Stockholms
universitet

The sampled protracted speciation model in macro-evolution: Calculating its probability density and a comparison with two other models

Siyamak Kaffashi

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2012:5
Matematisk statistik
Maj 2012

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Master Thesis **2012:5**
http://www.math.su.se

The sampled protracted speciation model in macro-evolution: Calculating its probability density and a comparison with two other models

Siyamak Kaffashi[*]

May 2012

Abstract

There have been several models suggested to model macro-evolution. Recently, it has been shown that in the birth-death process, diversified sampling is a better fit to empirical data (phylogenetic trees) than complete sampling and other sampling methods. This implies that taxon sampling (species sampling) has a large impact on estimating the diversification parameters. Besides, the protracted speciation model has been suggested to explain the observed slowdowns in the lineages through time plots. The protracted speciation model assumes that speciation events do not happen instantaneously, and it takes some time for a new born species (incipient species) to become a good species. The incipient lineages with dead good parent are also considered as good in the protracted speciation model. The likelihood functions have been derived for the diversified sampling and the random sampling models. Deriving the likelihood function for the protracted speciation model has faced some difficulties. In the present thesis, we try to calculate the probability density for the protracted speciation model by applying a method that has been suggested to calculate the probability of the Binary State Speciation Extinction (BiSSE) model. Unfortunately, we cannot provide an analytical closed form solution of the probability density function. Thus, we suggest the sampled protracted speciation model based on the sampling probability for good and incipient species, and explain the calculation of the probability density, using BiSSE approach. We compare the sampled protracted speciation model with the diversified and random sampling models, using simulated trees and empirical data. In this thesis, we show that the inference on the simulated trees is more accurate when the simulation and the inference methods are the same. We also show the effect of the transition (speciation completion) rate on the inferred parameters of the sampled protracted speciation model. The thesis also illustrates that the sampling probability of the incipient species does not affect the estimated parameters.

[*]Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: Siyamak.kaffashi@gmail.com. Supervisor: Sebastian Höhna.

Acknowledgements

I would like to sincerely thank my supervisor, Sebastian Höhna, for guiding me throughout this Master thesis with his great knowledge, patience and enthusiasm. I also would like to thank my family and friends who encouraged and supported me during my work on this thesis.

Contents

1 Introduction

The history of life has been always one of the essential questions of the human. In the 19^{th} century, the theory of Evolution by Charles Darwin changed biology dramatically; and Phylogenetics, which infers evolutionary trees from genetic data (Nei and Kumar, 2000), has influenced evolutionary biology drastically. In Phylogenetics, estimation of diversification rates has always been of interest. Modeling the birth-death process can be used to estimate the speciation and extinction rates (Nee et al., 1994). Speciation occurs when a species splits into two new species, and extinction happens when a species becomes extinct. λ and μ are the rates of the speciation events and the extinction events, respectively. Different methods have been proposed to model speciation and extinction events in phylogenetic trees. The constant rate birth-death process (Kendall, 1948) is a stochastic process that has been frequently used to model speciation and extinction events in phylogenetic trees. One species, in the past, gives birth to a new species with the speciation rate, λ, or dies with the extinction rate μ. The process of speciation and extinction produces a phylogenetic tree that includes both the species which has gone extinct and the species that survive to the present time (see Figure 1, left). These trees are called complete trees. The trees that only include the survived lineages are called extant or reconstructed trees (see Figure 1, right). Nee et al. (1994) showed that the diversification rates can be estimated from extant trees. Höhna et al. (2011) inferred the speciation and the extinction rates from incomplete trees. They compared the random sampling scheme, the cluster sampling scheme and the diversified sampling scheme. Sampling means that only a subsample of the extant species is used to reconstruct the phylogenetic tree. Höhna et al. (2011) showed that the diversified sampling model often fits the data significantly better than the other sampling schemes.

Etienne and Rosindell (2012) suggested the protracted speciation model based on the fact that speciation events do not happen instantaneously. The protraction of speciation events yields two types of species, namely the good species and the incipient species. After a speciation event, the new born species is an incipient species and it takes some time that the speciation completes and the new species becomes good (see Figure 2). As Etienne and Rosindell (2012) defined, in the protracted speciation model, only good species and the incipient species with extinct good parents are considered in the reconstructed tree. This model justifies the slow down observed in the lineages through time plots. There are five parameters defined for this model; the birth rates for the good and incipient species (λ_g and λ_i), the death rates for the good and the incipient species (μ_g and μ_i), and the transition rate (or the speciation completion rate) for incipient species becoming good species, λ_{ig} (the good species cannot become incipient). The incipient species with dead good parent creates some difficulties to calculate the probability density of the tree in the protracted speciation model. Thus, there has been no likelihood function derived for this model. Here, we applied the method of Maddison et al.(2007) to approximate the probability density. Maddison et al. (2007) developed the Binary State Speciation and Extinction Model (BiSSE) that uses the numerical integration to solve some differential equations that contain the probability density of the trees.

Unfortunately, applying the BiSSE method, the incipient species with dead parent that are included in the reconstructed tree caused some difficulties and prevented us from calculating the probability density. Therefore, in this thesis we suggest the new model based on the sampling probability of the good species (ρ_g) and the sampling probability of the incipient species (ρ_i) to overcome the problem. We call this new model, which is an extension of the protracted speciation model of Etienne and Rosindell (2011), the sampled protracted speciation model. In the sampled protracted speciation model, the incipient species, with either good or incipient parent species, have the same probability to be included in the reconstructed tree. To be more close to the original protracted speciation model, in this thesis, we assumed that the sampling probability for the good species is equal to one. To calculate the probability density of the sampled protracted speciation model, we used the BiSSE method.

Empirical reconstructed trees have shown a slowdown in diversification (McPeek(2008) and Phillimore and Price (2008)). Under the pure-birth process, diversification increases log-linearly; under the birth-death process the diversification increases log-linearly at the early phase and much steeper close to the present (which is called the "pull-of-the-present" effect). Both processes cannot explain the observed slowdown in diversification. Other researchers advocated for diversity dependent rates (e.g. Etiene et al. (2012)) whereas we consider it as an artefact of incomplete sampling (e.g. Höhna et al. (2011)) in combination with protracted speciation (Etiene and Rosindell, 2011)

By using the probability densities of the three models, we found ML estimates of the model parameters for both the simulated trees and the empirical data set. The results suggest that the parameter estimates are more accurate when the simulation procedure and the inference procedure are the same. Moreover, the results show that the increase of the transition rate causes the diversification rates for the good species decrease and the diversification rates for the incipient species increase.

After the introduction, we describe three models and point out the likelihood functions of the random sampling and diversified sampling models, and we explain the calculation of the probability density for the sampled protracted speciation model. In section 4, we describe the results for the simulations and the empirical data set. Section 5 is the discussion and finally, in the appendix we point out the unsuccessful attempts to calculate the density of the protracted speciation model.

2 Aims of the study

◇ Finding the probability density for the sampled protracted speciation model:

There has been not any likelihood function for the protracted speciation model, here, we try to calculate the probability density for the sampled speciation model by using Binary State Speciation and Extinction (BiSSE) method based on numerical integration.

◇ Comparing the sampled protracted speciation model with random sampling and diversified sampling models of the birth-death process:

We simulate trees under each model and we do the inference applying all three models on each. We also fit three models to the empirical data.

3 Models

3.1 The Birth-Death Process

In Phylogenetics, one of the methods, which often has been applied to model speciation and extinction events among species through time, is the constant-rate birth-death process (BDP)(Kendall, 1948). The Birth-death process is a continuous-time stochastic process specified by birth rates (speciation rates) and death rates (extinction rates). The process begins with one lineage at time t_0 in past and moves forward to the present time. The number of lineages in the process can vary through time since a linage can either branch into two lineages (gives birth to another lineage) or go extinct (dies). In the phylogenetic tree (τ), the probability that a speciation event happens in a short time interval Δt is $\lambda \Delta t$, and likewise, the probability that a lineage goes extinct in Δt is $\mu \Delta t$ (Rannala and Yang, 1996). Parameter λ is the rate of speciation (birth rate) and parameter μ is the rate of extinction (death rate) in the birth-death process. In order for a process to survive, it has to be assumed that the birth rate is greater than the death rate ($\lambda > \mu$). Here, we also assume that all speciation and extinction events in the process are independent. The number of lineages at the present time (n) and the time of divergence (t) are random variables under the model with the distributions specified by parameters λ and μ. In the complete phylogenetic tree, the time until the next event follows an exponential distribution with rate $n(\lambda + \mu)$.

The birth-death process shows an accelerating increase in the number of lineages through time. From the beginning, the lineages in a reconstructed phylogeny accumulate with rate $\lambda - \mu$ (it is called net diversification rate). However, the rate of accumulation changes from $\lambda - \mu$ to λ as it approaches very recent past. This change is called pull of the present (Stadler, 2011).

In Phylogenetics, studying the millions of years in the past is not possible easily, so we reconstruct the past on the basis of the molecular data from fossil records or living species. Studies on the basis of living species help us to understand speciation and extinction process for the large number of groups without a good fossil record.(Stadler, 2011).

The birth-death process produces complete phylogenies with both extant and extinct species (Figure 1; left). The extant species are the lineages which survive to the present time. Molecular data are rarely available for extinct species; thus, we have to infer phylogenetic trees from extant lineages. In 1994, Nee et al. proposed an inference method based on reconstructed phylogenies, i.e. phylogenies with only extant species in which all extinct species are deleted. A reconstructed tree is also called extant tree (Figure 1; right). We can estimate the speciation and extinction rates from both complete and incomplete trees (Höhna, 2011). Usually, in phylogenetic studies data is not entirely available for all species of a clade; thus, including sampling helps us to analyze the process more precisely. As I mentioned above, λ (speciation rate) and μ (extinction rate) are the parameters and t_0 is the initiation time or time of origin of the birth-death process. We take m as the number of extant species observed at the present time. We define $\rho = n/m$, the sampling fraction, as another parameter of the birth-death process. Here, n $(2 \leq n \leq m)$ is the number of observed sampled lineages out of m extant species. The sample fraction (ρ) is considered as a fixed parameter in the study (Höhna et al., 2011).

In order to obtain the probability distribution of a tree, we need to condition on some aspect of the tree. It could be either the number of species or the age of the tree. Since we usually do not have any information about the time of origin of the tree (t_0), it is very common to condition on sampling n species out of the m extant species, and assume a uniform prior distribution for the time of origin on $(0, \infty)$ (Höhna, 2011). When we condition on the number of sampled species, we assume that we have n species at the present time and construct the tree according to these n species. When we condition on the time of the origin, we assume that there is one species in a specified time in the past, and it evolves through time to the present with some number of extant species. The second approach, as we believe, is more natural so, in the remainder, we assume that we have some information about the time of the tree. In practice, we usually have data on the branching point of the first two ancestral lineages; then, the information we have is more related to the time of the most recent common ancestor of the sampled tree (t_1) than the time of origin (t_0). This time is also called crown age of the tree. In this case, instead of assuming a uniform prior distribution for the time of origin (t_0), we condition on the time of the most common recent ancestor, t_1.

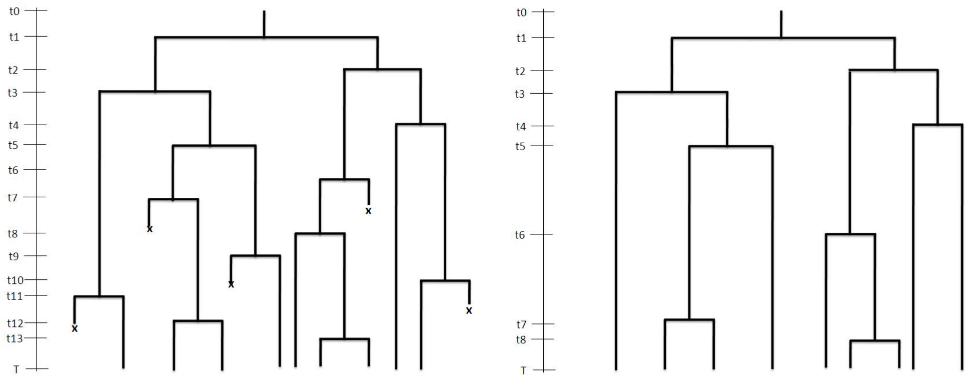

Figure 1: Complete tree (left) and corresponding reconstructed tree (right)

3.1.1 Random Sampling

In the constant-rate birth-death process conditioned on the time of the most recent common ancestor(t_1), we have m lineages as the number of extant species at the present time. There are two scenarios applicable for random sampling (Stadler, 2009). Under $n - sampling$, we take n lineages out of m extant species, uniformly at random, with a fixed sampling fraction ($m = n/samplingfraction$). Applying $\rho - sampling$, each extant species is sampled with a fixed sampling probability ρ. In this thesis, we apply the second approach of random sampling.

Likelihood Function

Nee et al.(1994) derived the probability density of the reconstructed tree conditioned on the time of the most recent common ancestor and the probability density of the random sample of the tree conditioned on the time of the most recent common ancestor, t_1.

$$f(\tau|t_1) = \lambda^{n-2}(1 - u_{t_1})^2 \prod_{i=2}^{n-1} P(t_i) \prod_{i=2}^{n-1}(1 - u_{t_i})$$

where

$$u_t = \frac{\lambda(1-e^{-(\lambda-\mu)t})}{\lambda-\mu e^{-(\lambda-\mu)t}}$$

and

$$P(t) = \frac{\rho(\lambda-\mu)}{\rho\lambda+(\lambda(1-\rho)-\mu)e^{-(\lambda-\mu)t}}$$

is the probability that a single lineage alive at time t has not gone extinct until the present time and has some sampled descendants. (ρ is the sampling probability)

This is derived from the probability that a single lineage alive at time t has some decendent and not has become extinct by the present time (Kendall, 1948) ($\rho = 1$)

3.1.2 Diversified Sampling

One of the methods of taking a sample of a phylogenetic tree is diversified sampling scheme. Höhna et al.(2011) show that diversified sampling fit the data better than complete, cluster and random sampling. In diversified sampling, we sample the n most diversified lineages out of m extant lineages in a phylogenetic tree. We assume the sample fraction is fixed ($\rho = n/m$) and m is the total number of sampled and non-sampled extant species. These n lineages are the most distant lineages in the tree. It means that the sum of edge lengths that connect these sampled lineages is maximized. We can obtain a diversified sample by choosing the first $n-1$ born lineages in the tree. (Höhna, 2011)

Likelihood Function

Höhna (2011) derived the probability density function for diversified sampling conditioned on the time of the most recent common ancestor of the sample tree as

$$f_{Ds}(\tau|t_1) = \tbinom{m-1}{n-1}(\tfrac{\lambda P_0(t_1)}{\mu})^{m-2}\tfrac{p_1(t_1)^2}{(1-p_0(t_1))^2}F(t_{n-1}|t_1)^{m-n}\prod_{i=2}^{n-1} f(t_i|t_1)$$

where

$$p_0(t) = \tfrac{\mu(1-e^{-(\lambda-\mu)t})}{\lambda-\mu e^{-(\lambda-\mu)t}}$$

is the probability that a linage goes extinct before time t in a full extant tree, and

$$p_1(t) = \tfrac{(\lambda-\mu)^2 e^{-(\lambda-\mu)t}}{(\lambda-\mu e^{-(\lambda-\mu)t})^2}$$

is the probability that a linage leaves exactly one descendent after time t in a full extant tree (Kendall, 1949), and

$$f(s|t_1) = \mu\tfrac{p_1(s)}{p_0(t)}$$

is the density function of a speciation event happened at time s ($0 \leq s \leq t_1$), and

$$F(s|t_1) = \tfrac{p_0(s)}{p_1(t)}$$

is its distribution function (Thompson, 1975).

3.2 The Protracted Speciation Model

As we said before, the constant rate birth-death process shows an increase in the number of lineages; this is called pull of the present. The pure birth process also shows an increase in the number of lineages, named the log-linear increase. On the contrary, the plot of lineages through time in the phylogenetic trees often shows a remarkable slowdown toward the present (Phillimore & Price (2008) and McPeek (2008)). This phenomenon cannot be explained by the constant rate birth-death process. There have been two explanations proposed for this slowdown in the very recent time in the lineages through time plot. First, taking a small sample from the actual phylogeny may cause the slowdown (Nee et al., 1994). Second, Purvis et al.(2009) proposed another explanation for the phenomenon implying the slowdown is because of the non-constancy of speciation and extinction rates on different levels of trees.

Etienne and Rosindell (2012) offered a new model to explain the slowdown based on the fact that a speciation event does not happen instantaneously. This is called the protracted speciation model. In the protracted speciation model, there are two types of species involved in the model. These two types are *good* species and *incipient* species. In a protracted speciation model, both, good and incipient species can speciate and go extinct. When a good species or an incipient species speciates, the new born species is an incipient species. The incipient species can transform into good species after some time; it is called that the speciation completes. Rosindell et al.(2010) defined the protracted speciation which considered a fixed time for the speciation completion. In the protracted speciation model, a rate of completion (transition) had been defined (Etienne and Rosindell, 2012). Therefore, five parameters are involved in this model. Speciation rates for good species and incipient species, extinction rates for good species and incipient species and one rate of speciation completion (or the rate of transition from an incipient species to a good species). In the protracted speciation model, all good extant species are involved in the reconstructed phylogeny. Besides, the extant incipient species that had a good extinct parent are considered as good species, so they are involved in the reconstructed tree(Etienne and Rosindell, 2012). Figure 2 shows a complete protracted speciation model and its related reconstructed tree.

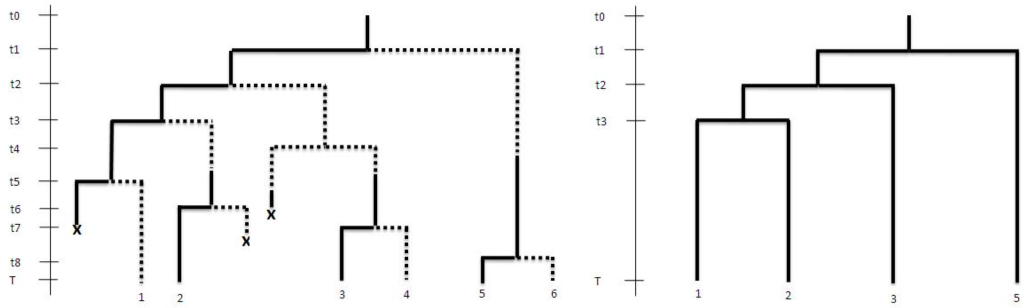

Figure 2: Complete protracted speciation model (left) and the reconstructed tree (right)

3.2.1 Likelihood Derivation

Finding the probability density function for the protracted speciation model

In the protracted speciation model, incipient species with a good extinct parent (like species 1 in Fig.2), which are considered as good species, cause some difficulties to find an expression for the likelihood of the protracted speciation model with extinction. Therefore, there is no probability density found for the general form of the protracted speciation model yet (Ettiene et al., 2012). Maddison et al. (2007) applied a new method based on numerical integration to calculate the likelihood for binary state speciation extinction models (BiSSE). In order to assess the effect of a binary state character on the birth and death rates using maximum likelihood method, Maddison et al. (2007) proposed a new method to calculate the probability for a phylogenetic tree and a binary state character under a simple model of birth-death process. The model involves six parameters including two speciation rates (one for each character state), two extinction rates, and two transition rates. (from state 0 to state 1, and vice versa). Calculations start from the tips of the tree with known character states and move backward in time to the root of the tree, considering all possible scenarios (all possible combinations of events) along branches and at nodes, yields the probability of the tree when the root is reached. Here, we tried to develop the probability calculation method of BiSSE to derive the probability density function for the protracted speciation model. Nevertheless, applying BiSSE method, we also could not manage to find the likelihood expression, exactly, due to the abovementioned problem. We tried two different approaches to calculate the probability that a species evolve to the present time according to the protracted speciation model. We point out these approaches and the related difficulties briefly in the appendix.

3.3 The Sampled Protracted Speciation Model

As we mentioned, in the protracted speciation model, all good species and the incipient species with a good extinct parent are observed in the reconstructed tree; the incipient species with a good extinct parent cause the difficulties which prevent us from deriving the likelihood function. To overcome this problem, here, we suggest an alternative model to the protracted speciation model which is the sampled protracted speciation model. In the sampled protracted speciation model, we define two sampling probabilities, ρ_g and ρ_i, for good species and incipient species, respectively. In this model, all good extant species have a chance to be observed in the reconstructed tree equal to the sampling probability ρ_g; however, in the protracted model, all of the good species are included, hence, we assume ρ_g being equal to 1. In a similar way, all extant incipient species (regardless of the parent state) have a chance to be observed in the reconstructed tree equal to the sampling probability ρ_i. As Figure 3 shows, a sampled reconstructed tree can include extant incipient species with observed good parent (lineages 6).

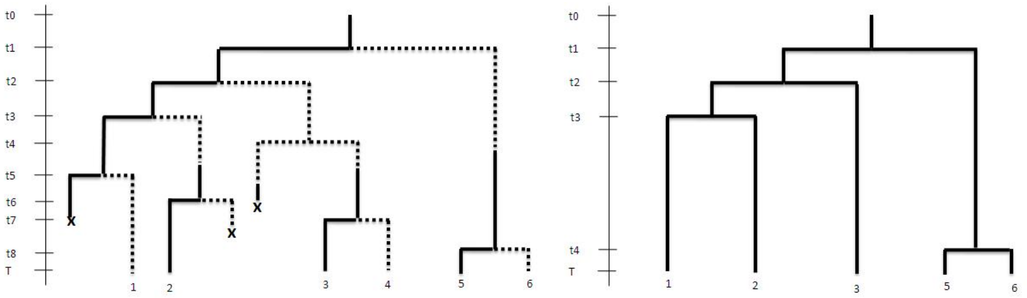

Figure 3: Complete protracted speciation model (left) and a sampled reconstructed tree (right)

3.3.1 Likelihood Derivation

Finding the probability density function for the sampled protracted speciation model

To calculate the probability of a lineage evolving to the present time in a sampled tree of the protracted speciation model, we apply a modification of BiSSE method. In the BiSSE method, a state change is not allowed at the speciation events while in our approach we know that when a good species gives birth to a new species, the state change is inevitable. Besides,

since in this approach, the transition from a good state to an incipient state is not allowed (a good species cannot become an incipient one), we assume the rate of the transition from a good state to an incipient state is fixed, and it is equal to zero whereas in the binary state speciation extinction model, states are exchangeable.
Like the BiSSE method, we want to calculate the probability of the tree by considering all possible scenarios on all branches and nodes of the reconstructed tree. We start from the tips in the tree at the present time ($t = 0$) and move backward in time to the origin of the tree ($t = T$). We condition on that the character states (good or incipient species) are known for all terminal taxa (tips). We calculate the probability of a lineage at time t that survives to the present time (the probability that the species at time t has some observed descendants in our reconstructed tree). We start from the tips of the tree and calculate this probability. Since we are required to cover all scenarios and events which might have happened towards the origin of the tree, in order to calculate the probability of survival of a lineage at time t, we are supposed to calculate the probability of a new born lineage which has gone extinct in the complete tree or the extant lineages which are not sampled in the reconstructed tree. Along a branch, we move backward in small time intervals and update the probabilities, step by step, until a node is reached. At each node, we have two arriving branches which form a new branch. We calculate the probabilities at the top of the new branch according to the two corresponding branches. Likewise, we perform our calculations to the origin of the tree. Once we reach the origin, we have gained the probability for the whole tree. We define four probabilities according to the different scenarios in the tree. These probabilities are:

$P_{Ng}(t)$: The probability that a good lineage at time t, through node N, has some sampled descendant(s) at the present time in the reconstructed tree.

$P_{Ni}(t)$: The probability that an incipient lineage at time t, through node N, has some sampled descendant(s) at the present time in the reconstructed tree.

$O_g(t)$: The probability that a good lineage at time t has no observed descendant at the present time in the reconstructed tree (the lineage either does not have any sampled descendant in the reconstructed tree at the present time or becomes extinct by the present time).

$O_i(t)$: The probability that an incipient linage at time t has no observed descendant at the present time in the reconstructed tree (the lineage either does not have any sampled descendant in the reconstructed tree at the present time or becomes extinct by the present time).

Including the sampling probabilities, we have seven parameters in our model. The parameters of the sampled protracted speciation model are:

λ_g : Speciation rate for good species.

λ_i : Speciation rate for incipient species.

μ_g : Extinction rate for good species.

μ_i : Extinction rate for incipient species.

λ_{ig} : Transition rate from incipient species to good species.

ρ_g : The sampling probability for good species.

ρ_i : The sampling probability for incipient species.

Calculations along a branch

Assuming that we know $P_{Ng}(t), P_{Ni}(t), O_g(t)$ and $O_i(t)$, we calculate the corresponding probabilities at time $t + \mathit{\Delta} t$ ($P_{Ng}(t + \mathit{\Delta} t), P_{Ni}(t + \mathit{\Delta} t), O_g(t + \mathit{\Delta} t)$ and $O_i(t + \mathit{\Delta} t)$), simultaneously. During the time interval $\mathit{\Delta} t$, different stories might have happened. We define $\mathit{\Delta} t$ small enough that only one event can happen in each interval. Figure 4 and Figure 5 show the alternative scenarios in a small interval for a good lineage and an incipient lineage at time $t + \mathit{\Delta} t$, respectively. Different scenarios for not observed lineages are treated below.

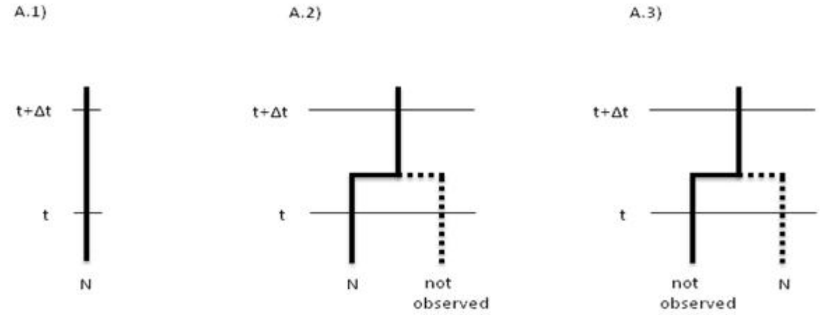

Figure 4: Alternative scenarios for a good lineage at time $t + \mathit{\Delta} t$ that has some sampled descendant(s) at the present time in the reconstructed tree through node N.

For $t > 0$ we have

A.1) No extinction, no speciation

$P_{A.1}(t+\mathit{\Delta} t) = (1-\mu_g \mathit{\Delta} t)(1-\lambda_g \mathit{\Delta} t)P_{Ng}(t)$

A.2) No extinction, speciation, the incipient lineage either survives to the present time and is not sampled or goes extinct by the present time

$P_{A.2}(t+\mathit{\Delta} t) = (1-\mu_g \mathit{\Delta} t)(\lambda_g \mathit{\Delta} t)P_{Ng}(t)O_i(t)$

A.3) No extinction, speciation, the good lineage either survives to the present time and is not sampled or goes extinct by the present time

$P_{A.3}(t+\mathit{\Delta} t) = (1-\mu_g \mathit{\Delta} t)(\lambda_g \mathit{\Delta} t)O_g(t)P_{Ni}(t)$

$$\begin{aligned} P_{Ng}(t+\mathit{\Delta} t) &= P_{A.1}(t+\mathit{\Delta} t) + P_{A.2}(t+\mathit{\Delta} t) + P_{A.3}(t+\mathit{\Delta} t) \\ &= (1-\mu_g \mathit{\Delta} t)(1-\lambda_g \mathit{\Delta} t)P_{Ng}(t) \\ &+ (1-\mu_g \mathit{\Delta} t)(\lambda_g \mathit{\Delta} t)P_{Ng}(t)O_i(t) \\ &+ (1-\mu_g \mathit{\Delta} t)(\lambda_g \mathit{\Delta} t)O_g(t)P_{Ni}(t) \end{aligned}$$

We assume that $\mathit{\Delta} t$ is a very small interval, so the multiple events within t are not possible. Then, for simplification, we can drop $\mathit{\Delta} t^2$, $\mathit{\Delta} t^3$, that represent multiple events probabilities. We have

$$P_{Ng}(t+\mathit{\Delta} t) = [1-(\lambda_g+\mu_g)\mathit{\Delta} t]P_{Ng}(t) + \lambda_g \mathit{\Delta} t P_{Ng}(t)O_i(t) + \lambda_g \mathit{\Delta} t O_g(t)P_{Ni}(t)$$

$$\frac{P_{Ng}(t+\mathit{\Delta} t) - P_{Ng}(t)}{\mathit{\Delta} t} = -(\lambda_g+\mu_g)P_{Ng}(t) + \lambda_g P_{Ng}(t)O_i(t) + \lambda_g O_g(t)P_{Ni}(t)$$

$$\frac{dP_{Ng}}{dt} = -(\lambda_g+\mu_g)P_{Ng}(t) + \lambda_g P_{Ng}(t)O_i(t) + \lambda_g O_g(t)P_{Ni}(t) \tag{1}$$

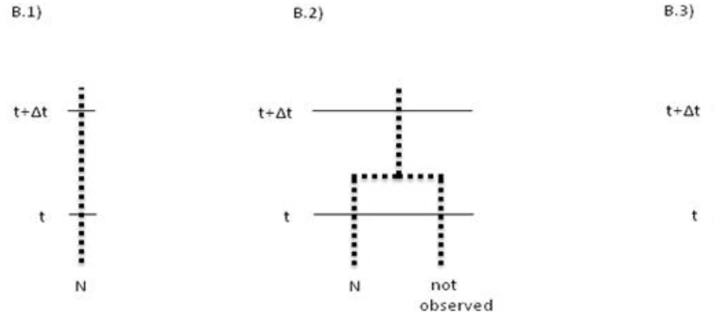

Figure 5: Alternative scenarios for an incipient lineage at time $t + \varDelta t$ that has some sampled descendant(s) at the present time in the reconstructed tree through node N.

For $t > 0$ we have

B.1) No extinction, no speciation, no transition

$P_{B.1}(t + \varDelta t) = (1 - \mu_i \varDelta t)(1 - \lambda_i \varDelta t)(1 - \lambda_{ig} \varDelta t) P_{Ni}(t)$

B.2) No extinction, speciation, no transition, and one incipient lineage either survives to the present time and is not sampled or goes extinct by the present time

$P_{B.2}(t + \varDelta t) = (1 - \mu_i \varDelta t)(\lambda_i \varDelta t)(1 - \lambda_{ig} \varDelta t) P_{Ni}(t) O_i(t)$

B.3) No extinction, no speciation, transition

$P_{B.3}(t + \varDelta t) = (1 - \mu_i \varDelta t)(1 - \lambda_i \varDelta t)(\lambda_{ig} \varDelta t) P_{Ng}(t)$

The total probability of an incipient species being included in the reconstructed trees is:

$$\begin{aligned} P_{Ni}(t + \varDelta t) &= P_{B.1}(t + t) + 2P_{B.2}(t + \varDelta t) + P_{B.3}(t + \varDelta t) \\ &= (1 - \mu_i \varDelta t)(1 - \lambda_i \varDelta t)(1 - \lambda_{ig} \varDelta t) P_{Ni}(t) \\ &+ 2(1 - \mu_i \varDelta t)(\lambda_i \varDelta t)(1 - \lambda_{ig} \varDelta t) P_{Ni}(t) O_i(t) \\ &+ (1 - \mu_i \varDelta t)(1 - \lambda_i \varDelta t)(\lambda_{ig} \varDelta t) P_{Ng}(t) \end{aligned}$$

$$P_{Ni}(t + \varDelta t) = [1 - (\mu_i + \lambda_i + \lambda_{ig})\varDelta t] P_{Ni}(t) + 2\lambda_i \varDelta t P_{Ni}(t) O_i(t) + \lambda_{ig} \varDelta t P_{Ng}(t)$$

$$\frac{P_{Ni}(t + \varDelta t) - P_{Ni}(t)}{\varDelta t} = -(\mu_i + \lambda_i + \lambda_{ig}) P_{Ni}(t) + 2\lambda_i P_{Ni}(t) O_i(t) + \lambda_{ig} P_{Ng}(t)$$

$$\frac{dP_{Ni}}{dt} = -(\mu_i + \lambda_i + \lambda_{ig}) P_{Ni}(t) + 2\lambda_i P_{Ni}(t) O_i(t) + \lambda_{ig} P_{Ng}(t) \tag{2}$$

We cannot solve the equations (1) and (2) analytically. So, given $O_g(t)$ and $O_i(t)$ which we will explain later in this section, we use numerical integration to find $P_{Ng}(t)$ and $P_{Ni}(t)$ at the bottom of each branch. If N is a good tip, the initial conditions ($t = 0$) are $P_{Ng}(t) = \rho_g$ and $P_{Ni}(t) = 0$, and if N is an incipient tip, the initial conditions ($t = 0$) are $P_{Ng}(t) = 0$ and $P_{Ni}(t) = \rho_i$.

$$P_{Ng}(t) = \begin{cases} P_{A.1}(t) + P_{A.2}(t) + P_{A.3}(t) & if \quad t > 0 \\ \begin{cases} \rho_g & \text{if } N \text{ is a good tip} \\ 0 & \text{otherwise} \end{cases} & if \quad t = 0 \end{cases}$$

$$P_{Ni}(t) = \begin{cases} P_{B.1}(t) + 2P_{B.2}(t) + P_{B.3}(t) & if \quad t > 0 \\ \begin{cases} \rho_i & \text{if } N \text{ is an incipient tip} \\ 0 & \text{otherwise} \end{cases} & if \quad t = 0 \end{cases}$$

Calculations at nodes

Once we calculate the probability of a branch coming from node N and, finally, when we reach node A, another branch coming from node M also has reached node A. The probability at node A and the corresponding branch is the product of probabilities of branches N and M at time t_A. Figure 6 shows the alternative scenarios at a node.

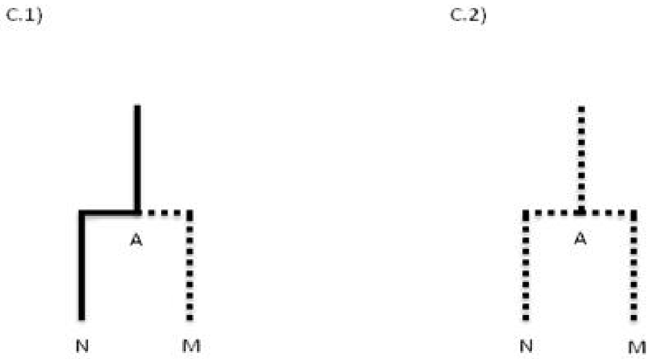

Figure 6: Alternative scenarios for a speciation event at time t_A in node A

At node A, we have

C.1) Speciation of a good species at time t_A (node A in the reconstructed tree)

$P_{C.1}(t_A) = \frac{1}{2}[P_{Ng}(t_A)P_{Mi}(t_A) + P_{Mg}(t_A)P_{Ni}(t_A)]\lambda_g$

C.2) Speciation of an incipient species at time t_A (node A in the reconstructed tree)

$P_{C.2}(t_A) = P_{Ni}(t_A)P_{Mi}(t_A)\lambda_i$

Since we do not know the state of node A, for a branch starting form node A the initial probabilities (at time t_A) are

$$P_{Ag}(t) = \begin{cases} P_{C.1}(t_A) & \text{if } A \text{ is a good species} \\ 0 & otherwise \end{cases}$$

$$P_{Ai}(t) = \begin{cases} P_{C.2}(t_A) & \text{if } A \text{ is an incipient species} \\ 0 & otherwise \end{cases}$$

Probabilities for the not observed species (Omitted species)

In order to consider all scenarios along a branch, we should also consider the lineages which are not observed. They either went extinct before the present time or, due to sampling probabilities, they were not sampled. Figure 7 and 8 show the alternative scenarios in a small interval for a good lineage and an incipient lineage which are not observed in the reconstructed tree.

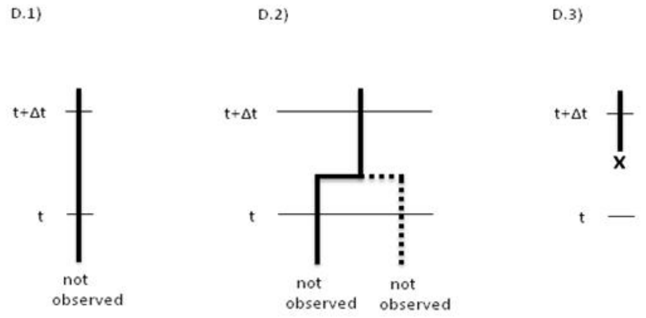

Figure 7: Alternative scenarios for a good lineage at time $t + \Delta t$ which has no observed descendant at the present time in the reconstructed tree

For $t > 0$

D.1) No extinction, no speciation, the good lineage either survives to the present time and is not sampled or goes extinct by the present time

$P_{D.1}(t + \Delta t) = (1 - \mu_g \Delta t)(1 - \lambda_g \Delta t) O_g(t)$

D.2) No extinction, speciation, none of the lineages is observed in the reconstructed tree (either because they survive to the present time and are not sampled or go extinct by the present time)

$P_{D.2}(t + \Delta t) = (1 - \mu_g \Delta t)(\lambda_g \Delta t) O_g(t) O_i(t)$

D.3) Extinction in Δt

$P_{D.3}(t + \Delta t) = \mu_g \Delta t$

The total probability of a good species having no observed descendants is:

$$\begin{aligned} O_g(t + \Delta t) &= P_{D.1}(t + \Delta t) + P_{D.2}(t + \Delta t) + P_{D.3}(t + \Delta t) \\ &= (1 - \mu_g \Delta t)(1 - \lambda_g \Delta t) O_g(t) \\ &+ (1 - \mu_g \Delta t)(\lambda_g \Delta t) O_g(t) O_i(t) \\ &+ \mu_g \Delta t \end{aligned}$$

$$O_g(t + \Delta t) = [1 - (\mu_g + \lambda_g)\Delta t] O_g(t) + \lambda_g \Delta t O_g(t) O_i(t) + \mu_g \Delta t$$

$$\frac{O_g(t + \Delta t) - O_g(t)}{\Delta t} = -(\mu_g + \lambda_g) O_g(t) + \lambda_g O_g(t) O_i(t) + \mu_g$$

$$\frac{dO_g}{dt} = -(\mu_g + \lambda_g) O_g(t) + \lambda_g O_g(t) O_i(t) + \mu_g \qquad (3)$$

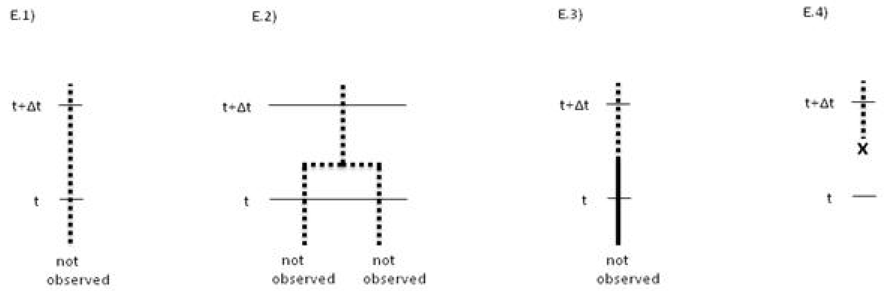

Figure 8: Alternative scenarios for an incipient lineage at time $t + \Delta t$ which has no observed descendant at the present time in the reconstructed tree

For $t > 0$

E.1) No extinction, no speciation, no transition, and the incipient lineage either survives to the present time and not is sampled or goes extinct by the present time

$P_{E.1}(t + \Delta t) = (1 - \mu_i \Delta t)(1 - \lambda_i \Delta t)(1 - \lambda_i g \Delta t) O_i(t)$

E.2) No extinction, speciation, no transition, and none of the lineages is observed in the reconstructed tree (either because they survive to the present time and are not sampled or go extinct by the present time)

$P_{E.2}(t + \Delta t) = (1 - \mu_i \Delta t)(\lambda_i \Delta t)(1 - \lambda_{ig} \Delta t) O_i(t) O_i(t)$

E.3) No extinction, no speciation, transition and the good lineage survives to the present but it is not included in the reconstructed tree (not sampled)

$P_{E.3}(t + \Delta t) = (1 - \mu_i \Delta t)(1 - \lambda_i \Delta t)(\lambda_{ig} \Delta t) O_g(t)$

E.4) Extinction in Δt

$P_{E.4}(t + \Delta t) = \mu_i \Delta t$

$$\begin{aligned} O_i(t + \Delta t) &= P_{E.1}(t + \Delta t) + P_{E.2}(t + \Delta t) + P_{E.3}(t + \Delta t) + P_{E.4}(t + \Delta t) \\ &\quad + (1 - \mu_i \Delta t)(1 - \lambda_i \Delta t)(1 - \lambda_i g \Delta t) O_i(t) \\ &\quad + (1 - \mu_i \Delta t)(\lambda_i \Delta t)(1 - \lambda_{ig} \Delta t) O_i(t) O_i(t) \\ &\quad + (1 - \mu_i \Delta t)(1 - \lambda_i \Delta t)(\lambda_{ig} \Delta t) O_g(t) \\ &\quad + \mu_i \Delta t \end{aligned}$$

$$O_i(t + \Delta t) = [1 - (\mu_i + \lambda_i + \lambda_{ig}) \Delta t] O_i(t) + \lambda_i \Delta t O_i(t) O_i(t) + \lambda_{ig} \Delta t O_g(t) + \mu_i \Delta t$$

$$\frac{O_i(t+\mathit{\Delta} t)-O_i(t)}{\mathit{\Delta} t} = -(\mu_i+\lambda_i+\lambda_{ig})O_i(t)+\lambda_i O_i(t)O_i(t)+\lambda_{ig}O_g(t)+\mu_i$$

$$\frac{dO_i}{dt} = -(\mu_i+\lambda_i+\lambda_{ig})O_i(t)+\lambda_i O_i(t)O_i(t)+\lambda_{ig}O_g(t)+\mu_i \tag{4}$$

Like equations (1) and (2), we cannot solve the equations (3) and (4) analytically. So, simultaneously, we use numerical integration to find $O_g(t)$ and $O_i(t)$ in the tree. The initial conditions ($t = 0$) are $O_g(t) = 1 - \rho_g$ and $O_i(t) = 1 - \rho_i$.

$$O_g(t) = \begin{cases} P_{D.1}(t) + P_{D.2}(t) + P_{D.3}(t) & if \quad t > 0 \\ 1 - \rho_g & if \quad t = 0 \end{cases}$$

$$O_i(t) = \begin{cases} P_{E.1}(t) + P_{E.2}(t) + P_{E.3}(t) + P_{E.4}(t) & if \quad t > 0 \\ 1 - \rho_i & if \quad t = 0 \end{cases}$$

4 Results

4.1 Simulations

Observed relations between parameters of the sampled protracted speciation model

We simulated 100 trees under each different parameter combination to find a relation between the parameters of the sampled protracted speciation model (Table 1). we observed that when

$\#G_c$ = Number of good species in the complete tree

$\#I_c$ = Number of incipient species in the complete tree

$\#G_s$ = Number of good species in the sampled tree

$\#I_s$ = Number of incipient species in the sampled tree

and,

$\lambda_g = \lambda_i$ and $\mu_g = \mu_i$

we have

$$\frac{\#G_c}{\#I_c} \approx \frac{\lambda_{ig}}{\lambda_g}$$

hence,

$$\frac{\#G_s}{\#I_s} \approx \frac{\rho_g \lambda_{ig}}{\rho_i \lambda_g}$$

Table 1: Mean values of $\frac{\#G_c}{\#I_c}$ and $\frac{\#G_s}{\#I_s}$ under the alternative parameter combinations

t_1	$\lambda_g = \lambda_i$	$\mu_g = \mu_i$	λ_{ig}	ρ_g	ρ_i	$\frac{\#G_c}{\#I_c}$	$\frac{\lambda_{ig}}{\lambda_g}$	$\frac{\#G_s}{\#I_s}$	$\frac{\rho_g\lambda_{ig}}{\rho_i\lambda_g}$
10	0.8	0.4	0.4	1	0.08	0.508	0.5	6.630	6.25
10	0.8	0.4	0.6	0.5	0.08	0.779	0.75	5.242	4.68
10	0.8	0.4	2	1	0.08	2.634	2.5	32.622	31.25
6	2	1	1.5	1	1	0.763	0.75	0.763	0.75
6	2	1	0.5	0.8	0.3	0.247	0.25	0.670	0.66
6	2	1	5	0.5	0.8	2.534	2.5	1.609	1.543

Inference on the simulated trees

In this part, first, we simulated 200 random sampled trees (sampling probability $\rho = \frac{1}{2}$), conditioned on the time that the most recent common ancestor speciates ($t_1 = 10Myr$). Besides, we simulated 200 trees, conditioned on t_1, and took diversified samples (sampling fraction $=\frac{1}{2}$). We did the simulations for the random sampled trees and the diversified sampled trees with software package TreeSim (Stadler, 2010) in R. In the random sampled trees and diversified sampled trees, speciation rate(s) and extinction rate(s) for the simulated trees were $\lambda = 0.8$ and $\mu = 0.4$ respectively. In addition, for the sampled protracted speciation trees, we wrote R code to simulate 200 trees conditioned on t_1. In the sampled protracted speciation model, we assumed that the birth and death rates for both good and incipient species are equal.$(\lambda_g = \lambda_i = 0.8,\ \mu_g = \mu_i = 0.8)$. We also assumed the transition rate equal to $\lambda_{ig} = 0.5$. Our goal is to simulate trees under the sampled protracted speciation model with parameters close to the original protracted speciation model. Therefore, in order for the sampled protracted speciation model to be more analogous to the protracted speciation model, we assumed that all good species are sampled, so $\rho_g = 1$. In order to find an appropriate ρ_i, using Monte Carlo approximation method, we simulated 1000 trees under the protracted speciation model ($\lambda_g = \lambda_i = 0.8$, $\mu_g = \mu_i = 0.8$ and $\lambda_{ig} = 0.5$) and found the average proportion of the incipient species became good equal to $\rho_i = 0.08$. For each type of these simulated trees, we estimated model parameters, the speciation rate λ and the extinction rate μ for the random sampling model and diversified sampling model, and the speciation rate of good lineages λ_g, the speciation rate of incipient lineages λ_i, the extinction rate of good lineages μ_g, and the extinction rate of incipient lineages μ_i for the sampled protracted speciation model. We also calculated the corresponding net diversification rates $(\lambda - \mu)$ and the relative extinction rates $(\frac{\mu}{\lambda})$.

We estimated the model parameters by applying maximum likelihood estimation method using the probability functions, introduced in previous chapters, for random sampling and diversified sampling models, and we performed numerical integration method for the sampled protracted speciation model.

In order to find ML estimates of parameters, we used the Hill-climbing algorithm (Nelder and Mead, 1965) implemented in the stats package of R (R. Development Core Team, 2011). For each simulated tree we estimated birth rate and death rate in random sampling and diversified sampling models. In the sampled protracted speciation model, as we mentioned, the aim is to be more close to the original protracted speciation model. As well as this, since the parameter space was 7 dimensional, it is not possible to find the precise maximum likelihood estimates with the little simulated data, so we assumed that ρ_g, ρ_i and λ_{ig} are fixed and equal to the values that the trees were simulated under; hence, we only estimated birth rates and death rates that are prominent in this study. However, when we work with the empirical data, we do not have such information, so while applying the sampled protracted speciation model on the simulated trees under the random and diversified schemes, we assumed that λ_{ig} and ρ_i are unknown, and they needed to be estimated.

Table 2: Mean estimates and standard deviation of the parameters of the random sampling model under three different simulation data ($t_1 = 10Myr$). True parameters were: $\lambda = 0.8$, $\mu = 0.4$ for the random and diversified sampled trees and $\lambda_g = \lambda_i = 0.8$, $\mu_g = \mu_i = 0.4$, $\lambda_{ig} = 0.5$, $\rho_g = 1$ (fixed), and $\rho_i = 0.08$ for the sampled protracted speciation trees

Estimates \Simulation	Random Sampling	Diversified Sampling	S.P.Speciation
$\widehat{\lambda}$	0.8514 [0.097]	0.4406 [0.113]	0.3912 [0.207]
$\widehat{\mu}$	0.4247 [0.117]	0.0047 [0.023]	0.0031 [0.014]
$\widehat{\frac{\mu}{\lambda}}$	0.4962 [0.067]	0.0100 [0.050]	0.0112 [0.046]
$\widehat{\lambda - \mu}$	0.4267 [0.064]	0.4359 [0.114]	0.3893 [0.098]

The result implies that when the simulation method and the analyze model are the same, estimated parameters are closer to the true values. Moreover, the estimations in the sampled protracted model are not as close to the true values as they are in the random sampling and diversified sampling models. We can also observe a remarkable bias in the estimations of the relative extinction rates and the net diversification rates in the sampled protracted model. The standard deviations of the parameter estimation in the sampled protracted speciation model are relatively larger than the standard deviation of the parameter estimations in the two other models.

By simulating three different types of phylogenetic trees, for each we fitted all three models and estimated the parameters. Table 2, shows the mean and standard deviations of the parameter estimations of random sampling model on three types of simulated data (the random sampled trees, the diversified sampled trees and the trees simulated under the sampled protracted speciation). The result implies that the parameter estimation of the random sampling model on the random sampled trees are close to the true values. It also shows that random sampling model underestimates the parameters while the data is the diversified sample and the sampled protracted speciation data.

When the three types of data are analyzed by the diversified sampling model, Table 3 shows that the parameters are estimated more correctly when the data is the diversified sample. The diversified sampling model overestimated the parameters when the data is simulated differently. Tables 4 shows the result of the sampled protracted speciation model

Table 3: Mean estimates and standard deviation of the parameters of the diversified sampling model under three different simulation data ($t_1 = 10Myr$). True parameters were: $\lambda = 0.8$, $\mu = 0.4$ for the random and diversified sampled trees and $\lambda_g = \lambda_i = 0.8$, $\mu_g = \mu_i = 0.4$, $\lambda_{ig} = 0.5$, $\rho_g = 1$ (fixed), and $\rho_i = 0.08$ for the sampled protracted speciation trees

Estimates \Simulation	Random Sampling	Diversified Sampling	S.P.Speciaition
$\widehat{\lambda}$	1.8438 [0.981]	0.7816 [0.048]	1.4228 [0.496]
$\widehat{\mu}$	1.6532 [1.063]	0.3852 [0.060]	1.2299 [0.573]
$\widehat{\frac{\mu}{\lambda}}$	0.8346 [0.187]	0.4924 [0.064]	0.8271 [0.154]
$\widehat{\lambda - \mu}$	0.1905 [0.136]	0.3963 [0.054]	0.1929 [0.112]

fitted on the three different types of simulated data. The result implies that the sampled protracted speciation model overestimates the speciation rate and the extinction rate for good species and it underestimates the parameters for the incipient species. The estimated values of the protracted speciation sample are more close to the true values.

Table 4: Mean estimates and standard deviation of the parameters of the sampled protracted speciation model under three different simulation data ($t_1 = 10Myr$). True parameters were: $\lambda = 0.8$, $\mu = 0.4$ for the random and diversified sampled trees and $\lambda_g = \lambda_i = 0.8$, $\mu_g = \mu_i = 0.4$, $\lambda_{ig} = 0.5$, $\rho_g = 1$ (fixed), and $\rho_i = 0.08$ for the sampled protracted speciation trees

Estimates \Simulation	Random Sampling	Diversified Sampling	S.P.Speciaition
$\widehat{\lambda_g}$	1.3606 [0.6524]	1.4043 [0.613]	0.9433 [0.388]
$\widehat{\lambda_i}$	0.5533 [0.117]	0.5776 [0.162]	0.7015 [0.132]
$\widehat{\mu_g}$	0.4818 [0.449]	0.8015 [0.475]	0.4145 [0.261]
$\widehat{\mu_i}$	0.3176 [0.109]	0.3384 [0.183]	0.3677 [0.140]
$\widehat{\lambda_{ig}}$	0.2960 [0.131]	0.5350 [0.335]	fixed =0.5
$\widehat{\rho_i}$	0.3450 [0.101]	0.5403 [0.173]	fixed =0.08
$\widehat{\frac{\mu_g}{\lambda_g}}$	0.3624 [0.184]	0.5862 [0.245]	0.4293 [0.126]
$\widehat{\frac{\mu_i}{\lambda_i}}$	0.5914 [0.164]	0.5841 [0.212]	0.5292 [0.189]
$\widehat{\lambda_g - \mu_g}$	0.8788 [0.503]	0.6027 [0.531]	0.5288 [0.207]
$\widehat{\lambda_i - \mu_i}$	0.2357 [0.157]	0.2391 [0.143]	0.3338 [0.152]

Furthermore, we studied the effects of the transition rate change and the change of sampling probability of the incipient species in the sampled protracted model. Firstly, assuming that $t_1 = 10Myr$, $\rho_g = 1$, $\rho_i = 0.08$ are fixed, we simulated 100 trees under different values of the transition rate ($0 < \lambda_{ig} \leq 15$), and we estimated the birth rates and the death rates. Figure 9.a illustrates that while the transition rate increases, the birth rate and death rate for the good species decrease and the birth rate and the death rate for the incipient species increase. The result implies that decrease in the birth rate of the good species is dramatically intensive when the transition rate is increasing form 0 to 2. Besides, Figure 9.a shows that when the transition rate increases, it does not affect the net diversification rate ($\lambda - \mu$) for both good and incipient species while the relative extinction rate ($\frac{\mu}{\lambda}$) for both good and incipient species noticeably increases. Secondly, assuming that $t_1 = 10Myr$, $\rho_g = 1$, $\lambda_{ig} = 0.5$ are fixed, we simulated 100 trees under different values of the sampling probability for the incipient species($0 \leq \rho_i \leq 1$) and estimated the birth rates and the death rates. Figure 9.b shows that there is no noticeable effect on the parameter estimations while the sampling probability of the incipient species changes from 0 to 1.

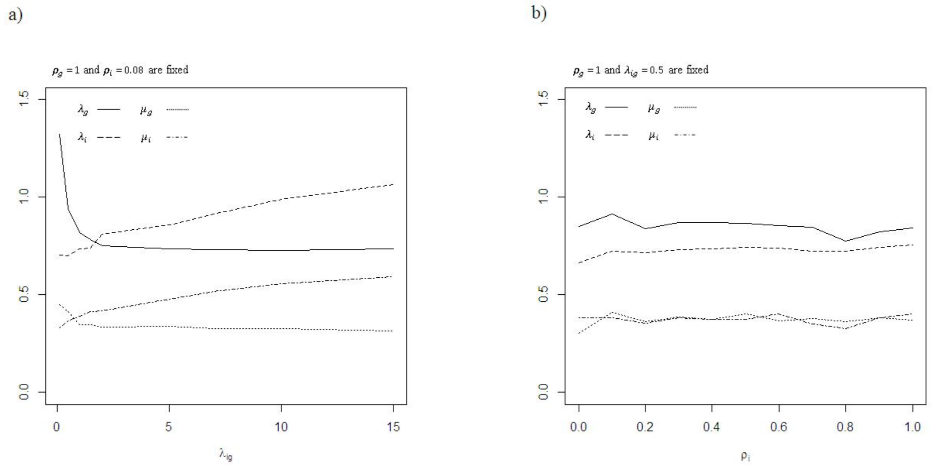

Figure 9: The effect of the transition rate (λ_{ig}) change on the parameter estimations (a) and the effect of changing the sampling probability of incipient species on the parameter estimations (b)

4.2 Empirical Data

For this part of the study, we chose a data set of the avian families. The tree includes seven families (Tesia, Cettia, Abroscopus, Tickellia, Orthotomus, Urosphena, and Hemitesia). Each family contains not only the species (29 species), but also subspecies. The tree contains 48 tips (subspecies) and 47 internal nodes. We fitted all three models to the data set. The only information we have is one species is missing in the tree. So it means we know that the sampling probability for the good species in the sampled protracted speciation model is $\rho_g = 0.97$. Since we did not have any more information about the sampling probability and sampling fraction, for the random sampling model and diversified sampling model, we estimated them as well as the diversification rates. In the sampled protracted speciation model, we assigned randomly one subspecies among the subspecies of a species as good and took the other subspecies as incipient species. We also estimated the transition rate and the sampling probability of the incipient species. Since the length of the edges on the tree are too short (the height of the tree is 0.1378085), we scaled the tree and considered the whole length equal to t_1=13.78085. Therefore, the estimated rates are not per million but they are relative rates. Table 5 shows the results.

Table 5: Paramater estimates of the three models on the avian data set

Parameters / Models	Random Sampling	Diversified Sampling	S.P. Speciation
$\widehat{\lambda}$	0.570	0.596	$\widehat{\lambda_g}= 0.440$ $\widehat{\lambda_i} = 0.385$
$\widehat{\mu}$	0.469	0.501	$\widehat{\mu_g} = 0.140$ $\widehat{\mu_i}= 0.371$
$\widehat{\frac{\mu}{\lambda}}$	0.823	0.842	$\widehat{\frac{\mu_g}{\lambda_g}} = 0.318$ $\widehat{\frac{\mu_i}{\lambda_i}} = 0.962$
$\widehat{\lambda - \mu}$	0.101	0.095	$\widehat{\lambda_g - \mu_g} = 0.300$ $\widehat{\lambda_i - \mu_i} = 0.015$
$\widehat{\rho}$	0.964	0.978	$\widehat{\rho_i} = 0.853$ $\widehat{\lambda_{ig}} = 0.741$

The result shows that the sampling probabilities for both the random sampling model and the diversified sampling model are estimated so large, that is why the parameter estimates for these two models are almost similar. Besides, it shows that the extinction rate of the incipient species is higher than the extinction rate of the good species. Moreover, the estimated sampling probability for the incipient species is smaller than the sampling probability for the good species. Finally, the result shows that the estimated transition rate (λ_{ig}) is larger than the estimated speciation and extinction rates, so the transition is relatively fast, but it is not instantaneous.

5 Discussion

In this thesis, we suggested the sampled protracted speciation model to calculate the probability density of the protracted speciation model based on sampling probabilities, and we compared it with the random sampling model and the diversified sampling model in the birth-death process in two parts of the simulations and the empirical data. Firstly, we found from the simulation part that, in all three models, the parameters are estimated more accurately under the model that the data were simulated by. For instance, when the data is simulated under the diversified sampling model or the sampled protracted speciation model, applying the random sampling model for parameter estimation causes underestimation of the speciation rate, extinction rate and the relative extinction rate. Besides, when the data is simulated under the random sampling model or the sampled protracted speciation model, applying the diversified sampling model for parameter estimation causes overestimation of the speciation rate, extinction rate and the relative extinction rate and underestimation of the net diversification rate. Moreover, in the sampled protracted speciation model, when the data is simulated under the same model with the transition rate $\lambda_{ig} = 0.5$, the speciation rate and the extinction rate for the good species are slightly overestimated and the speciation rate and the extinction rate for the incipient species are slightly underestimated. If we increase the transition rate, contrariwise, the diversification rates for the good species are slightly underestimated and the diversification rates for the incipient species are overestimated. Besides, when the data is simulated under the diversified sampling model or the random sampling model, applying the sampled protracted speciation model results a considerable overestimation of the diversification rates for the good species and underestimation of the diversification rates for the incipient species. As we showed in Figure 9, the transition rate increase has a considerable effect on the parameter estimation in the sampled protracted speciation model while the sampling probability of the incipient species has no noticeable effect on the parameter estimation.

The empirical data we chose contained both species and subspecies. We believe that the sampled protracted speciation model is a better model to fit for this kind of data. We considered one subspecies of each species as good and the others as incipient species. The random sampling and the diversified sampling models parameter estimations are almost the same. On the other hand, the sampled protracted speciation model estimated the parameters for the good and the incipient species quite differently. We know that most of the tips are incipient subspecies and toward the root we have mostly good species. The higher death rates and correspondingly the higher relative death rate for the incipient species ($\mu_i = 0.371$, $\frac{\mu_i}{\lambda_i} = 0.962$), in comparison with the good species ($\mu_g = 0.140$, $\frac{\mu_g}{\lambda_g} = 0.318$), implies that the speciation events for the incipient subspecies are mostly toward the tips.

The sampled protracted speciation model has seven parameters, optimizing the probability density by using the Hill-Climbing technique and finding the accurate estimates of the parameters is a very sensible process. We fixed three parameters of the model to, first, be more close to the original protracted speciation model, and second, find the more accurate maximizer estimates of the diversification rates.

Future Work

The change of the sampling probabilities for the good and the incipient species and their effect on the estimated parameters might be a further work to do. Besides, determining whether or not the trees with both species and subspecies can be analysed better with the sampled protracted speciation model can only be achieved through further empirical studies.

6 References

Etienne, R. S., Rosindell, J. (2012). Prolonging the Past Counteracts the Pull of the Present: Protracted Speciation Can Explain Observed Slowdowns in Diversification. *Syst Biol* (2012) 61(2): 204-213

Höhna, S., Stadler, T., Ronquist, F., and Britton, T. (2011) Inferring speciation and extinction rates under different species sampling schemes. *Mol Biol Evol* (2011) 28(9): 2577-2589.

Kendall, D. G. (1948). On the generalized "birth-and-death process". *Ann. Math. Statist.*, 19:1-15.

Maddison, W. P., Midford, P. E. and Otto, S. P. (2007). Estimating a Binary Character's Effect on Speciation and Extinction. *Syst Biol* (2007) 56(5): 701-710 doi:10.1080/10635150701607033.

McPeek, M. A. 2008. The ecological dynamics of clade diversification and community assembly. *American Naturalist* 172:E270E284.

Nee, S. C., May, R. M., and Harvey, P. (1994). The reconstructed evolutionary process. *Philos. Trans. Roy. Soc. London Ser. B,* 344:305-311.

Nei, M. and Kumar, S.(2000) Molecular Evolution and Phylogenetics. *Oxford University Press, New York.*

Nelder, J. and Mead, R. (1965). A simplex method for function minimization. *The computer journal*, 7(4):308.

Phillimore, A. B. and Price, T. D. (2008). Density-dependent cladogenesis in birds. *PLoS Biol*, 6:e71

Purvis, A., C. D. L. Orme, N. H. Toomey, and P. N. Pearson. 2009. Temporal patterns in diversification rates. chap. 15, Pages 278300 in Speciation and Patterns of Diversity (R. Butlin, D. Schluter, and J. Bridle, eds.). *Cambridge University Press.*

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.* ISBN 3-900051-07-0.

Rannala, B. and Yang, Z. (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of Molecular Evolution*, 43:304-311.

Rosindell, J., S. J. Cornell, S. P. Hubbell, and R. S. Etienne. 2010. Protracted speciation revitalizes the neutral theory of biodiversity. *Ecology Letters* 13:716727.

Stadler, T. (2009). On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *Journal of Theoretical Biology*, 261:58-66.

Stadler, T. (2010). TreeSim: Simulating trees under the birth-death model. *R package version 1.0.*

Stadler, T (2011). Inferring speciation and extinction processes from extant species. *dataPNAS* 2011 108 (39) 16145-16146; published ahead of print September 19, 2011, $doi: 10.1073/pnas.1113242108$

Thompson, E. A. (1975). Human evolutionary trees. *Cambridge University Press.*

7 Appendix

Unsuccessful approaches to calculate the probability density of the protracted speciation model

As we mentioned in the protracted speciation model section, there are some difficulties for deriving the probability that an extant species evolves to the present time for the protracted speciation model applying BiSSE method. In this section, we explain these difficulties. The reason why we could not calculate the whole probability of the reconstructed tree was some problems occurred due the rule which is defined in the protracted speciation model. The rule implies that the incipient species with good extinct parents are also observed as a good species in the reconstructed tree. There are two ways to face this rule. First, we can say the incipient species survives to the present time as an incipient species, and it is observed in the reconstructed tree because of its extinct good parent (Figure 10- left). Second, we can say once the good parent becomes extinct, the incipient species becomes good species (Figure 10 - Right). In both of these perspectives, we had some difficulties which prevented us to calculate the likelihood. The problems are explained in this part, briefly. According to the problem, we tried four approaches, considered different scenarios and defined corresponding probabilities, as we did in the sampled protracted speciation model, for each.

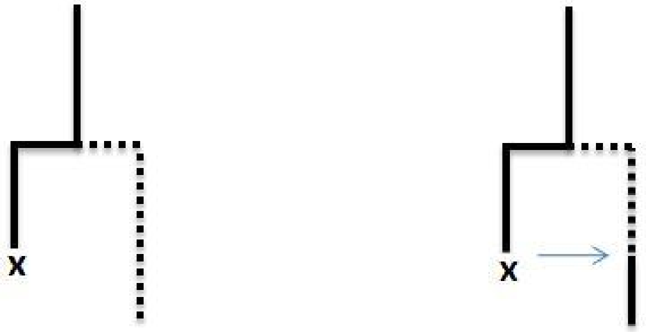

Figure 10: The incipient species with good extinct parents are also observed as a good species in the reconstructed tree

First approach is when we consider the incipient species survives to the present time as an incipient species, and it is observed in the reconstructed tree because of its extinct good parent (Figure 10- left). If we apply BiSSE approach and start from tips and move backward in time to the root of the reconstructed tree, as Figure 11 illustrates, we will see that in the reconstructed tree we have a state change from a good lineage to an incipient one. This transition is not anticipated by the model and its parameters; furthermore, we do not know the exact time of this event (the spaciation which has result death of the good parent). We just know that, definitely, it has happened some time on the external branch. Therefore, we decided to change the approach.

In order to overcome this problem, in the second approach, we decided to start from the root of the reconstructed tree and move forward in time, and considering all possible scenarios, calculate the whole probability of the tree. In this approach, since we do not know the state of the tips (good or incipient with an extinct good parent) and, even if we knew the state, we would not know the time of the split which finally results the death of good species (the parent), we have the same problem and we are not able to calculate the probability, either.

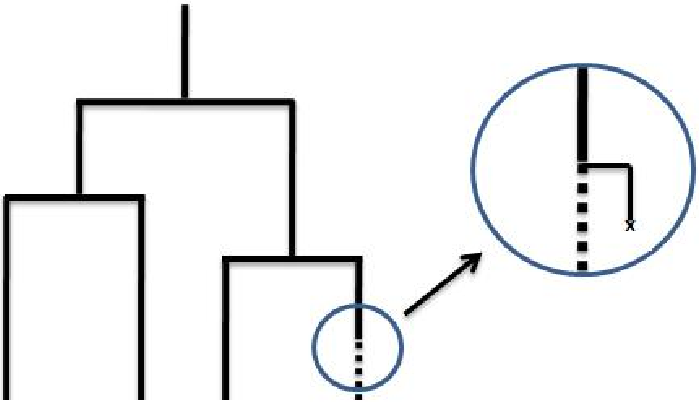

Figure 11: A part of a reconstructed tree including an observed incipient species with a good extinct parent. Here we assume that the incipietn species with the good parent survive a an incipient (Figure 10, left)

In the third approach, when we assume that once the good parent becomes extinct, the incipient species becomes good species (Figure 10 - Right), we applied the BiSSE model and moved backward in time. While moving backward on an external branch which used to be an incipient species and becomes good due to the dead good parent, firstly, the extinction happens on the other branch which is not included in the reconstructed tree (there is no information about the time of death); and secondly, before we reach the time of the split that results the dead good parent, the transition caused by this death happens. Thus, we cannot calculate the probability of the branch moving backward in time. Finally, in the forth approach, we assumed the same as the third approach (Figure 10 - Right); however, we decided to move forward in time to calculate the probability of the reconstructed tree. In this approach we do not have the matter of time order (when we move forward the split happens earlier than the death of good parent). We see that there are two scenarios, shown in Figure 12, happening. First, the incipient species does not become good itself before the time of extinction of the good parent (Figure 12 - A), and second, the incipient lineage becomes good, independently and earlier from the extinction time of its good parent (Figure 12 - B).

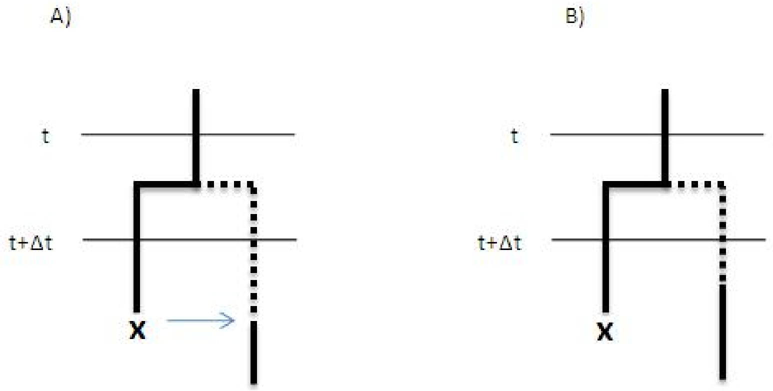

Figure 12: The incipient species might become good species either, independently from the extinction time of its parent (A) or due to and at the time of the extinction of its good parent

By Monte Carlo approximation we can find the fraction for happening of either of these scenarios. In order to calculate the probability of the scenario in Figure 12-A, we looked at the critical event as if the good species survives and at the time of its extinction the incipient one goes extinct (Figure 13). Since after a speciation event on the reconstructed tree, the lineages are not good species immediately and this scenario can happens anywhere on the tree, we can expect that on a branch the state changes several times.

Unfortunately, we found that this is not a successful way to find the likelihood, either. In the forward in time approach, we start form the root of the reconstructed tree and consider all possible events along branches and reach the tips of the tree. At the end, we calculate the likelihood as

$L = P_{(Noevent)}(\text{On all edges of the reconstructed tree})\ \lambda^{n-1}$

Where n is the number of extant species and λ is the speciation rate in the reconstructed tree. The speciation rate in the reconstructed tree, λ, could be either of λ_g (the speciation rate for good species) or λ_i (the speciation rate for incipient species) in each node.

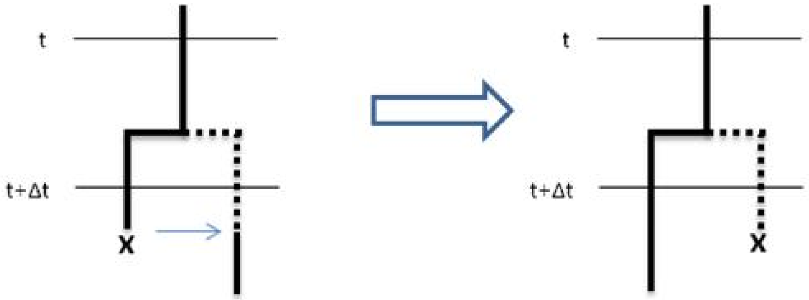

Figure 13: To calculate the probability it is assumed that at the extinction time of good lineage, the incipient one becomes extinct and the good one survives instead

In this definition we missed the point that states at nodes are different and this means the transitions happen on branches as events. If we knew the state of each node, and so we knew the number of independent and spontaneous transitions, we could not count the number of transitions cause by the rule which has described above and calculate the time of being in the good and incipient states in the tree. The problems that explained here briefly made us come up with the idea of the sampled protracted speciation model presented in this thesis.