What fruits can we get from this tree?

Laudanno, Giovanni

DOI:
10.33612/diss.155031292

Chapter 1

Introduction

Introduction

Why don't you use an individual-based model?

– Popular wisdom

1.1 Biological background

1.1.1 Evolution

Organisms evolve over time. This process is mainly driven by stochastic events and selection, and its effects can be observed after many generations. This introduces one of the greatest challenges in evolutionary biology: understanding what are the processes that lead to the diversity we observe in nature today. The task is extremely challenging for two reasons. The first reason is that evolutionary processes are (usually) extremely slow, possibly spanning over periods of tens or hundreds of million years. As a consequence it is impossible to reconstruct a process just by collecting data over time. There are intrinsic limitations to claims we can make because almost all the information we can exploit come from the present or, in the best case scenario, from the recent past, because fossil data are often not available. The second reason is that evolution is dominated by randomness. This is certainly one of the most fascinating aspects of it. Organisms evolve because, at each new generation, offspring's genetics can slightly differ with respect to their parents'. This is due to random mutations occurring in a population at each new generation. Such mutations can be selected for or not according to the interaction with the environment and with other individuals, as in the case of sexual selection. If selection plays a role then alleles that express advantageous phenotypic traits increase their frequencies in the population over time, at the expense of the less fit. Even when selection is not involved we can still have the allele distribution change over time due to pure random drift, where neutral genes propagate at each generation moved by sheer stochastic forces. In both cases randomness plays a central role. If we imagine rewinding an evolutionary process and running

it all over again the outcomes could be completely different from what we observe today. The way we deal with this is to choose an inference approach based on stochastic models.

1.1.2 Speciation

When a group of individuals is able to interbreed and produce fertile offspring they are said to belong to the same species, according to the biological species concept. It might happen that two portions of the same population of individuals belonging to the same species could become reproductively isolated with respect to each other. When this happens the individuals will interbreed only with individuals from the same sub-group. With time the allele distributions of the two populations will start to diverge until the individuals from one group will not be able to produce offspring anymore with the individuals from the other group. When this occurs the two populations belong, by the biological species concept, to different species. The process of forming two new species from one is called cladogenesis. It is one form of speciation which is, in general, the process that leads to the formation of new species. The other speciation mode is called anagenesis, and it refers to phyletic evolution within the same lineage.

There are several reasons for the cladogenetic process to occur. They are categorized in four major variants in nature. When two populations become geographically isolated (as for the rise of geographical barriers, like mountains or rivers) speciation is called "allopatric". When the process is driven by geographical isolation but one of the two populations has much smaller size with respect to the other one (as when a small group of colonizers migrates to an island) the process is called "peripatric". On the contrary, "sympatric" speciation occurs when the two sub-populations coexist in the same habitat but they become reproductively isolated for other reasons. There might be several drivers for sympatric speciations. One example is the case of sexual selection (Higashi, Takimoto, and Yamamura, 1999) as in Cichlid fish populations in African lakes (Allender et al., 2003). The last one, called "parapatric" speciation, also involves two sub-populations coexisting in the same area but the subpopulations are not randomly mixed as in the sympatric case; gene flow is therefore limited, but not zero as in the allopatric case. Speciation can also be induced artificially by humans, as in the case of lab experiments conducted on Drosophila Melanogaster or in the case of domesticated species like those belonging to the genus citrus (Wu et al., 2018).

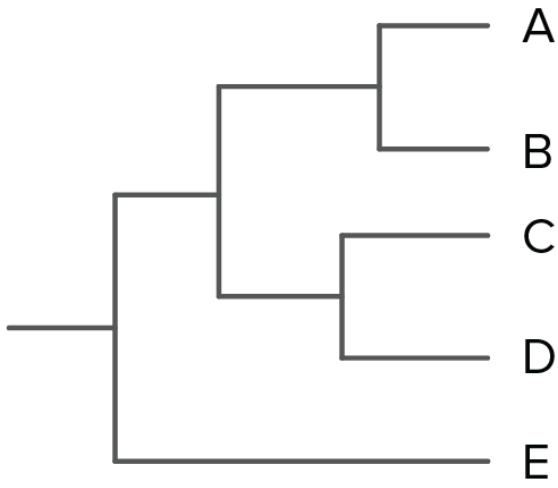

Figure 1.1: A phylogeny.

1.2 Phylogenetic Inference

A phylogenetic tree is an efficient way to summarize the phylogenetic history of a group of organisms. A tree is an oriented network (see Fig. 1.1) where each branch represents a lineage, each node a cladogenetic event and each tip an extant species. Phylogenetic trees tell us a story about the processes that lead to the diversity that we can observe today. The main problem is that, despite the botanical name, they cannot actually be found in nature. They are objects constructed by biologists from currently available genetic data and morphological information. Such information are obtained (mostly) from extant organisms and, to a lesser extent, from fossil records (when possible). Historically parsimony was invoked to construct the tree: the tree was the simplest pattern of divergence that could explain the observed data (Farris, 1970; Fitch, 1971). However, in the past few decades, the rapid technological developments in computational power have lead to the development of new statistical tools that allow us to reconstruct trees in a more sophisticated way (Felsenstein, 1978). Some of the most popular techniques rely on a Bayesian framework. This is the case for BEAST (Drummond and Bouckaert, 2015; Bouckaert et al., 2014; Bouckaert et al., 2019), RevBayes (Höhna et al., 2016) and MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). The general methodology employed by all these tools al-

ways starts from an alignment of inheritable sequences (such as DNA or proteins) from different individuals belonging to several species, aiming to reconstruct the phylogenetic history building progressively from present to past. In case of DNA, for example, one or more loci can be taken into consideration. The loci must be located in regions that are not subject to selection as species tree construction methods always assume neutral mutations. Using neutral markers in the mitochondrial DNA (mtDNA) is a common choice, because mtDNA mutates at higher rates compared to nuclear DNA (Moore, 1995), enhancing differences among different species over time. Moreover it is maternally inherited and therefore does not suffer from recombination. However, the mitochondrial genome is relatively short and may not contain sufficient non-coding DNA. Therefore nuclear DNA is also used.

However, to reconstruct a phylogeny data is not enough. In fact, it is also necessary to provide additional information about what we believe could be the possible mechanism that actually generates the tree.

1.2.1 The Bayesian framework

The main ingredients of the Bayesian framework are three: the "prior", the "likelihood" and the "posterior".

Our prior knowledge is summarized by a "prior distribution". For example we can hypothesise that the process is generated by some model M. Models usually have parameters, for example to describe the rates at which allowed events can occur in the process according to the model. We can refer to them as θ_M. We can therefore assign a joint probability to each model and set of parameter values $P(M, \theta_M)$.

Next there is the "likelihood function". In its generality, it is defined as the probability of observing some data, given a model M (and its parameters θ_M)

$$\mathscr{L}_{M,\theta_M}(D) = P(D|M, \theta_M) \tag{1.2.1}$$

The combination of a prior and a likelihood yields a "posterior distribution"

$$P(M, \theta_M|D) = KP(D|M, \theta_M)P(M, \theta_M) \tag{1.2.2}$$

where K is a normalization constant. The posterior can be read as the credibility we give to our model, given the experimental evidence. In this sense, it becomes clear why such quantity is pivotal in the inference process, as the goal is to find the best explanation for the process that generated the available data.

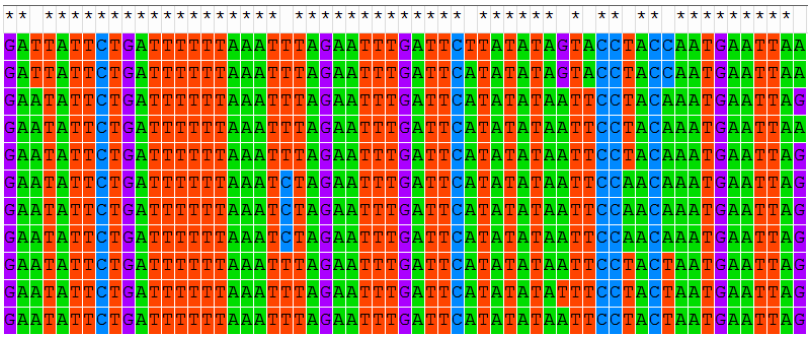

Figure 1.2: A sequence alignment.

1.2.2 Bayesian framework in phylogenetics

Here we explain how the the Bayesian framework is applied in phylogenetics. Most of the concepts are taken from Lemey, Salemi, and Vandamme (2009) (chapters 4 and 6), which summarizes in an excellent way Drummond and Bouckaert (2015) and Felsenstein (1981).

For the sake of convenience, in the following we will always refer to a generic tree having n_s tips and $n_s - 1$ internal nodes. The total information content of the tree is given by the pair $(R, \mathbf{t})$, where R contains the information about tree topology and $\mathbf{t}$ is the vector of its branching times. When the Bayesian framework is applied to phylogenetics the starting point, i.e. the experimental evidence, consists of heritable characters (DNA, RNA, proteins). We will label the collection of these characters as D. The characters are used to create an alignment composed of n_s sequences (one for each tree tip) of length n_c. The length n_c is the length of the character sequence sampled from each species.

$$D = \begin{bmatrix} d_{1,1} & d_{1,2} & \dots & d_{1,n_c} \\ d_{2,1} & d_{2,2} & \dots & d_{2,n_c} \\ \vdots & \vdots & \vdots & \vdots \\ d_{n_s,1} & d_{n_s,2} & \dots & d_{n_s,n_c} \end{bmatrix} \tag{1.2.3}$$

Each of the $d_{i,j}$ can be in each of the possible states defined by the heritable character considered. In the following, for the sake of simplicity, we will always refer to DNA, but the same reasoning could be applied to proteins or other kind of data (even not biological, as in the case of cultural evolution or the evolution of languages, see Bouckaert et al., 2012 or Bouckaert, Bowern, and Atkinson, 2018).

For each species those characters can be subject to random substitutions (e.g., character A becoming T) over time. The goal is to reconstruct backwards in time the process, driven by substitutions, that lead to the diversification of such alignments. At the end we will obtain a tree (or, more precisely, a posterior distribution of trees) that explains the phylogenetic history of data in the best possible way. To calculate the posterior probability distribution of phylogenies we use 1.2.2:

$$\begin{aligned} P(T,\theta_T,\theta_S|D) &= KP(D|T,\theta_T,\theta_S)P(T,\theta_T)P(\theta_S) \\ &= KP(D|T,\theta_S)P(T,\theta_T)P(\theta_S) \\ &= KP(D|T,\theta_S)P(T|\theta_T)P(\theta_T)P(\theta_S) \end{aligned} \tag{1.2.4}$$

where T is the tree, θ_T are the tree parameters, S is a nucleotide substitution model and θ_S are the substitution model's parameters. The term $P(T,\theta_T)$ summarizes our prior knowledge about the tree. The term $P(\theta_S)$ is instead the substitution model's prior. From the first to second line we remove the dependency on model parameters in the likelihood as they do not affect directly the probability of the observed characters.

1.2.3 The likelihood

The likelihood $P(D|T,\theta_T,\theta_S)$ expresses the probability to observe the alignment D at the tips, given the tree and the substitution model. Alignment data can be seen as a collection of data for each tip, therefore $D = \{d_{i,j}\}$ with $i = 1,\dots,n_s$ and $j = 1,\dots,n_c$, as in 1.2.3. For the i-th tip, the j-th character in the alignment can be represented as a vector of probabilities for each of the possible states. In case of DNA the collection of these states, also called "alphabet", is simply $\alpha = (A,C,G,T)$.

We can assign a likelihood to the entire phylogeny starting from the data at the tips using the "pruning algorithm" introduced by Felsenstein (see Felsenstein, 1981 and Felsenstein, 1973).

We can define the likelihood vector $\mathbf{L}^{i,j}$ of length n_α in such a way that each of its components represents the likelihood for a different character state: $\mathbf{L}^{i,j} = (L^{i,j}(A),L^{i,j}(C),L^{i,j}(G),L^{i,j}(T))$.

The likelihood at the tip i for character j is

$$L_z^{i,j} = \begin{cases} 1, & \text{if } z = d_{i,j} \\ 0, & \text{otherwise} \end{cases} \tag{1.2.5}$$

The algorithm proceeds using the given substitution model to calculate the likelihood at the parent node, given the likelihood vectors of the two children nodes and the branch lengths.

Each substitution model needs to satisfy 3 properties:

1. The process is memoryless, meaning that probability of future events depends only on the present state and not on past ones;
2. The matrix Q_S is constant over time;
3. The relative frequencies of characters, π_z with $z \in \alpha$, are at equilibrium.

Each substitution model is associated with an $n_\alpha \times n_\alpha$ matrix Q_S, where n_α is the length of the alphabet or, in other words, the number of possible states. In the case of DNA, for example, we have $n_\alpha = 4$. Each matrix entry represents the transition probability per unit time, i.e. the transition rates, from each of the 4 states to all the others. Entries on the main diagonal represent the probabilities per unit time of remaining in the same state (i.e. to have no transition at all in a time interval dt).

The simplest substitution model is the one by Jukes, Cantor, et al. (1969) (also known as JC69), in which all the transitions have equal probabilities per unit time. Its associated matrix is

$$Q_{JC69} = \mu \begin{bmatrix} -\frac{3}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & -\frac{3}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & -\frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & -\frac{3}{4} \end{bmatrix} \tag{1.2.6}$$

Here μ is the only parameter of the model which expresses the rate at which mutations occurs. Substitution models with greater complexity can also be used as in the case of HKY85 (Hasegawa, Kishino, and Yano, 1985), TN93 (Tamura and Nei, 1993) or GTR (Tavaré, 1986).

The rate matrix Q_S defines a system of ODE that can be solved to obtain the dynamics of $\mathbf{L}^{i,j}$

$$\frac{d\mathbf{L}^{i,j}(t)}{dt} = Q_S \mathbf{L}^{i,j}(t) \tag{1.2.7}$$

Its solution at time t_b, given the initial conditions at t_a, is given by

$$\mathbf{L}^{i,j}(t_b) = \Pi(t_a, t_b)\mathbf{L}^{i,j}(t_a) \tag{1.2.8}$$

where $\Pi(t_a,t_b)$ is a $n_\alpha \times n_\alpha$ matrix defined by the matrix exponential formula

$$\Pi(t_a,t_b) = e^{Q_S(t_b-t_a)} := \sum_{k=0}^{\infty} \frac{[Q_S(t_b-t_a)]^k}{k!} \tag{1.2.9}$$

From these we can calculate back one by one the likelihood at the internal nodes. The $z-$th component of the likelihood of the internal node i giving rise to species o_1 and o_2 is

$$L_z^{i,j} = \Big[\sum_{x\in\alpha} \Pi_{z,x}(t_i,t_{o_1}) L_x^{o_1,j}\Big] \cdot \Big[\sum_{y\in\alpha} \Pi_{z,y}(t_i,t_{o_2}) L_y^{o_2,j}\Big] \tag{1.2.10}$$

The process can be re-iterated up to the crown node. From the likelihood at the crown we can finally compute

$$P(D|T,\theta_T,\theta_S) = \prod_{j=1}^{n_c} \sum_{z\in\alpha} \pi_z L_z^{2n_s-1,j} \tag{1.2.11}$$

where π_z is the equilibrium frequency for the character z. The final likelihood is obtained as the product of the terms for each character taken separately, as they are usually assumed to be independent from each other.

1.2.4 The tree prior

Using the definition of conditional probability we can rewrite the prior

$$P(T,\theta_T) = P(T|\theta_T)P(\theta_T) \tag{1.2.12}$$

The quantity $P(T|\theta_T)$ is called "tree prior" within the Bayesian context. This is the name we will use in chapter 5.

This quantity can also be seen as a likelihood if we see the phylogeny as data, according to eq. 1.2.1. This function is the core of chapters 2, 3 and 4, where it will be referred to simply as "likelihood function". From now on, we will always refer to this when talking about likelihood.

1.3 Likelihood modeling

When talking about likelihood we refer to a function that expresses the probability for a diversification model and its parameters to explain a diversification process that results in the tree we observe

$$\mathscr{L}_{M,\theta_M}(D) = P(T|M,\theta_M). \tag{1.3.1}$$

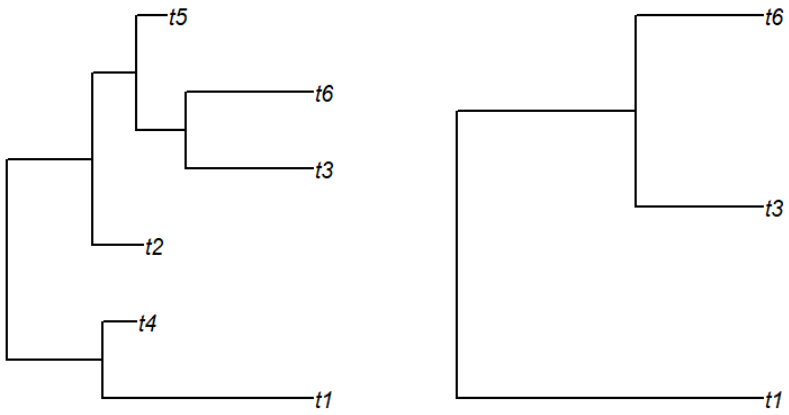

Figure 1.3: On the left panel a complete phylogeny. On the right panel a reconstructed phylogeny obtained by pruning extinct species from the complete one.

Outside of the Bayesian framework, likelihood maximization could be exploited in order to: (1) estimate the best parameters for the given model to obtain relevant information about the phylogeny; (2) perform model selection among two or more candidate models using tools like Akaike Information Criterion (AIC, Akaike, 1998) to establish which is the most likely process that generated the data.

1.3.1 Complete tree vs reconstructed tree

When calculating the likelihood for a phylogeny it is important to keep in mind that usually phylogenies we have access to do not show extinct species. Observed phylogenies are the result of pruning complete trees from extinct species. The result of this operation is called reconstructed tree. The difference between reconstructed and complete trees is shown in Fig. 1.3.

1.3.2 The Birth-Death model

We can interpret a phylogeny as the realization of a birth-death (BD) process, where each "birth" is a speciation event and each "death" is an extinction event. The symbols λ and μ will denote the per-species speciation and extinction rates,

respectively. They can be functions of time. The process can be described by the master equation for the variable $P_n(t)$ representing the probability of having n species at the time t

$$\begin{aligned}\frac{dP_n(t)}{dt} &= \lambda(t)(n-1)P_{n-1}(t) + \mu(t)(n+1)P_{n+1}(t) \\ &\quad - (\lambda(t)+\mu(t))\,nP_n(t).\end{aligned} \tag{1.3.2}$$

Its solution has been provided by Kendall, 1948b

$$\begin{aligned}P_0(t) &= 1-p(t), \\ P_n(t) &= p(t)(1-u(t))\,u(t)^{n-1}, \qquad \text{if } n>0,\end{aligned} \tag{1.3.3}$$

where

$$\begin{aligned}p(t) &= \frac{1}{1+\int_0^t \mu(t)\exp[\rho(t,\tau)]d\tau} \\ u(t) &= 1-p(t)\exp[\rho(0,t)] \\ \rho(t,\tau) &= \int_t^\tau [\mu(s)-\lambda(s)]ds\end{aligned} \tag{1.3.4}$$

This solution has been used by Nee, May, and Harvey, 1994 to assign likelihoods to phylogenies. They show that, if phylogenetic branches are independent from each other, the tree can be "broken", meaning that each branch can be considered separately from all the others. The branch originating at time t_i contributes by a factor $P_1(t_p-t_i)$ to the likelihood, as reported in 1.3.3. The likelihood for the entire phylogeny, conditioned on its survival, is thus proportional to the product of a P_1 term for each of the branches

$$\mathscr{L}(\lambda,\mu|\mathbf{t}) = \lambda^{n_s-2}(n_s-1)!\frac{\prod_{i=1}^{n_s}P_1(t_p-t_i)}{(1-P_0(t_p-t_c))^2}. \tag{1.3.5}$$

The factor λ^{n_s-2} comes from each of the speciations occurring on the n_s-2 internal nodes (excluding the crown, because we condition on the crown event) of the phylogeny. The factor $(n_s-1)!$ takes into account the number of possible topologies compatible with the set of branching times $\mathbf{t}=t_1,t_2,\ldots,t_{n_s}$ (here the first two time points $t_1=t_2$ represent the time at which the crown species are created, i.e. the crown age t_c). The division by $(1-P_0(t_p-t_c))^2$ conditions the likelihood on the survival of the crown species.

The result from Nee et al. is extremely important because their likelihood is analytical (and therefore extremely fast to compute) and it provides a solid and

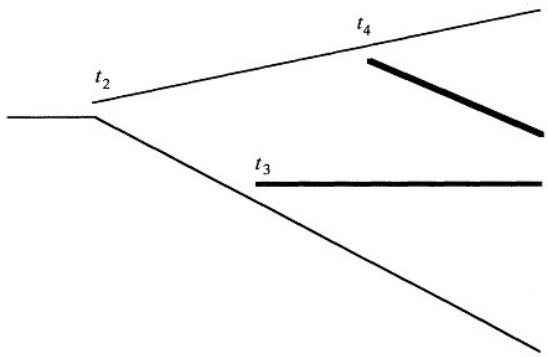

Figure 1.4: If branches are independent we can assign a likelihood to a phylogeny by "breaking the tree" and assigning a likelihood factor separately to each individual branch (Ref. Nee, May, and Harvey, 1994).

simple reference model that can be very useful for developing and testing new models. A prime example of this is represented by the model we present in chapter 4, which heavily relies on Nee et al.'s framework.

1.3.3 The Q-framework

The Q-framework has been originally developed to deal with diversity-dependent processes (Etienne et al., 2012). The process is described, as in the case of the BD model, by an ODE set

$$\begin{aligned}\frac{dQ_m^k(t)}{dt} &= \mu_{k+m+1}(m+1)Q_{m+1}^k(t) + \lambda_{k+m-1}(m-1+2k)Q_{m-1}^k(t) \\ &\quad -(\lambda_{m+k}+\mu_{m+k})(m+k)Q_m^k(t), \quad \forall m>0, \\ \frac{dQ_0^k(t)}{dt} &= \mu_{k+1}Q_1^k(t) - (\lambda_k+\mu_k)kQ_0^k(t), \qquad \text{if } m=0. \end{aligned} \tag{1.3.6}$$

where the quantity $Q_m^k(t)$ is the probability for the process to be compatible with the phylogeny up to time t, having k visible species in the phylogeny as well as additional m species that cannot be observed because they are not sampled or they go extinct before the present. We will not explain the model in full detail here, because it is extensively described in chapter 2, where the full procedure to calculate the likelihood is provided.

This framework entails a fundamental difference with the standard Nee et al. approach. In fact within the Q-framework it is possible to keep track, at each

time point, of the probability distribution with respect to any number of missing species m. This is possible because the quantity Q is directly used to calculate the likelihood, whereas in the Nee et al.'s approach the likelihood is combined by breaking the tree, combining independent elements from each branch. In the literature, such feature has allowed for the development of diversity-dependent models for single clades as well as multiple clades colonizing an island in DAISIE (Valente, Phillimore, and Etienne, 2015). Also in this thesis this approach has been heavily employed, as in chapters 2 and 3.

1.3.4 Simulating phylogenies

A BD process can be easily simulated by exploiting the Doob-Gillespie algorithm (Gillespie, 1976; Gillespie, 1977). Simulated processes start at the crown age, with only the two crown species present in the species pool. Each iteration of the algorithm consists of three steps:

1. A waiting time for the next even is sampled from an exponential distribution

$$P(\Delta t = X) = r_{tot} \mathrm{e}^{-r_{tot} X} \tag{1.3.7}$$

   where r_{tot} is the total rate, which in the case of the standard BD model is just $r_{tot} = n(\lambda + \mu)$;

2. An event is sampled from the pool of all the possible events. In case of a BD process

$$P(e = e_\lambda) = \frac{n\lambda}{r_{tot}}, \quad P(e = e_\mu) = \frac{n\mu}{r_{tot}}; \tag{1.3.8}$$

3. The pool and the running time are updated. In a BD process, when a speciation (extinction) occurs the parent (extinct) species is sampled from the current pool. In case of a speciation a new species is added to the pool and the topology is affected by the parent species. In case of extinction the extinct species is removed from the pool. The time is updated just by adding Δt to the current time.

The routine stops when the time reaches or exceeds the present time. This algorithm has been used in chapters 3 and 4 to simulate datasets of phylogenies.

1.3.5 Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) is a technique that can be used to infer relevant information from phylogenies. By maximizing the likelihood in the model's parameter space we can obtain the parameter values that best explain the data, according to the chosen model. Writing a likelihood function is usually not an easy task when dealing with phylogenies, because they are very complex objects, composed of continuous components (the branches) as well as discrete ones (the nodes). Furthermore, as mentioned before, we always deal with incomplete data, as information about extinct species is usually not available. For this reason, before operating an MLE on actual data, it is generally considered good practice to assess the efficacy of the likelihood function on a dataset of simulated data. Applying the procedure to data generated according to known parameters, we can effectively assess whether the model is able to recover the original parameters. The procedure is shown in Fig. 1.5.

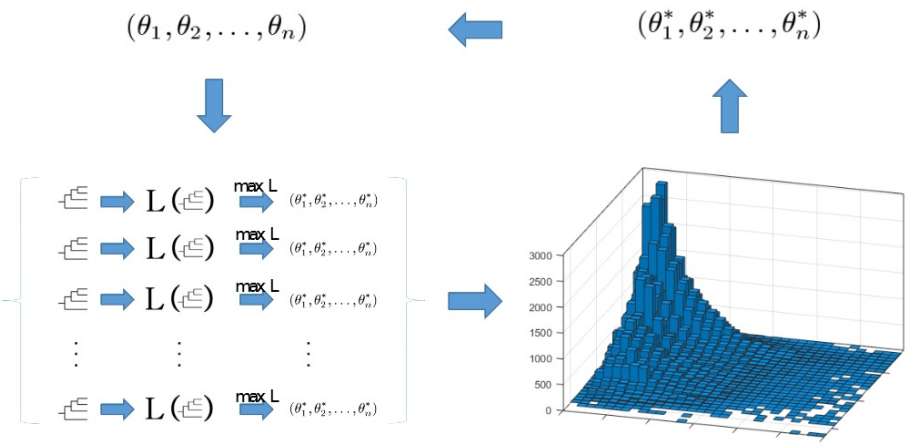

Figure 1.5: Assessment of the model using MLE. (Top-left) The procedure starts for a set of known parameters. (Bottom-left) The parameters are used to simulate a dataset of trees, from which parameters are estimated through maximum likelihood. (Bottom-right) We collect all the MLE results in a distribution. (Top-right) The median of the distribution is computed and compared to the original values.

We validated the models in chapters 3 and 4 using this procedure.

1.4 Thesis outline

In this thesis we discuss inference in phylogenetics through likelihood modeling. The main element around which each chapter revolves is always a likelihood model.

In chapter 2 we provide a stronger mathematical foundation to the Q-framework. This was needed as in the original publication the model had only been proven numerically. We instead prove the equivalence of the likelihood with several current models, in the cases where this is possible. We did so because we need the framework for the next two chapters.

In chapter 3 we developed a new mathematical model to describe fast-paced diversification processes. We did so by adding a new (diversity-dependent) event alongside standard speciation and extinction ones. This new event is driven by the environment and it could potentially lead to multiple speciations at the same time. The code to run the model is wrapped in a R package (Laudanno, 2020a) and is publicly available at `https://github.com/Giappo/mbd`.

In chapter 4 we highlight the problem with some of currently available models dealing with rate shifts occurring on a single lineage within a phylogeny. This has been applied, for example, to the case of the diversification process where the development of a key innovation allows a specific lineage to escape competition with other species in the clade (Etienne et al., 2012; Etienne and Haegeman, 2012). We develop a new likelihood formula in the contexts of both the P-framework (for the diversity-independent process) and the Q-framework (for the diversity-dependent one). We test the model maximizing the likelihood on datasets of simulated phylogenies. The code to run the model is wrapped in a R package (Laudanno, 2020b) and is publicly available at `https://github.com/Giappo/sls`.

In chapter 5 we present an R package called `pirouette`, to assess the error made in Bayesian inference when using standard inference models. It does so by taking an input phylogeny and using it to simulate an alignment. Then the alignment is passed to BEAST which uses an inference model to yield a posterior distribution of trees. The inference error is estimated by comparing the posterior with the original tree using a given tree statistic. The R code to run `pirouette` is publicly available at `https://github.com/richelbilderbeek/pirouette`.