g(z)+\nu(z))dz} \\
\mathcal{L}^1_{I,1}(t_2) &= \mathcal{L}^0_{I,1}(t_1)e^{-\int_{t_1}^{t_2}g(z)dz}\left(1 - e^{-\int_{t_1}^{t_2}\nu(z)dz}\right) + \mathcal{L}^1_{I,1}(t_1)e^{-\int_{t_1}^{t_2}g(z)dz}
\end{aligned}$$

B.4.4 Likelihood at a branching point

We compute here the likelihood of a tree N at a branching point between two subtrees A and B. Figure B.8 illustrates different situations leading to a phylogenetic tree N composed of two subtrees A and B.

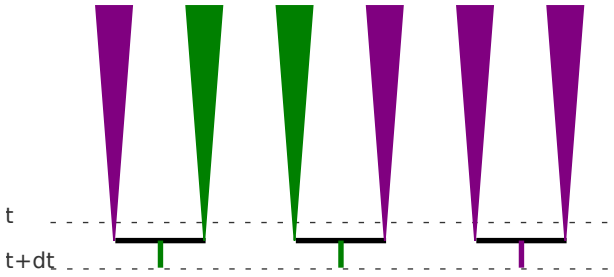


Figure B.8 – All events at a node point leading to either a type 0 or a type 1 tree.

A type 0 tree is obtained if there is a branching event at time t, and either (see Fig. B.8) :

— (A is of type 0) and (B is of type 1) or

— (A is of type 1) and (B is of type 0)

Hence it follows :

$$\mathcal{L}^0_{N,f}(t) = g\left(\mathcal{L}^0_{A,f}\mathcal{L}^1_{B,f} + \mathcal{L}^1_{A,f}\mathcal{L}^0_{B,f}\right) \tag{B.10}$$

A type 1 tree is obtained if there is a branching event at time t, (A is of type 1) and (B is of type 1).

Hence it follows :

$$\mathcal{L}^1_{N,f} = g\mathcal{L}^1_{A,f}\mathcal{L}^1_{B,f} \tag{B.11}$$

B.4.5 Peeling algorithm implementation

We compute the likelihood of a tree, recursively computing the likelihoods of subtrees from the root to the tips, using expressions (B.10-B.11) at node points and expressions (B.4-B.5) and (B.8-B.9) for internal lineages, and finally using expressions (B.4-B.5) and (B.6-B.7) for the tip lineages of the phylogeny.

Note that it is necessary to compute both likelihoods of type 0 and type 1 trees at the same time as they are coupled in the differential equations. The resulting likelihood of the tree X is $\mathcal{L}_{X,f} = \mathcal{L}^0_{X,f} + \mathcal{L}^1_{X,f}$.

The algorithm is written in Python and R, and is available from the authors upon request.

Appendix C

Paper Appendix : A unifying comparative phylogenetic framework including traits coevolving across interacting lineages

We provide here the full, uncut, Online Appendix to our paper (Manceau et al., 2017) presented in chapter 4.
This content is also available on Dryad under the following doi : *https://doi.org/10.5061/dryad.52636*.

Contents of the chapter

C.1 Derivation of the distribution in a general setting

C.1.1 The distribution of trait values is Gaussian

Recall that a vector is Gaussian if all linear combination of its components follows a normal distribution. We will thus show by induction that all linear combinations of the traits follow a normal distribution.

The process of trait evolution starts either at the stem root with a vector of size d defined by the initial conditions $X_{\tau_0} = {}^{tr}(X_0^1, ...X_0^d)$, or at the crown root with a vector of size $2d$ defined by the initial conditions: $X_{\tau_0} = {}^{tr}(X_0^1, ...X_0^d, X_0^1..., X_0^d)$, or at any other step, provided the initial conditions are Gaussian by assumption.

Now, assume that X_{τ_i} is a Gaussian vector.

Then, $\forall t \in (\tau_i, \tau_{i+1})$, after integration we have the following closed expression for the value of the process X_t.

$$X_t = e^{-tA_i}\left(e^{\tau_i A_i}X_{\tau_i} + \int_{\tau_i}^{t} e^{sA_i}a_i(s)ds + \int_{\tau_i}^{t} e^{sA_i}\Gamma_i(s)dW_s\right) \tag{C.1}$$

Moreover, we have, for any deterministic function Φ (Gardiner et al., 1985),

$$\int_{t_n}^{t} \Phi_s dW_s \sim \mathcal{N}\left(0, \int_{t_n}^{t} \Phi_s {}^{tr}\Phi_s ds\right)$$

Hence, X_t is a linear combination of Gaussian vectors, which makes it a Gaussian vector.

Last, suppose that at time τ_{i+1}, the jth branch splits, in which case the vector grows. All linear combinations of the components of X_t at time τ_{i+1}^- have a normal distribution. And the d additional components added at time τ_{i+1} belong to the components at time τ_{i+1}^-. It follows that all linear combinations of the new vector still have a normal distribution.

C.1.2 Integrating the evolution of the distribution through each epoch

Still assuming that we know the (Gaussian) distribution of X_{τ_i} at the beginning of an epoch (τ_i, τ_{i+1}), a few more lines allow us to provide a closed formula for the distribution of X_t at all time $t \in (\tau_i, \tau_{i+1})$. Indeed, using Equation (C.1), and the fact that, if X and Y are two independent Gaussian vectors with expectation vectors respectively m_X and m_Y and covariance matrices respectively Σ_X and Σ_Y , then:

$$\begin{aligned} DX + d &\sim \mathcal{N}\left(Dm_X + d\ ,\ D\Sigma_X {}^{tr}D\right) \\ X + Y &\sim \mathcal{N}\left(m_X + m_Y\ ,\ \Sigma_X + \Sigma_Y\right) \end{aligned}$$

It thus follows that, $\forall t \in [\tau_i, \tau_{i+1}]$,

$$m_t = e^{(\tau_i - t)A_i}m_{\tau_i} + \int_{\tau_i}^{t} e^{(s-t)A_i}a_i(s)ds \tag{4.4a}$$

$$\Sigma_t = \left(e^{(\tau_i - t)A_i}\right)\Sigma_{\tau_i}{}^{tr}\left(e^{(\tau_i - t)A_i}\right) + \int_{\tau_i}^{t}\left(e^{(s-t)A_i}\Gamma_i(s)\right){}^{tr}\left(e^{(s-t)A_i}\Gamma_i(s)\right)ds \tag{4.4b}$$

Applying these equations for $t = \tau_{i+1}$ thus gives the distribution of the trait vector at time τ_{i+1} , which is the result stated in Equations (4.4a, 4.4b) in the main text.

Remark that, unless one of the very first branches immediately dies at the beginning of the process at a fixed initial condition, the density of the tip distribution has support in $\mathbb{R}^{nd}$. One can check that Σ_t stays positive definite (implying that $\det \Sigma_t \neq 0$), even when some Γ_i are not positive definite (except the first one).

C.1.3 Evolution of the distribution through ODE resolution

The expectation and covariance formulae provided in Equations (4.4a, 4.4b) require to deal with an integral which is not always straightforward to compute. Alternatively, one can prefer to take the derivative of this expression, get a set of ODEs verified by the expectation and covariance elements through each epoch, and subsequently integrate the ODE system. We show now another way to derive this set of ODEs.

First, we write the stochastic differential equation on any epoch (τ_i, τ_{i+1}) and for each trait k, which is given in the most general setting by:

$$dX_t^{(k)} = \left(a_i^{(k)}(t) - \sum_{m=1}^{n_t d} A_i^{(k,m)} X_t^{(m)} \right) dt + \sum_{m=1}^{n_t d} \Gamma_i^{(k,m)}(t) dW_t^{(m)}$$

Itô's formula (Gardiner et al., 1985) then gives us:

$$\begin{aligned} d\left(X_t^{(k)} X_t^{(l)}\right) = & \; X_t^{(k)} dX_t^{(l)} + X_t^{(l)} dX_t^{(k)} + d < X_t^{(k)}, X_t^{(l)} > \\ = & \left(a_i^{(l)}(t) X_t^{(k)} - \sum_{m=1}^{n_t d} A_i^{(l,m)} X_t^{(m)} X_t^{(k)} \right) dt + \sum_{m=1}^{n_t d} \Gamma_i^{(l,m)}(t) X_t^{(k)} dW_t^{(m)} \\ + & \left(a_i^{(k)}(t) X_t^{(l)} - \sum_{m=1}^{n_t d} A_i^{(k,m)} X_t^{(m)} X_t^{(l)} \right) dt + \sum_{m=1}^{n_t d} \Gamma_i^{(k,m)}(t) X_t^{(l)} dW_t^{(m)} \\ + & \sum_{m=1}^{n_t d} \Gamma_i^{(l,m)}(t) \Gamma_i^{(k,m)}(t) dt \end{aligned}$$

Taking the expectation, it follows that

$$\begin{aligned} \frac{d}{dt}\mathbb{E}\left(X_t^{(k)} X_t^{(l)}\right) = & \; a^{(l)}(t)\mathbb{E}\left(X_t^{(k)}\right) + a_i^{(k)}(t)\mathbb{E}\left(X_t^{(l)}\right) \\ - & \sum_{m=1}^{n_t d} A_i^{(l,m)} \mathbb{E}\left(X_t^{(m)} X_t^{(k)}\right) - \sum_{m=1}^{n_t d} A_i^{(k,m)} \mathbb{E}\left(X_t^{(m)} X_t^{(l)}\right) \\ + & \sum_{m=1}^{n_t d} \Gamma_i^{(l,m)}(t) \Gamma_i^{(k,m)}(t) \end{aligned}$$

In the same fashion, we get

$$\frac{d}{dt}\mathbb{E}(X_t^{(k)}) = a_i^{(k)}(t) - \sum_{m=1}^{n_t d} A_i^{(k,m)} \mathbb{E}\left(X_t^{(m)}\right) \tag{4.5a}$$

This leads to

$$\begin{aligned} \frac{d}{dt}\left(\mathbb{E}(X_t^{(k)})\mathbb{E}(X_t^{(l)})\right) = & \; \mathbb{E}(X_t^{(l)}) \frac{d}{dt}\mathbb{E}(X_t^{(k)}) + \mathbb{E}(X_t^{(k)}) \frac{d}{dt}\mathbb{E}(X_t^{(l)}) \\ = & \; a_i^{(k)}(t)\mathbb{E}\left(X_t^{(l)}\right) - \sum_{m=1}^{n_t d} A_i^{(k,m)} \mathbb{E}\left(X_t^{(m)}\right) \mathbb{E}\left(X_t^{(l)}\right) \\ + & \; a_i^{(l)}(t)\mathbb{E}\left(X_t^{(k)}\right) - \sum_{m=1}^{n_t d} A_i^{(l,m)} \mathbb{E}\left(X_t^{(m)}\right) \mathbb{E}\left(X_t^{(k)}\right) \end{aligned}$$

Putting together these different parts gives us the ODE satisfied by all covariances:

$$\frac{d}{dt}\mathrm{Cov}\left(X_t^{(k)}, X_t^{(l)}\right) = \frac{d}{dt}\left(\mathbb{E}\left(X_t^{(k)}X_t^{(l)}\right) - \mathbb{E}(X_t^{(k)})\mathbb{E}(X_t^{(l)})\right)$$
$$= -\sum_{m=1}^{n_t d}\left[A_i^{(k,m)}\mathrm{Cov}\left(X_t^{(m)}, X_t^{(l)}\right) + A_i^{(l,m)}\mathrm{Cov}\left(X_t^{(m)}, X_t^{(k)}\right) - \Gamma_i^{(l,m)}(t)\Gamma_i^{(k,m)}(t)\right] \tag{4.5b}$$

Note that in a vectorial formalism with the expectation vector m and covariance matrix Σ, these sets of ODEs can be written equivalently as follows

$$\frac{dm_t}{dt} = a_i(t) - A_i m_t \tag{C.2}$$
$$\frac{d\Sigma_t}{dt} = -A_i\Sigma_t - {}^{tr}\Sigma_t{}^{tr}A_i + \Gamma_i{}^{tr}\Gamma_i \tag{C.3}$$

C.2 Distribution for some models without interactions between lineages

C.2.1 Distribution of classic univariate models

We present in this section how previously known results of analytic tip distribution of univariate models fit in, and can be rediscovered with, our framework. Results are summarized in Table C.1.

The scheme is identical for each model:

i) Reduce Equations (4.4a, 4.4b) or (4.5a, 4.5b) according to the model.

ii) Look for an analytical solution at any time τ_i, by calculating manually the expectations and covariances at $\tau_1, \tau_2, \tau_3, ...$

iii) Prove by induction that the analytical solution holds at any time τ_i.

We call $t_{k,l}$ the time of the most recent common ancestor to lineages k and l, and $t_{k,k}$ the death time of lineage k, equal to T if it survives until present (see Fig. C.1). We further note $\mathbb{1}_{k \text{ alive}}(t)$ the quantity that equals one if lineage k is alive at time t and zero otherwise, and $\mathbb{1}_{k=l}$ that equals one if $k = l$ and zero otherwise. Last, $t_1 \wedge t_2$ stands for the minimum of the two values t_1 and t_2.

The unity vector (vector full of 1) is denoted by V, I refers to the identity matrix (diagonal matrix with diagonal values equal to 1), and U refers to the unity matrix (matrix full of 1). Their size is the same as the size of the vector of traits X_t considered. Considering non-ultrametric trees including fossils amounts to replacing vector V and matrices I and U by their homologs V_{alive}, I_{alive} and U_{alive}, where the subscript specifies that the vector and matrices have 0 on lines and columns corresponding to lineages that are extinct in the given epoch.

Code	m_0	Σ_0	$(m_T)^{(k)}$	$(\Sigma_T)^{(k,l)}$
BM	m_0	v_0	$m_0 + bt_{k,k}$	$v_0 + \sigma^2 t_{k,l}$
OU	θ	0	θ	$\frac{\sigma^2}{2\psi} e^{-\psi(t_{k,k}+t_{l,l}-2t_{k,l})} \left(1 - e^{-2\psi t_{k,l}}\right)$
OU	θ	$\frac{\sigma^2}{2\psi}$	θ	$\frac{\sigma^2}{2\psi} e^{-\psi(t_{k,k}+t_{l,l}-2t_{k,l})}$
ACDC	m_0	v_0	m_0	$v_0 + \frac{\sigma_0^2}{2r}(e^{2rt_{k,l}} - 1)$
DD	m_0	v_0	m_0	$v_0 + \sigma_0^2 \sum_{j=0}^{N-1} e^{2rn_{\tau_j}} (\tau_{j+1} - \tau_j) \mathbb{1}_{t_{k,l} > \tau_j}$

Table C.1 – Analytic tip distribution for models without interactions between traits or lineages. We recall that $t_{k,l}$ is the absolute time of the most recent common ancestor to lineages k and l, and $t_{k,k}$ is the death time of lineage k, equal to T if it survives until present.

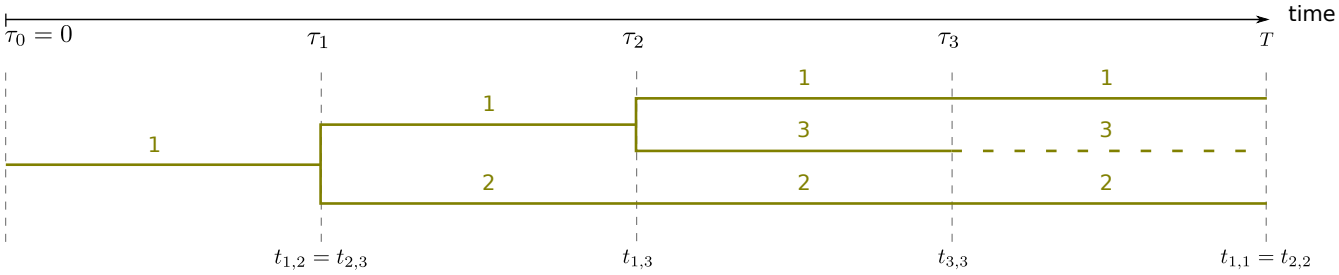


Figure C.1 – Formalism used in analytic formulae presented in Table C.1.

Brownian Motion (BM)

We show how to get the well-known expression of the distribution of a trait evolving under BM, on non-necessarily ultrametric trees. We take $a = bV_{\text{alive}}$, $A = 0$ and $\Gamma = \sigma I_{\text{alive}}$, i.e. the process follows the equation:

$$dX_t = bV_{\text{alive}}dt + \sigma I_{\text{alive}}dW_t$$

Equations (4.4a) and (4.4b) lead to the following recurrence formulae driving the law of X_t through each epoch $[\tau_i, \tau_{i+1})$:

$$\begin{aligned}\mathbb{E}(X_t) &= \mathbb{E}(X_{\tau_i}) + b(t-\tau_i)V_{\text{alive}}\\ \text{Var}(X_t) &= \text{Var}(X_{\tau_i}) + \sigma^2(t-\tau_i)I_{\text{alive}}\end{aligned}$$

Alternatively, Equations (4.5a) and (4.5b) lead to the following recurrence formulae driving the law of X_t through each epoch $[\tau_i, \tau_{i+1})$:

$$\begin{aligned}\frac{d}{dt}\mathbb{E}(X_t^{(k)}) &= b\mathbb{1}_{\text{k alive}}(t)\\ \frac{d}{dt}\text{Cov}\left(X_t^{(k)}, X_t^{(l)}\right) &= \sigma^2\mathbb{1}_{k=l}\mathbb{1}_{\text{k alive}}(t)\end{aligned}$$

We can show by induction on i that for any i the expectation and covariance matrix at time τ_i are such that, for any (k, l):

$$\mathbb{E}(X_{\tau_i}^{(k)}) = \mathbb{E}(X_0) + b(t_{k,k} \wedge \tau_i) \tag{C.4}$$

$$\text{Cov}\left(X_{\tau_i}^{(k)}, X_{\tau_i}^{(l)}\right) = \text{Var}(X_0) + \sigma^2(t_{k,l} \wedge \tau_i) \tag{C.5}$$

Indeed, we verify Equations (C.4, C.5) at step $i = 1$.

Now, suppose Equations (C.4, C.5) hold at step n. Using either Equations (4.4a, 4.4b) or (4.5a, 4.5b), we get:

$$\begin{aligned}\mathbb{E}(X_{\tau_{n+1}^-}^{(k)}) &= \mathbb{E}(X_0) + b(t_{k,k} \wedge \tau_{n+1})\\ \text{Cov}\left(X_{\tau_{n+1}^-}^{(k)}, X_{\tau_{n+1}^-}^{(l)}\right) &= \text{Var}(X_0) + \sigma^2(t_{k,l} \wedge \tau_{n+1})\end{aligned}$$

If τ_{n+1} is a death time of a lineage, Equations (C.4, C.5) are verified at step $n+1$.

If τ_{n+1} is a branching time, we verify that the new lineage inherits the expectation and covariances of its mother, as well as the same coalescence times with other lineages. It also follows that Equations (C.4, C.5) are verified at step $n+1$.

Finally, by induction, we get the tip distribution:

$$\begin{aligned}\mathbb{E}(X_T^{(k)}) &= \mathbb{E}(X_0) + bt_{k,k}\\ \text{Cov}\left(X_T^{(k)}, X_T^{(l)}\right) &= \text{Var}(X_0) + \sigma^2 t_{k,l}\end{aligned}$$

Ornstein-Uhlenbeck (OU)

We can get another well-known distribution for a trait evolving under an Ornstein-Uhlenbeck process on a tree. We take $a = \psi\theta V_{\text{alive}}$, $A = \psi I_{\text{alive}}$ and $\Gamma = \sigma I_{\text{alive}}$, i.e. the process follows the equation:

$$dX_t = (\psi\theta V_{\text{alive}} - \psi I_{\text{alive}} X_t)dt + \sigma I_{\text{alive}} dW_t$$

Expressions (4.4a) and (4.4b) simplify into the following recurrence formulae:

$$\mathbb{E}(X_t) = e^{-\psi(t-\tau_i)I_{\text{alive}}}\left(\mathbb{E}(X_{\tau_i}) - \theta V_{\text{alive}}\right) + \theta V_{\text{alive}}$$
$$\text{Var}(X_t) = e^{-2\psi(t-\tau_i)I_{\text{alive}}}\left(\text{Var}(X_{\tau_i}) - \frac{\sigma^2}{2\psi}I_{\text{alive}}\right) + \frac{\sigma^2}{2\psi}I_{\text{alive}}$$

Alternatively, here again, one can prefer to apply Equations (4.5a) and (4.5b):

$$\frac{d}{dt}\mathbb{E}(X_t^{(k)}) = \psi \mathbb{1}_{k\text{alive}}(t)\left(\theta - \mathbb{E}\left(X_t^{(k)}\right)\right)$$
$$\frac{d}{dt}\text{Cov}\left(X_t^{(k)}, X_t^{(l)}\right) = -\psi(\mathbb{1}_{k\text{ alive}}(t) + \mathbb{1}_{l\text{ alive}}(t))\text{Cov}\left(X_t^{(k)}, X_t^{(l)}\right) + \sigma^2 \mathbb{1}_{k=l}$$

We can show by induction that for any epoch i, the expectation and covariance matrix at time τ_i are such that, for all (k, l):

$$\mathbb{E}(X_{\tau_i}^{(k)}) = \theta + e^{-\psi(t_{k,k}\wedge\tau_i)}\left(\mathbb{E}(X_0) - \theta\right) \tag{C.6}$$
$$\text{Cov}\left(X_{\tau_i}^{(k)}, X_{\tau_i}^{(l)}\right) = e^{-\psi(t_{k,k}\wedge\tau_i + t_{l,l}\wedge\tau_i - 2(t_{k,l}\wedge\tau_i))}\left[\frac{\sigma^2}{2\psi} + e^{-2\psi(t_{k,l}\wedge\tau_i)}\left(\text{Var}(X_0) - \frac{\sigma^2}{2\psi}\right)\right] \tag{C.7}$$

Indeed, we verify Equations (C.6, C.7) at step $i = 0$.

Now, suppose Equations (C.6, C.7) hold at step n. Using either Equations (4.4a, 4.4b) or (4.5a, 4.5b), we get:

$$\mathbb{E}(X_{\tau_{n+1}^-}^{(k)}) = \theta + e^{-\psi(t_{k,k}\wedge\tau_{n+1})}\left(\mathbb{E}(X_0) - \theta\right)$$
$$\text{Cov}\left(X_{\tau_{n+1}^-}^{(k)}, X_{\tau_{n+1}^-}^{(l)}\right) = e^{-\psi(t_{k,k}\wedge\tau_{n+1} + t_{l,l}\wedge\tau_{n+1} - 2(t_{k,l}\wedge\tau_{n+1}))}\left[\frac{\sigma^2}{2\psi} + e^{-2\psi(t_{k,l}\wedge\tau_{n+1})}\left(\text{Var}(X_0) - \frac{\sigma^2}{2\psi}\right)\right]$$

If τ_{n+1} is a death time of a lineage, Equations (C.6, C.7) are verified at step $n + 1$.

If τ_{n+1} is a branching time, we verify that the new lineage inherits the expectation and covariances of its mother, as well as the same coalescence times with other lineages. It also follows that Equations (C.6, C.7) are verified at step $n + 1$.

Finally, by induction, we get the tip distribution:

$$\mathbb{E}(X_T^{(k)}) = \theta + e^{-\psi t_{k,k}}\left(\mathbb{E}(X_0) - \theta\right)$$
$$\text{Cov}\left(X_T^{(k)}, X_T^{(l)}\right) = e^{-\psi(t_{k,k} + t_{l,l} - 2t_{k,l})}\left[\frac{\sigma^2}{2\psi} + e^{-2\psi t_{k,l}}\left(\text{Var}(X_0) - \frac{\sigma^2}{2\psi}\right)\right]$$

Two classes of initial distributions are typically considered in the literature:

i) If we consider a process starting at $X_0 = \theta$ (i.e. with $\mathbb{E}(X_0) = \theta$ and $\text{Var}(X_0) = 0$), we get the following expectation vector m_T and covariance matrix Σ_T at the tips:

$$m_T = {}^{tr}(\theta, \theta, ..., \theta) \quad \text{and} \quad \Sigma_T = \frac{\sigma^2}{2\psi}\Upsilon_1$$
$$\text{where } \Upsilon_1 = \left[e^{-\psi(t_{k,k} + t_{l,l} - 2t_{k,l})}\left(1 - e^{-2\psi t_{k,l}}\right)\right]_{1\le k,l\le K}$$

ii) When $\psi > 0$, if we consider a process starting under its stationary distribution (i.e. $\mathbb{E}(X_0) = \theta$ and $\mathrm{Var}(X_0) = \frac{\sigma^2}{2\psi}$), it simplifies into the following expectation vector and covariance matrix:

$$m_T = {}^{tr}(\theta, \theta, ..., \theta) \quad \text{and} \quad \Sigma_T = \frac{\sigma^2}{2\psi}\Upsilon_2$$
$$\text{where } \Upsilon_2 = \left[e^{-\psi(t_{k,k}+t_{l,l}-2t_{k,l})}\right]_{1\leq k,l\leq K}$$

ACDC (accelerating or decelerating rate)

In the ACDC process, the rate of phenotypic evolution varies exponentially through time, with $a = 0$, $A = 0$ and $\Gamma = \sigma_0 e^{rt} I_{\text{alive}}$ (here, $r > 0$). The process follows the equation:

$$dX_t = \sigma_0 e^{rt} I_{\text{alive}} dW_t$$

Here again, we can simplify Equations (4.4a, 4.4b) or (4.5a, 4.5b). With Equations (4.4a, 4.4b), we get the following recurrence formulae driving the law of X_t through each epoch (τ_i, τ_{i+1}):

$$\mathbb{E}(X_t) = \mathbb{E}(X_{\tau_i})$$
$$\mathrm{Var}(X_t) = \mathrm{Var}(X_{\tau_i}) + \frac{\sigma_0^2}{2r}\left(e^{2rt} - e^{2r\tau_i}\right) I_{\text{alive}} dt$$

We can show by induction that for any i, the expectation and covariance matrix at time τ_i are such that, for any (k, l):

$$\mathbb{E}(X^{(k)}_{\tau_i}) = \mathbb{E}(X_0) \tag{C.8}$$
$$\mathrm{Cov}\left(X^{(k)}_{\tau_i}, X^{(l)}_{\tau_i}\right) = \mathrm{Var}(X_0) + \frac{\sigma_0^2}{2r}\left(e^{2r(t_{k,l}\wedge\tau_i)} - 1\right) \tag{C.9}$$

Indeed, we verify Equations (C.8, C.9) at step $i = 0$.

Now, suppose Equations (C.8, C.9) hold at step n. Using either Equations (4.4a, 4.4b) or (4.5a, 4.5b), we get:

$$\mathbb{E}(X^{(k)}_{\tau_{n+1}^-}) = \mathbb{E}(X_0)$$
$$\mathrm{Cov}\left(X^{(k)}_{\tau_{n+1}^-}, X^{(l)}_{\tau_{n+1}^-}\right) = \mathrm{Var}(X_0) + \frac{\sigma_0^2}{2r}\left(e^{2r(t_{k,l}\wedge\tau_{n+1})} - 1\right)$$

If τ_{n+1} is a death time of a lineage, Equations (C.8, C.9) are verified at step $n + 1$.

If τ_{n+1} is a branching time, we verify that the new lineage inherits the expectation and covariances of its mother, as well as the same coalescence times with other lineages. It also follows that Equations (C.8, C.9) are verified at step $n + 1$.

Finally, by induction, we get the tip distribution:

$$\mathbb{E}(X^{(k)}_T) = \mathbb{E}(X_0)$$
$$\mathrm{Cov}\left(X^{(k)}_T, X^{(l)}_T\right) = \mathrm{Var}(X_0) + \frac{\sigma_0^2}{2r}\left(e^{2rt_{k,k}} - 1\right)$$

ACDC and OU processes lead to the same present-time distributions on ultrametric trees

This has been shown previously in Uyeda et al. 2015. More precisely, OU is equivalent to a model with accelerating rates at present, and only on ultrametric phylogenies.

Looking at expressions of expectations and covariance matrices under ACDC and OU with initial conditions $X_0 = \theta$, we see that we can choose parameters such that we get the exact same distribution. First take $\mathbb{E}(X_0) = \theta$: the two expectation vectors are identical. Moreover, we can choose parameters such that the covariance matrices are equal:

$$\begin{aligned} & \frac{\sigma^2}{2\psi} e^{-2\psi(T-t_{k,l})}\left(1-e^{-2\psi t_{k,l}}\right) = \frac{\sigma_0^2}{2r}\left(e^{2rt_{k,l}}-1\right) \\ \iff & \frac{\sigma^2}{2\psi} e^{-2\psi T}\left(e^{2\psi t_{k,l}}-1\right) = \frac{\sigma_0^2}{2r}\left(e^{2rt_{k,l}}-1\right) \\ \iff & r=\psi \quad \text{and} \quad \sigma_0^2 = \sigma^2 e^{-2\psi T} \end{aligned}$$

Note that this no longer holds on non-ultrametric trees, neither with different initial conditions on the OU.

Diversity-Dependent (DD)

In the DD process, the rate of phenotypic evolution is fixed at the base of the tree and varies exponentially with the number of lineages in the reconstructed phylogeny, with $a = 0$, $A = 0$ and $B(t) = \sigma_0 e^{rn_t} I_{\text{alive}}$. The process follows the equation:

$$dX_t = \sigma_0 e^{rn_t} I_{\text{alive}} dW_t$$

Equations (4.4a, 4.4b) lead to the following recurrence formulae driving the law of X_t through each epoch (τ_i, τ_{i+1}):

$$\begin{aligned} \mathbb{E}(X_t) &= \mathbb{E}(X_{\tau_i}) \\ \text{Var}(X_t) &= \text{Var}(X_{\tau_i}) + \sigma_0^2 e^{2rn_{\tau_i}}(t-\tau_i) I_{\text{alive}} \end{aligned}$$

Note that, alternatively, one can again prefer to apply Equations (4.5a, 4.5b).

We can then show by induction that for any i, the expectation and covariance matrix at time τ_i are such that, for any (k, l):

$$\mathbb{E}(X_{\tau_i}^{(k)}) = \mathbb{E}(X_0) \tag{C.10}$$

$$\text{Cov}\left(X_{\tau_i}^{(k)}, X_{\tau_i}^{(l)}\right) = \text{Var}(X_0) + \sigma_0^2 \sum_{j=0}^{i-1} e^{2rn_{\tau_j}}(\tau_{j+1}-\tau_j)\mathbb{1}_{t_{k,l}>\tau_j} \tag{C.11}$$

Indeed, we verify Equations (C.10, C.11) at step $i = 0$.

Now, suppose Equations (C.10, C.11) hold at step n. Using either Equations (4.4a, 4.4b) or (4.5a, 4.5b), we get:

$$\begin{aligned} \mathbb{E}(X_{\tau_{n+1}^-}^{(k)}) &= \mathbb{E}(X_0) \\ \text{Cov}\left(X_{\tau_{n+1}^-}^{(k)}, X_{\tau_{n+1}^-}^{(l)}\right) &= \text{Var}(X_0) + \sigma_0^2 \sum_{j=0}^{n} e^{2rn_{\tau_j}}(\tau_{j+1}-\tau_j)\mathbb{1}_{t_{k,l}>\tau_j} \end{aligned}$$

If τ_{n+1} is a death time of a lineage, Equations (C.10, C.11) are verified at step $n+1$.

If τ_{n+1} is a branching time, we verify that the new lineage inherits the expectation and covariances of its mother, as well as the same coalescence times with other lineages. It also follows that Equations (C.10, C.11) are verified at step $n+1$.

Finally, by induction, we get the tip distribution at present time $\tau_N = T$:

$$\mathbb{E}(X_T^{(k)}) = \mathbb{E}(X_0)$$
$$\mathrm{Cov}\left(X_T^{(k)}, X_T^{(l)}\right) = \mathrm{Var}(X_0) + \sigma_0^2 \sum_{j=0}^{N-1} e^{2rn_{\tau_j}} (\tau_{j+1} - \tau_j) \mathbb{1}_{t_{k,l} > \tau_j}$$

C.2.2 Distribution of classic multivariate models

The same methodology applies to classic multivariate models that incorporate interactions between traits within lineages but not between lineages. In our formalism, for all i, A_i and Γ_i are block diagonal, with $d \times d$ blocks on the diagonal corresponding to the traits within each lineage. We call these blocks respectively A^* and Γ^*. Moreover, the vector a_i is the repetition of identical sequences a^* of d elements.

Writing the matrix products in Equations (4.4a, 4.4b) provides us with $d \times d$ blocks that behave identically during each epoch. Indeed, we can use:

$$m_{\tau_i}^{*(k)} = \begin{pmatrix} \mathbb{E}(X_{\tau_i}^{(k,1)}) \\ \mathbb{E}(X_{\tau_i}^{(k,2)}) \\ \vdots \\ \mathbb{E}(X_{\tau_i}^{(k,d)}) \end{pmatrix} \text{ and } \Sigma_{\tau_i}^{*(k,l)} = \begin{pmatrix} \mathrm{Cov}\left(X_{\tau_i}^{(k,1)}, X_{\tau_i}^{(l,1)}\right) & \mathrm{Cov}\left(X_{\tau_i}^{(k,1)}, X_{\tau_i}^{(l,2)}\right) & \dots & \mathrm{Cov}\left(X_{\tau_i}^{(k,1)}, X_{\tau_i}^{(l,d)}\right) \\ \mathrm{Cov}\left(X_{\tau_i}^{(k,2)}, X_{\tau_i}^{(l,1)}\right) & \mathrm{Cov}\left(X_{\tau_i}^{(k,2)}, X_{\tau_i}^{(l,2)}\right) & \dots & \mathrm{Cov}\left(X_{\tau_i}^{(k,2)}, X_{\tau_i}^{(l,d)}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}\left(X_{\tau_i}^{(k,d)}, X_{\tau_i}^{(l,1)}\right) & \mathrm{Cov}\left(X_{\tau_i}^{(k,d)}, X_{\tau_i}^{(l,2)}\right) & \dots & \mathrm{Cov}\left(X_{\tau_i}^{(k,d)}, X_{\tau_i}^{(l,d)}\right) \end{pmatrix}$$

In which case Equations (4.4a, 4.4b) lead to the recurrence formulae:

$$\begin{aligned} m_{\tau_{i+1}}^{*(k)} &= e^{(\tau_i - \tau_{i+1})\mathbb{1}_{k \text{ alive}}(\tau_{i+1})A^*} m_{\tau_i}^{*(k)} + \mathbb{1}_{k \text{ alive}}(\tau_{i+1}) \int_{\tau_i}^{\tau_{i+1}} e^{(s-\tau_{i+1})A^*} a^*(s) ds \\ \Sigma_{\tau_{i+1}}^{*(k,l)} &= e^{(\tau_i - \tau_{i+1})\mathbb{1}_{k \text{ alive}}(\tau_{i+1})A^*} \Sigma_{\tau_i}^{*(k,l)tr} \left(e^{(\tau_i - \tau_{i+1})\mathbb{1}_{l \text{ alive}}(\tau_{i+1})A^*}\right) \\ &\quad + \mathbb{1}_{k=l} \int_{\tau_i}^{\tau_{i+1}} \left(e^{(s-\tau_{i+1})A^*}\Gamma^*\right)^{tr} \left(e^{(s-\tau_{i+1})A^*}\Gamma^*\right) ds \end{aligned}$$

We can then prove by induction that for any epoch i and any pair of lineages (k,l)

$$m_{\tau_i}^{*(k)} = e^{-\tau_i \wedge t_{k,k} A^*} m_0^* + \int_0^{\tau_i \wedge t_{k,k}} e^{(s - \tau_i \wedge t_{k,k})A^*} a^*(s) ds \tag{C.12}$$
$$\Sigma_{\tau_i}^{*(k,l)} = e^{-\tau_i \wedge t_{k,k} A^*} \Sigma_0^{*tr} \left(e^{-\tau_i \wedge t_{l,l} A^*}\right) + \int_0^{t_{k,l} \wedge \tau_i} \left(e^{-\tau_i \wedge t_{k,k} A^*}\Gamma^*\right)^{tr} \left(e^{-\tau_i \wedge t_{l,l} A^*}\Gamma^*\right) ds \tag{C.13}$$

Indeed, we verify Equations (C.12, C.13) at step $i = 0$.

Now, suppose Equations (C.12, C.13) hold at step i. Using Equations (4.4a, 4.4b), we get:

$$
\begin{aligned}
m^{*(k)}_{\tau_{i+1}^-} &= e^{(\tau_i-\tau_{i+1})\mathbb{1}_{k \text{ alive}}(\tau_i)A^*} m^{*(k)}_{\tau_i} + \mathbb{1}_{k \text{ alive}}(\tau_i) \int_{\tau_i}^{\tau_{i+1}} e^{(s-\tau_{i+1})A^*} a^*(s) ds \\
&= e^{(\tau_i-\tau_{i+1})\mathbb{1}_{k \text{ alive}}(\tau_i)A^*} e^{-\tau_i \wedge t_{k,k} A^*} m^*_0 + \int_0^{\tau_i \wedge t_{k,k}} e^{(\tau_i-\tau_{i+1})\mathbb{1}_{k \text{ alive}}(\tau_i)A^*} e^{(s-\tau_i\wedge t_{k,k})A^*} a^*(s) ds \\
&\quad + \mathbb{1}_{k \text{ alive}}(\tau_i) \int_{\tau_i}^{\tau_{i+1}} e^{(s-\tau_{i+1})A^*} a^*(s) ds \\
&= e^{-\tau_{i+1}\wedge t_{k,k} A^*} m^*_0 + \int_0^{\tau_{i+1}\wedge t_{k,k}} e^{(s-\tau_{i+1}\wedge t_{k,k})A^*} a^*(s) ds
\end{aligned}
$$

as well as:

$$
\begin{aligned}
\Sigma^{*(k,l)}_{\tau_{i+1}^-} &= e^{(\tau_i-\tau_{i+1})\mathbb{1}_{k \text{ alive}}(\tau_{i+1})A^*} \Sigma^{*(k,l)}_{\tau_i} tr\left(e^{(\tau_i-\tau_{i+1})\mathbb{1}_{l \text{ alive}}(\tau_{i+1})A^*}\right) \\
&\quad + \mathbb{1}_{k=l} \int_{\tau_i}^{\tau_{i+1}} \left(e^{(s-\tau_{i+1})A^*}\Gamma^*\right) tr\left(e^{(s-\tau_{i+1})A^*}\Gamma^*\right) ds \\
&= e^{(\tau_i-\tau_{i+1})\mathbb{1}_{k \text{ alive}}(\tau_{i+1})A^*} e^{-\tau_i\wedge t_{k,k}A^*} \Sigma^{*tr}_0 \left(e^{-\tau_i\wedge t_{l,l}A^*}\right) tr\left(e^{(\tau_i-\tau_{i+1})\mathbb{1}_{l \text{ alive}}(\tau_{i+1})A^*}\right) \\
&\quad + \int_0^{t_{k,l}\wedge\tau_i} e^{(\tau_i-\tau_{i+1})\mathbb{1}_{k \text{ alive}}(\tau_{i+1})A^*} \left(e^{-\tau_i\wedge t_{k,k}A^*}\Gamma^*\right) tr\left(e^{-\tau_i\wedge t_{l,l}A^*}\Gamma^*\right) tr\left(e^{(\tau_i-\tau_{i+1})\mathbb{1}_{l \text{ alive}}(\tau_{i+1})A^*}\right) ds \\
&\quad + \mathbb{1}_{k=l} \int_{\tau_i}^{\tau_{i+1}} \left(e^{(s-\tau_{i+1})A^*}\Gamma^*\right) tr\left(e^{(s-\tau_{i+1})A^*}\Gamma^*\right) ds \\
&= e^{-\tau_{i+1}\wedge t_{k,k}A^*} \Sigma^{*tr}_0 \left(e^{-\tau_{i+1}\wedge t_{l,l}A^*}\right) + \int_0^{t_{k,l}\wedge\tau_{i+1}} \left(e^{-\tau_{i+1}\wedge t_{k,k}A^*}\Gamma^*\right) tr\left(e^{-\tau_{i+1}\wedge t_{l,l}A^*}\Gamma^*\right) ds
\end{aligned}
$$

If τ_{i+1} is a death time of a lineage, Equations (C.12, C.13) are verified at step $i+1$.

If τ_{i+1} is a branching time, we verify that the new lineage inherits the expectation and covariances of its mother, as well as the same coalescence times with other lineages. It also follows that Equations (C.12, C.13) are verified at step $i+1$.

Finally, by induction, we get the tip distribution:

$$
\begin{aligned}
m^{*(k)}_T &= e^{-t_{k,k}A^*} m^*_0 + \int_0^{t_{k,k}} e^{(s-t_{k,k})A^*} a^*(s) ds \\
\Sigma^{*(k,l)}_T &= e^{-t_{k,k}A^*} \Sigma^{*tr}_0 \left(e^{-t_{l,l}A^*}\right) + \int_0^{t_{k,l}} \left(e^{-t_{k,k}A^*}\Gamma^*\right) tr\left(e^{-t_{l,l}A^*}\Gamma^*\right) ds
\end{aligned}
$$

OU-BM model

As a first illustration, consider a model with $d = 3$ traits with equation during each epoch and on each lineage k as follows:

$$
\begin{aligned}
dX^{(k,1)}_t &= \psi\left(b_1 + b_2 X^{(k,2)}_t + b_3 X^{(k,3)}_t - X^{(k,1)}_t\right) dt + \sigma_1 dW^{(k,1)}_t \\
dX^{(k,2)}_t &= \sigma_2 dW^{(k,2)}_t \\
dX^{(k,3)}_t &= \sigma_3 dW^{(k,3)}_t
\end{aligned}
$$

These equations describe the evolution of two independent traits evolving following a BM (traits 2 and 3), and one trait following an OU with optimal trait value given by a linear combination of traits 2 and 3. Its main interest is to infer the dependence of one trait to two other independent traits on a

phylogeny. Knowing the distribution at the beginning of a given epoch, we use Equations (4.4a, 4.4b) to compute the distribution at the end of the epoch.

A is block-diagonal with the following blocks A^*:

$$A^* = \begin{pmatrix} 1 & -b_2 & -b_3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Writing $\Delta = s - \tau_{i+1}$, it follows that $e^{\Delta A_i}$ is block diagonal with 3×3 elements given by:

$$e^{\Delta A^*} = \begin{pmatrix} e^{\Delta} & -b_2\left(e^{\Delta}-1\right) & -b_3\left(e^{\Delta}-1\right) \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Moreover, Γ_i is block-diagonal with diagonal blocks:

$$\Gamma^* = \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{pmatrix}$$

The matrix product $(e^{\Delta A_i}\Gamma_i)^{tr}(e^{\Delta A_i}\Gamma_i)$ is thus block-diagonal with 3×3 blocks:

$$\begin{pmatrix} (\sigma_1^2 + b_2^2\sigma_2^2 + b_3^2\sigma_3^2)e^{2\Delta} - 2(b_2^2\sigma_2^2 + b_3^2\sigma_3^2)e^{\Delta} + (b_2^2\sigma_2^2 + b_3^2\sigma_3^2) & -b_2\sigma_2^2(e^{\Delta}-1) & -b_3\sigma_3^2(e^{\Delta}-1) \\ -b_2\sigma_2^2(e^{\Delta}-1) & \sigma_2^2 & 0 \\ -b_3\sigma_3^2(e^{\Delta}-1) & 0 & \sigma_3^2 \end{pmatrix}$$

These matrices can be used to compute $m_T^{*(k)}$ and $\Sigma_T^{*(k,l)}$, with the help of Equations (C.12, C.13).

OU-OU model

Consider now a model with $d = 2$ traits with equation during each epoch and on each lineage k given by:

$$\begin{aligned} dX_t^{(k,1)} &= \psi\left(b_1 + b_2 X_t^{(k,2)} - X_t^{(k,1)}\right)dt + \sigma_1 dW_t^{(k,1)} \\ dX_t^{(k,2)} &= \psi\left(b_3 - X_t^{(k,2)}\right)dt + \sigma_2 dW_t^{(k,2)} \end{aligned}$$

These equations describe the evolution of one trait evolving following an OU (trait 2), and one trait following an OU with optimal trait value given by an affine transformation of trait 2. Its main interest is to infer the dependence of one trait to another trait on a phylogeny. Knowing the distribution at the beginning of a given epoch, we use Equations (4.4a, 4.4b) to compute the distribution at the end of the epoch.

A_i is block diagonal, with the following 2×2 blocks A^*:

$$A^* = \begin{pmatrix} 1 & -b_2 \\ 0 & 1 \end{pmatrix}$$

Again, writing $\Delta = s - \tau_{i+1}$, it follows that $e^{\Delta A_i}$ is block diagonal with 2×2 elements given by:

$$e^{\Delta A^*} = \begin{pmatrix} e^{\Delta} & -b_2 \Delta e^{\Delta} \\ 0 & e^{\Delta} \end{pmatrix}$$

Moreover, Γ_i is diagonal with repeated values:

$$\Gamma^* = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}$$

The matrix product $(e^{\Delta A_i}\Gamma_i)^{tr}(e^{\Delta A_i}\Gamma_i)$ is thus block-diagonal with 2*2 blocks:

$$\begin{pmatrix} \sigma_1^2 e^{2\Delta} + b_2^2 \Delta^2 \sigma^2 e^{2\Delta} & -b_2 \sigma_2^2 \Delta e^{2\Delta} \\ -b_2 \sigma_2^2 \Delta e^{2\Delta} & \sigma_2^2 e^{2\Delta} \end{pmatrix}$$

These matrices can be used to compute $m_T^{*(k)}$ and $\Sigma_T^{*(k,l)}$, with the help of Equations (C.12, C.13).

C.3 Distribution for some models with interactions between lineages

C.3.1 Distribution with a constant, A symmetric, and $\Gamma = \sigma I$

When $\Gamma = \sigma I$ and A is symmetric, Equations (4.4a, 4.4b) become:

$$\mathbb{E}(X_t) = e^{(\tau_i - t)A_i}\mathbb{E}(X_{\tau_i}) + \int_{\tau_i}^{t} e^{(s-t)A_i} a_i(s) ds$$
$$\mathrm{Var}(X_t) = \left(e^{(\tau_i - t)A_i}\right) \mathrm{Var}(X_{\tau_i})^{tr} \left(e^{(\tau_i - t)A_i}\right) + \sigma^2 \int_{\tau_i}^{t} e^{2(s-t)A_i} ds$$

If A_i is symmetric with coefficients in $\mathbb{R}$, it can be diagonalized by orthogonal passage matrices: we can exhibit a matrix Q verifying ${}^{tr}QA_iQ = \Lambda_i$ is diagonal and $Q^{-1} = {}^{tr}Q$.

$$\mathbb{E}(X_t) = Qe^{(\tau_i - t)\Lambda_i}{}^{tr}Q\mathbb{E}(X_{\tau_i}) + Q\left(\int_{\tau_i}^{t} e^{(s-t)\Lambda_i} ds\right){}^{tr}Qa_i$$
$$\mathrm{Var}(X_t) = Qe^{\Lambda_i(\tau_i - t)}{}^{tr}Q\mathrm{Var}(X_{\tau_i})Qe^{(\tau_i - t)\Lambda_i}{}^{tr}Q + \sigma^2 Q\left(\int_{\tau_i}^{t} e^{2(s-t)\Lambda_i} ds\right){}^{tr}Q$$

This is the expression that we need for the numerical integration, in particular, of the phenotype matching model.

Note that with A diagonalizable but not symmetric, Equations (4.4a, 4.4b) can also be reduced, but the transposition of A is no longer A, and it does not lead exactly to the same expression.

C.3.2 The phenotype matching (PM) model

We consider here the phenotype matching model introduced in Nuismer and Harmon (2014), with the following equation describing the evolution of any trait k through each epoch:

$$dX_t^{(k)} = \psi\left(\theta - X_t^{(k)}\right)dt + S\left(\left(\frac{1}{n_t}\sum_{l=1}^{n_t} X_t^{(l)}\right) - X_t^{(k)}\right)dt + \sigma dW_t^{(k)}$$

We introduce the line vector u, with value u_j that equals 1 if lineage j is alive, and 0 otherwise. In order to use our framework, we further want to express the model in the form given by Equation (4.2). This is achieved by taking:

$$a_i = \psi\theta^{tr}u$$
$$A_i = (\psi + S)\mathrm{diag}(u) - \frac{S}{u^{tr}u}{}^{tr}uu$$
$$\Gamma_i = \sigma\mathrm{diag}(u)$$

where $\mathrm{diag}(u)$ is the diagonal matrix with diagonal elements the elements of the vector u.

First, the tip distribution can be computed using the general algorithm that numerically resolves the set of ODEs given in Equations (4.5a, 4.5b). Second, the PM model falls within the class of models studied in the previous section, that is, with a symmetric A matrix. The tip distribution can thus be numerically computed faster using this reduction.

We describe here a third (and faster) way to derive the tip distribution. It is based on an analytical reduction of Equations (4.4a, 4.4b) that is specific to the PM model.

Remark that $\mathrm{diag}(u)$ and ${}^{tr}uu$ commute, leading to the following calculus,

$$
\begin{aligned}
e^{(\tau_i-\tau_{i+1})A_i} &= e^{(\tau_i-\tau_{i+1})((\psi+S)\mathrm{diag}(u)-\frac{S}{u^{tr}u}{}^{tr}uu)} \\
&= e^{(\tau_i-\tau_{i+1})(\psi+S)\mathrm{diag}(u)} e^{-(\tau_i-\tau_{i+1})\frac{S}{u^{tr}u}{}^{tr}uu} \\
&= \mathrm{diag}\left(e^{(\tau_i-\tau_{i+1})(\psi+S)u}\right)\left(\sum_{k\geq 0}\frac{\left(\frac{-(\tau_i-\tau_{i+1})S}{u^{tr}u}\right)^k ({}^{tr}uu)^k}{k!}\right)
\end{aligned}
$$

Where e^w is the line vector with elements e^{w_j}. Further, remark that for any $k \geq 1$,

$$
\begin{aligned}
({}^{tr}uu)^k &= ({}^{tr}uu)({}^{tr}uu)({}^{tr}uu)...({}^{tr}uu) \\
&= {}^{tr}u(u^{tr}u)(u^{tr}u)...(u^{tr}u)u \\
&= (u^{tr}u)^{k-1}({}^{tr}uu)
\end{aligned}
$$

For simplicity, we will write in the following $\Delta = \tau_i - \tau_{i+1}$, leading us to

$$
\begin{aligned}
e^{\Delta A_i} &= \mathrm{diag}\left(e^{(\psi+S)\Delta u}\right)\left(I + \sum_{k\geq 1}\frac{\left(\frac{-S\Delta}{u^{tr}u}\right)^k (u^{tr}u)^{k-1}({}^{tr}uu)}{k!}\right) \\
&= \mathrm{diag}\left(e^{(\psi+S)\Delta u}\right)\left(I + \frac{1}{u^{tr}u}\left(\sum_{k\geq 1}\frac{(-(\tau_i-\tau_{i+1})S)^k}{k!}\right){}^{tr}uu\right) \\
&= \mathrm{diag}\left(e^{(\psi+S)\Delta u}\right)\left(I + \frac{1}{u^{tr}u}\left(e^{-S\Delta}-1\right){}^{tr}uu\right) \\
&= \mathrm{diag}\left(e^{(\psi+S)\Delta u}\right) + \frac{1}{u^{tr}u}\mathrm{diag}\left(e^{-S\Delta}e^{(\psi+S)\Delta u}\right){}^{tr}uu - \frac{1}{u^{tr}u}\mathrm{diag}\left(e^{(\psi+S)\Delta u}\right){}^{tr}uu \\
&= \mathrm{diag}\left(e^{(\psi+S)\Delta u}\right) + \frac{1}{u^{tr}u}(e^{\psi\Delta}-e^{(\psi+S)\Delta})\,{}^{tr}uu
\end{aligned} \tag{C.14}
$$

Where the last equality is due to the product by ${}^{tr}u$, allowing to forget the cases where $u_j = 0$ in the exponential.

We further need to compute

$$
\begin{aligned}
\int_{\tau_i}^{\tau_{i+1}} e^{(s-\tau_{i+1})A_i}a_i ds &= \psi\theta\int_{\tau_i}^{\tau_{i+1}} e^{\psi(s-\tau_{i+1})}ds\ {}^{tr}u \\
&= \theta\left(1-e^{\psi\Delta}\right){}^{tr}u
\end{aligned} \tag{C.15}
$$

We thus get $m_{\tau_{i+1}^-}$ with the help of Equations (C.14) and (C.15).

Now, in order to simplify Equation (4.4b), remark that A_i and Γ_i are symmetric, and so are $e^{\Delta A_i}$ and $e^{\Delta A_i}\Gamma_i$. Moreover, Γ_i is diagonal, and commutes with any other matrix, leading to,

$$
\Sigma_{\tau_{i+1}^-} = e^{\Delta A_i}\Sigma_{\tau_i}e^{\Delta A_i} + \int_{\tau_i}^{\tau_{i+1}} e^{2(s-\tau_{i+1})A_i}\Gamma_i\Gamma_i ds
$$

The first term can be computed thanks to Equation (C.14). For the second one, remark that ${}^{tr}uu\ \mathrm{diag}(u) = {}^{tr}uu$, thus leading to

$$\int_{\tau_i}^{\tau_{i+1}} e^{2(s-\tau_{i+1})A_i}\Gamma_i\Gamma_i ds = \sigma^2 \int_{\tau_i}^{\tau_{i+1}} e^{2(\psi+S)(s-\tau_{i+1})} ds \ \mathrm{diag}(u)$$
$$+ \frac{\sigma^2}{u^{tr}u} \int_{\tau_i}^{\tau_{i+1}} \left(e^{2\psi(s-\tau_{i+1})} - e^{2(\psi+S)(s-\tau_{i+1})} \right) ds \ {}^{tr}uu \ \mathrm{diag}(u)$$
$$= \sigma^2 \frac{(1-e^{2(\psi+S)\Delta})}{2(\psi+S)} \ \mathrm{diag}(u) + \frac{\sigma^2}{u^{tr}u} \left(\frac{1-e^{2\psi\Delta}}{2\psi} - \frac{1-e^{2(\psi+S)\Delta}}{2(\psi+S)} \right) \ {}^{tr}uu \quad \text{(C.16)}$$

We thus get $\Sigma_{\tau_{i+1}^-}$ with the help of Equations (C.14) and (C.16).

C.3.3 The phenotype matching (PM) model with biogeography

In this section we describe ways to compute the tip distribution under the PM model, taking into account the biogeography (that is, species interact only when they co-occur in the same localities). We consider a fixed number of islands N_I. Matrix U gives us the presence/absence of lineages in the distinct islands, with element u_{ij} that equals 1 if lineage j is present on island i and zero otherwise. Vector S gives the strength of interaction on each island. The model states that the trait of lineage j evolves through phenotype matching with all species that are sympatric:

$$dX_t^{(j)} = \psi \left(\theta - X_t^{(j)} \right) dt + \sum_{i=1}^{N_I} S_i u_{ij} \left(\frac{\sum_{l=1}^{n} u_{il} X_t^{(l)}}{\sum_{l=1}^{n} u_{il}} - X_t^{(j)} \right) dt + \sigma dW_t^{(j)}$$

Take for example 5 lineages evolving on 3 distinct islands with the following U matrix on a given epoch:

$$U = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

This means that species number 1 is present on island 2 and 3, species number 2 is only present on island 1, and so on... Said differently, we see that species number 3 interacts on island 1 with species 2, and on island 2 with species 1 and 4. Our species traits are driven by the following equations:

$$dX_t^{(1)} = \left(\psi \left(\theta - X_t^{(1)} \right) + S_2 \left(\frac{X_t^{(1)} + X_t^{(3)} + X_t^{(4)}}{3} - X_t^1 \right) + S_3 \left(\frac{X_t^{(1)} + X_t^{(5)}}{2} - X_t^1 \right) \right) dt + \sigma dW_t^{(1)}$$
$$dX_t^{(2)} = \left(\psi \left(\theta - X_t^{(2)} \right) + S_1 \left(\frac{X_t^{(2)} + X_t^{(3)}}{2} - X_t^2 \right) \right) dt + \sigma dW_t^{(2)}$$
$$dX_t^{(3)} = \left(\psi \left(\theta - X_t^{(3)} \right) + S_1 \left(\frac{X_t^{(2)} + X_t^{(3)}}{2} - X_t^3 \right) + S_2 \left(\frac{X_t^{(1)} + X_t^3 + X_t^{(4)}}{3} - X_t^3 \right) \right) dt + \sigma dW_t^{(3)}$$
$$dX_t^{(4)} = \left(\psi \left(\theta - X_t^{(4)} \right) + S_2 \left(\frac{X_t^{(1)} + X_t^3 + X_t^{(4)}}{3} - X_t^4 \right) \right) dt + \sigma dW_t^{(4)}$$
$$dX_t^{(5)} = \left(\psi \left(\theta - X_t^{(5)} \right) + S_3 \left(\frac{X_t^{(1)} + X_t^{(5)}}{2} - X_t^5 \right) \right) dt + \sigma dW_t^{(5)}$$

It thus follows that the vectorial equation can be written:

$$dX_t = \left(\begin{pmatrix} \psi\theta \\ \psi\theta \\ \psi\theta \\ \psi\theta \\ \psi\theta \end{pmatrix} - \begin{pmatrix} \psi + \frac{2}{3}S_2 + \frac{1}{2}S_3 & 0 & -\frac{S_2}{3} & -\frac{S_2}{3} & -\frac{S_3}{2} \\ 0 & \psi + \frac{1}{2}S_1 & -\frac{S_1}{2} & 0 & 0 \\ -\frac{S_2}{3} & -\frac{S_1}{2} & \psi + \frac{1}{2}S_1 + \frac{2}{3}S_2 & -\frac{S_2}{3} & 0 \\ -\frac{S_2}{3} & 0 & -\frac{S_2}{3} & \psi + \frac{2}{3}S_2 & 0 \\ -\frac{S_1}{2} & 0 & 0 & 0 & \psi + \frac{1}{2}S_1 \end{pmatrix} X_t \right) dt + \sigma dW_t$$

Provided no island is empty, the model can be written in our framework with $a = \psi\theta V$, $\Gamma = \sigma I$, and, finally, A which is the matrix with elements:

$$(A)_{jj} = \psi + \sum_{i=1}^{N_I} S_i u_{ij} (1 - \frac{1}{\sum_{l=1}^{n} u_{il}})$$
$$(A)_{jk} = -\sum_{i=1}^{N_I} S_i u_{ij} u_{ik} \frac{1}{\sum_{l=1}^{n} u_{il}}$$

Matrix A is symmetric, and we can thus use the developments presented in Appendix C.3.1 to speed up the computation time.

Nonetheless, a better analytical reduction can be derived when islands are exclusive, meaning that species are allowed to occur on one island only. Under this assumption, matrix $U^T U$ is diagonal with element $(U^T U)_{ii}$ being the number of lineages belonging to island i. We now introduce the line vector r, of size N_I, full of ones. For simplicity, we also write in the following $\Delta = \tau_i - \tau_{i+1}$. With these notations, and provided no island is empty, the model can be written under our framework with:

$$\begin{aligned} a_i &= \psi\theta\ {}^T(rU) \\ A_i &= \ \mathrm{diag}((\psi r + S)U) - {}^T U \mathrm{diag}(S)(U^T U)^{-1} U \\ \Gamma_i &= \sigma \mathrm{diag}(rU) \end{aligned}$$

As for the one island case, we can speed up the computation of the exponential by remarking that:

$$\begin{aligned} e^{\Delta A_i} &= e^{\Delta\ \mathrm{diag}((\psi r + S)U)} e^{-\Delta\ {}^T U \mathrm{diag}(S)(U^T U)^{-1} U} \\ &= e^{\Delta\ \mathrm{diag}((\psi r + S)U)} \sum_{k \geq 0} \frac{(-\Delta^T U \mathrm{diag}(S)(U^T U)^{-1} U)^k}{k!} \end{aligned}$$

We then observe that:

$$\begin{aligned} &(-\Delta^T U \mathrm{diag}(S)(U^T U)^{-1} U)^k \\ =&(-\Delta^T U \mathrm{diag}(S)(U^T U)^{-1} U)(-\Delta^T U \mathrm{diag}(S)(U^T U)^{-1} U)...(-\Delta^T U \mathrm{diag}(S)(U^T U)^{-1} U) \\ =&{}^T U(-\Delta \mathrm{diag}(S))(U^T U)^{-1}(U^T U)(-\Delta \mathrm{diag}(S))(U^T U)^{-1}(U^T U)...(U^T U)(-\Delta \mathrm{diag}(S))(U^T U)^{-1} U \\ =&{}^T U(-\Delta \mathrm{diag}(S))^k (U^T U)^{-1} U \end{aligned}$$

Thus leading to the following expression:

$$
\begin{aligned}
e^{\Delta A_i} &= e^{\Delta \text{diag}((\psi r+S)U)} \left(I + \sum_{k\geq 1} \frac{(-\Delta^T U \text{diag}(S)(U^TU)^{-1}U)^k}{k!} \right) \\
&= \text{diag}(e^{\Delta(\psi r+S)U}) \left(I + {}^TU \left(\sum_{k\geq 1} \frac{(-\Delta \text{diag}(S))^k}{k!} \right) (U^TU)^{-1}U \right) \\
&= \text{diag}(e^{\Delta(\psi r+S)U}) \left(I + {}^TU \left(\text{diag}(e^{-\Delta S}) - I \right) (U^TU)^{-1}U \right) \\
&= \text{diag}(e^{\Delta(\psi r+S)U}) \left(I - {}^TU(U^TU)^{-1}U \right) + \text{diag}(e^{\Delta(\psi r+S)U})^T U \text{diag}(e^{-\Delta S})(U^TU)^{-1}U \\
&= \text{diag}(e^{\Delta(\psi r+S)U}) \left(I - {}^TU(U^TU)^{-1}U \right) + \text{diag}(e^{\Delta(\psi r+S)U}) \text{diag}(e^{-\Delta SU}) \, {}^TU(U^TU)^{-1}U \\
&= \text{diag}(e^{\Delta(\psi r+S)U}) \left(I - {}^TU(U^TU)^{-1}U \right) + \text{diag}(e^{\Delta \psi r U}) \, {}^TU(U^TU)^{-1}U
\end{aligned} \tag{C.17}
$$

Where the second to last line holds under the assumption that each species belong to at most one island.

We further need to compute

$$
\begin{aligned}
\int_{\tau_i}^{\tau_{i+1}} e^{(s-\tau_{i+1})A_i} a_i ds &= \psi\theta \int_{\tau_i}^{\tau_{i+1}} \text{diag}(e^{(s-\tau_{i+1})\psi r U}) ds \, {}^TU^T r \\
&= \psi\theta \int_{\tau_i}^{\tau_{i+1}} e^{(s-\tau_{i+1})\psi} ds \, {}^TU^T r \\
&= \theta \left(1 - e^{\psi\Delta} \right) \, {}^TU^T r
\end{aligned} \tag{C.18}
$$

We thus get $m_{\tau_{i+1}^-}$ with the help of Equations (C.17) and (C.18).

We now turn to the reduction of the variance expression. Remark first that A_i and Γ_i are symmetric, and so are $e^{\Delta A_i}$ and $e^{\Delta A_i}\Gamma_i$. Moreover, Γ_i is diagonal, and commutes with $e^{\Delta A_i}$, leading to:

$$
\Sigma_{\tau_{i+1}^-} = e^{\Delta A_i} \Sigma_{\tau_i} e^{\Delta A_i} + \int_{\tau_i}^{\tau_{i+1}} e^{2(s-\tau_{i+1})A_i} \Gamma_i \Gamma_i ds
$$

The first term can be computed thanks to equation (C.17). For the second one we get

$$
\begin{aligned}
\int_{\tau_i}^{\tau_{i+1}} e^{2(s-\tau_{i+1})A_i} \Gamma_i \Gamma_i ds &= \sigma^2 \int_{\tau_i}^{\tau_{i+1}} e^{2(s-\tau_{i+1})\text{diag}(r(\psi I+S)U)} ds \left(I - {}^TU(U^TU)^{-1}U \right) \, \text{diag}(rU) \\
&\quad + \sigma^2 \int_{\tau_i}^{\tau_{i+1}} e^{2(s-\tau_{i+1})\psi \, \text{diag}(rU)} ds \, {}^TU(U^TU)^{-1}U \, \text{diag}(rU) \\
&= \sigma^2 \int_{\tau_i}^{\tau_{i+1}} \text{diag}(e^{2(s-\tau_{i+1})(\psi r+S)U}) ds \left(\text{diag}(rU) - {}^TU(U^TU)^{-1}U \right) \\
&\quad + \sigma^2 \int_{\tau_i}^{\tau_{i+1}} \text{diag}(e^{2(s-\tau_{i+1})\psi r U}) ds \, {}^TU(U^TU)^{-1}U
\end{aligned} \tag{C.19}
$$

At the end, we get $\Sigma_{\tau_{i+1}^-}$ with the help of Equations (C.17) and (C.19).

C.3.4 The generalist matching mutualism (GMM) model

We recall the model formulation here. Assume that we rank first the n_1 plant traits, before the n_2 butterfly traits in the X vector. Traits evolve following the equation:

$$\forall k \in \{1, ..., n_1\},\ \ dX_t^{(k)} = S\left(d_1 + \frac{1}{n_2}\sum_{l=n_1+1}^{n_1+n_2} X_t^{(l)} - X_t^{(k)}\right)dt + \sigma dW_t^{(k)}$$
$$\forall l \in \{n_1+1, ..., n_1+n_2\},\ \ dX_t^{(l)} = S\left(d_2 + \frac{1}{n_1}\sum_{k=1}^{n_1} X_t^{(k)} - X_t^{(l)}\right)dt + \sigma dW_t^{(l)}$$

In the general framework formulation, this leads to:

$$a(t) = {}^{tr}(Sd_1, ..., Sd_1, Sd_2, ..., Sd_2)$$
$$A = \begin{pmatrix} S & 0 & \dots & 0 & \frac{-S}{n_2} & \dots & \dots & \frac{-S}{n_2} \\ 0 & \ddots & \ddots & \vdots & \vdots & & & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots & & & \vdots \\ 0 & \dots & 0 & \ddots & \frac{-S}{n_2} & \dots & \dots & \frac{-S}{n_2} \\ \frac{-S}{n_1} & \dots & \dots & \frac{-S}{n_1} & \ddots & 0 & \dots & 0 \\ \vdots & & & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & & & \vdots & \vdots & \ddots & \ddots & 0 \\ \frac{-S}{n_1} & \dots & \dots & \frac{-S}{n_1} & 0 & \dots & 0 & S \end{pmatrix}$$
$$\Gamma = \sigma I$$

We would like to be able to compute the expectation and variance easily during each epoch. We thus want to reduce Equations (4.4a, 4.4b). For simplicity, we will write in the following $\Delta = \tau_i - \tau_{i+1}$. With some work, we can find the generic element of the matrix $e^{\Delta A}$.

First, we decompose $A = S(I + Z)$, where I is the identity matrix, and Z is made of two blocks with elements $\frac{-1}{n_2}$ and $\frac{-1}{n_1}$. I and Z commute, meaning that:

$$e^{\Delta A} = e^{\Delta S(I+Z)} = e^{\Delta SI}e^{\Delta SZ} = e^{\Delta S}e^{\Delta SZ}$$

Moreover, we can find by induction the generic element of the matrix Z^k, as presented in Figure (C.2).

We then use this to find the generic element of the matrix $e^{\Delta SZ} = \sum_{k\geq 0}\frac{S^k\Delta^k Z^k}{k!} = I + \sum_{k\geq 1}\frac{S^k\Delta^k Z^k}{k!}$. We recall that the odd and even parts of the exponential are:

$$e^{\lambda} - e^{-\lambda} = \sum_{k\geq 0}\frac{\lambda^k}{k!} - \sum_{k\geq 0}\frac{(-1)^k\lambda^k}{k!} = 2\sum_{k\geq 0}\frac{\lambda^{2k+1}}{(2k+1)!}$$
$$\text{and}\ \ e^{\lambda} + e^{-\lambda} = 2\sum_{k\geq 0}\frac{\lambda^{2k}}{(2k)!}$$

Then, matrices $e^{\Delta SZ}$ and $e^{\Delta A}$ are composed of four distinct blocks, which expressions are shown in Figure C.3.

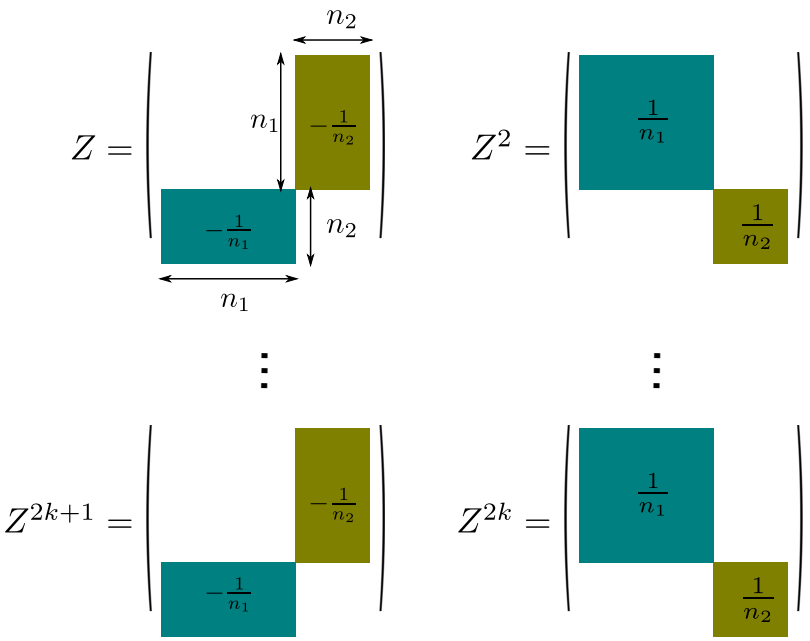


Figure C.2 – Generic element of the matrix $Z^k,\ \forall k \in \mathbb{N}^*$.

We thus got the main element from which we can derive the expectation vector $m_{\tau_{i+1}^-}$:

$$
\begin{aligned}
m_{\tau_{i+1}^-} &= e^{\Delta A_i} m_{\tau_i} + \int_{\tau_i}^{\tau_{i+1}} e^{(s-\tau_{i+1})A_i} a_i(s) ds \\
&= e^{\Delta A_i} m_{\tau_i} + \int_{\tau_i}^{\tau_{i+1}} \begin{pmatrix}
Sd_1 e^{S(s-\tau_{i+1})} + Sd_1 \frac{e^{2S(s-\tau_{i+1})} - 2e^{S(s-\tau_{i+1})} + 1}{2} + Sd_2 \frac{1 - e^{2S(s-\tau_{i+1})}}{2} \\
\vdots \\
Sd_1 e^{S(s-\tau_{i+1})} + Sd_1 \frac{e^{2S(s-\tau_{i+1})} - 2e^{S(s-\tau_{i+1})} + 1}{2} + Sd_2 \frac{1 - e^{2S(s-\tau_{i+1})}}{2} \\
Sd_2 e^{S(s-\tau_{i+1})} + Sd_1 \frac{1 - e^{2S(s-\tau_{i+1})}}{2} + Sd_2 \frac{e^{2S(s-\tau_{i+1})} - 2e^{S(s-\tau_{i+1})} + 1}{2} \\
\vdots \\
Sd_2 e^{S(s-\tau_{i+1})} + Sd_1 \frac{1 - e^{2S(s-\tau_{i+1})}}{2} + Sd_2 \frac{e^{2S(s-\tau_{i+1})} - 2e^{S(s-\tau_{i+1})} + 1}{2}
\end{pmatrix} ds
\end{aligned}
$$

$$e^{\Delta SZ} = I + \sum_{k\geq 1} \frac{S^k \Delta^k Z^k}{k!} = I + \left(\begin{array}{cc} \frac{e^{S\Delta} + e^{-S\Delta} - 2}{2n_1} & \frac{e^{-S\Delta} - e^{S\Delta}}{2n_2} \\ \frac{e^{-S\Delta} - e^{S\Delta}}{2n_1} & \frac{e^{S\Delta} + e^{-S\Delta} - 2}{2n_2} \end{array} \right)$$

$$e^{\Delta A} = e^{\Delta S} e^{\Delta SZ} = e^{\Delta S} I + \left(\begin{array}{cc} \frac{\alpha}{n_1} & \frac{\beta}{n_2} \\ \frac{\beta}{n_1} & \frac{\alpha}{n_2} \end{array} \right)$$

where $\alpha := \dfrac{e^{2S\Delta} - 2e^{S\Delta} + 1}{2}$

$\beta := \dfrac{1 - e^{2S\Delta}}{2}$

Figure C.3 – Generic elements of matrices $e^{\Delta SZ}$ and $e^{\Delta A}$.

$$m_{\tau_{i+1}^-} = e^{\Delta A_i} m_{\tau_i} + \int_{\tau_i}^{\tau_{i+1}} \begin{pmatrix} S\frac{d_1+d_2}{2} + S\frac{d_1}{2} e^{2S(s-\tau_{i+1})} - S\frac{d_2}{2} e^{2S(s-\tau_{i+1})} \\ \vdots \\ S\frac{d_1+d_2}{2} + S\frac{d_1}{2} e^{2S(s-\tau_{i+1})} - S\frac{d_2}{2} e^{2S(s-\tau_{i+1})} \\ S\frac{d_1+d_2}{2} - S\frac{d_1}{2} e^{2S(s-\tau_{i+1})} + S\frac{d_2}{2} e^{2S(s-\tau_{i+1})} \\ \vdots \\ S\frac{d_1+d_2}{2} - S\frac{d_1}{2} e^{2S(s-\tau_{i+1})} + S\frac{d_2}{2} e^{2S(s-\tau_{i+1})} \end{pmatrix} ds$$

$$= e^{\Delta A_i} m_{\tau_i} + \begin{pmatrix} -S\frac{d_1+d_2}{2}\Delta + \frac{d_1}{4}(1 - e^{2S\Delta}) - \frac{d_2}{4}(1 - e^{2S\Delta}) \\ \vdots \\ -S\frac{d_1+d_2}{2}\Delta + \frac{d_1}{4}(1 - e^{2S\Delta}) - \frac{d_2}{4}(1 - e^{2S\Delta}) \\ -S\frac{d_1+d_2}{2}\Delta - \frac{d_1}{4}(1 - e^{2S\Delta}) + \frac{d_2}{4}(1 - e^{2S\Delta}) \\ \vdots \\ -S\frac{d_1+d_2}{2}\Delta - \frac{d_1}{4}(1 - e^{2S\Delta}) + \frac{d_2}{4}(1 - e^{2S\Delta}) \end{pmatrix}$$

We now turn to the derivation of the covariance matrix, which requires simplifying:

$$\int_{\tau_i}^{\tau_{i+1}} \left(e^{(s-\tau_{i+1})A_i}\Gamma_i(s)\right)^{tr}\left(e^{(s-\tau_{i+1})A_i}\Gamma_i(s)\right) ds = \sigma^2 \int_{\tau_i}^{\tau_{i+1}} \left(e^{(s-\tau_{i+1})A_i}\right)^{tr}\left(e^{(s-\tau_{i+1})A_i}\right) ds$$

The expression of this last matrix is given in Figure C.4.

$$\sigma^2 \int_{\tau_i}^{\tau_{i+1}} \left(e^{(s-\tau_{i+1})A_i}\right)^{tr}\left(e^{(s-\tau_{i+1})A_i}\right) ds$$

$$= \sigma^2 \int_{\tau_i}^{\tau_{i+1}} \begin{pmatrix} e^{S(s-\tau_{i+1})}I + \begin{pmatrix} \frac{\alpha}{n_1} & \frac{\beta}{n_2} \\ \frac{\beta}{n_1} & \frac{\alpha}{n_2} \end{pmatrix} \end{pmatrix} \begin{pmatrix} e^{S(s-\tau_{i+1})}I + \begin{pmatrix} \frac{\alpha}{n_1} & \frac{\beta}{n_1} \\ \frac{\beta}{n_2} & \frac{\alpha}{n_2} \end{pmatrix} \end{pmatrix} ds$$

$$= \sigma^2 \int_{\tau_i}^{\tau_{i+1}} \begin{pmatrix} e^{2S(s-\tau_{i+1})}I + \left(\frac{1}{n_1}+\frac{1}{n_2}\right) \begin{pmatrix} \frac{e^{4S(s-\tau_{i+1})}-2e^{2S(s-\tau_{i+1})}+1}{4} & \frac{1-e^{4S(s-\tau_{i+1})}}{4} \\ \frac{1-e^{4S(s-\tau_{i+1})}}{4} & \frac{e^{4S(s-\tau_{i+1})}-2e^{2S(s-\tau_{i+1})}+1}{4} \end{pmatrix} \end{pmatrix} ds$$

$$= \sigma^2 \begin{pmatrix} \frac{1-e^{2S\Delta}}{2S}I + \left(\frac{1}{n_1}+\frac{1}{n_2}\right) \begin{pmatrix} \frac{1-e^{4S\Delta}}{16S} - \frac{1-e^{2S\Delta}}{4S} - \frac{\Delta}{4} & \frac{e^{4S\Delta}-1}{16S} - \frac{\Delta}{4} \\ \frac{e^{4S\Delta}-1}{16S} - \frac{\Delta}{4} & \frac{1-e^{4S\Delta}}{16S} - \frac{1-e^{2S\Delta}}{4S} - \frac{\Delta}{4} \end{pmatrix} \end{pmatrix}$$

Figure C.4 – Generic elements of matrices that help us compute the covariance matrix of the distribution.

C.4 Simulation and Inference

We do not give any new result in this Appendix section. Instead, we present the ways we implemented numerically simulations and inferences for all models described in the paper. These have been previously described in a number of papers.

C.4.1 Numerical methods for simulating data

Simulating the whole trajectory of the process

We use the Euler-Maruyama scheme, which works like the Euler scheme for ODEs, but with the addition of a small Gaussian random variable at each time step (Gardiner et al., 1985). We discretize each epoch (τ_i, τ_{i+1}) with a mesh Δ_t. We consider m standard Gaussian vectors of dimension nd: $(U_j)_{j=1}^m$. We approximate our SDE on this interval in the following way:

$$Y_0 = X_0$$
$$Y_{\tau_i+m\Delta_t} = Y_{\tau_i+(m-1)\Delta_t} + (a_i(\tau_i + (m-1)\Delta_t) - A_i Y_{\tau_i+(m-1)\Delta_t})\Delta_t + \Gamma(\tau_i + (m-1)\Delta_t)\sqrt{\Delta_t}U_m$$

When a branching occurs, the values of the process on the splitting branch are duplicated at the end of the vector Y. We then iterate this operation from the root up to present time.

This simulation allows us to get the whole trajectory of the process on the tree, which can mainly be used to produce pictures as in Figure C.5, and eventually get a useful intuition on the process. However, we rarely use the whole trajectories, because observed data are only composed of tip trait values.

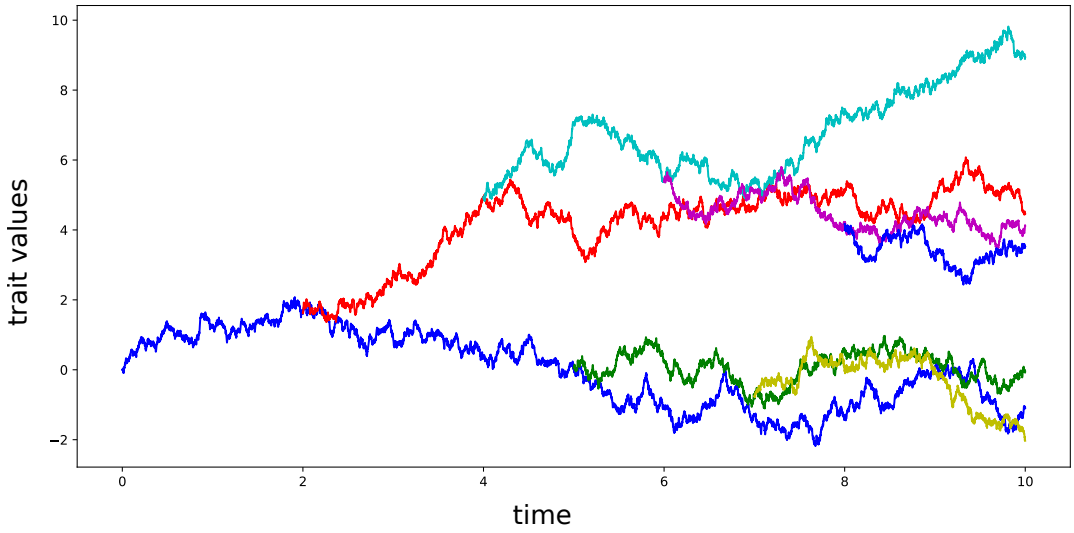


Figure C.5 – Evolution of a Brownian phenotypic trait along a tree, following the SDE: $dX_t = \sigma I dW_t$.

Simulating values of the process at the tips only

This second simulation protocol allows us to simulate the process values at the tips only. Suppose that we know the vector m of expectations and the covariance matrix Σ at the tips of the tree.

We then simply simulate numerically a Gaussian vector with law:

$$X_{t_f} \sim \mathcal{N}(m\ ,\ \Sigma)$$

This is by far the quickest way to get the tip values. However, as the inference protocol relies on the use of the same vector of expectations and covariance matrix, one may prefer to use the other simulation protocols to test the consistency between simulation and inference. In case there is an issue with the derivation of the tip distribution, there would be a discrepancy between simulations and inferences.

C.4.2 Parameter inference

Parameter inference principle

We consider here that we know the topology of the true phylogeny with K tips, its branch lengths, and the state of d phenotypic traits at the tip, denoted by $\mathcal{X}$.

We assume any model of phenotypic evolution relying on linear SDEs, with vector of parameters p. We can compute the expectation m_p and the covariance Σ_p of the process X at tree tips, which law is then: $X \sim \mathcal{N}(m_p, \Sigma_p)$. Recall from Appendix C.1.2 that Σ_p is positive definite in most cases, and is thus theoretically non-singular. However, one must be cautious with numerical implementations, as numerical approximations might still lead to 'numerically non-invertible' matrices. Here, we assume that the variance matrix is invertible, and the density of the vector X is:

$$\forall x \in \mathbb{R}^{Kd}, \quad f(x) = \frac{1}{\sqrt{(2\pi)^{Kd}\det(\Sigma_p)}} e^{-\frac{1}{2}tr(x-m_p)\Sigma_p^{-1}(x-m_p)}$$

We can thus write the likelihood of the observed phenotypic traits as,

$$\begin{aligned}\mathcal{L}(p) &= f(\mathcal{X}|p) \\ &= \frac{1}{\sqrt{(2\pi)^{Kd}\det(\Sigma_p)}} e^{-\frac{1}{2}tr(\mathcal{X}-m_p)\Sigma_p^{-1}(\mathcal{X}-m_p)}\end{aligned}$$

The maximum likelihood estimators (MLE) are the parameter values that maximize the likelihood function, that is,

$$\hat{p} = \underset{p}{\operatorname{argmax}}\ \mathcal{L}(p)$$

Equivalently, we can minimize the following function,

$$-\ln(\mathcal{L}(p)) = \frac{1}{2}Kd\ln(2\pi) + \frac{1}{2}\ln(\det(\Sigma_p)) + \frac{1}{2}tr(\mathcal{X}-m_p)\Sigma_p^{-1}(\mathcal{X}-m_p)$$

or, removing the constants,

$$U(p) = \ln(\det(\Sigma_p)) + {}^{tr}(\mathcal{X}-m_p)\Sigma_p^{-1}(\mathcal{X}-m_p)$$

Analytical derivation of the MLE

Among all models described in the paper, only the BM model allows the analytic derivation of the MLE estimators. Take for illustration a BM model without drift starting with $(m_0, v_0) = (0, 0)$. According to Table C.1, the expectation m and covariance matrix Σ at the tips are $m = 0$ and $\Sigma = \sigma^2 T$, where matrix T has element $T^{(k,l)} = t_{k,l}$.

We get the MLE $\widehat{\sigma}$ by looking analytically for the minimum of U,

$$
\begin{aligned}
U(\sigma) &= \ln(\det(\sigma^2 T)) + {}^{tr}\mathcal{X}\frac{T^{-1}}{\sigma^2}\mathcal{X} \\
&= \ln\det T + 2n\ln\sigma + \frac{1}{\sigma^2}{}^{tr}\mathcal{X}T^{-1}\mathcal{X} \\
\frac{dU}{d\sigma} &= \frac{2n}{\sigma} - \frac{2}{\sigma^3}{}^{tr}\mathcal{X}T^{-1}\mathcal{X}
\end{aligned}
$$

Thus leading to,

$$
\widehat{\sigma}^2 = \frac{1}{n}{}^{tr}\mathcal{X}T^{-1}\mathcal{X}
$$

Speeding up the ML estimation by reducing the dimension of the parameter space

Maximizing the likelihood can take a long time, especially when the dimension of the parameter space is large. It can thus be interesting to make assumptions that lower the number of parameters, when this is biologically tolerable. Examples include,

— starting an OU process with $m_0 = \theta$,

— considering no root variance, $v_0 = 0$,

— starting a PM model with $m_0 = \theta$ (in which case we easily show that the expectation remains θ in all lineages),

— putting $\psi = 0$ in the PM model.

In many models (e.g. BM, OU, ACDC, PM with $m_0 = \theta$...), distinct sets of parameters p_1 and p_2 are involved in the computation of m and Σ, and the expectation vector m can be expressed as $m = Cp_1$. In this case, at a given p_2, we can analytically get the parameters p_1 maximizing $\ln(\mathcal{L}(p_1, p_2))$,

$$
\frac{\partial}{\partial p_1}U(p_1, p_2) = 0 \iff \frac{d}{dp_1}{}^{tr}(\mathcal{X} - Cp_1)\Sigma_{p_2}^{-1}(\mathcal{X} - Cp_1) = 0
$$

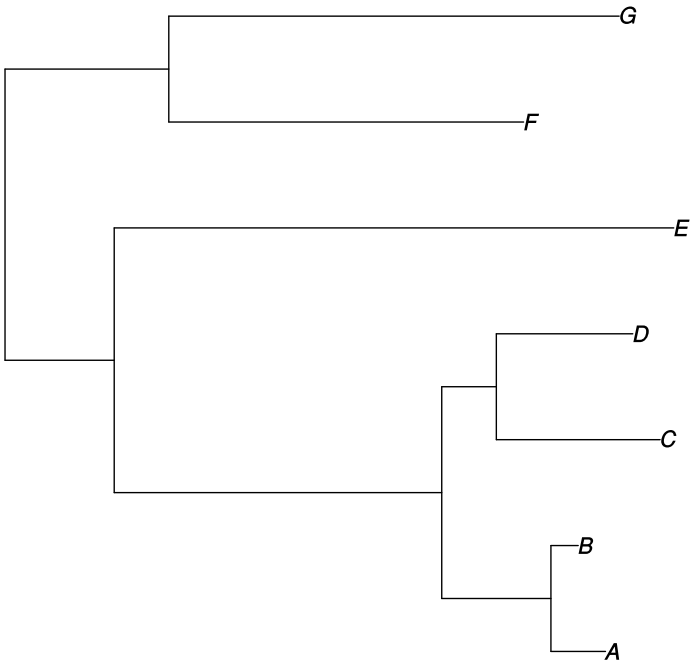
Doing so, we get the same formula as in (Hansen, 1997; Butler and King, 2004), i.e. $\widehat{p_1} = \left({}^{tr}C_1\Sigma_{p_2}^{-1}C_1\right)^{-1}{}^{tr}C_1\Sigma_{p_2}^{-1}X$.

C.5 Tutorial: using the RPANDA code to study trait coevolution

The aim of this section is to describe the R code associated to our framework. We describe the class `PhenotypicModel`, we show how to manipulate the different methods included in the class, we illustrate their use around a simple (non-ultrametric) tree, and we finally explain how to use our codes to write new models fitting the framework.

We first need to load useful R packages, along with our codes, and a small, non-ultrametric, tree.

```
In [219]: source("Loading.R")
          newick <- "((((A:1,B:0.5):2,(C:3,D:2.5):1):6,E:10.25):2,(F:6.5,G:8.25):3):1;"
          tree <- read.tree(text=newick)
          plot(tree)
```



C.5.1 The 'PhenotypicModel' class

Our code is structured around one main R class that we called 'PhenotypicModel', which is intended to mimic the framework that we proposed in the main text. Each object of the 'PhenotypicModel' encompasses informations on the tree, on the parameters of the model, on the starting values, and, finally, on the collection of (a_i, A_i, Γ_i) for all epochs.

Loading a pre-defined model

Because we wanted this code both to be user-friendly and to serve as an illustration of what can be written within this framework, we implemented all models in main Table 4.1 in a generic constructor

`createModel`, in the file 'ModelBank.R', that takes for arguments the tree and the name of the required model.

Available models include:

BM Brownian Motion model with linear drift.
Starts with two lineages having the same value $X_0 \sim \mathcal{N}(m_0, v_0)$.
One trait in each lineage, all lineages evolving independently after branching following the equation.

$$dX_t^{(i)} = ddt + \sigma dW_t^{(i)}$$

BM_from0 Same as above, but starting with two lineages having the same value $X_0 \sim \mathcal{N}(0, 0)$.

BM_from0_driftless Same as above, but with $d = 0$.

OU Ornstein-Uhlenbeck model.
Starts with two lineages having the same value $X_0 \sim \mathcal{N}(m_0, v_0)$.
One trait in each lineage, all lineages evolving independently after branching, following the equation:

$$dX_t^{(i)} = \psi(\theta - X_t)dt + \sigma dW_t^{(i)}$$

OU_from0 Same as above, but starting with two lineages having the same value $X_0 \sim \mathcal{N}(0, 0)$.

ACDC ACcelerating or DeCelerating model.
Starts with two lineages having the same value $X_0 \sim \mathcal{N}(m_0, v_0)$.
One trait in each lineage, all lineages evolving independently after branching, following the equation:

$$dX_t^{(i)} = \sigma_0 e^{rt} dW_t^{(i)}$$

DD Diversity-Dependent model.
Starts with two lineages having the same value $X_0 \sim \mathcal{N}(m_0, v_0)$.
One trait in each lineage, all lineages evolving independently after branching, following the equation:

$$dX_t^{(i)} = \sigma_0 e^{rn_t} dW_t^{(i)}$$

PM Phenotype Matching model.
Starts with two lineages having the same value $X_0 \sim \mathcal{N}(m_0, v_0)$.
One trait in each lineage, all lineages evolving then non-independently following the expression:

$$dX_t^{(i)} = \psi\left(\theta - X_t^{(i)}\right) + S\left(\frac{1}{n}\sum_{k=1}^{n} X_t^{(k)} - X_t^{(i)}\right) + \sigma dW_t^{(i)}$$

PM_OUless Simplified Phenotype Matching model.
Starts with two lineages having the same value $X_0 \sim \mathcal{N}(m_0, v_0)$.
One trait in each lineage, all lineages evolving then non-independently following the expression:

$$dX_t^{(i)} = S\left(\frac{1}{n}\sum_{k=1}^{n} X_t^{(k)} - X_t^{(i)}\right) + \sigma dW_t^{(i)}$$

To get a first glimpse at 'PhenotypicModel' objects, we first create two such objects. The first one is a Brownian Motion (BM), the second one is an Ornstein-Uhlenbeck process (OU). Note that both models include m_0 and v_0 as parameters.

```
In [220]: modelBM <- createModel(tree, 'BM')
          modelOU <- createModel(tree, 'OU')
```

Access to the content of the model

The function `show` is intended to give basic information on a specific 'PhenotypicModel' object, whereas the `print` function displays full information.

```
In [221]: show(modelBM)

*****************************************************************
*** Object of Class PhenotypicModel ***
*** Name of the model : [1] "BM"
*** Parameters of the model : [1] "m0"    "v0"    "d"     "sigma"
*** Description : Brownian Motion model with linear drift.
Starts with two lineages having the same value X_0 ~ Normal(m0,v0).
One trait in each lineage, all lineages evolving independently after branching.
dX_t = d dt + sigma dW_t
*** Periods : the model is cut into  13  parts.
For more details on the model, call : print(PhenotypicModel)
*****************************************************************

In [222]: print(modelOU)

*****************************************************************
*** Object of Class PhenotypicModel ***
*** Name of the model : [1] "OU"
*** Parameters of the model : [1] "m0"    "v0"    "psi"   "theta" "sigma"
*** Description : Ornstein-Uhlenbeck model.
Starts with two lineages having the same value X_0 ~ Normal(m0,v0).
One trait in each lineage, all lineages evolving independently after branching.
dX_t = psi(theta- X_t) dt + sigma dW_t
*** Epochs : the model is cut into  13  parts.
 [1]  0.00  2.00  3.00  8.00  9.00  9.50 10.00 10.50 11.00 11.25 11.50 12.00
[13] 12.25
*** Lineages branching (to be copied at the end of the corresponding period) :
 [1] 1 1 2 1 5 2 1 7 1 4 6 5 3
*** Positions of the new trait at the end of each period :
 [1] 2 3 4 5 6 0 7 0 0 0 0 0 0
*** Initial condition :
function (params)
return(list(mean = c(params[1]), var = matrix(c(params[2]))))
<environment: 0x9617460>
*** Vectors a_i, A_i, Gamma_i on each period i :
function (i, params)
{
    vectorU <- getLivingLineages(i, eventEndOfPeriods)
    vectorA <- function(t) return(params[3] * params[4] * vectorU)
    matrixGamma <- function(t) return(params[5] * diag(vectorU))
    matrixA <- params[3] * diag(vectorU)
    return(list(a = vectorA, A = matrixA, Gamma = matrixGamma))
}
<environment: 0x9617460>
*** Constraints on the parameters :
function (params)
return(params[2] >= 0 && params[5] >= 0 && params[3] != 0)
<environment: 0x9617460>
*** Defaut parameter values : [1] 0 0 1 0 1
*** Tip labels :
[1] "A" "B" "C" "D" "E" "F" "G"
*** Tip labels for simulations :
```

```
[1] "A" "F" "E" "G" "C" "D" "B"
*****************************************************************
```

List of class attributes

The latter command gave us some insight into how a `PhenotypicModel` is defined. It has the following list of attributes:

name a name,

paramsNames the names of all parameters,

comment a short description,

period the vector of times at which successive branching and death of lineages occur,

numbersCopy vector containing the lineage number which branches or dies at the end of each period,

numbersPaste vector containing the lineage number in which a daughter lineage is placed at the end of each period (zero if the end of the period corresponds to a death),

initialCondition a function of the parameters giving the initial mean and variance of the gaussian process at the root of the tree,

aAGamma the functions corresponding to $a_i(t)$, A_i, and $\Gamma_i(t)$ that define the evolution of the process on each period, depending on parameters,

constraints a function of the parameters giving the definition range,

params0 a vector of defaut parameter values.

Each of these attributes can be accessed and changed through the use of the following syntax.

```
In [223]: modelBM['name']
```

```
Out[223]: 'BM'
```

```
In [224]: modelBM['paramsNames']
```

```
Out[224]: 'm0'  'v0'  'd'  'sigma'
```

```
In [225]: modelOU['paramsNames'] <- c("mean0", "var0", "selectionStrength", "equilibrium",
          "noise")
          show(modelOU)
```

```
*****************************************************************
*** Object of Class PhenotypicModel ***
*** Name of the model : [1] "OU"
*** Parameters of the model : [1] "mean0"             "var0"              "selectionStrength"
[4] "equilibrium"       "noise"
*** Description : Ornstein-Uhlenbeck model.
Starts with two lineages having the same value X_0 ~ Normal(m0,v0).
One trait in each lineage, all lineages evolving independently after branching.
dX_t = psi(theta- X_t) dt + sigma dW_t
*** Periods : the model is cut into  13  parts.
For more details on the model, call : print(PhenotypicModel)
*****************************************************************
```

However, changes must be made cautiously, in order to keep a coherent model. For example, changing 'paramsNames' for a shorter vector would not be authorized, but other deleterious actions could work and lead to issues with methods associated to `PhenotypicModel` objects.

```
In [226]: modelOU['paramsNames'] <- c("mean0", "var0")
```

```
Error in validityMethod(as(object, superClass)): [PhenotypicModel : validation]
There should be the same number of defaut parameters and parameter names.
```

C.5.2 Methods associated to the 'PhenotypicModel' class

All 'PhenotypicModel' objects are associated to methods intended to do the basic operations that we need to do with models of trait evolution, i.e.,

i) simulate tip trait data,

ii) compute the likelihood of tip trait data,

iii) fit the model to tip trait data.

Simulating tip trait data

The method `simulateTipData` works for any `PhenotypicModel` object. We simply give it the model and the set of parameters and it returns a realisation of the process (tip data).

```
In [227]: dataBM <- simulateTipData(modelBM, c(0,0,0,1))
          dataBM
```

```
*** Simulation of tip trait values ***
Simulates step-by-step the whole trajectory, but returns only the tip data.
Computation time : 0.3909395 secs
```

Out[227]:

A	-2.71863
F	1.043329
E	0.665404
G	-3.440327
C	0.272335
D	-0.7023421
B	-2.010951

A third, optional, argument, changes the behaviour of the method.

— "method=1": first computes the tip distribution at present, before drawing a realization of this distribution,

— "method=2": simulates step-by-step the whole trajectory of the process, plots the trajectories through time, and returns the tip data.

— "method=3": (default) simulates step-by-step the whole trajectory of the process, before returning only the tip data.

```
In [228]: dataOU <- simulateTipData(modelOU, c(0,0,1,5,1), method=1)
          dataOU
```

```
*** Simulation of tip trait values ***
Computes the tip distribution, and returns a simulated dataset drawn in this distribution.
Computation time : 0.0009741783 secs
```

Out[228]:

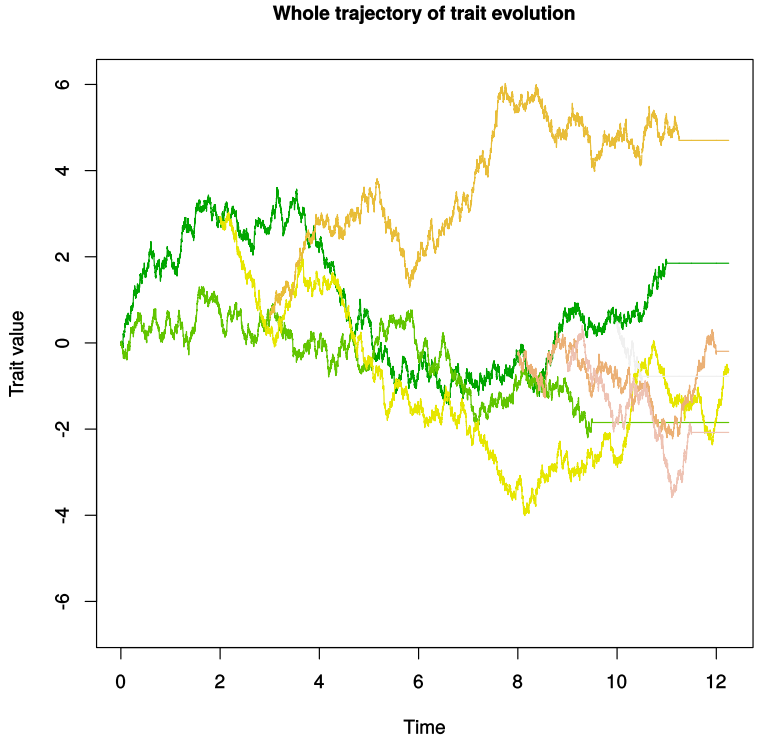
A	4.179412
B	5.776153
C	4.984526
D	4.480901
E	5.693471
F	4.636019
G	5.815942

```
In [229]: simulateTipData(modelBM, c(0,0,0,1), method=2)
```

```
*** Simulation of tip trait values ***
Simulates step-by-step the whole trajectory, plots it, and returns tip data.
Computation time : 0.479032 secs
```

Out[229]:

A	1.850113
F	-1.846854
E	-0.6321431
G	4.701758
C	-0.1940776
D	-2.077116
B	-0.7752916



Getting the distribution of the model under a given set of parameters

The method `getTipDistribution` computes the mean vector m and variance matrix Σ such that, under the model, the tip trait data X follows $\mathcal{N}(m, \Sigma)$.

The related method `getDataLikelihood` returns the -ln(likelihood) of a given data set under the model, with a given set of parameters.

```
In [230]: getTipDistribution(modelBM, c(0,0,1,1))
```

Out[230]:

$mean

A	11
B	10.5
C	12
D	11.5
E	12.25
F	9.5
G	11.25

$Sigma

	A	B	C	D	E	F	G
A	11	10	8	8	2	0	0
B	10.0	10.5	8.0	8.0	2.0	0.0	0.0
C	8	8	12	9	2	0	0
D	8.0	8.0	9.0	11.5	2.0	0.0	0.0
E	2.00	2.00	2.00	2.00	12.25	0.00	0.00
F	0.0	0.0	0.0	0.0	0.0	9.5	3.0
G	0.00	0.00	0.00	0.00	0.00	3.00	11.25

```
In [231]: getDataLikelihood(modelBM, dataBM, c(0,0,1,1))
```

Out[231]: 36.0510113479088

Maximum likelihood estimation of parameters

The method `fitTipData` uses the latter two methods to find the set of parameters that minimizes -ln(likelihood) for a given model, on a given data set. We can apply this method to simulated datasets, and compare the maximum likelihood estimators with the parameters used in the simulation.

Note that this function accepts a third, optional, parameter, that is the starting vector 'params0' given to optimize the likelihood. If no value is specified, the function takes the attribute 'params0' in the `PhenotypicModel` object.

```
In [232]: fitTipData(modelBM, dataBM)
```

```
*** Fit of tip trait data ***
Finds the maximum likelihood estimators of the parameters,
returns the likelihood and the inferred parameters.
**WARNING** : This function uses the standard R optimizer "optim".
It may not always converge well.
Please double check the convergence by trying
distinct parameter sets for the initialisation.
Computation time : 0.02105212 secs
```

Out[232]:

$value 13.3539168672421

$inferredParams *m0* 0.112360024529455

v0 4.3703974585017e-08

d -0.0733871266399529

sigma 0.64761762031608

```
In [233]: fitTipData(modelOU, dataOU)
```

```
*** Fit of tip trait data ***
Finds the maximum likelihood estimators of the parameters,
returns the likelihood and the inferred parameters.
**WARNING** : This function uses the standard R optimizer "optim".
It may not always converge well.
Please double check the convergence by trying
distinct parameter sets for the initialisation.
Computation time : 0.2915776 secs
```

Out[233]:

$value 7.5162883379935

$inferredParams *mean0* 13.5665180225751

var0 1.6815664554916e-05

selectionStrength 0.648513938633288

equilibrium 5.05532921748184

noise 0.766630199120977

It doesn't seem quite good, but it also seems like the choice in the starting parameters m_0, v_0 has a bad influence. As presented in Online Appendix C.4.2, in many models (e.g. BM, OU, ACDC, PM with $m_0 = \theta \dots$), distinct sets of parameters p_1 and p_2 are involved in the computation of m and Σ, and the expectation vector m can be expressed as $m = Cp_1$. In particular, many models verify $m = {}^{tr}(m_0, m_0, ...m_0)$. When this is the case, the fit of tip data can be improved and speeded up by using the third parameter of the function `GLSstyle=TRUE`.

```
In [234]: fitTipData(modelBM, dataBM, GLSstyle=TRUE)
          fitTipData(modelOU, dataOU, GLSstyle=TRUE)
```

```
*** Fit of tip trait data ***
Finds the maximum likelihood estimators of the parameters,
returns the likelihood and the inferred parameters.
**WARNING** : This function uses the standard R optimizer "optim".
It may not always converge well.
Please double check the convergence by trying
distinct parameter sets for the initialisation.
Computation time : 0.03260899 secs
```

Out[234]:

$value 13.5302740469078

$inferredParams *m0* -0.00550320295296933

v0 2.28469756397133e-07

d -0.313019528308928

sigma 0.663621107698308

```
*** Fit of tip trait data ***
Finds the maximum likelihood estimators of the parameters,
returns the likelihood and the inferred parameters.
Computation time : 0.1760004 secs
```

Out[234]:

$value 7.82305350777471

$inferredParams *mean0* 5.10957361631891

var0 3.36222531349288e-05

selectionStrength 1.87722870245168

equilibrium -1.98889519193151

noise 1.91905948952067

With so few data in hand, we could also prefer to consider directly models starting with $(m_0, v_0) = (0, 0)$. We create two new models 'BM_from0' and 'OU_from0' with the subtle difference that $(m_0, v_0) = (0, 0)$ and the models thus retain respectively only two and three parameters.

These two models are included in the 'ModelBank' file.

```
In [235]: modelBMfromZero <- createModel(tree, 'BM_from0')
          modelBMfromZero['paramsNames']
```

Out[235]: 'd' 'sigma'

```
In [236]: modelOUfromZero <- createModel(tree, 'OU_from0')
          modelOUfromZero['paramsNames']
```

Out[236]: 'psi' 'theta' 'sigma'

```
In [237]: fitTipData(modelBMfromZero, dataBM)
```

```
*** Fit of tip trait data ***
Finds the maximum likelihood estimators of the parameters,
returns the likelihood and the inferred parameters.
**WARNING** : This function uses the standard R optimizer "optim".
It may not always converge well.
Please double check the convergence by trying
distinct parameter sets for the initialisation.
Computation time : 0.01061678 secs
```

Out[237]:

$value 13.3540474589618

$inferredParams *d* -0.0633929373190768

sigma 0.647501517840828

The `fitTipData` function uses the `optim` function available in R to maximize the likelihood. This optimizer is widely used to fit phenotypic models, but is known to sometimes converge on a local optima rather than the maximum likelihood. It is thus important to assess the sensitivity of the solution to the choice of the initial parameter values before drawing conclusions.

Finally, the functions `getTipDistribution`, `simulateTipData` and `fitTipData` all have a last optional argument, called `v` for "verbose mode". With `v=TRUE`, the functions gives informations in the console, whereas with `v=FALSE` the function remains silent.

C.5.3 Toward an in-depth understanding of the code structure

This section can be skipped if you are not interested in using this framework to build your own model. Otherwise, it is worth understanding how the different models relate to each others.

Relationships between the different classes of models

The superclass, for which all the above-mentionned functions are defined, is the `PhenotypicModel` class. When a model is only known as a `PhenotypicModel`, the method that computes the tip distribution, namely `getTipDistribution` is the most general one. It thus computes the distribution by resolving numerically the ODE system presented in main text Equations (4.5a, 4.5b), which can take a lot of time.

However, faster algorithms are available to compute the tip distribution under specific models (see e.g. analytical tip distribution formulas in Table C.1). This is the rationale to define subclasses:

PhenotypicBM For the Brownian model.

PhenotypicOU For the Ornstein-Uhlenbeck model.

PhenotypicACDC For the Accelerating/Decelerating model.

PhenotypicDD For the Diversity-Dependent model.

PhenotypicPM For the Phenotype-Matching model.

PhenotypicGMM For the Generalist Matching Mutualism model.

PhenotypicADiag Models for which, $\forall i, A_i$ is symmetric and $\Gamma_i = \sigma I$.

For each of these subclasses, an other, more appropriated, function `getTipDistribution` has been written. `PhenotypicModels` which are also `PhenotypicOU`, will preferentially use methods defined for `PhenotypicOU` when they exist.

Application: three different ways to define an OU

In the `createModel` function, the keyword 'OU' constructs a model in the class `PhenotypicOU`. In this class, the function `getTipDistribution` uses the analytical formula show in Online Appendix C.2.1 to speed up the computation of m and Σ.

Alternatively, the keyword 'OUbis' defines the exact same model, but as an instance of the class `PhenotypicADiag`. Thus, the function `getTipDistribution` uses the reduction show in Online Appendix C.3.1 to compute m and Σ.

Last, the keyword 'OUter' still defines the exact same model, but as an instance of the class `PhenotypicModel`. Thus, the function `getTipDistribution` uses the resolution of the ODE system to compute m and Σ.

The following lines of code show that the function returns the same value with the three different methods, but do not take the same amount of time.

```
In [240]: modelOU <- createModel(tree, 'OU')
          modelOUbis <- createModel(tree, 'OUbis')
          modelOUter <- createModel(tree, 'OUter')
          params <- c(0,0,0.2,1,2)

In [241]: getTipDistribution(modelOU, params, v=TRUE)

*** Computation of tip traits distribution through the analytical formula for an OU process ***
Computation time : 0.000497818 secs

Out[241]:
```

$mean

A	0.8891968
B	0.8775436
C	0.909282
D	0.8997412
E	0.9137064
F	0.8504314
G	0.8946008

$Sigma

	A	B	C	D	E	F	G
A	9.8772266	7.2724966	2.3654513	2.6142280	0.1171813	0.0000000	0.0000000
B	7.2724966	9.8500442	2.6142280	2.8891687	0.1295054	0.0000000	0.0000000
C	2.36545128	2.61422796	9.91770253	3.23775807	0.09593997	0.00000000	0.00000000
D	2.6142280	2.8891687	3.2377581	9.8994816	0.1060301	0.0000000	0.0000000
E	0.11718135	0.12950541	0.09593997	0.10603007	9.92553417	0.00000000	0.00000000
F	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	9.7762923	0.3657529
G	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.3657529	9.8889100

```
In [242]: getTipDistribution(modelOUbis, params, v=TRUE)
```

```
*** Computation of tip traits distribution through integrated formula ***
(Method working for models with a constant, A diagonalizable, and Gamma constant)
Computation time : 0.002770185 secs
```

Out[242]:

$mean

A	0.8891968
F	0.8504314
E	0.9137064
G	0.8946008
C	0.909282
D	0.8997412
B	0.8775436

$Sigma

	A	F	E	G	C	D	B
A	9.8772266	0.0000000	0.1171813	0.0000000	2.3654513	2.6142280	7.2724966
F	0.0000000	9.7762923	0.0000000	0.3657529	0.0000000	0.0000000	0.0000000
E	0.11718135	0.00000000	9.92553417	0.00000000	0.09593997	0.10603007	0.12950541
G	0.0000000	0.3657529	0.0000000	9.8889100	0.0000000	0.0000000	0.0000000
C	2.36545128	0.00000000	0.09593997	0.00000000	9.91770253	3.23775807	2.61422796
D	2.6142280	0.0000000	0.1060301	0.0000000	3.2377581	9.8994816	2.8891687
B	7.2724966	0.0000000	0.1295054	0.0000000	2.6142280	2.8891687	9.8500442

```
In [243]: getTipDistribution(modelOUter, params, v=TRUE)
```

```
*** Computation of tip traits distribution through ODE resolution ***
(Method working for any model)
Computation time : 0.01829243 secs
```

Out[243]:

$mean

A	0.8891984
F	0.8504309
E	0.9137081
G	0.8946024
C	0.9092837
D	0.8997429
B	0.8775447

$Sigma

	A	F	E	G	C	D	B
A	9.8772243	0.0000000	0.1171837	0.0000000	2.3654834	2.6142593	7.2725143
F	0.0000000	9.7762896	0.0000000	0.3657561	0.0000000	0.0000000	0.0000000
E	0.11718371	0.00000000	9.92553239	0.00000000	0.09594306	0.10603262	0.12950776
G	0.0000000	0.3657561	0.0000000	9.8889077	0.0000000	0.0000000	0.0000000
C	2.36548343	0.00000000	0.09594306	0.00000000	9.91769978	3.23780799	2.61425810
D	2.6142593	0.0000000	0.1060326	0.0000000	3.2378080	9.8994793	2.8891973
B	7.2725143	0.0000000	0.1295078	0.0000000	2.6142581	2.8891973	9.8500418

In [244]:
```
dataOU <- simulateTipData(modelOU, c(0,0,0.2,1,2))
fitTipData(modelOU, dataOU)
fitTipData(modelOUbis, dataOU)
fitTipData(modelOUter, dataOU)
```

```
*** Simulation of tip trait values ***
Simulates step-by-step the whole trajectory, but returns only the tip data.
Computation time : 0.2363398 secs
*** Fit of tip trait data ***
Finds the maximum likelihood estimators of the parameters,
returns the likelihood and the inferred parameters.
**WARNING** : This function uses the standard R optimizer "optim".
It may not always converge well.
Please double check the convergence by trying
distinct parameter sets for the initialisation.
Computation time : 0.1814284 secs
```

Out[244]:

$value 15.0174906724384

$inferredParams *m0* -26.3722559360675

v0 0.111663973605588

psi 0.0973609295443122

theta 14.9673044542728

sigma 1.12338425846849

```
*** Fit of tip trait data ***
Finds the maximum likelihood estimators of the parameters,
returns the likelihood and the inferred parameters.
**WARNING** : This function uses the standard R optimizer "optim".
It may not always converge well.
Please double check the convergence by trying
```

```
distinct parameter sets for the initialisation.
Computation time : 0.7557919 secs
```

Out[244]:

$value 15.0174906724384

$inferredParams *m0* -26.3722559360675

v0 0.111663973605588

psi 0.0973609295443122

theta 14.9673044542728

sigma 1.12338425846849

```
*** Fit of tip trait data ***
Finds the maximum likelihood estimators of the parameters,
returns the likelihood and the inferred parameters.
**WARNING** : This function uses the standard R optimizer "optim".
It may not always converge well.
Please double check the convergence by trying
distinct parameter sets for the initialisation.
Computation time : 6.088683 secs
```

Out[244]:

$value 15.0174914969285

$inferredParams *m0* -26.3722559360675

v0 0.111663973605588

psi 0.0973609295443122

theta 14.9673044542728

sigma 1.12338425846849

Focusing on the computation time, it is quite easily seen how interesting it can be to do some more analytical work and write more appropriated `getTipDistribution` functions. Still, the defaut function written for the superclass `PhenotypicModel` should always work.

Using the framework to define a new model

We illustrate here how the current code can be used to numerically study a specific model that has not been implemented elsewhere. We focus here on the implementation of the 'GMM' model described in the main text, explaining step by step the following procedure, that is generalizable to any model:

i) we identify what the periods are,

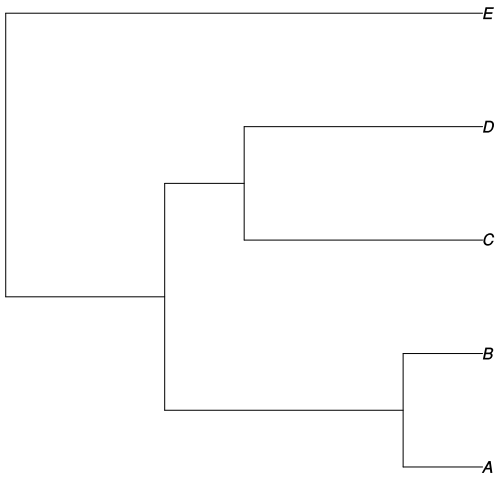
ii) we write the model in a vectorial form on each period,

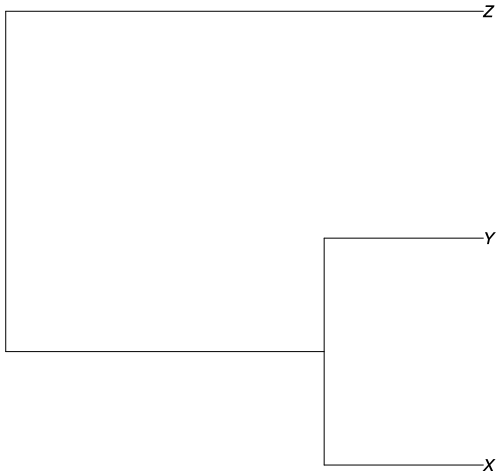
iii) we implement it naively first,

iv) we make analytical developments to speed up the computation time, and subsequently introduce a new class more appropriated to this model.

For simplicity, we implement GMM for two ultrametric trees here. In our example, the two trees will be:

```
In [245]: newick1 <- "(((A:1,B:1):3,(C:3,D:3):1):2,E:6);"
          tree1 <- read.tree(text=newick1)
          plot(tree1)
          newick2 <- "((X:1.5,Y:1.5):3,Z:4.5);"
          tree2 <- read.tree(text=newick2)
          plot(tree2)
```





The first step consists in implementing a function `endOfPeriodsGMM(tree1, tree2)`, which takes as input two trees (the trees corresponding to our two interacting clades), and returns:

— the list of successive branching times (τ_i) (vector `periods`),

— information on which branch gives birth at that time (vector `copy`),

— the number assigned to the newly created branch at that time (vector `paste`),

— the number of lineages in clade 1 and 2 at each time (vectors `nLineages1` and `nLineages2`),

— the label of tips at the end (vector `labeling`).

For example, our function, called on the two preceding trees, returns:

```
In [246]: endOfPeriodsGMM(tree1, tree2)
```

Out[246]:

$periods 0 1.5 2 3 4.5 5 6

$copy 1 3 1 3 5 1 0

$paste 2 4 3 4 7 5 0

$nLineages1 2 2 3 4 4 5 0

$nLineages2 1 2 2 2 3 3 0

$labeling 'A' 'E' 'C' 'D' 'B' 'X' 'Z' 'Y'

The second step now consists in writing the model in the vectorial form required in the framework, during each epoch i. The form of the a, A and Γ matrices is shown in Online Appendix C.3.4, and depends on the number of lineages in the two clades during each epoch.

We introduce the constructor `createModelCoevolution(tree1, tree2)`, which is a function that takes as input two ultrametric trees corresponding to the two clades, and returns an object of class `PhenotypicModel`. It relies on the central function `aAGamma` that defines the collection of (a_i, A_i, Γ_i) during each epoch.

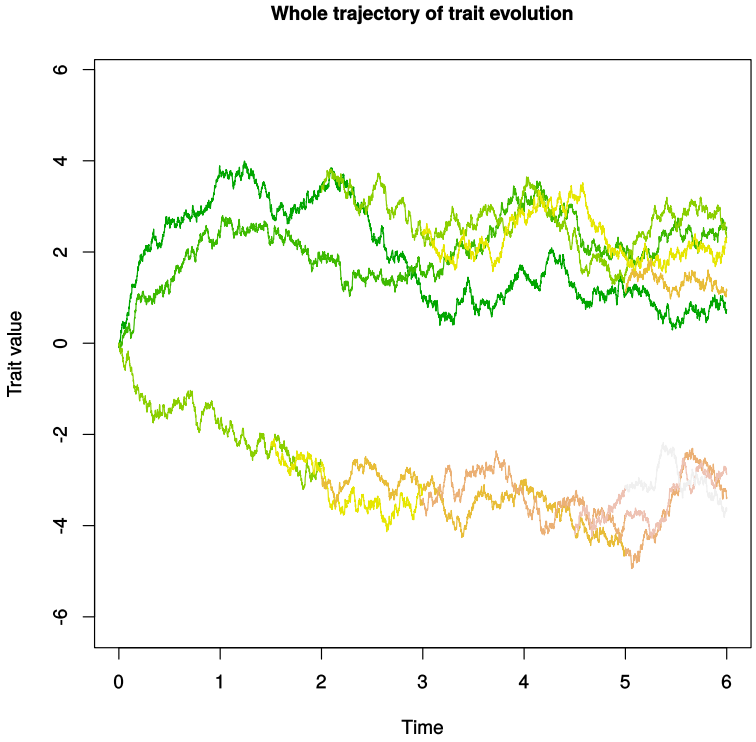
This first version of the GMM implementation allows us to simulate tip data, to get the tip distribution under any parameter set, and to fit tip data.

```
In [248]: modelGMMbis <- createModelCoevolution(tree1, tree2, keyword="GMMbis")
          modelGMMbis
```

```
Out[248]:
*********************************************************************
*** Object of Class PhenotypicModel ***
*** Name of the model : [1] "GMMbis"
*** Parameters of the model : [1] "m0"    "v0"    "d1"    "d2"    "S"     "sigma"
*** Description : Generalist Matching Mutualism model.
Starts with 3 or 4 lineages having the same value X_0 ~ Normal(m0,v0).
One trait in each lineage, all lineages evolving then non-independtly
according to the GMM expression.
*** Periods : the model is cut into  7  parts.
For more details on the model, call : print(PhenotypicModel)
*********************************************************************
```

```
In [249]: dataGMM <- simulateTipData(modelGMMbis, c(0,0,5,-5, 1, 1), method=2)
```

```
*** Simulation of tip trait values ***
Simulates step-by-step the whole trajectory, plots it, and returns tip data.
Computation time : 0.319762 secs
```




```
In [250]: getTipDistribution(modelGMMbis, c(0,0,5,-5,0.5,1))
```

Out[250]:

$mean

A	2.493801
E	2.493801
C	2.493801
D	2.493801
B	2.493801
X	-2.493801
Z	-2.493801
Y	-2.493801

$Sigma

	A	E	C	D	B	X	Z	Y
A	2.196011	1.171214	1.215844	1.215844	1.563892	1.399735	1.341619	1.399735
E	1.171214	2.141458	1.172730	1.172730	1.171214	1.337713	1.279597	1.337713
C	1.215844	1.172730	2.199045	1.248832	1.215844	1.379237	1.321122	1.379237
D	1.215844	1.172730	1.248832	2.199045	1.215844	1.379237	1.321122	1.379237
B	1.563892	1.171214	1.215844	1.215844	2.196011	1.399735	1.341619	1.399735
X	1.399735	1.337713	1.379237	1.379237	1.399735	2.200083	1.190366	1.423215
Z	1.341619	1.279597	1.321122	1.321122	1.341619	1.190366	2.158430	1.190366
Y	1.399735	1.337713	1.379237	1.379237	1.399735	1.423215	1.190366	2.200083

```
In [251]: fitTipData(modelGMMbis, dataGMM, c(0,0,5,-5,1,1))
```

```
*** Fit of tip trait data ***
Finds the maximum likelihood estimators of the parameters,
returns the likelihood and the inferred parameters.
**WARNING** : This function uses the standard R optimizer "optim".
It may not always converge well.
Please double check the convergence by trying
distinct parameter sets for the initialisation.
Computation time : 3.728739 secs
```

Out[251]:

$value 6.61385667009296

$inferredParams *m0* 0.00512480151380221

v0 2.69680996514239e-05

d1 5.03536962882004

d2 -5.83517142115953

S 0.231631941480316

sigma 0.361942471141108

However, this first implementation relies on the `PhenotypicModel` class, which uses the method `getTipDistribution` that solves the ODE system through each epoch, and thus takes time.

The analytical reduction presented in Online Appendix C.3.4 can also be implemented. To this end, we create a new class named `PhenotypicGMM`, associated with an other function `getTipDistribution`. Using these developments allows us to compute more rapidly the tip distribution under the model.

```
In [252]: modelGMM <- createModelCoevolution(tree1, tree2, keyword="GMM")
          modelGMM
```

Out[252]:

```
*********************************************************************
*** Object of Class PhenotypicModel ***
*** Name of the model : [1] "GMM"
*** Parameters of the model : [1] "m0"    "v0"    "d1"    "d2"    "S"     "sigma"
*** Description : Generalist Matching Mutualism model.
Starts with 3 or 4 lineages having the same value X_0 ~ Normal(m0,v0).
One trait in each lineage, all lineages evolving then non-independtly
according to the GMM expression.
*** Periods : the model is cut into  7  parts.
For more details on the model, call : print(PhenotypicModel)
*********************************************************************
```

```
In [253]: getTipDistribution(modelGMM, c(0,0,5,-5,0.5,1), v=TRUE)
          getTipDistribution(modelGMMbis, c(0,0,5,-5,0.5,1), v=TRUE)
```

```
*** Analytical computation of tip traits distribution ***
(Method working for the GMM model only)
Computation time : 0.0008528233 secs
```

Out[253]:

$mean

A	2.493803
E	2.493803
C	2.493803
D	2.493803
B	2.493803
X	-2.493803
Z	-2.493803
Y	-2.493803

$Sigma

	A	E	C	D	B	X	Z	Y
A	2.196010	1.171213	1.215843	1.215843	1.563890	1.399736	1.341620	1.399736
E	1.171213	2.141459	1.172730	1.172730	1.171213	1.337713	1.279597	1.337713
C	1.215843	1.172730	2.199045	1.248832	1.215843	1.379238	1.321122	1.379238
D	1.215843	1.172730	1.248832	2.199045	1.215843	1.379238	1.321122	1.379238
B	1.563890	1.171213	1.215843	1.215843	2.196010	1.399736	1.341620	1.399736
X	1.399736	1.337713	1.379238	1.379238	1.399736	2.200083	1.190366	1.423213
Z	1.341620	1.279597	1.321122	1.321122	1.341620	1.190366	2.158430	1.190366
Y	1.399736	1.337713	1.379238	1.379238	1.399736	1.423213	1.190366	2.200083

```
*** Computation of tip traits distribution through ODE resolution ***
(Method working for any model)
Computation time : 0.01734638 secs
```

Out[253]:

$mean

A	2.493801
E	2.493801
C	2.493801
D	2.493801
B	2.493801
X	-2.493801
Z	-2.493801
Y	-2.493801

$Sigma

	A	E	C	D	B	X	Z	Y
A	2.196011	1.171214	1.215844	1.215844	1.563892	1.399735	1.341619	1.399735
E	1.171214	2.141458	1.172730	1.172730	1.171214	1.337713	1.279597	1.337713
C	1.215844	1.172730	2.199045	1.248832	1.215844	1.379237	1.321122	1.379237
D	1.215844	1.172730	1.248832	2.199045	1.215844	1.379237	1.321122	1.379237
B	1.563892	1.171214	1.215844	1.215844	2.196011	1.399735	1.341619	1.399735
X	1.399735	1.337713	1.379237	1.379237	1.399735	2.200083	1.190366	1.423215
Z	1.341619	1.279597	1.321122	1.321122	1.341619	1.190366	2.158430	1.190366
Y	1.399735	1.337713	1.379237	1.379237	1.399735	1.423215	1.190366	2.200083

Appendix D

Appendix to the relaxed molecular clock with spikes

We provide here a short appendix to the work in progress presented in the last chapter 5 of the thesis, related to the model of molecular evolution with episodes of rapid cladogenetic accumulation of mutations (spikes).

We present first details necessary to fully understand the initialization and the movement proposal of the MCMC presented in the main text, and second two Monte-Carlo approaches to compute the likelihood of an alignment at the tips of the tree.

Contents of the chapter

D.1 Expected number of substitutions on each branch

We explain in this section how to derive:

i) $\mathbb{E}(D_{w_1,w_2})$ (without knowledge), and

ii) $\mathbb{E}(D_{w_1,w_2} \mid \mathcal{A})$ (conditional on the observed alignment).

These two quantities are key to the initialization function g and movement proposal q of the MCMC described in the main text (chapter 5). The function g is also the main ingredient of the Importance Sampling algorithm that we provide hereafter.

D.1.1 Without any knowledge

The first quantity can be obtained by first computing the probability, for one nucleotide, to be in distinct states in two nodes w_1 and w_2 separated by a branch of length t. Under JC69 for example, with a substitution rate α, it leads to :

$$p_{\text{diff}} = \frac{3}{4}\left(1 - e^{-4\alpha t}\right)$$

Because the m nucleotides in the sequence are independent, the number of nucleotides having distinct states in w_1 and w_2 follows a binomial distribution $\mathcal{B}(m, p_{\text{diff}})$. Its expectation is thus $\mathbb{E}(D_{w_1,w_2}) = mp_{\text{diff}}$ and its variance is $\text{Var}(D_{w_1,w_2}) = mp_{\text{diff}}(1 - p_{\text{diff}})$.

D.1.2 Conditional on the observed alignment

The second quantity requires more work, and is obtained following an idea presented by Friedman et al. (2002). We perform the usual *pruning algorithm* but instead of stopping the algorithm once the root of the tree is attained, we add one depth-traversal of the tree to compute at each node a second conditional likelihood, as if we came from the other part of the tree. More precisely, we compute at each node w:

$$L_w = \left(\mathbb{P}((X_f)_{f\in\mathcal{F}(w)} \mid X_w = i)\right)_{i\in\{A,T,G,C\}}$$
$$L_{w\downarrow} = \left(\mathbb{P}((X_f)_{f\in\mathcal{F}(\rho)\setminus\mathcal{F}(w)} \mid X_w = i)\right)_{i\in\{A,T,G,C\}}$$

These two quantities need to be computed at each node of the tree. The probability of the trajectories of the Markov Chain on an internal branch separating w_1 and w_2, of length t and displaying n_S spikes, can then be expressed as:

$$\begin{aligned}
\mathbb{P}\left(X_{w_1} = i, X_{w_2} = j \mid (X_f)_{f\in\mathcal{F}(\rho)}\right) &= \mathbb{P}\left(X_{w_2} = j \mid X_{w_1} = i, (X_f)_{f\in\mathcal{F}(\rho)}\right)\ \mathbb{P}\left(X_{w_1} = i \mid (X_f)_{f\in\mathcal{F}(\rho)}\right) \\
&= \mathbb{P}\left(X_{w_2} = j \mid X_{w_1} = i, (X_f)_{f\in\mathcal{F}(w_2)}\right)\ \mathbb{P}\left(X_{w_1} = i \mid (X_f)_{f\in\mathcal{F}(\rho)}\right) \\
&= \frac{\mathbb{P}\left(X_{w_2} = j, (X_f)_{f\in\mathcal{F}(w_2)} \mid X_{w_1} = i\right)}{\mathbb{P}\left((X_f)_{f\in\mathcal{F}(w_2)} \mid X_{w_1} = i\right)}\ \frac{\mathbb{P}\left(X_{w_1} = i, (X_f)_{f\in\mathcal{F}(\rho)}\right)}{\mathbb{P}\left((X_f)_{f\in\mathcal{F}(\rho)}\right)}
\end{aligned}$$

The first part is simplified as:

$$\begin{aligned}
\frac{\mathbb{P}\left(X_{w_2} = j, (X_f)_{f\in\mathcal{F}(w_2)} \mid X_{w_1} = i\right)}{\mathbb{P}\left((X_f)_{f\in\mathcal{F}(w_2)} \mid X_{w_1} = i\right)} &= \frac{\mathbb{P}\left((X_f)_{f\in\mathcal{F}(w_2)} \mid X_{w_2} = j\right)\mathbb{P}\left(X_{w_2} = j \mid X_{w_1} = i\right)}{\mathbb{P}\left((X_f)_{f\in\mathcal{F}(w_2)} \mid X_{w_1} = i\right)} \\
&= \frac{P(n_S,t)_{ij} L_{w_2}(j)}{\left(P(n_S,t)L_{w_2}\right)(i)}
\end{aligned}$$

While the second part is decomposed as:

$$\frac{\mathbb{P}\left(X_{w_1}=i,(X_f)_{f\in\mathcal{F}(\rho)}\right)}{\mathbb{P}\left((X_f)_{f\in\mathcal{F}(\rho)}\right)} = \frac{\mathbb{P}\left(X_{w_1}=i\right)\mathbb{P}\left((X_f)_{f\in\mathcal{F}(w_1)}\mid X_{w_1}=i\right)\mathbb{P}\left((X_f)_{f\in\mathcal{F}(\rho)\setminus\mathcal{F}(w_1)}\mid X_{w_1}=i\right)}{\mathbb{P}\left((X_f)_{f\in\mathcal{F}(\rho)}\right)}$$
$$= \frac{\pi(i)L_{w_1}(i)L_{w_1\downarrow}(i)}{\pi(L_{w_1}\cdot L_{w_1\downarrow})}$$

Where the symbol $\cdot$ stands for the entrywise (also called Hadamard) product. Note that the denominator is computed as if the tree were rooted at node w_1. We thus finally obtain the following equation:

$$\mathbb{P}\left(X_{w_1}=i,X_{w_2}=j\mid(X_f)_{f\in\mathcal{F}(\rho)}\right) = \frac{P(n_S,t)_{ij}L_{w_2}(j)}{\left(P(n_S,t)L_{w_2}\right)(i)}\,\frac{\pi(i)L_{w_1}(i)L_{w_1\downarrow}(i)}{\pi(L_{w_1}\cdot L_{w_1\downarrow})}$$

This equation can be used to compute the desired expectation on each branch:

$$\mathbb{E}(D_{w_1,w_2}\mid\mathcal{A}) = \sum_{k=1}^{m}\sum_{i\neq j}\mathbb{P}\left(X_{w_1}^{(k)}=i,X_{w_2}^{(k)}=j\mid(X_f^{(k)})_{f\in\mathcal{F}(\rho)}\right)$$

D.2 Likelihood of a present-day alignment

In this appendix section, we provide a Monte-Carlo and an Importance Sampling procedure allowing us to compute the likelihood of a present-day alignment. The idea consists in integrating the likelihood conditional on the spike positions, over the spike positions. Spikes would thus be considered to be hidden random variables, and the optimization could allow to estimate ν,α,κ as parameters of the model.

As we did not come with an efficient optimization procedure yet, this section remains in appendix. We are currently planning to adapt an expectation-maximization algorithm (see section 1.3.2) to optimize the likelihood function.

Note that these two algorithms could also be used in case we need to compute Bayes factors later for model selection purposes. The probability of the alignment would be computed by integrating over spike positions and parameter values (the parameters being considered as random variables in a Bayesian approach).

D.2.1 A Monte-Carlo approach

The likelihood of the observed alignment can be written as:

$$l(\mathcal{A}) = \sum_{\mathcal{S}} l(\mathcal{A}|\mathcal{S})f(\mathcal{S}) = \mathbb{E}_{\mathcal{S}\sim f}\left(l(\mathcal{A}|\mathcal{S})\right)$$

The first idea consists in drawing a sequence of iid realisations $\mathcal{S}_i$ under f, using algorithm 9. For each of these, we are able to compute $l(\mathcal{A}|\mathcal{S}_i)$ using algorithm 12. The likelihood we are looking for is then the mean of these conditional likelihoods on a *large number* of realisations (see discussion on Monte-Carlo integration in section 1.3).

In practice, as soon as the alignment length m rises, we can compute $\ln l(\mathcal{A}|\mathcal{S}_i)$, but we cannot numerically represent their value after exponentiation. We overcome this issue by using the following subterfuge to estimate the likelihood of the alignment:

Algorithm 14 *(Monte-Carlo approach to estimate* $\ln l(\mathcal{A})$*)*

i) Draw n_0 *realisations* $\mathcal{S}_i$ *under law* f*. Only store :*

$$ll_0 := \max_i \ln l(\mathcal{A}|\mathcal{S}_i)$$

ii) Draw n realisations $\mathcal{S}_j$ under law f. Compute :

$$\begin{aligned}\ln\left(\frac{1}{n}\sum_{j=1}^{n} e^{\ln l(\mathcal{A}|\mathcal{S}_j)}\right) &= -\ln n + \ln\left(\sum_{j=1}^{n} e^{ll_0+\ln l(\mathcal{A}|\mathcal{S}_j)-ll_0}\right)\\ &= -\ln n + ll_0 + \ln\left(\sum_{j=1}^{n} e^{\ln l(\mathcal{A}|\mathcal{S}_j)-ll_0}\right)\end{aligned}$$

The first part searches for a value ll_0 ensuring that all other likelihood values will be *on an appropriate scale*. If n_0 is not large enough, the value of ll_0 might not be large enough and there is a possibility to get during the second step a value $e^{\ln l(\mathcal{A}|\mathcal{S}_j)-ll_0} = \infty$ due to numerical limitations. In practice, $n_0 = 50$ was sufficient in all tests that we performed.

D.2.2 An Importance Sampling approach

We seek to describe a law g on the same space as f. We would further need to be able to simulate under g and evaluate it quickly on any realisation $\mathcal{S}$. Finally, we would like g to give an important weight to values $\mathcal{S}$ such that $l(\mathcal{A}|\mathcal{S})$ is large, and to give non-zero weight to any realisation $\mathcal{S}$ such that $f(\mathcal{S}) > 0$.

We actually built the function g described in section 5.3.2 exactly as to satisfy the criteria needed for this Importance Sampling procedure. For simplicity, we then used it in the MCMC algorithm presented in the main text.

Provided we already know this function g, we can compute:

$$\begin{aligned}l(\mathcal{A}) &= \sum_{\mathcal{S}} l(\mathcal{A}|\mathcal{S})\frac{f(\mathcal{S})}{g(\mathcal{S})}g(\mathcal{S})\\ &= \mathbb{E}_{\mathcal{S}\sim g}\left(l(\mathcal{A}|\mathcal{S})\frac{f(\mathcal{S})}{g(\mathcal{S})}\right)\end{aligned}$$

We thus draw iid simulations $\mathcal{S}_i$ under g, and compute the mean, over these realisations, of the quantity written above. If g is sufficiently well tuned, we get a chance to lower the variance of the estimate of $l(\mathcal{S})$, as compared to the previously proposed Monte-Carlo approach. Less iterations and faster convergence would thus be obtained (see discussion on Importance Sampling in section 1.3).

The complete Importance Sampling (IS) procedure is described hereafter :

Algorithm 15 *(IS to estimate $\ln l(\mathcal{A})$)*

i) Compute $\mathbb{E}(D_{w_1,w_2} \mid \mathscr{A} = \mathcal{A})$, use it to derive x and $b(x)$, on each branch of the tree.

ii) Draw n_0 simulations $\mathcal{S}_i$ under law g. Store only the following value :

$$ll_0 := \max_i \ \ln l(\mathcal{A}|\mathcal{S}_i) + \ln f(\mathcal{S}_i) - \ln g(\mathcal{S}_i)$$

iii) Draw n simulations $\mathcal{S}_j$, under law g. Compute :

$$\begin{aligned}\ln\left(\frac{1}{n}\sum_{j=1}^{n} e^{\ln l(\mathcal{A}|\mathcal{S}_j)+\ln f(\mathcal{S}_i)-\ln g(\mathcal{S}_i)}\right) &= -\ln n + \ln\left(\sum_{j=1}^{n} e^{ll_0+\ln l(\mathcal{A}|\mathcal{S}_j)+\ln f(\mathcal{S}_i)-\ln g(\mathcal{S}_i)-ll_0}\right)\\ &= -\ln n + ll_0 + \ln\left(\sum_{j=1}^{n} e^{\ln l(\mathcal{A}|\mathcal{S}_j)+\ln f(\mathcal{S}_i)-\ln g(\mathcal{S}_i)-ll_0}\right)\end{aligned}$$

Appendix E

Appendix Paper : Empirical application of a model of phenotypic evolution including competition among lineages

Building on a model called *Matching Competition* introduced by Nuismer and Harmon (2014), Drury et al. (2016) designed a study aimed at assessing the importance of interspecies competition in driving phenotypic differentiation over long timescales.

The Matching Competition model is an example of a model that fits well within the framework presented in chapter 4. It indeed describes the evolution of a trait which would be, on each lineage, repulsed from the mean trait of all other lineages. We provided guidance in this study to fit the model on a trait dataset of *Anolis* lizards.

The original article is reproduced in the following pages.

Syst. Biol. 65(4):700–710, 2016

DOI:10.1093/sysbio/syw020
Advance Access publication March 9, 2016

Estimating the Effect of Competition on Trait Evolution Using Maximum Likelihood Inference

JONATHAN DRURY*, JULIEN CLAVEL, MARC MANCEAU, AND HÉLÈNE MORLON

Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS, Inserm, Ecole Normale Supérieure, PSL Research University, F-75005 Paris, France
**Correspondence to be sent to: IBENS, 46 rue d'Ulm, F-75005 Paris, France; E-mail: drury@biologie.ens.fr*

Received 31 July 2015; reviews returned 24 February 2016; accepted 1 March 2016
Associate Editor: Luke Harmon

Abstract.—Many classical ecological and evolutionary theoretical frameworks posit that competition between species is an important selective force. For example, in adaptive radiations, resource competition between evolving lineages plays a role in driving phenotypic diversification and exploration of novel ecological space. Nevertheless, current models of trait evolution fit to phylogenies and comparative data sets are not designed to incorporate the effect of competition. The most advanced models in this direction are diversity-dependent models where evolutionary rates depend on lineage diversity. However, these models still treat changes in traits in one branch as independent of the value of traits on other branches, thus ignoring the effect of species similarity on trait evolution. Here, we consider a model where the evolutionary dynamics of traits involved in interspecific interactions are influenced by species similarity in trait values and where we can specify which lineages are in sympatry. We develop a maximum likelihood based approach to fit this model to combined phylogenetic and phenotypic data. Using simulations, we demonstrate that the approach accurately estimates the simulated parameter values across a broad range of parameter space. Additionally, we develop tools for specifying the biogeographic context in which trait evolution occurs. In order to compare models, we also apply these biogeographic methods to specify which lineages interact sympatrically for two diversity-dependent models. Finally, we fit these various models to morphological data from a classical adaptive radiation (Greater Antillean *Anolis* lizards). We show that models that account for competition and geography perform better than other models. The matching competition model is an important new tool for studying the influence of interspecific interactions, in particular competition, on phenotypic evolution. More generally, it constitutes a step toward a better integration of interspecific interactions in many ecological and evolutionary processes. [Adaptive radiation; *Anolis*; community phylogenetics; interspecific competition; maximum likelihood; phylogenetic comparative methods; trait evolution.]

Interactions between species can be strong selective forces. Indeed, many classical evolutionary theories assume that interspecific competition has large impacts on fitness. Character displacement theory (Brown and Wilson 1956; Grant 1972; Pfennig and Pfennig 2009), for example, posits that interactions between species, whether in ecological or social contexts, drive adaptive changes in phenotypes. Similarly, adaptive radiation theory (Schluter 2000) has been a popular focus of investigators interested in explaining the rapid evolution of phenotypic disparity (Grant and Grant 2002; Losos 2009; Mahler et al. 2013; Weir and Mursleen 2013), and competitive interactions between species in a diversifying clade are a fundamental component of adaptive radiations (Schluter 2000; Losos and Ricklefs 2009; Grant and Grant 2011).

Additionally, social interactions between species, whether in reproductive (Gröning and Hochkirch 2008; Pfennig and Pfennig 2009) or agonistic (Grether et al. 2009, 2013) contexts, are important drivers of changes in signal traits used in social interactions. Several evolutionary hypotheses predict that geographical overlap with closely related taxa should drive divergence in traits used to distinguish between conspecifics and heterospecifics (e.g., traits involved in mate recognition; Wallace 1889; Fisher 1930; Dobzhansky 1940; Mayr 1963; Gröning and Hochkirch 2008; Ord and Stamps 2009; Ord et al. 2011). Moreover, biologists interested in speciation have often argued that interspecific competitive interactions are important drivers of divergence between lineages that ultimately leads to reproductive isolation. Reinforcement (Dobzhansky 1937, 1940), for example, is often thought to be an important phase of speciation (Grant 1999; Coyne and Orr 2004; Rundle and Nosil 2005; Pfennig and Pfennig 2009) wherein selection against hybridization leads to a reduction in interspecific mate competition as a result of concomitant divergence in traits involved in mate recognition.

In addition to the importance of interspecific competition in driving phenotypic divergence between species, competitive interactions are also central to many theories of community assembly, which posit that species with similar ecologies exclude each other from the community (Elton 1946). In spite of the importance of interspecific competition to these key ecological and evolutionary theories, the role of competition in driving adaptive divergence and species exclusion from ecological communities has been historically difficult to measure (Losos 2009), because both trait divergence and species exclusion resulting from competition between lineages during their evolutionary history have the effect of eliminating competition between those lineages at the present (i.e., the contemporary distribution of traits hold a signature of the "ghost of competition past," Connell 1980). Community phylogeneticists have aimed to solve part of this conundrum by analyzing the phylogenetic structure of local communities: assuming that phylogenetic similarity between two species is a good proxy for their ecological similarity, competitive interactions are considered to have been more important in shaping communities comprising phylogenetically

(and, therefore, ecologically) distant species (Webb et al. 2002; Cavender-Bares et al. 2009). However, there is an intrinsic contradiction in this reasoning, because using phylogenetic similarity as a proxy for ecological similarity implicitly (or explicitly) assumes that traits evolved under a Brownian model of trait evolution, meaning that species interactions had no effect on trait divergence (Kraft et al. 2007; Cavender-Bares et al. 2009; Mouquet et al. 2012; Pennell and Harmon 2013).

More generally, and despite the preponderance of classical evolutionary processes that assume that interspecific interactions have important fitness consequences, existing phylogenetic models treat trait evolution within a lineage as independent from traits in other lineages. For example, in the commonly used Brownian motion (BM) and Ornstein–Uhlenbeck models of trait evolution (Cavalli-Sforza and Edwards 1967; Felsenstein 1988; Hansen and Martins 1996), once an ancestor splits into two daughter lineages, the trait values in those daughter lineages do not depend on the trait values of sister taxa. Some investigators have indirectly incorporated the influence of interspecific interactions by fitting models where evolutionary rates at a given time depend on the diversity of lineages at that time (e.g., the "diversity-dependent" models of Mahler et al. 2010; Weir and Mursleen 2013). While these models capture some parts of the interspecific processes of central importance to evolutionary theory, such as the influence of ecological opportunity, they do not explicitly account for trait-driven interactions between lineages, as trait values in one lineage do not vary directly as a function of trait values in other evolving lineages.

Recently, Nuismer and Harmon (2015) proposed a model where the evolution of a species' trait depends on other species' traits. In particular, they consider a model, which they refer to as the model of phenotype matching, where the probability that an encounter between two individuals has fitness consequences declines as the phenotypes of the individuals become more dissimilar. The consequence of the encounter on fitness can be either negative if the interaction is competitive, resulting in character divergence (matching competition, e.g., resource competition), or positive if the interaction is mutualistic, resulting in character convergence (matching mutualism, e.g., Müllerian mimicry). Applying Lande's formula (Lande 1976) and given a number of simplifying assumptions—importantly that all lineages evolve in sympatry and that variation in competitors' phenotypes does not strongly influence the outcome of competition—this model yields a simple prediction for the evolution of a population's mean phenotype.

Here, we develop inference tools for fitting a simple version of the matching competition model (i.e., the phenotype matching model of Nuismer and Harmon incorporating competitive interactions between lineages) to combined phylogenetic and trait data. We begin by showing how to compute likelihoods associated with this model. Next, we use simulations to explore the statistical properties of maximum likelihood (ML) estimation of the matching competition model (parameter estimation as well as model identifiability). While the inclusion of interactions between lineages is an important contribution to quantitative models of trait evolution, applying the matching competition model to an entire clade relies on the assumption that all lineages in the clade are sympatric. However, this assumption will be violated in most empirical cases, so we also developed a method for incorporating data on the biogeographical overlap between species for the matching competition model. We also implemented these biogeographical tools for the linear and exponential diversity-dependent trait models of Weir and Mursleen (2013), wherein the evolutionary rate at a given time in a tree varies as a function of the number of lineages in the reconstructed phylogeny at that time (see also Mahler et al. 2010), so that rates vary only as a function of the number of sympatric lineages.

We then fit the model to data from a classical adaptive radiation: Greater Antillean *Anolis* lizards (Harmon et al. 2003; Losos 2009). Many lines of evidence support the hypothesis that resource competition is responsible for generating divergence between species in both habitat use (e.g., Pacala and Roughgarden 1982) and morphology (Schoener 1970; Williams 1972; see review in Losos 1994). Thus, we can make an *a priori* prediction that model comparison will uncover a signature of competition in morphological traits that vary with habitat and resource use. Given the well-resolved molecular phylogeny (Mahler et al. 2010, 2013) and the relatively simple geographical relationships between species (i.e., many species are restricted to single islands, Rabosky and Glor 2010; Mahler and Ingram 2014), the Greater Antillean *Anolis* lizards provide a good test system for exploring the effect of competition on trait evolution using the matching competition model.

Methods

Likelihood Estimation of the Matching Competition Model

We consider the evolution of a quantitative trait under the matching competition model of Nuismer and Harmon (2015) wherein trait divergence between lineages will be favored by selection. We make the assumption that the outcome of competitive interactions is similar between all members of an evolving clade rather than sensitive to pairwise phenotypic similarity (i.e., that α in equation 1 of Nuismer and Harmon 2015 is small). This assumption is crucial, as it ensures that the evolution of a population's mean phenotype is given by a linear model (equation S38 in Nuismer and Harmon 2015). Importantly, this implies that the expected distribution of trait values on a given phylogeny follows a multivariate normal distribution (Manceau et al., manuscript in preparation), as is the case for classical models of quantitative trait evolution (Hansen and Martin 1996; Harmon et al. 2010; Weir and

Mursleen 2013). In our current treatment of the model, we remove stabilizing selection to focus on the effect of competition (see "Discussion" section). Under these two simplifying assumptions, the mean trait value for lineage i after an infinitesimally small time step dt is given by (equation S38 in Nuismer and Harmon 2015 with $\psi=0$):

$$z_i(t+dt)=z_i(t)+S(\mu(t)-z_i(t))dt+\sigma dB_i \quad (1)$$

where $z_i(t)$ is the mean trait value for lineage i at time t, $\mu(t)$ is the mean trait value for the entire clade at time t, S measures the strength of interaction (more intense competitive interactions are represented by larger negative values), and drift is incorporated as BM σdB_i with mean $=0$ and variance $=\sigma^2 dt$. Note that when $S=0$ or $n=1$ (i.e., when a species is alone), this model reduces to BM. Under the model specified by equation (1), if a species trait value is greater (or smaller) than the trait value average across species in the clade, the species' trait will evolve toward even larger (or smaller) trait values. Since trait values of extinct lineages were likely similar to trait values of lineages surviving to the present, we assume that the mean clade value (and thus, the outcome of competitive interactions) is not greatly influenced by extinction. We directly assess the impact of extinction on parameter estimation below and discuss the strengths and limitations of this formulation of the matching competition model in the "Discussion" section.

Given that the expected distribution of trait values on a phylogeny under the matching competition model specified in equation (1) follows a multivariate normal distribution, it is entirely described with its expected mean vector (made of terms each equal to the character value at the root of the tree) and variance–covariance matrix. Nuismer and Harmon (2015) provide the system of ordinary differential equations describing the evolution of the variance and covariance terms through time (their equations 10b and 10c). These differential equations can be integrated numerically from the root to the tips of phylogenies to compute expected variance–covariance matrices for a given set of parameter values and the associated likelihood values given by the multivariate normal distribution.

Additionally, to relax the assumption that all of the lineages in a clade coexist sympatrically, we included a term to specify which lineages co-occur at any given time-point in the phylogeny, which can be inferred, for example, by biogeographical reconstruction. We define piecewise constant coexistence matrices $\mathbf{A}$, where $\mathbf{A}_{i,j}$ equals 1 at time t if i and j are sympatric at that time, and 0 otherwise (Fig. 1). The evolution of the trait value for lineage i is then given by:

$$z_i(t+dt)=z_i(t)+S\left(\left(\frac{1}{n_i}\sum_{l=1}^{n}\mathbf{A}_{i,l}z_l(t)\right)-z_i(t)\right)dt+\sigma dB_i \quad (2)$$

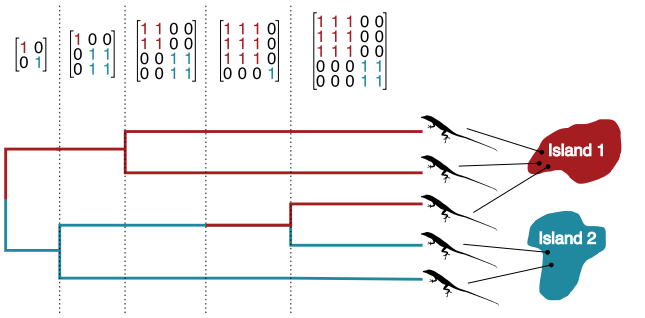


FIGURE 1. Illustration of geography matrices (defined for each lineage at every node and after each dispersal event inferred, e.g., by stochastic mapping) delineating which lineages interact in sympatry in an imagined phylogeny. These matrices were used to identify potentially interacting lineages for the matching competition and both diversity-dependent models of character evolution (see equations (3–5) in the main text). *Anolis* outline from http://phylopic.org courtesy of Sarah Werning, licensed under Creative Commons. (http://creativecommons.org/licenses/by/3.0/).

where $n_i=\sum_{j=1}^{n}\mathbf{A}_{ij}$ is the number of lineages interacting with lineage i at time t (equal to the number n of extant lineages in the reconstructed phylogeny at time t if all species are sympatric) such that trait evolution is only influenced by sympatric taxa. When a species is alone, $\mathbf{A}_{i,i}=1$, all other $\mathbf{A}_{i,j}=0$, $n_{i,i}=1$, and thus equation (2) reduces to the Brownian model.

We show (Appendix S1 in the Supplementary Material available on Dryad at http://dx.doi.org/10.5061/dryad.d670p) that the corresponding system of ordinary differential equations describing the evolution of the variance and covariance terms through time is:

$$\frac{dv_{i,i}}{dt}=-\frac{2S(n_i-1)}{n_i}v_{i,i}+\frac{2S}{n_i}\left(\sum_{\substack{l=1\\(l\neq i)}}^{n}\mathbf{A}_{i,l}v_{l,i}\right)+\sigma^2 \quad (3a)$$

$$\frac{dv_{i,j}}{dt}=-S\left(\frac{n_i-1}{n_i}+\frac{n_j-1}{n_j}\right)v_{i,j}+\frac{S}{n_i}\sum_{\substack{k=1\\k\neq i}}^{n}\mathbf{A}_{i,k}v_{k,j}+\frac{S}{n_j}\sum_{\substack{l=1\\l\neq j}}^{n}\mathbf{A}_{j,l}v_{l,i} \quad (3b)$$

where $v_{i,i}$ is the variance for each species i at time t and $v_{i,j}$ is the covariance for each species pair i, j at time t. Using numerical integration, we solve this system of ordinary differential equations from the root of the tree to the tips in order to calculate the values of the variance–covariance matrix expected under the model for a given phylogeny and set of parameter values. Specifically, equations (3a) and (3b) dictate how the variance and covariance values change through time along the branches of the tree; at a given branching event, the variance and covariance values associated with the two daughter species are simply inherited from those of the ancestral species. With the expected variance–covariance matrix at present, we calculate the likelihood for the model using the likelihood function

for a multivariate normal distribution (e.g., Harmon et al. 2010). Then, using standard optimization algorithms, we identify the ML values for the model parameters. The matching competition model has three free parameters: σ^2, S,and the ancestral state z_0 at the root. As with other models of trait evolution, the ML estimate for the ancestral state is computed through GLS using the estimated variance–covariance matrix (Grafen 1989; Martins and Hansen 1997).

We used the ode function in the R package deSolve (Soetaert et al. 2010) to perform the numerical integration of the differential equations using the "lsoda" solver, and the Nelder–Mead algorithm implemented in the optim function to perform the ML optimization. Codes for these analyses are freely available on GitHub (https://github.com/hmorlon/PANDA) and included in the R package RPANDA (Morlon et al. 2016).

Incorporating Geography into Diversity-Dependent Models

Using the same geography matrix **A** described above for the matching competition model (Fig. 1), we modified the diversity-dependent linear and exponential models of Weir and Mursleen (2013) to incorporate biological realism into the models, because ecological opportunity is only relevant within rather than between biogeographical regions. The resulting variance–covariance matrices, **V**, of these models have the elements:

$$\mathbf{V}_{ij} = \sum_{m=2}^{M} (\sigma_0^2 + bn_i)\left(\max\left(s_{ij} - t_{m-1}, 0\right) - \max\left(s_{ij} - t_m, 0\right)\right) \quad (4)$$

for the diversity-dependent linear model, and

$$\mathbf{V}_{ij} = \sum_{m=2}^{M} (\sigma_0^2 \times e^{rn_i})\left(\max\left(s_{ij} - t_{m-1}, 0\right) - \max\left(s_{ij} - t_m, 0\right)\right) \quad (5)$$

for the diversity-dependent exponential model, where σ_0^2 is the rate parameter at the root of the tree, b and r are the slopes in the linear and exponential models, respectively, s_{ij} is the shared path length of lineages i and j from the root of the phylogeny to their common ancestor, n_i is the number of sympatric lineages (as above) between times t_{m-1} and t_m (where t_1 is 0, the time at the root, M rep resents the time at the tips, and thus t_M is the total length of the tree) (Weir and Mursleen 2013). When b or $r = 0$, these models reduce to BM. For the linear version of the model, we constrained the ML search such that the term $(\sigma_0^2 + bn_i)$ in equation (3) $\geqslant 0$ to prevent the model from having negative evolutionary rates at any t_m. Since the right-hand parts of equations (4) and (5) become 0 after lineages i and j split, the covariance of lineages i and j is simply the variance accumulated during the time between the root of the tree and their most recent common ancestor.

Simulation-based Analysis of Statistical Properties of the Matching Competition Model

To verify that the matching competition model can be reliably fit to empirical data, we simulated trait data sets to estimate its statistical properties (i.e., parameter estimation and identifiability using AICc). For all simulations, we began by first generating 100 pure-birth trees using TreeSim (Stadler 2014). To determine the influence of the number of tips in a tree, we ran simulations on trees of size $n = 20$, 50, 100, and 150. We then simulated continuous trait data sets by applying the matching competition model recursively from the root to the tip of each tree (Paradis 2012), following equation (1), assuming that all lineages evolved in sympatry. For these simulations, we set $\sigma^2 = 0.05$ and systematically varied S (-1.5, -1, -0.5, -0.1, or 0). Finally, we fit the matching competition model to these data sets using the ML optimization described above.

To determine the ability of the approach to accurately estimate simulated parameter values, we first compared estimated parameters to the known parameters used to simulate data sets under the matching competition model (S and σ^2). We also quantified the robustness of these estimates in the presence of extinction by estimating parameters for data sets simulated on birth–death trees; in addition, we compared the robustness of the matching competition model to extinction to that of the diversity-dependent models. These two latter sets of analyses are described in detail in the Supplementary Appendix 2 available on Dryad.

To assess the ability to correctly identify the matching competition model when it is the generating model, we compared the fit (measured by AICc, Burnham and Anderson 2002) of this model to other commonly used trait models on the same data (i.e., data simulated under the matching competition model). Specifically, we compared the matching competition model to (i) BM, (ii) Ornstein–Uhlenbeck/single-stationary peak model (OU; Hansen and Martin 1996), (iii) exponential time-dependent (TD_{exp}, i.e., the early burst model, or the ACDC model with the rate parameter set to be negative, Blomberg et al. 2003; Harmon et al. 2010), (iv) linear time-dependent evolutionary rate (TD_{in}, Weir and Mursleen 2013), (v) linear rate diversity-dependent (DD_{lin}, Mahler et al. 2010; Weir and Mursleen 2013), and (vi) exponential rate diversity-dependent (DD_{exp}, Weir and Mursleen 2013). These models were fitted using Geiger (Harmon et al. 2008) when available there (BM, OU, TD_{exp}, TD_{in}), or using our own codes, available in RPANDA (Morlon et al. 2016) when they were not available in Geiger (DD_{lin}, DD_{exp}). With the exception of TD_{exp}, which we restricted to have decreasing rates through time since the accelerating rates version of the model is unidentifiable from OU (Uyeda et al. 2015), we did not restrict the ML search for the parameters in TD_{in} or DD models.

We assessed the identifiability of other trait models against the matching competition model by calculating the fit of this model to data sets simulated under the same trait models mentioned above. For BM and OU

models, we generated data sets from simulations using parameter values from the appendix of Harmon et al. (2010) scaled to a tree of length 400 (BM, $\sigma^2 = 0.03$; OU, $\sigma^2 = 0.3$, $\alpha = 0.06$). For both the linear and exponential versions of the time- and diversity-dependent models, we simulated data sets with starting rates of $\sigma^2 = 0.6$ and ending rates of $\sigma^2 = 0.01$, declining with a slope determined by the model and tree (e.g., for time-dependent models, the slope is a function of the total height of the tree; for the TD_{exp} model, these parameters result in a total of 5.9 half-lives elapsing from the root to the tip of the tree, Slater and Pennell 2014). In another set of simulations, we fixed the tree size at 100 tips and varied parameter values to determine the effect of parameter values on identifiability (see "Results" section). As above, we calculated the AICc for all models for each simulated data set.

Finally, to understand how removing stabilizing selection from the likelihood of the matching competition model affects our inference in the presence of stabilizing selection, we simulated data sets with both matching competition and stabilizing selection on 100 tip trees, across a range of parameter space ($S = -1$, -0.5, and 0, $\alpha = 0.05$, 0.5, and 5, holding σ^2 at 0.05). We fit BM, OU, and matching competition models to these simulated data sets. All simulations were performed using our own codes, available in RPANDA (Morlon et al. 2016).

Fitting the Matching Competition Model of Trait Evolution to Caribbean Anolis Lizards

To determine whether the matching competition model is favored over models that ignore interspecific interactions in an empirical system where competition likely influenced character evolution, we fit the matching competition model to a morphological data set of adult males from 100 species of Greater Antillean *Anolis* lizards and the time calibrated, maximum clade credibility tree calculated from a Bayesian sample of molecular phylogenies (Mahler et al. 2010, 2013; Mahler and Ingram 2014). We included the first four size-corrected phylogenetic principal components from a set of 11 morphological measurements, collectively accounting for 93% of the cumulative variance explained (see details in Mahler et al. 2013). Each of these axes is readily interpretable as a suite of specific morphological characters (see "Discussion" section), and together, the shape axes quantified by these principal components describe the morphological variation associated with differences between classical ecomorphs in Caribbean anoles (Williams 1972). In addition to the matching competition model, we fit the six previously mentioned models (BM, OU, TD_{exp}, TD_{lin}, DD_{exp}, and DD_{lin}) separately to each phylogenetic PC axis in the *Anolis* data set.

For the matching competition model and diversity-dependent models, to determine the influence of uncertainty in designating clades as sympatric and allopatric, we fit the model for each trait using 101 sets of geography matrices (i.e., **A** in equations (1), (2), and (3), see Fig. 1): one where all lineages were set as sympatric, and the remaining 100 with biogeographical reconstructions from the output of the make.simmap function in phytools (Revell 2012). To simplify the ML optimization, we restricted S to take negative values while fitting the matching competition model including the biogeographical relationships among taxa (i.e., we forced the optimization algorithm to only propose S values $\leqslant 0$).

RESULT

Statistical Properties of the Matching Competition Model

Across a range of S values, ML optimization returns reliable estimates of parameter values for the matching competition model (Fig. 2). As the number of tips increases, so does the reliability of ML parameter values (Fig. 2). Parameter estimates remain reliable in the presence of extinction, unless the extinction fraction is very large (i.e., $\geqslant 0.6$; Supplementary Appendix 2 available on Dryad). When data sets are simulated under the matching competition model, model selection using AICc generally picks the matching competition model as the best model (Fig. 3, Supplementary Fig. S1 available on Dryad); the strength of this discrimination depends on both the S value used to simulate the data and the size of the tree (Fig. 3, Supplementary Fig. S1 available on Dryad). For example, when $S = -0.1$, the matching competition model often has a higher AICc than BM, largely due to the fact that the BM model has one less parameter.

Simulating data sets under BM, OU, DD_{exp}, and DD_{lin} generating models, we found that in most scenarios, and in most parameter space, these models are distinguishable from the matching competition model (Fig. 4a,b,e,f, Supplementary Fig. S2 available on Dryad). As with the matching competition model, the ability to distinguish between models using AICc generally increases with increasing tree sizes (Fig. 4) and with increasing magnitude of parameter values (Supplementary Fig. S2 available on Dryad). When character data were simulated under a TD_{lin} model of evolution, the matching competition and/or the diversity-dependent models tended to have lower AICc values than the TD_{lin} model, especially among smaller trees (Fig. 4d). For data generated under a TD_{exp} model, model selection always favored the matching competition model over the TD_{exp} model (Fig. 4c).

Though the current implementation of the ML tools for the matching competition do not incorporate stabilizing selection, simulating data sets with both matching competition and stabilizing selection reveals that as the strength of stabilizing selection increases relative to the strength of competition (i.e., α as increases relative to S), AICc model selection shifts from favoring

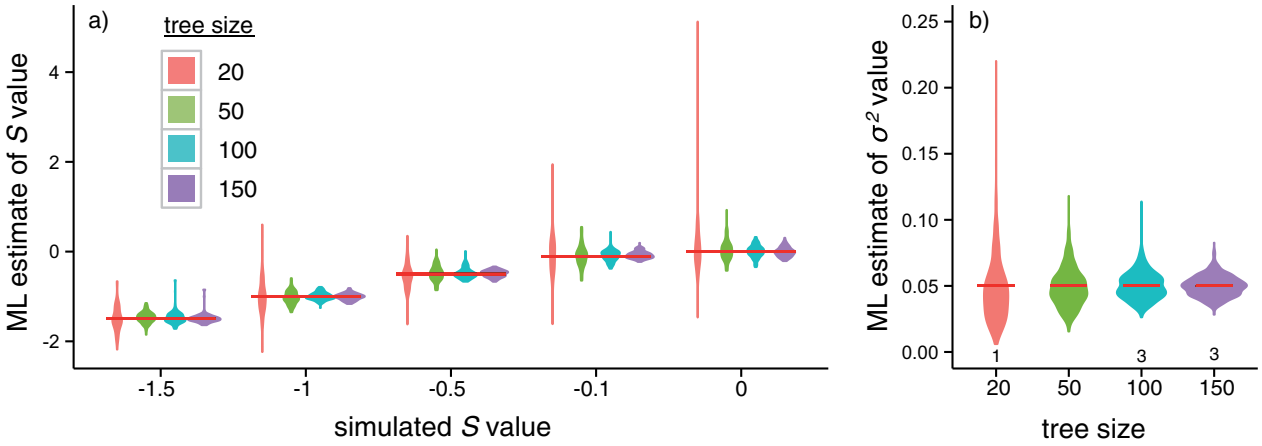


FIGURE 2. Parameter estimation under the matching competition model. As tree size increases and/or the magnitude of competition increases (i.e., the S parameter in the matching competition model becomes more negative), so does the accuracy of ML parameter estimates of (a) S ($n = 100$ for each tree size and S value combination; red horizontal lines indicate the simulated Svalue) and (b) σ^2 ($n = 500$ for each tree size; red horizontal lines indicate the simulated value). In a small number of cases (7/2000), the ML estimate for σ^2 was unusually large (> 0.25), and we removed these rare cases for plotting. The numbers below the violin plots in (b) show the number of outliers removed for each tree size.

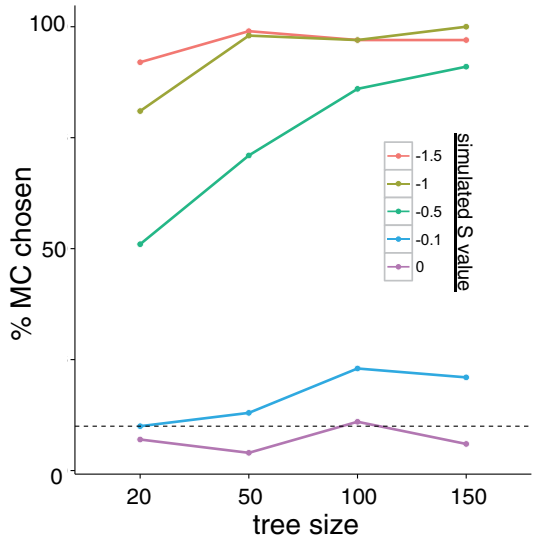


FIGURE 3. AICc support for data sets simulated under the matching competition (MC) model increases with tree size and with increasing levels of competition (i.e., increasingly negative S values). The dotted line denotes 10%.

the matching competition model (under large S, small α scenarios) to favoring the OU model (under small S, large α scenarios) (Supplementary Fig. S3 available on Dryad). Likewise, ML increasingly underestimates the value of S as the value of α increases (Supplementary Fig. S4 available on Dryad).

Competition in Greater Antillean Anolis *Lizards*

For the first four phylogenetic principal components describing variation in *Anolis* morphology, we found that models that incorporate species interactions fit the data better than models that ignore them (Table 1). PC1, which describes variation in hindlimb/hindtoe length (Mahler et al. 2013), is fit best by the matching competition model. PC2, which describes variation in body size (snout vent length) is fit best by the linear diversity-dependent model. PC3, which describes variation in forelimb/foretoe length, and PC4, which describes variation in lamellae number are fit with mixed support across the models included, but with models incorporating species interactions providing the best overall fits.

Additionally, for every PC axis, the best-fit models were ones that incorporated the geographic relationships among species in the tree, and these conclusions were robust to uncertainty in ancestral reconstructions of sympatry (Table 1).

DISCUSSION

The inference methods we present here represent an important new addition to the comparative trait analysis toolkit. Whereas previous models had not accounted for the influence of trait values in other lineages on character evolution, the matching competition model takes these into account. Furthermore, extending both the matching competition model and two diversity-dependent trait evolution models to incorporate geographic networks of sympatry further extends the utility and biological realism of these models.

We found that the matching competition model has increasing AICc support and accuracy of parameter estimation with increasing tree sizes and competition strength. We also found that, for most of the generating models we tested, AICc-based model selection does not tend to erroneously select the matching competition model (i.e., these models are identifiable from the

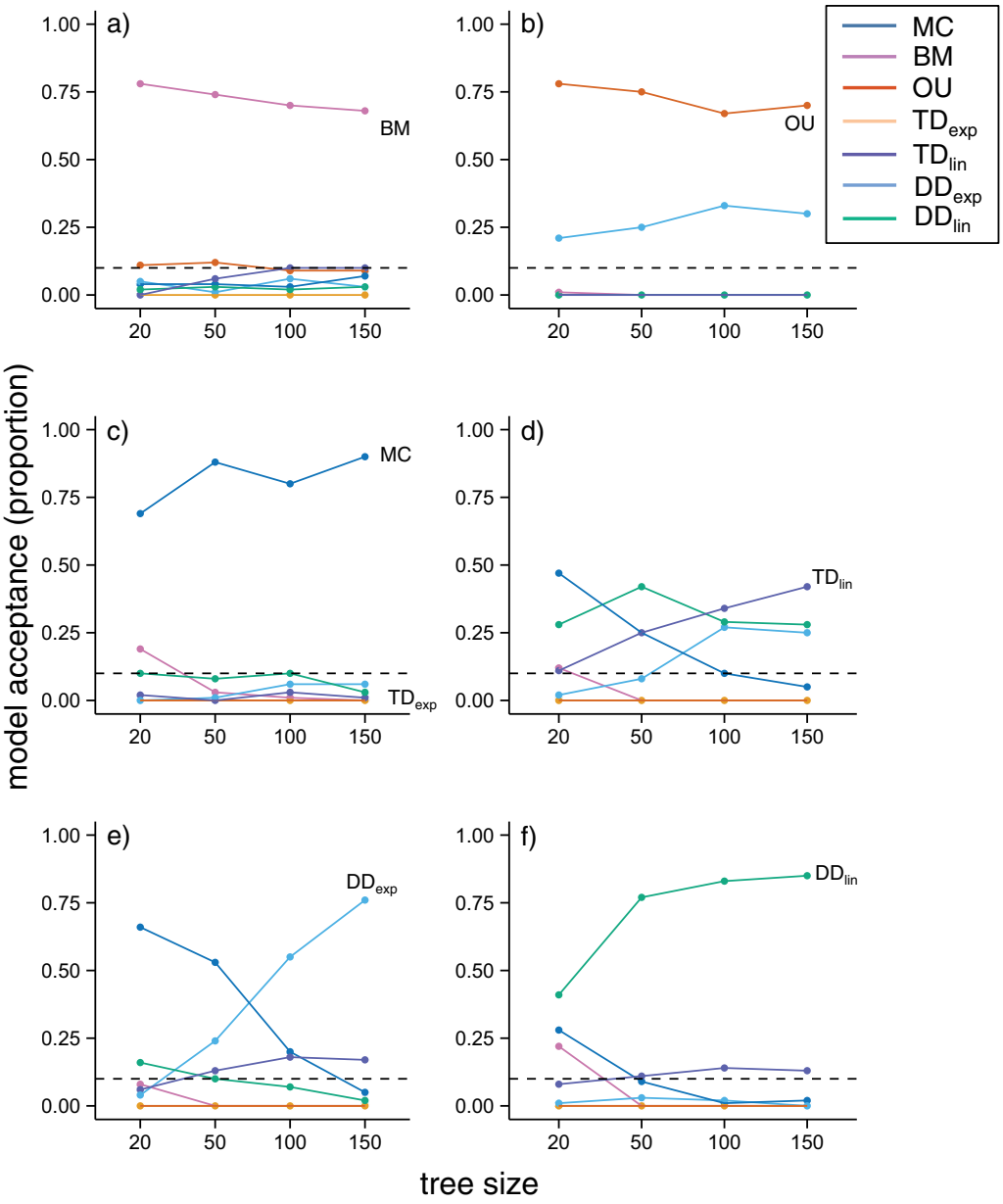


FIGURE 4. Identifiability simulation results for the matching competition (MC) model. When the generating model is either (a) BM, (b) OU, (e) DD_{exp} (for larger trees) or (f) DD_{lin}, the generating model is largely favored by model selection. However, both (c) TD_{exp} and (d) TD_{in} (for smaller trees) are erroneously rejected as the generating model. The dotted lines denote 10%.

matching competition model). As with all other models, the statistical properties of the matching competition model will depend on the size and shape of a particular phylogeny as well as specific model parameter values. Future investigators can employ other approaches, such as phylogenetic Monte Carlo and posterior predictive simulations directly on their empirical trees (Boettiger et al. 2012; Slater and Pennell 2014), to assess the confidence they can have in their results.

We did, however, find that data generated under time-dependent models were often fit better by models that incorporate interspecific interactions (i.e., density-dependent and matching competition models) (Fig. 4c,d). This was especially true for the TD_{exp} model, often referred to as the early-burst model—the matching competition model nearly always fit data generated under the TD_{exp} model better than the TD_{exp} model (Fig. 4c). We do not view this as a major limitation of the model for two reasons. First, the TD_{exp} model is known to be statistically difficult to estimate on neontological data alone (Harmon et al. 2010; Slater et al. 2012a; Slater and Pennell 2014). Second, and more importantly, time-dependent models are not process-based models, but rather incorporate time since the root of a tree as a proxy for ecological opportunity or available niche space (Harmon et al. 2010; Mahler et al. 2010). The matching

TABLE 1. Comparison of model fits for the first four phylogenetic PC axes of a morphological data set of Greater Antillean anoles

Trait	Model	k	σ^2	b	r	S	$\ln(\mathcal{L})$	ΔAICc	Akaike weights
pPC1	BM	2	0.0033	—	—	—	−13.68	21.36 (0.19)	<0.01
	OU	3	0.0033	—	—	—	−13.68	—	—
	TDexp	3	0.0324	—	−0.068	—	−5.20	6.51 (0.19)	0.03
	TDlinear	3	0.0113	−0.019	—	—	−4.88	5.89 (0.19)	0.04
	DDexp	3	0.0184	—	−0.028	—	−4.37	4.87 (0.19)	0.06 (0.004)
	DDexp + GEO	3	0.0087 (1.48E-5)	—	−0.043 (7.29E-5)	—	−8.00	12.05 (0.19)	<0.01
	DDlin	3	0.0089	−0.00008	—	—	−4.89	5.91 (0.19)	0.04
	DDlin + GEO	3	0.0060 (6.25E-6)	−0.00011 (1.24E-7)	—	—	−8.23	12.49 (0.18)	<0.01
	MC_{sym}	3	0.0010	—	—	−0.037	−3.67	2.96 (0.19)	0.16 (0.01)
	MC + GEO		0.0010 (6.75E-6	—	—	**−0.038 (0.00017)**	**−1.94(0.10)**	**0 (0.031)**	**0.68 (0.02)**
pPC2	BM	2	0.0027	—	—	—	−4.69	9.64 (0.036)	0.01
	OU	3	0.0027	—	—	—	−4.69	—	—
	TDexp	3	0.0046	—	−0.014	—	−4.30	10.99 (0.036)	<0.01
	TDlinear	3	0.0047	−0.011	—	—	−4.23	10.85 (0.036)	<0.01
	DDexp	3	0.0041	—	−0.006	—	−4.27	10.94 (0.036)	<0.01
	DDexp + GEO	3	0.0068 (1.36E-5)	—	−0.039 (9.72E-5)		0.51 (0.024)	1.37 (0.015)	0.33 (0.002)
	DDlin	3	0.0042	−0.00002	—	—	−4.21	10.82 (0.036)	<0.01
	DDlin + GEO	**3**	**0.0054 (4.24E-6)**	**−0.00010 (9.17E-8)**	**—**	**—**	**1.20 (0.017)**	**0 (0)**	**0.64 (0.001)**
	MC_{sym}	3	0.0021	—	—	−9.9e-3	−3.95	9.79 (0.036)	<0.01
	MC + GEO	3	0.0018 (2.44E-6)	—	—	−0.015 (4.67E-5)	−2.94	8.30 (0.047)	0.01
pPC3	BM	2	0.0010	—	—	—	45.57	2.56 (0.021)	0.09 (0.0003)
	OU	3	0.0010	—	—	—	45.57	—	—
	TDexp	3	0.0020	—	−0.019	—	46.30	3.22 (0.021)	0.06 (0.0002)
	TDlinear	3	0.0019	−0.013	—	—	46.41	3.02 (0.021)	0.07 (0.0003)
	DDexp	3	0.0017	—	−0.008	—	46.40	3.02 (0.021)	0.07 (0.0003)
	DDexp + GEO	3	0.0015 (1.48E-6)	—	−0.017 (3.72E-5)	—	46.79 (0.006)	2.24 (0.024)	0.10 (0.0006)
	DDlin	3	0.0017	−0.000009	—	—	46.46	2.90 (0.021)	0.08 (0.0003)
	DDlin + GEO	3	0.0014 (8.76E-7)	−0.000016 (2.96E-8)	—	—	46.68 (0.005)	2.44 (0.023)	0.09 (0.0005)
	MC_{sym}	3	0.0007	—	—	−0.012	46.75	2.33 (0.021)	0.10 (0.0004)
	MC + GEO	**3**	0.0006 (6.15E-7)	**—**	**—**	**−0.017 (3.38E-5)**	**47.91 (0.011)**	**0 (0)**	**0.32 (0.002)**
pPC4	BM	2	0.0006	—	—	—	69.07	2.50 (0.016)	0.06 (0.0002)
	OU	3	0.0006	—	—	—	69.07	—	—
	TDexp	3	0.0015	—	−0.025	—	70.55	1.66 (0.016)	0.09 (0.0003)
	TDlinear	3	0.0012	−0.013	—	—	70.45	1.86 (0.016)	0.08 (0.0003)
	DDexp	3	0.0012	—	−0.010	—	70.52	1.73 (0.016)	0.08 (0.0003)
	DDexp + GEO	3	0.0010 (1.18E-6)	—	−0.020 (4.38E-5)	—	71.28 (0.011)	0.13 (0.020)	0.18 (0.001)
	DDlin	3	0.0011	−0.000006	—	—	70.39	1.99 (0.016)	0.07 (0.0002)
	DDlin + GEO	3	0.0009 (5.77E-7)	−0.000009 (1.98E-8)	—	—	70.78 (0.008)	1.12 (0.016)	0.11 (0.0006)
	MC	3	0.0004	—	—	−0.015	71.1	0.57 (0.016)	0.15 (0.0005)
	MC + GEO	**3**	0.0004 (4.21E-7)	**—**	**—**	**−0.016 (3.56E-5)**	**71.34 (0.009)**	**0 (0.012)**	**0.19 (0.001)**

Notes: Models run incorporating geography matrices are indicated by "+ GEO," and models with the lowest AICc for each trait are shaded and written in bold text. Parameter values presented follow the nomenclature of equations (2–4) in the main text, and k represents the number of parameters estimated for each model. Note that TD_{exp} is the ACDC model (or the early-burst model when $r < 0$). OU model weights were excluded because the ML estimates of α equaled 0 for all PC axes, and thus the OU model was equivalent to BM. Median (standard error) of parameter estimates, ΔAICc values, and Akaike weights are presented for fits across 100 sampled stochastic maps of *Anolis* biogeography (standard errors are omitted for Akaike weights < 0.05).

competition and density-dependent models explicitly account for the interspecific competitive interactions that time-dependent models purport to model, thus we argue that these process-based models are more biologically meaningful than time-dependent models (Moen and Morlon 2014).

We did not incorporate stabilizing selection in our model. Preliminary analyses suggested that S and α are not identifiable (though their sum may be), as competition and stabilizing selection operate in opposite directions. As a result, when trait data are simulated with simultaneous stabilizing selection and matching competition, the strength of competition is underestimated. In addition, which model is chosen by model selection depends on the ratio of the strength of attraction toward an optimum to the strength of competition, with Brownian model being selected at equal strengths (Supplementary Figs. S3, S4 available on Dryad). Given that many traits involved in competitive interactions are also likely to have been subject to stabilizing selection (i.e., extreme trait values eventually become targeted by negative selection), statistical inference under the matching competition model without stabilizing selection is likely to underestimate the true effect of competition on trait evolution. Future work aimed at directly incorporating stabilizing selection in the inference tool could provide a more accurate quantification of the effect of competition, although dealing with the nonidentifiability issue may require incorporating additional data such as fossils.

Because the matching competition model depends on the mean trait values in an evolving clade, ML estimation is robust to extinction, whereas the diversity-dependent models are less so (Supplementary Appendix

S2, Supplementary Figs. S5–S8 available on Dryad). Nevertheless, given the failure of ML to recover accurate parameter estimates of the matching competition model at high levels of extinction ($\mu : \lambda \geqslant 0.6$), we suggest that these models should not be used in clades where the extinction rate is known to be particularly high. In such cases, it would be preferable to modify the inference framework presented here to include data from fossil lineages (Slater et al. 2012a) by adapting the ordinary differential equations described in equations (3a) and (3b) for nonultrametric trees.

For all of the traits we analyzed in the Greater Antillean *Anolis* lizards, we found that models incorporating both the influence of other lineages and the specific geographical relationships among lineages were the most strongly supported models (though less strikingly for PC3 and PC4). Incorporating uncertainty in biogeographical reconstruction, which we encourage future investigators to do in general, demonstrated that these conclusions were robust to variation in the designation of allopatry and sympatry throughout the clade. We note that while stochastic mapping is reasonable for a group like Greater Antillean *Anolis* lizards, where species are found on single islands, more sophisticated biogeographical models should be used in most other cases (e.g., Ronquist and Sanmartín 2011; Landis et al. 2013; Matzke 2014). The matching competition model is favored in the phylogenetic principal component axis describing variation in relative hindlimb size. Previous research demonstrates that limb morphology explains between ecomorph variation in locomotive capabilities and perch characteristics (Losos 1990, 2009; Irschick et al. 1997), and our results suggest that the evolutionary dynamics of these traits have been influenced by the evolution of limb morphology in other sympatric lineages. These results support the assumption that interspecific interactions resulting from similarity in trait values are important components of adaptive radiations (Losos 1994; Schluter 2000), a prediction that has been historically difficult to test (Losos 2009, but see Mahler et al. 2010). In combination with previous research demonstrating a set of convergent adaptive peaks in morphospace to which lineages are attracted (Mahler et al. 2013), our results suggest that competition likely played an important role in driving lineages toward these distinct peaks. Because we expect the presence of selection toward optima to lead to underestimation of the S parameter in the matching competition model (Supplementary Figs. S3, S4 available on Dryad), we would have likely detected an even stronger effect of competition in the *Anolis* data set if we had included stabilizing selection. Recently, Uyeda and colleagues (2015) demonstrated that the use of principal components can bias inferences of trait evolution. We used BM-based phylogenetic PC axes here, which should reduce this potential bias (Revell 2009). We recognize that there is some circularity in assuming BM in order to compute phylogenetic PC axes before fitting other trait models to these axes; a general solution to address this circularity problem remains to be found (Uyeda et al. 2015). Uyeda and colleagues suggested that using phylogenetic PC axes sorts the traits according to specific models. In the Greater Antillean *Anolis* lizards, the first axes are easily interpretable as specific suites of traits relevant to competitive interactions, and our results suggest that competition played an important role in shaping the evolution of these traits.

The linear version of Nuismer and Harmon (2015) model (equation (1)) results from making the simplifying assumption that the outcome of competition is not highly sensitive to variation in sympatric competitors' phenotypes (i.e., that α in their equation 1, and as a result also S in our equations, are small). We used this version here, since currently available likelihood tools for trait evolution rely on the multivariate normal distribution, which is to be expected only for this linear form of the model. The current formulation (equation 1) corresponds to a scenario in which the rate of phenotypic evolution in a lineage gets higher as the lineage deviates from the mean phenotype, although character displacement theory, for example, posits that selection for divergence should be the strongest when species are most ecologically similar (Brown and Wilson 1956). Given this formulation of the model, large S values are not to be expected, and we indeed found relatively small S values when fitting the model to the *Anolis* data set. Investigators finding high S values should treat them with caution and consider enforcing bounds on the likelihood search. Nevertheless, the developments presented here provide an important new set of tools for investigating the impact of interspecific interactions on trait evolution, and researchers can perform posterior simulations to assess the realism of the resulting inference. Future development of likelihood-free methods, such as Approximate Bayesian Computation (Slater et al. 2012b; Kutsukake and Innan 2013), may be possible for fitting the version of the model in which the outcome of competitive interactions depends on distance in trait space.

We imagine that the matching competition model and biogeographical implementations of diversity-dependent models will play a substantial role in the study of interspecific competition. For example, by comparing the fits of the matching competition model with other models that do not include competitive interactions between lineages, biologists can directly test hypotheses that make predictions about the role of interspecific interactions in driving trait evolution. In other words, while the effect of competition has been historically difficult to detect (Losos 2009), it may be detectable in the contemporary distribution of trait values and their covariance structure (Hansen and Martins 1996; Nuismer and Harmon 2015). The ability to consider trait distributions among species that arise from a model explicitly accounting for the effect of species interactions on trait divergence is also an important step toward a more coherent integration of macroevolutionary models of phenotypic evolution in community ecology.

There are many possible extensions of the tools developed in this article. In the future, empirical applications of the model can be implemented with more complex geography matrices that are more realistic for mainland taxa (e.g., using ancestral biogeographical reconstruction, Ronquist and Sanmartín 2011; Landis et al. 2013), and can also specify degrees of sympatric overlap (i.e., syntopy). Additionally, the current version of the model is rather computationally expensive with larger trees (on a Mac laptop with a 2.6 GHz processor, ML optimization for the matching competition model takes several minutes for a tree with 50 tips and can take 30 minutes or longer on 100 tip trees). Further work developing an analytical solution to the model may greatly speed up the likelihood calculation and permit the inclusion of stabilizing selection.

The current form of the model assumes that the degree of competition is equal for all interacting lineages. Future modifications of the model, such as applications of stepwise AICc algorithms (Alfaro et al. 2009; Thomas and Freckleton 2012; Mahler et al. 2013) or reversible-jump Markov chain Monte Carlo (Pagel and Meade 2006; Eastman et al. 2011; Rabosky 2014; Uyeda and Harmon 2014), may be useful to either identify more intensely competing lineages or test hypotheses about the strength of competition between specific taxa. Improvements could also be made on the formulation itself of the evolution of a species' trait as a response to the phenotypic landscape in which the species occurs. Moreover, a great array of extensions will come from modeling species interactions not only within clades, but also among interacting clades, as in the case of coevolution in bipartite mutualistic or antagonistic networks, such as plant–pollinator or plant–herbivore systems.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.d670p.

Funding

This research was funded by the Agence Nationale de la Recherche [grant CHEX-ECOEVOBIO] and the European Research Council [grant 616419-PANDA] to H.M.

Acknowledgments

We thank J. Weir for providing R code for diversity-dependent models and E. Lewitus, O. Missa, F. Anderson, L. Harmon, and two anonymous reviewers for helpful comments on the manuscript.

References

Alfaro M.E., Santini F., Brock C., Alamillo H., Dornburg A., Rabosky D.L., Carnevale G., Harmon L.J. 2009. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. Proc. Natl. Acad. Sci. 106:13410–13414.

Blomberg S.P., Garland T., Ives A.R. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57:717–745.

Boettiger C., Coop G., Ralph P. 2012. Is your phylogeny informative? Measuring the power of comparative methods. Evolution 66: 2240–2251.

Brown W.L., Wilson E.O. 1956. Character displacement. Syst. Zool. 5:49.

Burnham, K., Anderson, D. 2002. Model selection and multimodel inference: a practical information-theoretic approach. New York: Springer.

Cavalli-Sforza L.L., Edwards A.W.F. 1967. Phylogenetic analysis. Models and estimation procedures. Am. J. Hum. Genet. 19: 233–257.

Cavender-Bares J., Kozak K.H., Fine P.V.A., Kembel S.W. 2009. The merging of community ecology and phylogenetic biology. Ecol. Lett. 12:693–715.

Connell, J. H. 1980. Diversity and the coevolution of competitors, or the ghost of competition past. Oikos 35:131–138.

Coyne J.A., Orr H.A. 2004. Speciation. Sunderland (MA): Sinauer Associates.

Dobzhansky T. 1937. Genetics and the origin of species. New York: Columbia University Press.

Dobzhansky T. 1940. Speciation as a stage in evolutionary divergence. Am. Nat. 74:312–321.

Eastman J.M., Alfaro M.E., Joyce P., Hipp A.L., Harmon L.J. 2011. A novel comparative method for identifying shifts in the rate of character evolution on trees. Evolution 65:3578–3589.

Elton C. 1946. Competition and the structure of ecological communities. J. Anim. Ecol. 15:54–68.

Felsenstein. 1988. Phylogenies and quantitative characters. Annu. Rev. Ecol. Evol. Syst. 19:445–471.

Fisher R.A. 1930. The genetical theory of natural selection. Oxford: Oxford University Press.

Grafen A. 1989. The phylogenetic regression. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 326:119–157.

Grant P.R. 1972. Convergent and divergent character displacement. Biol. J. Linn. Soc. 4:39–68.

Grant P.R. 1999. Ecology and Evolution of Darwin's finches. Princeton (NJ): Princeton University Press.

Grant P.R., Grant B.R. 2002. Adaptive radiation of Darwin's finches: recent data help explain how this famous group of Galápagos birds evolved, although gaps in our understanding remain. Am. Sci. 90:130–139.

Grant P.R., Grant B.R. 2011. How and Why Species Multiply: the Radiation of Darwin's Finches. Princeton, NJ: Princeton University Press.

Grether G.F., Anderson C.N., Drury J.P., Kirschel A.N.G., Losin N., Okamoto K., Peiman K.S. 2013. The evolutionary consequences of interspecific aggression. Ann. N. Y. Acad. Sci. 1289:48–68.

Grether G.F., Losin N., Anderson C.N., Okamoto K. 2009. The role of interspecific interference competition in character displacement and the evolution of competitor recognition. Biol. Rev. 84: 617–635.

Gröning J., Hochkirch A. 2008. Reproductive interference between animal species. Q. Rev. Biol. 83:257–282.

Hansen T.F., Martins E.P. 1996. Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. Evolution 50:1404–1417.

Harmon L.J., Losos J.B., Jonathan Davies T., Gillespie R.G., Gittleman J.L., Bryan Jennings W., Kozak K.H., McPeek M.A., Moreno-Roark F., Near T.J., Purvis A., Ricklefs R.E., Schluter D., Schulte J.A., Seehausen O., Sidlauskas B.L., Torres-Carvajal O., Weir J.T., Mooers A.T. 2010. Early bursts of body size and shape evolution are rare in comparative data. Evolution 64:2385–2396.

Harmon L.J., Schulte J.A., Larson A., Losos J.B. 2003. Tempo and mode of evolutionary radiation in iguanian lizards. Science 301: 961–964.

Harmon L.J., Weir J.T., Brock C.D., Glor R.E., Challenger W. 2008. GEIGER: investigating evolutionary radiations. Bioinformatics 24:129–131.

Irschick D.J., Vitt L.J., Zani P.A., Losos J.B. 1997. A comparison of evolutionary radiations in mainland and Caribbean Anolis lizards. Ecology 78:2191–2203.

Kraft N.J.B., Cornwell W.K., Webb C.O., Ackerly D.D. 2007. Trait evolution, community assembly, and the phylogenetic structure of ecological communities. Am. Nat. 170:271–283.
Kutsukake N., Innan H. 2013. Simulation-based likelihood approach for evolutionary models of phenotypic traits on phylogeny. Evolution 67:355–367.
Lande R. 1976. Natural selection and random genetic drift in phenotypic evolution. Evolution. 30:314–334.
Landis M.J., Matzke N.J., Moore B.R., Huelsenbeck J.P. 2013. Bayesian analysis of biogeography when the number of areas is large. Syst. Biol. 62:789–804.
Losos J.B. 1990. The evolution of form and function: morphology and locomotor performance in West Indian *Anolis* lizards. 44: 1189–1203.
Losos J.B. 1994. Integrative approaches to evolutionary ecology: *Anolis* lizards as model systems. Annu. Rev. Ecol. Syst. 25:467–493.
Losos J.B. 2009. Lizards in an evolutionary tree: ecology and adaptive radiation of Anoles. Los Angeles (CA): University of California Press.
Losos J.B., Ricklefs R.E. 2009. Adaptation and diversification on islands. Nature 457:830–836.
Mahler D.L., Ingram T. 2014. Phylogenetic comparative methods for studying clade-wide convergence. In: Garamszegi L., editor. Modern phylogenetic comparative methods and their application in evolutionary biology. New York: Springer. p. 425–450.
Mahler D.L., Ingram T., Revell L.J., Losos J.B. 2013. Exceptional convergence on the macroevolutionary landscape in island lizard radiations. Science 341:292–295.
Mahler D.L., Revell L.J., Glor R.E., Losos J.B. 2010. Ecological opportunity and the rate of morphological evolution in the diversification of greater Antillean anoles. Evolution 64:2731–2745.
Martins E.P., Hansen T.F. 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am. Nat. 149: 646–667.
Matzke, N. 2014. Model selection in historical biogeography reveals that founder-event speciation is a crucial process in island clades. Syst. Biol. 63:951–970.
Mayr E. 1963. Animal species and evolution. Cambridge (MA): Harvard University Press.
Moen D., Morlon H. 2014. Why does diversification slow down? Trends Ecol. Evol. 29:190–197.
Morlon H, Lewitus E, Condamine FL, Manceau M., Clavel, J., Drury, J. 2016. RPANDA: an R package for macroevolutionary analyses on phylogenetic trees. Methods Ecol. Evol. doi:10.1111/2041-210X.12526.
Mouquet N., Devictor V., Meynard C.N., Munoz F., Bersier L.-F., Chave J., Couteron P., Dalecky A., Fontaine C., Gravel D., Hardy O.J., Jabot F., Lavergne S., Leibold M., Mouillot D., Münkemüller T., Pavoine S., Prinzing A., Rodrigues A.S.L., Rohr R.P., Thébault E., Thuiller W. 2012. Ecophylogenetics: advances and perspectives. Biol. Rev. 87:769–785.
Nuismer S.L., Harmon L.J. 2015. Predicting rates of interspecific interaction from phylogenetic trees. Ecol. Lett. 18:17–27.
Ord T.J., King L., Young A.R. 2011. Contrasting theory with the empirical data of species recognition. Evolution 65:2572–2591.
Ord T.J., Stamps J.A. 2009. Species identity cues in animal communication. Am. Nat. 174:585–593.
Pacala S., Roughgarden J. 1982. Resource partitioning and interspecific competition in two two-species insular *Anolis* lizard communities. Science 217:444–446.
Pagel M., Meade A. 2006. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. Am. Nat. 167:808–825.
Paradis E. 2012. Analysis of phylogenetics and evolution with R. New York: Springer.
Pennell M.W., Harmon L.J. 2013. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. Ann. N. Y. Acad. Sci. 1289:90–105.
Pfennig K.S., Pfennig D.W. 2009. Character displacement: ecological and reproductive responses to a common evolutionary problem. Q. Rev. Biol. 84:253–276.
Rabosky D.L. 2014. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. PLoS One 9:e89543.
Rabosky D.L., Glor R.E. 2010. Equilibrium speciation dynamics in a model adaptive radiation of island lizards. Proc. Natl. Acad. Sci. 107:22178–22183.
Revell L.J. 2009. Size-correction and principal components for interspecific comparative studies. Evolution 63:3258–3268.
Revell L.J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. 3:217–223.
Ronquist F., Sanmartín I. 2011. Phylogenetic methods in biogeography. Annu. Rev. Ecol. Evol. Syst. 42:441–464.
Rundle H.D., Nosil P. 2005. Ecological speciation. Ecol. Lett. 8:336–352.
Schluter D. 2000. The ecology of adaptive radiation. Oxford: Oxford University Press.
Schoener T.W. 1970. Size patterns in West Indian *Anolis* lizards. II. Correlations with the sizes of particular sympatric species-displacement and convergence. Am. Nat. 104:155–174.
Slater G.J., Harmon L.J., Alfaro M.E. 2012a. Integrating fossils with molecular phylogenies improves inference of trait evolution. 66:3931–3944.
Slater G.J., Harmon L.J., Wegmann D., Joyce P., Revell L.J., Alfaro M.E. 2012b. Fitting models of continuous trait evolution to incompletely sampled comparative data using approximate Bayesian computation. Evolution 66:752–762.
Slater G.J., Pennell M.W. 2014. Robust regression and posterior predictive simulation increase power to detect early bursts of trait evolution. Syst. Biol. 63:293–308.
Soetaert K., Petzoldt T., Setzer R.W. 2010. Solving differential equations in R: package deSolve. J. Stat. Softw. 33:1–25.
Stadler T. 2014. TreeSim: simulating trees under the birth-death model. R package Version 2.1. http://CRAN.R-project.org/package=TreeSim.
Thomas G.H., Freckleton R.P. 2012. MOTMOT: models of trait macroevolution on trees. Methods Ecol. Evol. 3:145–151.
Uyeda J.C., Caetano D.S., Pennell M.W. 2015. Comparative analysis of principal components can be misleading. Syst. Biol. 64:677–689.
Uyeda J.C., Harmon L.J. 2014. A novel Bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data. Syst. Biol. 63:902–918.
Wallace A.R. 1889. Darwinism. 2007 facs.: Cosimo, Inc.
Webb C.O., Ackerly D.D., McPeek M.A., Donoghue M.J. 2002. Phylogenies and community ecology. Annu. Rev. Ecol. Syst. 33: 475–505.
Weir J.T., Mursleen S. 2013. Diversity-dependent cladogenesis and trait evolution in the adaptive radiation of the auks (Aves: Alcidae). Evolution 67:403–416.
Williams E.E. 1972. The origin of faunas. Evolution of lizard congeners in a complex island fauna: a trial analysis. Evol. Biol. 6:47–89.

Bibliography

Aguilée, R., A. Lambert, and D. Claessen. 2011. Ecological speciation in dynamic landscapes. J. Evolution. Biol. 24:2663–2677. 53

Aldous, D. 1996. Probability distributions on cladograms. Pages 1–18 *in* Random discrete structures. Springer. 15

Aldous, D., M. Krikun, and L. Popovic. 2008. Stochastic models for phylogenetic trees on higher-order taxa. J. Math. Biol. 56:525–557. 53, 61

Aldous, D. J. 2001. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. Stat. Sci. Pages 23–34. 16

Aldous, D. J., M. A. Krikun, and L. Popovic. 2011. Five statistical questions about the tree of life. Syst. Biol. 60:318–328. 53, 61

Alexander, S. A. 2013. Infinite graphs in systematic biology, with an application to the species problem. Acta Biotheor. 61:181–201. 62

Alexander, S. A., A. de Bruin, and D. J. Kornet. 2015. An alternative construction of internodons: The emergence of a multi-level tree of life. B. Math. Biol. 77:23–45. 62

Alfaro, M. E., F. Santini, C. Brock, H. Alamillo, A. Dornburg, D. L. Rabosky, G. Carnevale, and L. J. Harmon. 2009. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. P. Natl. Acad. Sci. USA 106:13410–13414. 43, 65, 74

Archibald, J. D. 2009. Edward Hitchcock's pre-Darwinian (1840) 'Tree of Life'. J. Hist. Biol. 42:561–592. 2

Avise, J. C. and R. M. Ball. 1990. Principles of genealogical concordance in species concepts and biological taxonomy. Oxford Surv. Evol. Bio. 7:45–67. 53, 76

Bartoszek, K. 2014. Quantifying the effects of anagenetic and cladogenetic evolution. Math. Biosci. 254:42–57. 81

Bartoszek, K., S. Glémin, I. Kaj, and M. Lascoux. 2016. The Ornstein-Uhlenbeck process with migration: evolution with interactions. arXiv preprint arXiv:1607.07970 . 26

Bartoszek, K., J. Pienaar, P. Mostad, S. Andersson, and T. F. Hansen. 2012. A phylogenetic comparative method for studying multivariate adaptation. J. Theor. Biol. 314:204–215. 80, 81, 86, 88, 121

Bastide, P., M. Mariadassou, and S. Robin. 2016. Detection of adaptive shifts on phylogenies by using shifted stochastic processes on a tree. J. R. Stat. Soc. B 79:1067–1093. 47, 48, 98

Baum, D. A. 2009. Species as ranked taxa. Syst. Biol. 58:74–86. 53, 62

Beaulieu, J. M., D.-C. Jhwueng, C. Boettiger, and B. C. O'Meara. 2012. Modeling stabilizing selection: expanding the Ornstein–Uhlenbeck model of adaptive evolution. Evolution 66:2369–2383. 47, 80

Beaulieu, J. M. and B. C. O'meara. 2016. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. Syst. Biol. 65:583–601. 49

Behdenna, A., J. Pothier, S. S. Abby, A. Lambert, and G. Achaz. 2016. Testing for independence between evolutionary processes. Syst. Biol. 65:812–823. 121

Bickford, D., D. J. Lohman, N. S. Sodhi, P. K. Ng, R. Meier, K. Winker, K. K. Ingram, and I. Das. 2007. Cryptic species as a window on diversity and conservation. Trends Ecol. Evol. 22:148–155. 62

Blomberg, S. P. 2017. Beyond Brownian motion and the Ornstein-Uhlenbeck process: Stochastic diffusion models for the evolution of quantitative characters. bioRxiv . 45

Blomberg, S. P., T. Garland, and A. R. Ives. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57:717–745. 80

Blum, M. G. B. and O. François. 2006. Which random processes describe the tree of life ? A large scale study of phylogenetic tree imbalance. Syst. Biol. 55:685–691. 65, 69, 70

Bock, W. J. 2004. Species: the concept, category and taxon. J. Zool. Syst. Evol. Res. 42:178–190. 53

Bokma, F. 2002. Detection of punctuated equilibrium from molecular phylogenies. J. Evolution. Biol. 15:1048–1056. 47, 81, 98

Bokma, F. 2008. Detection of 'punctuated equilibrium' by Bayesian estimation of speciation and extinction rates, ancestral character states, and rates of anagenetic and cladogenetic evolution on a molecular phylogeny. Evolution 62:2718–2726. 81, 98

Bóna, M. 2011. A walk through combinatorics: an introduction to enumeration and graph theory. World scientific, singapore ed. 58, 124

Boucher, F. C. and V. Démery. 2016. Inferring bounded evolution in phenotypic characters from phylogenetic comparative data. Syst. Biol. 65:651–661. 25

Bromham, L. 2009. Why do species vary in their rate of molecular evolution? Biol. Lett. 5:401–404. 98

Brower, A. V. Z. 2004. Comment on 'Molecular phylogenies link rates of evolution and speciation' (ii). Science 303:173–173. 98

Brown, W. L. and E. O. Wilson. 1956. Character displacement. Syst. Zool. 5:49–64. 81

Butler, M. A. and A. A. King. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. Am. Nat. 164:683–695. 80, 83, 85, 98, 165

Champagnat, N. and A. Lambert. 2013. Splitting trees with neutral Poissonian mutations ii: Largest and oldest families. Stoch. Proc. Appl. 123:1368–1414. 67

Chave, J. and E. G. Leigh. 2002. A spatially explicit neutral model of beta-diversity in tropical forests. Theor. Popul. Biol. 62:153–168. 65

Clavel, J., L. Aristide, and H. Morlon. 2018. A penalized likelihood framework for high dimensional phylogenetic comparative methods and an application to new-world monkeys brain evolution. Syst. Biol. (under review) . 121

Clavel, J., G. Escarguel, and G. Merceron. 2015. mvMORPH: an R package for fitting multivariate evolutionary models to morphometric data. Methods Ecol. Evol. 6:1311–1319. 80

Clavel, J. and H. Morlon. 2017. Accelerated body size evolution during cold climatic periods in the cenozoic. Proc. Natl. Acad. Sci. USA. 114:4183–4188. 46, 85

Condamine, F. L., J. Rolland, and H. Morlon. 2013. Macroevolutionary perspectives to environmental change. Ecol. Lett. 16:72–85. 18, 44

Condit, R., N. Pitman, E. G. Leigh, J. Chave, J. Terborgh, R. B. Foster, P. Núñez, S. Aguilar, R. Valencia, G. Villa, H. C. Muller-Landau, E. Losos, and S. P. Hubbell. 2002. Beta-diversity in tropical forest trees. Science 295:666–669. 73

Conow, C., D. Fielder, Y. Ovadia, and R. Libeskind-Hadas. 2010. Jane: a new tool for the cophylogeny reconstruction problem. Algorithms Mol. Biol. 5:1–10. 95

Consortium, . G. P. et al. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56. 121

Coyne, J. A. and H. A. Orr. 2004. Speciation vol. 37. Sinauer Associates Sunderland, MA. 72

Cross, R. 2017. The inside story on 20,000 vertebrates. Science 357:742–743. 121

Dale, J., C. J. Dey, K. Delhey, B. Kempenaers, and M. Valcu. 2015. The effects of life history and sexual selection on male and female plumage colouration. Nature 527:367–370. 80

Darwin, C. 1859. On the origins of species by means of natural selection. 2, 3

Davies, T. J., A. P. Allen, L. Borda-de Água, J. Regetz, and C. J. Melián. 2011. Neutral biodiversity theory can explain the imbalance of phylogenetic trees but not the tempo of their diversification. Evolution 65:1841–1850. 45, 65, 72, 74

Dawkins, R. and J. R. Krebs. 1979. Arms races between and within species. Proc. R. Soc. B 205:489–511. 81

De Queiroz, K. 2007. Species concepts and species delimitation. Syst. Biol. 56:879–886. 53, 62

De Queiroz, K. and M. J. Donoghue. 1988. Phylogenetic systematics and the species problem. Cladistics 4:317–338. 53

Dobzhansky, T. G. 1937. Genetics and the origin of species. Columbia University Press, New York. 3

Dress, A., V. Moulton, M. Steel, and T. Wu. 2010. Species, clusters and the 'tree of life': A graph-theoretic perspective. J. Theor. Biol. 265:535–542. 62

Drury, J., J. Clavel, M. Manceau, and H. Morlon. 2016. Estimating the effect of competition on trait evolution using maximum likelihood inference. Syst. Biol. 65:700–710. 77, 81, 85, 88, 89, 94, 95, 110, 189

Dunne, J. A., R. J. Williams, N. D. Martinez, R. A. Wood, and D. H. Erwin. 2008. Compilation and network analyses of Cambrian food webs. PLoS Biol. 6:1–16. 95

Durrett, R. 2008. Probability models for DNA sequence evolution. Springer. 17, 53

Eastman, J. M., M. E. Alfaro, P. Joyce, A. L. Hipp, and L. J. Harmon. 2011. A novel comparative method for identifying shifts in the rate of character evolution on trees. Evolution 65:3578–3589. 80

Ehrlich, P. R. and P. H. Raven. 1964. Butterflies and plants: a study in coevolution. Evolution 18:586–608. 80, 81

Eldredge, N. and S. J. Gould. 1972. Models in Paleobiology chap. Punctuated equilibria: an alternative to phyletic gradualism, Pages 82–115. San francisco: Freeman cooper ed. Schopf, Thomas J.M. 98

Estes, S. and S. J. Arnold. 2007. Resolving the paradox of stasis: models with stabilizing selection explain evolutionary divergence on all timescales. Am. Nat. 169:227–244. 115

Etienne, R. S. and D. Alonso. 2005. A dispersal-limited sampling theory for species and alleles. Ecol. Lett. 8:1147–1156. 65, 74

Etienne, R. S., B. Haegeman, T. Stadler, T. Aze, P. N. Pearson, A. Purvis, and A. B. Phillimore. 2012. Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. Proc. R. Soc. B 279:1300–1309. 65, 74

Etienne, R. S., H. Morlon, and A. Lambert. 2014. Estimating the duration of speciation from phylogenies. Evolution 68:2430–2440. 56, 67

Etienne, R. S. and J. Rosindell. 2011. Prolonging the past counteracts the pull of the present: Protracted speciation can explain observed slowdowns in diversification. Syst. Biol. 61:204–213. 67, 72, 74

Felsenstein, J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. Am. J. Hum. Genet. 25:471–492. 4, 45, 80, 98, 117

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410. 39

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376. 13, 69, 106, 117

Fenster, C. B., W. S. Armbruster, P. Wilson, M. R. Dudash, and J. D. Thomson. 2004. Pollination syndromes and floral specialization. Annu. Rev. Ecol. Evol. Syst. 35:375–403. 81

Fitch, W. M. 1971. Toward defining the course of evolution: Minimum change for a specific tree topology. Syst. Zool. Pages 406–416. 38

Fitch, W. M. and J. J. Beintema. 1990. Correcting parsimonious trees for unseen nucleotide substitutions: the effect of dense branching as exemplified by ribonuclease. Mol. Biol. Evol. 7:438–443. 98

FitzJohn, R. G. 2010. Quantitative traits and diversification. Syst. Biol. 59:619–633. 48

Friedman, N., M. Ninio, I. Pe'er, and T. Pupko. 2002. A structural EM algorithm for phylogenetic inference. J. Comput. Biol. 9:331–353. 186

Fujisawa, T. and T. G. Barraclough. 2013. Delimiting species using single-locus data and the generalized mixed yule coalescent (GMYC) approach: a revised method and evaluation on simulated datasets. Syst. Biol. 62:707–724. 62, 116

Galtier, N. and J. Dutheil. 2007. Coevolution within and between genes. Pages 1–12 *in* Gene and Protein Evolution vol. 3. Karger Publishers. 121

Galtier, N., O. Gascuel, and A. Jean-Marie. 2005. Markov models in molecular evolution. Pages 3–24 *in* Statistical Methods in Molecular Evolution. Springer. 21

Gardiner, C. W. et al. 1985. Handbook of stochastic methods vol. 4. Springer Berlin. 84, 142, 143, 163

Gascuel, F., R. Ferrière, R. Aguilée, and A. Lambert. 2015. How ecology and landscape dynamics shape phylogenetic trees. Syst. Biol. 64:590–607. 53

Goldberg, E. E., L. T. Lancaster, and R. H. Ree. 2011. Phylogenetic inference of reciprocal effects between geographic range evolution and diversification. Syst. Biol. 60:451–465. 49

Golley, F. B., K. Petrusewicz, and L. Ryszkowski. 1975. Small mammals: their productivity and population dynamics vol. 5. Cambridge University Press, Cambridge. 73

Gould, S. J. 1989. Wonderful Life: The Burgess Shale and the Nature of History. Norton, new york ed. 120

Graham, C. H. and P. V. A. Fine. 2008. Phylogenetic beta diversity: Linking ecological and evolutionary processes across space in time. Ecol. Lett. 11:1265–1277. 57, 65

Hadfield, J. D., B. R. Krasnov, R. Poulin, and S. Nakagawa. 2014. A tale of two phylogenies: comparative analyses of ecological interactions. Am. Nat. 183:174–187. 95

Hadfield, J. D. and S. Nakagawa. 2010. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. J. Evolution. Biol. 23:494–508. 115

Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351. 45, 80, 98, 165

Hansen, T. F. and K. Bartoszek. 2012. Interpreting the evolutionary regression: the interplay between observational and biological errors in phylogenetic comparative studies. Syst. Biol. 61:413–425. 86

Hansen, T. F. and E. P. Martins. 1996. Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. Evolution 50:1404–1417. 80, 115

Hansen, T. F., J. Pienaar, and S. H. Orzack. 2008. A comparative method for studying adaptation to a randomly evolving environment. Evolution 62:1965–1977. 80, 86

Harmon, L. J., J. B. Losos, T. Jonathan Davies, R. G. Gillespie, J. L. Gittleman, W. Bryan Jennings, K. H. Kozak, M. A. McPeek, F. Moreno-Roark, T. J. Near, et al. 2010. Early bursts of body size and shape evolution are rare in comparative data. Evolution 64:2385–2396. 45, 46, 80

Harmon, L. J., J. T. Weir, C. D. Brock, R. E. Glor, and W. Challenger. 2008. Geiger: investigating evolutionary radiations. Bioinformatics 24:129–131. 80

Hayward, J. and T. R. Horton. 2014. Phylogenetic trait conservation in the partner choice of a group of ectomycorrhizal trees. Mol. Ecol. 23:4886–4898. 95

Heard, S. B. 1996. Patterns in phylogenetic tree balance with variable and evolving speciation rates. Evolution Pages 2141–2148. 74

Hennig, W. 1965. Phylogenetic systematics. Annu. Rev. Entomol. 10:97–116. 10, 53

Hey, J. 1992. Using phylogenetic trees to study speciation and extinction. Evolution 46:627–640. 73

Ho, L. S. T. and C. Ané. 2014. Intrinsic inference difficulties for trait evolution with Ornstein-Uhlenbeck models. Methods Ecol. Evol. 5:1133–1146. 94

Hohenlohe, P. A. and S. J. Arnold. 2008. MIPoD: a hypothesis-testing framework for microevolutionary inference from patterns of divergence. Am. Nat. 171:366–385. 115

Hoppe, A., S. Türpitz, and M. Steel. 2017. Species notions that combine phylogenetic trees and phenotypic partitions. arXiv preprint arXiv:1711.08145 . 110

Hubbell, S. P. 2001. The Unified Neutral Theory of Biodiversity and Biogeography. Princeton University Press, Princeton. 44, 51, 53, 54, 56, 63, 65

Hubbell, S. P. 2003. Modes of speciation and the lifespans of species under neutrality: a response to the comment of Robert E. Ricklefs. Oikos 100:193–199. 56

Hudson, R. R. and J. A. Coyne. 2002. Mathematical consequences of the genealogical species concept. Evolution 56:1557–1565. 60, 62

Ives, A. R. and H. C. J. Godfray. 2006. Phylogenetic analysis of trophic associations. Am. Nat. 168:1–14. 95

Ives, A. R., P. E. Midford, and T. Garland, Jr. 2007. Within-species variation and measurement error in phylogenetic comparative methods. Syst. Biol. 56:252–270. 115

Jabot, F. and J. Chave. 2009. Inferring the parameters of the neutral theory of biodiversity using phylogenetic information and implications for tropical forests. Ecol. Lett. 12:239–248. 53, 54, 56, 65, 74

Jhwueng, D.-C. and V. Maroulas. 2014. Phylogenetic Ornstein–Uhlenbeck regression curves. Stat. Probab. Lett. 89:110–117. 80

Jones, G. 2017. Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. J. Math. Biol. 74:447–467. 117

Jukes, T. H. and C. R. Cantor. 1969. Evolution of protein molecules. Mammalian protein metabolism 3:132. 4, 102

Karsenti, E., S. G. Acinas, P. Bork, C. Bowler, C. De Vargas, J. Raes, M. Sullivan, D. Arendt, F. Benzoni, J.-M. Claverie, M. Follows, G. Gorsky, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels-Lewis, U. Krzic, F. Not, H. Ogata, S. Pesant, E. G. Reynaud, C. Sardet, M. E. Sieracki, S. Speich, D. Velayoudon, J. Weissenbach, P. Wincker, and the Tara Oceans Consortium. 2011. A holistic approach to marine eco-systems biology. PLoS Biol.. 9:1–5. 121

Kendall, D. G. 1948. On the generalized 'birth-and-death' process. Ann. Math. Stat. 19:1–15. 4, 68, 99, 130

Khabbazian, M., R. Kriebel, K. Rohe, and C. Ané. 2016. Fast and accurate detection of evolutionary shifts in Ornstein–Uhlenbeck models. Methods Ecol. Evol. 7:811–824. 47, 98

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120. 102

Knowles, L. L. 2004. The burgeoning field of statistical phylogeography. J. Evolution. Biol. 17:1–10. 117

Kopp, M. 2010. Speciation and the neutral theory of biodiversity. Bioessays 32:564–570. 54

Kwok, R. B. H. 2011. Phylogeny, genealogy and the Linnaean hierarchy: a logical analysis. J. Math. Biol. 63:73–108. 62

Labra, A., J. Pienaar, and T. F. Hansen. 2009. Evolution of thermal physiology in Liolaemus lizards: adaptation, phylogenetic inertia, and niche tracking. Am. Nat. 174:204–220. 80

Lambert, A. and C. Ma. 2015. The coalescent in peripatric metapopulations. J. Appl. Probab. 52:538–557. 56

Lambert, A., H. Morlon, and R. S. Etienne. 2015. The reconstructed tree in the lineage-based model of protracted speciation. J. Math. Biol. 70:367–397. 56, 67

Lambert, A. and T. Stadler. 2013. Birth–death models and coalescent point processes: The shape and probability of reconstructed phylogenies. Theor. Popul. Biol. 90:113–128. 18, 65

Lambert, A. and M. Steel. 2013. Predicting the loss of phylogenetic diversity under non-stationary diversification models. J. Theor. Biol. 337:111–124. 65

Lanave, C., G. Preparata, C. Sacone, and G. Serio. 1984. A new method for calculating evolutionary substitution rates. J. Mol. Evol. 20:86–93. 21, 101

Lande, R. 1985. Expected time for random genetic drift of a population between stable phenotypic states. Proc. Natl. Acad. Sci. USA. 82:7641–7645. 115

Landis, M. and J. G. Schraiber. 2017. Punctuated evolution shaped modern vertebrate diversity. bioRxiv . 47, 98

Landis, M. J., J. G. Schraiber, and M. Liang. 2013. Phylogenetic analysis using Lévy processes: finding jumps in the evolution of continuous traits. Syst. Biol. 62:193–204. 47, 81, 98

Lartillot, N. 2013. A phylogenetic Kalman filter for ancestral trait reconstruction using molecular data. Bioinformatics 30:488–496. 28

Lartillot, N. and F. Delsuc. 2012. Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. Evolution 66:1773–1787. 118

Lartillot, N., M. J. Phillips, and F. Ronquist. 2016. A mixed relaxed clock model. Phil. Trans. R. Soc. B 371:20150132. 47, 108

Lartillot, N. and R. Poujol. 2011. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. Mol. Biol. Evol. 28:729–744. 46, 108, 118

Lepage, T., D. Bryant, H. Philippe, and N. Lartillot. 2007. A general comparison of relaxed molecular clock models. Mol. Biol. Evol. 24:2669–2680. 98, 106, 107

Liow, L. H., T. Reitan, and P. G. Harnik. 2015. Ecological interactions on macroevolutionary time scales: clams and brachiopods are more than ships that pass in the night. Ecol. Lett. 18:1030–1039. 82

Loeuille, N. and M. Loreau. 2005. Evolutionary emergence of size-structured food webs. Proc. Natl. Acad. Sci. USA. 102:5761–5766. 95

MacArthur, R. H. and E. O. Wilson. 1967. The Theory of Island Biogeography. Princeton University Press, Princeton. 65, 74

Maddison, W. P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536. 60, 116

Maddison, W. P. and L. L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55:21–30. 117

Maddison, W. P., P. E. Midford, and S. P. Otto. 2007. Estimating a binary character's effect on speciation and extinction. Syst. Biol. 56:701–710. 19, 48

Mahler, D. L., L. J. Revell, R. E. Glor, and J. B. Losos. 2010. Ecological opportunity and the rate of morphological evolution in the diversification of Greater Antillean anoles. Evolution 64:2731–2745. 80

Manceau, M. and A. Lambert. 2016. The species problem from the modeler's point of view. bioRxiv . 51, 123

Manceau, M., A. Lambert, and H. Morlon. 2015. Phylogenies support out-of-equilibrium models of biodiversity. Ecol. Lett. 18:347–356. 56, 61, 63, 129

Manceau, M., A. Lambert, and H. Morlon. 2017. A unifying comparative phylogenetic framework including traits coevolving across interacting lineages. Syst. Biol. 66:551–568. 77, 141

Martín González, A. M., B. Dalsgaard, D. Nogués-Bravo, C. H. Graham, M. Schleuning, P. K. Maruyama, S. Abrahamczyk, R. Alarcón, A. C. Araujo, F. P. Araújo, et al. 2015. The macroecology of phylogenetically structured hummingbird–plant networks. Glob. Ecol. Biogeogr. 24:1212–1224. 95

Martinez, N. D. 2006. Network evolution: exploring the change and adaptation of complex ecological systems over deep time. Pages 287–302 *in* Ecological Networks: Linking Structure to Dynamics in Food Webs (M. Pascual and J. Dunne, eds.). Oxford University Press. 95

Martins, E. P. 2004. Compare, version 4.6 b. computer programs for the statistical analysis of comparative data. 80

Mayden, R. L. 1997. A hierarchy of species concepts: the denouement in the saga of the species problem. *in* Species, the units of biodiversity. Claridge, M.F. and Dawah, H.A. and Wilson, M. R. 53, 62

McGill, B. J., B. A. Maurer, and M. D. Weiser. 2006. Empirical evaluation of neutral theory. Ecology 87:1411–1423. 65

McKinney, M. L. 1997. Extinction vulnerability and selectivity: Combining ecological and paleontological views. Annu. Rev. Ecol. Syst. 28:495–516. 65

McPeek, M. A. 2008. The ecological dynamics of clade diversification and community assembly. Am. Nat. 172:270–284. 65, 70, 74

Mehta, R. S., D. Bryant, and N. A. Rosenberg. 2016. The probability of monophyly of a sample of gene lineages on a species tree. Proc. Natl. Acad. Sci. USA. 113:8002–8009. 60

Melián, C. J., D. Alonso, S. Allesina, R. S. Condit, and R. S. Etienne. 2012. Does sex speed up evolutionary rate and increase biodiversity? PLoS Comput. Biol. 8:1–9. 56

Missa, O., C. Dytham, and H. Morlon. 2016. Understanding how biodiversity unfolds through time under neutral theory. Philos. T. Roy. Soc. B. 371. 53

Moen, D. and H. Morlon. 2014a. From dinosaurs to modern bird diversity: extending the time scale of adaptive radiation. PLoS Biol. 12:1–4. 80

Moen, D. and H. Morlon. 2014b. Why does diversification slow down ? Trends Ecol. Evol. 29:190–197. 74

Mooers, A. ., L. J. Harmon, M. G. B. Blum, D. H. J. Wong, and S. B. Heard. 2007. Some models of phylogenetic tree shape. *in* Reconstructing Evolution: New mathematical and computational advances. Oxford University Press, Oxford. 65

Morlon, H. 2014. Phylogenetic approaches for studying diversification. Ecol. Lett. 17:508–525. 42, 52, 65, 74

Morlon, H., E. Lewitus, F. L. Condamine, M. Manceau, J. Clavel, and J. Drury. 2015. RPANDA: an R package for macroevolutionary analyses on phylogenetic trees. Methods Ecol. Evol. 7:589–597. 69, 80, 92

Morlon, H., T. L. Parsons, and J. B. Plotkin. 2011a. Reconciling molecular phylogenies with the fossil record. P. Natl. Acad. Sci. USA 108:16327–16332. 44, 65, 74

Morlon, H., M. D. Potts, and J. B. Plotkin. 2010. Inferring the dynamics of diversification: A coalescent approach. PLoS Biol.. 8:1–13. 65, 66, 73

Morlon, H., D. W. Schwilk, J. A. Bryant, P. A. Marquet, A. G. Rebelo, C. Tauss, B. J. M. Bohannan, and J. L. Green. 2011b. Spatial patterns of phylogenetic diversity. Ecol. Lett. 14:141–149. 65, 73

Nee, S., P. H. Harvey, and A. O. Mooers. 1992. Tempo and mode of evolution revealed from molecular phylogenies. P. Natl. Acad. Sci. USA 89:8322–8326. 65

Nee, S. and R. M. May. 1997. Extinction and the loss of evolutionary history. Science 278:692–694. 65, 73

Nee, S., R. M. May, and P. H. Harvey. 1994. The reconstructed evolutionary process. Philos. T. Roy. Soc. B. 344:305–311. 18, 43

Nuismer, S. L. and L. J. Harmon. 2014. Predicting rates of interspecific interaction from phylogenetic trees. Ecol. Lett. 18:17–27. 81, 89, 94, 115, 154, 189

Nuismer, S. L., P. Jordano, and J. Bascompte. 2013. Coevolution and the architecture of mutualistic networks. Evolution 67:338–354. 95

O'Dwyer, J. P. and J. L. Green. 2010. Field theory for biogeography: A spatially explicit model for predicting patterns of biodiversity. Ecol. Lett. 13:87–95. 65

O'Meara, B. C., C. Ané, M. J. Sanderson, and P. C. Wainwright. 2006. Testing for different rates of continuous trait evolution using likelihood. Evolution 60:922–933. 80

Orr, H. A. 1995. The population genetics of speciation: the evolution of hybrid incompatibilities. Genetics 139:1805–1813. 56

Pagel, M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proc. R. Soc. B 255:37–45. 81

Pagel, M., C. Venditti, and A. Meade. 2006. Large punctuational contribution of speciation to evolutionary divergence at the molecular level. Science 314:119–121. 98

Papadopoulou, A. and L. L. Knowles. 2016. Toward a paradigm shift in comparative phylogeography driven by trait-based hypotheses. Proc. Natl. Acad. Sci. USA. 113:8018–8024. 122

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290. 12

Peichel, C. L. and D. A. Marques. 2017. The genetic and molecular architecture of phenotypic diversity in sticklebacks. Philos. T. Roy. Soc. B. 372. 98, 108

Pennell, M. W. and L. J. Harmon. 2013. An integrative view of phylogenetic comparative methods: Connections to population genetics, community ecology, and paleobiology. Ann. NY Acad. Sci. 1289:90–105. 57, 80

Pennell, M. W., L. J. Harmon, and J. C. Uyeda. 2014. Is there room for punctuated equilibrium in macroevolution? Trends Ecol. Evol. 29:23–32. 47, 98

Perrière, G. and C. Brochier-Armanet. 2010. Concepts et méthodes en phylogénie moléculaire. Springer. 22, 39, 41

Phillimore, A. B. and T. D. Price. 2008. Density-dependent cladogenesis in birds. PLoS Biol.. 6:483–489. 74

Pigot, A. L., I. P. F. Owens, and C. D. L. Orme. 2012. Speciation and extinction drive the appearance of directional range size evolution in phylogenies and the fossil record. PLoS Biol.. 10:1–9. 70

Pigot, A. L., A. B. Phillimore, I. P. F. Owens, and C. D. L. Orme. 2010. The shape and temporal dynamics of phylogenetic trees arising from geographic speciation. Syst. Biol. 59:660–673. 45, 65, 74

Poisot, T. and D. Stouffer. 2016. How ecological networks evolve. bioRxiv . 119

Price, S. A., S. S. Hopkins, K. K. Smith, and V. L. Roth. 2012. Tempo of trophic evolution and its impact on mammalian diversification. Proc. Natl. Acad. Sci. USA. 109:7008–7012. 48

Puigbò, P., Y. I. Wolf, and E. V. Koonin. 2013. Seeing the tree of life behind the phylogenetic forest. BMC Biol. 11:1–3. 62

Puillandre, N., A. Lambert, S. Brouillet, and G. Achaz. 2012. ABGD, automatic barcode gap discovery for primary species delimitation. Mol. Ecol. 21:1864–1877. 62, 116

Pybus, O. G. and P. H. Harvey. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. P. Roy. Soc. Lond. B. Bio. 267:2267–2272. 69

Pyron, R. A. 2011. Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. Syst. Biol. 60:466–481. 117

Pyron, R. A. and F. T. Burbrink. 2013. Phylogenetic estimates of speciation and extinction rates for testing ecological and evolutionary hypotheses. Trends Ecol. Evol. 28:729–736. 52

Quental, T. B. and C. R. Marshall. 2013. How the red queen drives terrestrial mammals to extinction. Science 341:290–292. 74

Quintero, I., P. Keil, W. Jetz, and F. W. Crawford. 2015. Historical biogeography using species geographical ranges. Syst. Biol. 64:1059–1073. 80

Rabosky, . D. L. 2014. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. PLoS One 9:1–15. 43, 65

Rabosky, D. L. and E. E. Goldberg. 2015. Model inadequacy and mistaken inferences of trait-dependent speciation. Syst. Biol. 64:340–355. 49

Rabosky, D. L. and I. J. Lovette. 2008. Density-dependent diversification in north american wood warblers. P. Roy. Soc. Lond. B. Bio. 275:2363–2371. 44, 65, 74

Rafferty, N. E. and A. R. Ives. 2013. Phylogenetic trait-based analyses of ecological networks. Ecology 94:2321–2333. 95

Rannala, B. and Z. Yang. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645–1656. 117

Regan, C. T. 1925. Organic evolution. Nature 116:398–401. 53

Reitan, T., T. Schweder, J. Henderiks, et al. 2012. Phenotypic evolution studied by layered stochastic differential equations. Ann. Appl. Stat. 6:1531–1551. 82

Revell, L. J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. 3:217–223. 88

Revell, L. J. and D. C. Collar. 2009. Phylogenetic analysis of the evolutionary correlation using likelihood. Evolution 63:1090–1100. 80

Richards, C. L., B. C. Carstens, and L. L. Knowles. 2007. Distribution modelling and statistical phylogeography: an integrative framework for generating and testing alternative biogeographical hypotheses. J. Biogeogr. 34:1833–1845. 122

Rolland, J., F. Jiguet, K. A. Jønsson, F. L. Condamine, and H. Morlon. 2014. Settling down of seasonal migrants promotes bird diversification. Proc. R. Soc. B 281. 48

Ronquist, F., S. Klopfstein, L. Vilhelmsen, S. Schulmeister, D. L. Murray, and A. P. Rasnitsyn. 2012. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. Syst. Biol. 61:973–999. 117

Ronquist, F. and I. Sanmartín. 2011. Phylogenetic methods in biogeography. Annu. Rev. Ecol. Evol. Syst. 42:441–464. 95

Rosenzweig, M. L. 1995. Species Diversity in Space and Time. Cambridge university press, Cambridge. 65

Rosindell, J., S. J. Cornell, S. P. Hubbell, and R. S. Etienne. 2010. Protracted speciation revitalizes the neutral theory of biodiversity. Ecol. Lett. 13:716–727. 45, 56, 67, 72, 74

Rosindell, J., L. J. Harmon, and R. S. Etienne. 2015. Unifying ecology and macroevolution with individual-based theory. Ecol. Lett. 18:472–482. 53

Rosindell, J., S. P. Hubbell, and R. S. Etienne. 2011. The unified neutral theory of biodiversity and biogeography at age ten. Trends Ecol. Evol. 26:340–348. 57

Rosindell, J. and A. B. Phillimore. 2011. A unified model of island biogeography sheds light on the zone of radiation. Ecol. Lett. 14:552–560. 74

Roux, C., C. Fraïsse, J. Romiguier, Y. Anciaux, N. Galtier, and N. Bierne. 2016. Shedding light on the grey zone of speciation along a continuum of genomic divergence. PLoS Biol.. 14:1–22. 116

Ruta, M., P. J. Wagner, and M. I. Coates. 2006. Evolutionary patterns in early tetrapods. I. rapid initial diversification followed by decrease in rates of character change. Proc. R. Soc. B 273:2107–2111. 80

Samadi, S. and A. Barberousse. 2006. The tree, the network, and the species. Biol. J. Linn. Soc. 89:509–521. 62

Sankoff, D. 1975. Minimal mutation trees of sequences. SIAM J. Appl. Math. 28:35–42. 38

Schluter, D. 2000. Ecological character displacement in adaptive radiation. Am. Nat. 156:4–16. 18

Seehausen, O., R. K. Butlin, I. Keller, C. E. Wagner, J. W. Boughman, P. A. Hohenlohe, C. L. Peichel, G.-P. Saetre, C. Bank, Ă. Brännström, et al. 2014. Genomics and the origin of species. Nat. Rev. Genet. 15:176–192. 98

Semple, C. and M. A. Steel. 2003. Phylogenetics vol. 24. Oxford university press, oxford ed. 6, 7, 10, 11

Simpson, G. G. 1944. Tempo and mode in evolution. Columbia university press, new york ed. 80

Slater, G. J. 2015. Iterative adaptive radiations of fossil canids show no evidence for diversity-dependent trait evolution. Proc. Natl. Acad. Sci. USA. 112:4897–4902. 80

Slater, G. J., L. J. Harmon, and M. E. Alfaro. 2012. Integrating fossils with molecular phylogenies improves inference of trait evolution. Evolution 66:3931–3944. 94

Sletvold, N., J. Trunschke, M. Smit, J. Verbeek, and J. Ăgren. 2016. Strong pollinator-mediated selection for increased flower brightness and contrast in a deceptive orchid. Evolution 70:716–724. 81

Sneath, P. H. A. 1976. Phenetic taxonomy at the species level and above. Taxon 25:437–450. 53

Solís-Lemus, C., L. L. Knowles, and C. Ané. 2014. Bayesian species delimitation combining multiple genes and traits in a unified framework. Evolution 69:492–507. 117

Stadler, T. 2011. Mammalian phylogeny reveals recent diversification rate shifts. P. Natl. Acad. Sci. USA 108:6187–6192. 44, 65

Stadler, T. 2013. Recovering speciation and extinction dynamics based on phylogenies. J. Evolution. Biol. 26:1203–1219. 42, 52

Steel, M. 2014. Tracing evolutionary links between species. Am. Math. Mon. 121:771–792. 15, 58

Sukumaran, J. and L. L. Knowles. 2017. Multispecies coalescent delimits structure, not species. Proc. Natl. Acad. Sci. USA. 114:1607–1612. 116

Thomas, G. H., J. A. Bright, and C. R. Cooney. 2016. Dataset: Mark my bird. Natural History Museum Data Portal (data.nhm.ac.uk) . 121

Thomas, G. H. and R. P. Freckleton. 2012. MOTMOT: models of trait macroevolution on trees. Methods Ecol. Evol. 3:145–151. 80

Thomas, G. H., R. P. Freckleton, and T. Székely. 2006. Comparative analyses of the influence of developmental mode on phenotypic diversification rates in shorebirds. Proc. R. Soc. B 273:1619–1624. 46, 80

Tolkoff, M. R., M. E. Alfaro, G. Baele, P. Lemey, and M. A. Suchard. 2017. Phylogenetic factor analysis. Syst. Biol. . 121

Uyeda, J. C., D. S. Caetano, and M. W. Pennell. 2015. Comparative analysis of principal components can be misleading. Syst. Biol. 64:677–689. 25, 88, 94, 149

Uyeda, J. C., T. F. Hansen, S. J. Arnold, and J. Pienaar. 2011. The million-year wait for macroevolutionary bursts. Proc. Natl. Acad. Sci. USA. 108:15908–15913. 47

Van Valen, L. 1973. A new evolutionary law. Evol. Theor. 1:1–30. 80, 81

Velasco, J. D. 2008. Species concepts should not conflict with evolutionary history, but often do. Stud. Hist. Philos. Biol. Biomed. Sci. 39:407–414. 56

Vences, M., J. M. Guayasamin, A. Miralles, and I. De La Riva. 2013. To name or not to name: Criteria to promote economy of change in Linnaean classification schemes. Zootaxa 3636:201–244. 62

Webb, C. O., D. D. Ackerly, M. M. A. , and M. J. Donoghue. 2002. Phylogenies and community ecology. Annu. Rev. Ecol. Syst. 33:475–505. 65

Webster, A. J., R. J. Payne, and M. Pagel. 2003. Molecular phylogenies link rates of evolution and speciation. Science 301:478–478. 98

Weiblen, G. D. 2004. Correlated evolution in fig pollination. Syst. Biol. 53:128–139. 81

Weir, J. T. and S. Mursleen. 2013. Diversity-dependent cladogenesis and trait evolution in the adaptive radiation of the auks (aves: Alcidae). Evolution 67:403–416. 46, 80

Wiens, J. J., D. D. Ackerly, A. P. Allen, B. L. Anacker, L. B. Buckley, H. V. Cornell, E. I. Damschen, J. T. Davies, J.-A. Grytnes, and S. P. Harrison. 2010. Niche conservatism as an emerging principle in ecology and conservation biology. Ecol. Lett. 13:1310–1324. 65

Witt, C. C. and R. T. Brumfield. 2004. Comment on 'molecular phylogenies link rates of evolution and speciation' (i). Science 303:173–173. 98

Yang, Z. and B. Rannala. 2010. Bayesian species delimitation using multilocus sequence data. Proc. Natl. Acad. Sci. USA. 107:9264–9269. 62, 116

Yule, G. U. 1925. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. Phil. Trans. R. Soc. Lond. B . 4

Zhang, C., T. Stadler, S. Klopfstein, T. A. Heath, and F. Ronquist. 2015. Total-evidence dating under the fossilized birth–death process. Syst. Biol. 65:228–249. 117, 122

Zhang, J., P. Kapli, P. Pavlidis, and A. Stamatakis. 2013. A general species delimitation method with applications to phylogenetic placements. Bioinformatics 29:2869–2876. 116

Zuckerkandl, E. and L. Pauling. 1962. Horizons in Biochemistry chap. Molecular disease, evolution and genetic heterogeneity. Academic Press, New York. 4, 98, 106