Expanding current theoretical models of diversification (provisional title)

Introductory Essay

Theoretical Research in Evolutionary Life Sciences (TRÊS)

University of Groningen

Supervisor:

Dr. Rampal S. Etienne

Author:

Giovanni Laudanno

Contents

Introduction

Why math?

> "I am convinced that purely mathematical construction enables us to find those concepts and those lawlike connections between them that provide the key to the understanding of natural phenomena." Albert Einstein 1933

What is a model? Model is some representation of reality, and it is the best way human culture, unable to read directly at the nature's "source code", developed to explain natural phenomena since the birth of modern science. Galileo, one of main founders of science as we know it today, used to say that nature is a big book but to be able to read it we must first learn the language and grasp the symbols in which it is written, in one word: mathematics. But why? A lot of answers could be given to this question but what I feel to answer here, in a very brief way, is because nature, per se, is deceptive. As everybody will probably know in fact, for centuries people have believed that every celestial body used to rotate around our planet, just because that was the everyday experience. The great paradigm shift from geocentric to heliocentric model was achieved by introducing a proper way to collect data as well as developing the right theoretical tools to read them, i.e. the invention of calculus, to eventually create a model and properly read nature (and not being deceived). Since then a lot of mathematical models were developed and they allowed us to systematize our knowledge, predict phenomena and understand their underlying principles.

So, why modeling in biology? The impact and the necessity of the mathematics in life sciences have grown enormously in recent decades as recently developed techniques gave us access to enormous quantities of data. Besides complexity of the processes demands the development of models to unveil underlying mechanisms. This is especially true when it comes to deal with evolutionary processes.

Possible points to develop here:

- how to collect data: developing of technology gave us access to new stuff (e.g. PCR invention). For example, in the 90's, since invention of the PCR molecular information starts to become preponderant as before mostly morphologic data were considered. That brought a lot of attention to the field in the academic world with more and more people working on that in the last years. All these studies, of course, demand also more and more theorethical support.

- As our understanding of the field improves in time, yet a lot of new questions arise. For example: in a given habitat is the metapopulation (basicly, the population of populations living there) being able to growth and fill all available niches, reaching a steady state? Or big events (huge floods, glaciation events, mountains arising etc) can occur before the reaching of the steady state? Why species richness is not geographically equally distributed? Why there is so much species' richness on the tropics zone respect to colder areas? Other big questions?

- Theoretical development of the field: for example Felzenstein 1981 applies likelihood approach to phylogenetic trees and changes the way people could speculate on such problems

- Models are usually used to infer parameters like speciation or extinction rates; but they are not only useful to study rates by themselves but they can also be used to create new models that requires additional informations such as time scales of diversification, influence of traits on the diversification process etc

- Something else?

Chapter 1

Background

1.1 Diversification models

1.1.1 Pure Birth Model

From fossils and genetic data we know that some process must be occurred to increase number of species in a given habitat, even though this can require millions of years. Given the required time interval the simplest possible explanation to give is claiming that there should be some average speciation rate (usually denoted as λ) bringing the total number of species to grow over time. This is the Yule's Model (often called simply "Pure birth model"). The assumption is straightforward: assuming a constant speciation rate, the more species you have the faster is the speciation process:

$$\frac{dN(t)}{dt} = \lambda N(t) \tag{1.1}$$

this very easy ODE (Ordinary Differential Equation) is solved by an exponential growth solution:

$$N(t) = N_0 e^{\lambda t} \tag{1.2}$$

It is also possible to define such model in terms of probability distributions $P_n(t)$ for each number of species present in the phylogeny at time t :

$$\frac{dP_n(t)}{dt} = (n-1)\lambda P_{n-1}(t) - n\lambda P_n(t) \tag{1.3}$$

where the two terms on the right hand side give, respectively, the probability to enter or going out from the state P_n. This kind of equation is known as master equation for a

Markov Process and is broadly used to describe several processes in a variety of fields like, among others, biology, physics or finance.

It is necessary to point out the main features of this process:

- this is a set of ODEs, due to the their shape, it is known beforehand their solutions have always an exponential behaviour;

- the state space is discrete (as the n index);

- state transitions can occur only between neighbouring states, $n \to n+1$ (for speciation), $n \to n-1$ (for extinction), i.e. no more than one event can occur in the time interval dt;

- what happens to P_n in the time interval dt depends only on adjacent time intervals (the process is "memoryless").

Then it is possible to evaluate the average (or estimated) number of species using these probability distributions:

$$N(t) = \sum_{n=1}^{\infty} nP_n(t) \tag{1.4}$$

to recover the same result.

This kind of model gives us possibility to run simulations to include stochasticity in the process and could be really useful as a check, due to its simplicity.

One may of course argue this kind of model to be quite too simplicistic. We know, though, that species can either speciate or going extinct; for this reason the first and natural expansion of this model is to take into account also an extinction rate.

1.1.2 Birth-Death models

Here, in the same fashion of the Yule's model, a new rate is introduced: extinction rate μ. Even though these rates could be time-dependent or diversity-dependant, standard birth-death models take into account only λ and μ as constants through all the process:

$$\frac{dP_n(t)}{dt} = (n-1)\lambda P_{n-1}(t) + (n+1)\mu P_{n+1}(t) - n(\lambda+\mu)P_n(t) \tag{1.5}$$

It obviously respects the same properties I listed for the pure birth case. To evaluate the average number of species in time you can follow the same procedure as before.

1.1.3 Kendall 1948 - An analytical solution for the standard Birth-Death model

Standard birth-death model has been analytically solved by Kendall[14] in 1948. To do that he used the generating function approach, defining:

$$\phi(z,t) = \sum_{n=-\infty}^{\infty} z^n P_n(t) \tag{1.6}$$

where $z \in [0,1]$ is an auxiliary variable. This approach is a standard technique for this kind of problem and allow to transform a set of infinite ODEs in just one PDE (Partial Differential Equation) for the unknown generating function $\phi(z,t)$, from which it is possible to come back to the solution of the original problem with simple operations. Multiplying each equation of the ODEs set by z^n and summing them all together is possible to get the PDE:

$$\frac{\partial \phi(z,t)}{\partial t} = (\lambda z - \mu)(z-1)\frac{\partial \phi(z,t)}{\partial z} \tag{1.7}$$

that could be solved which is in the standard Lagrangian form and could be solved through method of characteristics. From auxiliary equation:

$$-dt = \frac{dz}{(\lambda z - \mu)(z-1)} \tag{1.8}$$

along with initial condition (equivalent to $P_n(-T) = \delta_{n,1}$):

$$\phi(z,-T) = z \tag{1.9}$$

it is possible to get the general solution:

$$\phi(z,t) = \frac{(\lambda z - \mu)\Lambda(t) - \mu(z-1)\Lambda(-T)}{(\lambda z - \mu)\Lambda(t) - \lambda(z-1)\Lambda(-T)} \tag{1.10}$$

where $\Lambda(t) = e^{-(\lambda-\mu)t}$ (here my notation is slightly different from original Kendall's one on λ and Λ). Now it is possible to derive $P_n(t)$ for all n. For example, for the first 2 solutions:

$$P_0(t) = \phi(z,t)|_{z=0} = \frac{\mu\Lambda(-T) - \mu\Lambda(t)}{\lambda\Lambda(-T) - \mu\Lambda(t)} = \frac{e^{(\lambda-\mu)T} - e^{-(\lambda-\mu)t}}{\frac{\lambda}{\mu}e^{(\lambda-\mu)T} - e^{-(\lambda-\mu)t}} \tag{1.11}$$

$$P_1(t) = \frac{\partial}{\partial z}\phi(z,t)|_{z=0} = \frac{\Lambda(t)\Lambda(-T)(\lambda-\mu)^2}{[\mu\Lambda(t) - \lambda\Lambda(T)]^2} = [1 - P_0(t)] \cdot [1 - u_t] \tag{1.12}$$

where:

$$u_t = \frac{\lambda[\Lambda(-T) - \Lambda(t)]}{\lambda\Lambda(-T) - \mu\Lambda(t)} \tag{1.13}$$

$$1 - P_0(t) = \frac{(\lambda - \mu)\Lambda(-T)}{\lambda\Lambda(-T) - \mu\Lambda(t)} \tag{1.14}$$

it is also possible to prove that for any generic n, solutions may be expressed as a geometric series:

$$P_n(t) = [1 - P_0(t)] \cdot (1 - u_t)u_t^{n-1} \tag{1.15}$$

1.2 Likelihood approaches

For several models, included birth death, is possible to infer best values for parameters maximizing the likelihood function. A likelihood function (often used in its logarithmic form) measures the probability to have parameters θ given the data:

$$L = L(\theta|data) \tag{1.16}$$

Studying evolutionary processes usually data are phylogenies. These can be represented as trees and are composed by two main part: branching times and topology. Branching times give information about the position of nodes (or speciation events) along the tree; topology give us information about how nodes are connected together, regardless of branches lenghts.

So in this kind of approach is crucial to find a proper way to assign a probability to the model, given the data.

1.2.1 Nee et al. 1994 - A likelihood paradygm for constant rates

1.2.1.1 Understanding the process

Nee, May and Harvey in their 1994's article propose a way to unambiguously assign a likelihood to any diversification process having constant extinction and speciation rates. They define everything in terms of two functions:

$$u_t := \frac{\lambda(1 - exp(-(\lambda - \mu)t)}{\lambda - \mu exp(-(\lambda - \mu)t)} \tag{1.17}$$

$$P(t, T) := \frac{\lambda - \mu}{\lambda - \mu exp\{-(\lambda - \mu)(T - t)\}} \tag{1.18}$$

where $P(t, T)$ is the probability that a single lineage alive at time t has some descendants, i.e. has not gone extinct, at the later time T (Nee et al. 1994 [18]); u_t instead is the

probability that another lineage appears up to time t.

These two functions derive directly from Kendall's solutions (Kendall 1948[14]). In fact to get the same function you just have to redefine boundaries ($[-T, t] \to [0, t]$ for u_t and $[-T, t] \to [t, T]$ for $P(t, T)$, derived from Kendall's $1 - P_0(t)$).

Is possible to link these two functions to the solutions of master equations 1.5. For example, for a simple birth-death process which goes extinct before time t:

$$P_0(t) = 1 - P(0, t) \tag{1.19}$$

$$P_1(t) = P(0, t)(1 - u_t) \tag{1.20}$$

Even though this process is usually not so interesting (you can define finer processes like "a birth-death process that survives to T such that at least one of the i lineages existing at time t has some descendants at time T", but its form is way more complicated and it is not useful to clearly explain the procedure to follow), it allows to highlight how the reasoning works. In fact, given a solution $P_n(t)$ to the set of ODEs, here there are two (indipendent) equations from which is possible to derive explicitly $P(t, T)$ and u_t, even for non-standard birth-death models. It is not possible, though, to create this link (e.g. when diversity-dependence rates are present).

1.2.1.2 Building the Likelihood

From these two functions Nee et al. suggest a very neat way to assign a likelihood to a given tree. The drawback is that this can only work with "breakable" trees, i.e. a tree in which what happens in every branch does not depend on what happens along other branches (e.g. constant rates processes can be broken, diversity-dependent processes can not because a single speciation event with rate λ_n depends on all existent species in that time). If the tree is breakable for each branch and node a probability is assigned and the likelihood is computed as a general product of all these terms.

Thus Nee's likelihood is given as:

$$lik = (N-1)!\lambda^{N-2}(1 - u_{x_2})^2 \left[\prod_{i=3}^{N}(1 - u_{x_i})\right] \left[\prod_{i=3}^{N} P(t_i, T)\right] \tag{1.21}$$

where $x_i = T - t_i$. In this formula every birth event contributes with a $(i-1)\lambda P(t_i, T)$ factor and each branch (i.e. the amount of time in which a lineage do not give birth) contributes with a factor $(1 - u_{x_i})$. Note that crown species are already there, so there is

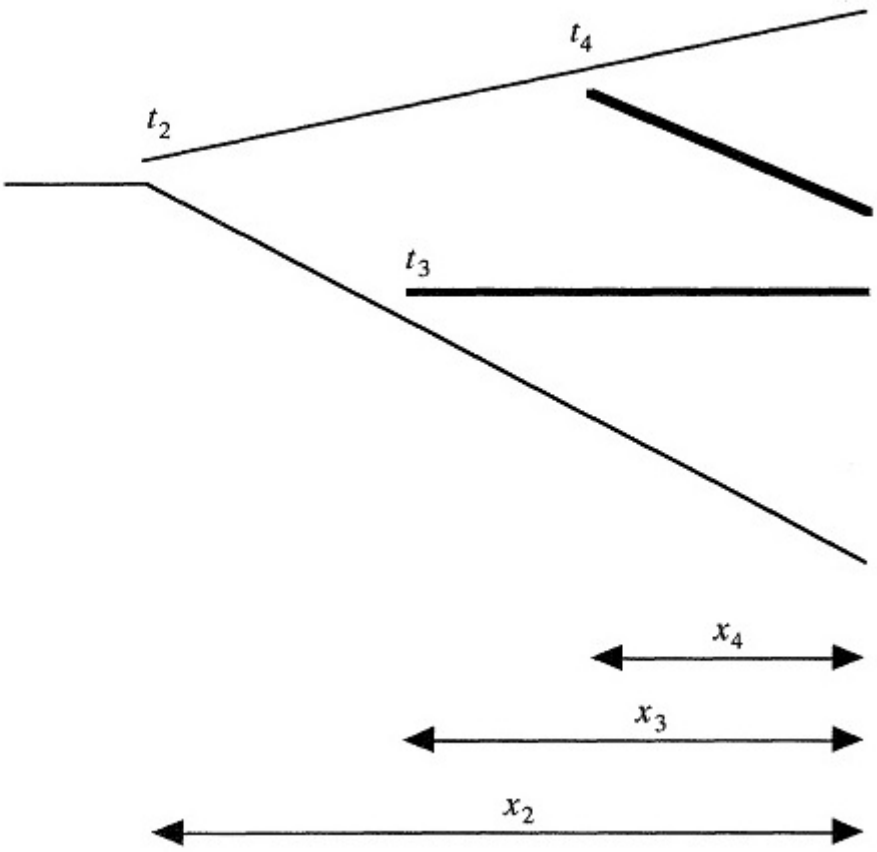

Figure 1.1 – A broken phylogenetic tree. Bold lines are daughter branches while x_i are the lenght of time elapsed between the nodes and the present day (from Nee et al. 1994 [18]).

no birth term for them, and they start to appear together in the phylogeny from x_2.

Even though this approach could present several limitations has the notable properties to provide an analytical (and precise) expression for the likelihood and thus it is suitable for checks on more general models.

1.2.2 Etienne et al. 2012 - How to deal with diversity-dependence

So Nee et al's method provides an elegant and efficient way to deal with "breakable tree" but what about more complex models, such as diversity-dependence ones? In this case building a likelihood might become a truly challenging problem.

As usual what happens along a branch depends on rates but, in this case, this actually depends on total number of species (actual and missing ones). For this reason "breaking" the tree is not a valid option and it was instead splitted considering all the possible infinitesimal

time slices to be eventually integrated. To do that, though, it becomes mandatory to define a new function to be integrated that could be directly linked to the tree likelihood.

Etienne et al.[5] managed to find a way to deal with this problem defining a new set of dynamic equations, close to the usual birth-death master equations1.5, whose solution is the probability $Q_n^k(t)$ that "a realization of the diversification process is consistent with the phylogeny up to time t and has n species at time t":

$$\frac{dQ_n^k(t)}{dt} = \mu_{k+n+1}(n+1)Q_{n+1}^k(t)+\lambda_{k+n-1}(n-1+2k)Q_{n-1}^k(t)-(\lambda_{k+n}+\mu_{k+n})nQ_n^k(t) \quad (1.22)$$

where k is the (fixed) number of species in phylogeny and n is the number of missing species to eventually be set to zero at the end. Importantly, it is also possible to take into account species extant at the present time t_p but missing in the phylogeny. It is also worth to point out that, even though the equation is formally very similar to the standard one, the biggest difference lies in the presence of the $2k$ in the λ-based gain term. This is due to the fact that the two sets of species k and n are distinct and thus a speciation event can occur in two different ways.

A set of equations is defined for each interval $[t_k, t_{k+1}]$ $\forall k \; in \; \underline{t}$, being $\underline{t}$ the whole collection of branching times. To link solutions from one time interval to the next one it is necessary to include proper operations on nodes up to the present time t_p.

This method follows the same phylosophy of Nee et al. as their likelihood is obtained by multiplying together components coming both from nodes and branches to eventually include all "fragments" of the tree. Besides the model itself was checked using constant rates and it has been proved that it gives back Nee et al. model's results, and it may therefore be considered a direct generalization of that.

In the following I will refer to this kind of model as the diversity-dependent Q model.

1.3 On influence of traits on diversification processes

In the last years a series of models of increasing complexity has been developed to include influence of traits on diversification process.

1.3.1 BiSSE - A binary state model

1.3.1.1 Dynamics along branches

The first article of series involving this topic is Maddison et al. 2007 [17] in which they introduced BiSSE model to study how and whether speciation and extinction rates may depend on the state of a particular binary-state character. Also in this case model is ruled by core differential equations (in this case non linear and coupled):

$$\frac{dD_{N0}}{dt} = -(\lambda_0 + \mu_0 + q_{01})D_{N0}(t) + q_{01}D_{N1}(t) + 2\lambda_0 E_0(t)D_{N0}(t) \tag{1.23}$$

$$\frac{dE_0}{dt} = \mu_0 - (\lambda_0 + \mu_0 + q_{01})E_0(t) + q_{01}E_1(t) + \lambda_0 E_0(t)^2 \tag{1.24}$$

There are also same equations for the presence of those traits flipping every 0 to 1 and viceversa. Probabilities$D_0(t)$ and $D_1(t)$ describe the chance that a lineage beginning at time t with state 0 or 1 would evolve into the clade like that observed to have descended from node N, close to the present. $E_0(t)$ is the probability that a lineage starting at time t in state 0 leaves no descendants at the present day ($E_1(t)$ does the same for state 1). These two equations are derived taking in consideration every possible outcome of these two functions along branches.

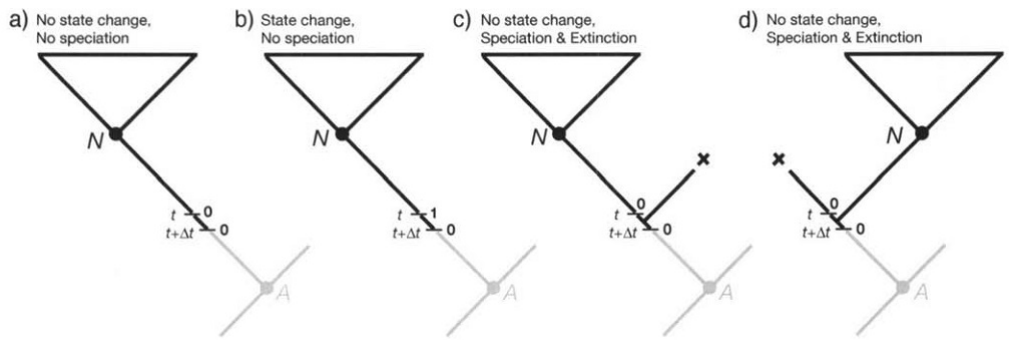

Figure 1.2 – Alternative scenarios by which a lineage with state 0 at time $t + \Delta t$ on the branch might yield clade descended from node N but no other living descendants (figure from Maddison et al. 2007 [17]).

Here it is possible to consider both speciation and extinction to be different in absence or presence of the trait while q_{01} and q_{10} are transition rates between states.

1.3.1.2 On nodes and root

As in Etienne et al.[5] to build a likelihood the solution of differential equations describe branch components of the tree while additional relations are required to express nodes' contributions. In these case, since integration is made backwards in time, they simply use for every speciation event in the phylogeny the relation:

$$D_{A0}(t_A) = D_{N0}(t_A)D_{M0}(t_A)\lambda_0 \tag{1.25}$$

$$D_{A1}(t_A) = D_{N1}(t_A)D_{M1}(t_A)\lambda_1 \tag{1.26}$$

where M is the sister node to N.

After the whole integration process, at the root, they end with D_{R0} and D_{R1} and it is necessary to combine them to get a likelihood. The way authors choose to do that is to sum them weightening each term with relative frequencies (at equilibrium) of species being in state 0 (or 1) defined as $x_0 = \frac{n_0}{n_0+n_1}$ (or $x_1 = \frac{n_1}{n_0+n_1}$), where these n_i dynamics are regulated by:

$$\frac{dn_0}{dt} = \lambda_0 n_0 - \mu_0 n_0 - q_{01} n_0 + q_{10} n_1 \tag{1.27}$$

$$\frac{dn_1}{dt} = \lambda_1 n_1 - \mu_1 n_1 - q_{10} n_1 + q_{01} n_0 \tag{1.28}$$

Leading to:

$$\frac{dx_0}{dt} = (\lambda_0 - \mu_0 - \lambda_1 + \mu_1)x_0(1 - x_0) - xq_{01} + (1 - x)q_{10} \tag{1.29}$$

which has to be put equal to zero to get the equilibrium frequency solution. So in the end they get:

$$L = \hat{x}_0 D_{R0} + \hat{x}_1 D_{R1} \tag{1.30}$$

To check consistency of the model it has been applied to several simulated trees.

1.3.2 Further models

In later years several models has been proposed to expand BiSSE, including continous values traits (QuaSSE - Fitzjohn 2010 [7]), multiple traits (MuSSE - Fitzjohn 2012 [8]), cladogenetic shifts on traits (ClaSSE - Goldberg & Igic [10]) or dealing with geographic

states (GeoSSE - Goldberg et al. [11]). All these models are included in the Diversitree package [8] and share the same structure as BiSSE model: numerical integration along branches to get $E(t)$ and $D(t)$ functions and operations on nodes dealing with traits to eventually get a likelihood function to maximize and test on simulated trees.

1.3.3 MuSSE - A multiple states model

In 2012 Fitzjohn expanded BiSSE to build a multi-trait model (MuSSE[8]). The process of integration is developed backwards in time as in BiSSE. The way to deal with the nodes at the branching times is basicly the same as in BiSSE but now each lineages contributes with a vector (indicated with index i).

1.3.3.1 Dynamics along branches

Equations are quite similar to BiSSE, but now there is one equation for each possible value of the considered trait:

$$\frac{dE_i(t)}{dt} = \mu_i - (\lambda_i + \mu_i + \sum_{j \neq i} q_{ij})E_i(t) + \lambda_i E_i(t)^2 + \sum_{j \neq i} q_{ij} E_j(t) \tag{1.31}$$

$$\frac{dD_{N,i}}{dt} = -(\lambda_i + \mu_i + \sum_{j \neq i} q_{ij})D_{N,i}(t) + 2\lambda_i E_i(t) D_{N,i}(t) + \sum_{j \neq i} q_{ij} D_{N,j}(t) \tag{1.32}$$

1.3.3.2 On nodes and root

On nodes the method is still the same:

$$D_{N',i}(t) = D_{N,i}(t) D_{M,i}(t) \lambda_i \tag{1.33}$$

At the root likelihood is calculated in the same way as BiSSE summing up all D_{Ri} with their weights.

1.3.3.3 Multiple traits simultaneously

With MuSSE it is also possible to analyse multiple traits at the same time. To do that it is necessary to define parameters in an appropriate way. For two traits for example it is possible to use a linear approach:

$$\lambda_{i,j} = \lambda_0 + \lambda_A X_A + \lambda_B X_B + \lambda_{AB} X_A X_B \tag{1.34}$$

$$\mu_{i,j} = \mu_0 + \mu_A X_A + \mu_B X_B + \mu_{AB} X_A X_B \tag{1.35}$$

in these expression X_A and X_B are indicator variables that are 1 when trait A and B (respectively) are in the "1" state , λ_0 (μ_0) is the "intercept" speciation (extinction) rate (if all traits are in state 0), λ_A and λ_B (μ_A and μ_B) are the "main effects" of traits A and B while λ_{AB} (μ_{AB}) is the interaction between those. For rates of transition instead:

$$q_{A01,j} = q_{A01,0} + q_{A01,B} X_B \tag{1.36}$$

where $q_{A01,0}$ is the intercept term and $q_{A01,B}$ is the main effect of trait B.

1.3.4 QuaSSE - A quantitative traits model

QuaSSE[7] is another BiSSE-like model in which the trait can span across a range of continous values. So the trait x can affect speciation $\lambda(x,t)$ and extinction $\mu(x,t)$ rates, that might be also function of time. The process is integrated backwards, like in the BiSSE model.

Here character evolution along lineages is modeled using a diffusion process, referring to this set of equations:

$$\Phi(x,t) = lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{-\infty}^{\infty} (z-x) g(z,t|x,t+\Delta t) dz \tag{1.37}$$

$$\sigma^2(x,t) = lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{-\infty}^{\infty} (z-x)^2 g(z,t|x,t+\Delta t) dz \tag{1.38}$$

$$0 = lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{-\infty}^{\infty} (z-x)^k g(z,t|x,t+\Delta t) dz \; k \geq 3 \tag{1.39}$$

where $g(z,t|x,t+\Delta t)$ is the transition probability density function for the diffusion process.

$\Phi$is the directional (drift) term while σ^2 is the diffusion term which is able to capture the stochastic elements of character evolution.

1.3.4.1 Core equations

In the same fashion of "Maddison et al. 2007[17]" all possible ways a lineage could either go extinct or following the observed clades are considered. This leads to a couple of core (partial) differential equations (for D and E) for the QuaSSE model, quite similar to the

old ones:

$$\frac{\partial D_N(x,t)}{\partial t} = \tag{1.40}$$

$$2\lambda(x)D_N(x,t)E(x,t) - (\lambda(t)+\mu(t))D_N(x,t) + \Phi(x,t)\frac{\partial D_N(x,t)}{\partial x} + \frac{\sigma(x,t)^2}{2}\frac{\partial^2 D_N(x,t)}{\partial x^2}$$

$$\frac{\partial E(x,t)}{\partial t} = \tag{1.41}$$

$$-\mu(x) + \lambda(x)E(x,t)^2 - (\lambda(t)+\mu(t))E(x,t) + \Phi(x,t)\frac{\partial E(x,t)}{\partial x} + \frac{\sigma(x,t)^2}{2}\frac{\partial^2 E(x,t)}{\partial x^2}$$

To solve these PDEs it is necessary to define properly the boundary conditions. Let's suppose that x spans a domain $[x_l, x_r]$. It is known that $E(x,0) = 0$ for every x and also that integration of $\int_{x_l}^{x_r} D_N(x,0) = 1$ so let's suppose a normal distribution for that $D_N = N(x_{obs}, \sigma_{obs})$. Besides all functions' derivatives $(\partial_x\mu(x), \partial_x\lambda(x), \partial_x E(x), \partial_x D_N(x))$ on domain borders x_l and x_r will be zero.

1.3.4.2 Calculations on nodes and root

At nodes between N and M clades everything is like BiSSE model:

$$D_{N'}(x,t) = D_N(x,t)D_M(x,t)\lambda(x) \tag{1.42}$$

At the root time t_R instead sums over D functions are replaced by an integration over all possible continous trait's values:

$$D_R = \int_{x_l}^{x_r} D_R(x,t_R)dx \tag{1.43}$$

this is equivalent to assign a flat prior to the character state at the root. Is it also possible (and this is what Fitzgerald actually does) to use the same distribution as weight:

$$D_R = \int_{x_l}^{x_r} D_R(x,t_R)\frac{D_R(x,t_R)}{\int_{x_l}^{x_r} D_R(y,t_R)dy}dx \tag{1.44}$$

Chapter 2

Research Questions

This chapter is NOT completed. It is composed, especially for RQ2-3-4, instead of a series of notes about how to formulate questions.

2.1 Can environment's cyclical changes drive allopatric speciation to huge local diversity?

IMPORTANT: ADD SOMETHING ABOUT THE FACT THAT ONE OF THE OPEN QUESTION IS TO UNDERSTAND IF SPECIATION PROCESS IS MORE FREQUENTLY ALLOPATRIC OR SIMPATRIC. IN FACT IF IT IS SYMPATRIC MORE SELECTION FORCES ARE INVOLVED.

Lake Tanganyika is estimated to be the second largest freshwater lake in the world by volume, and the second deepest as well as being the longest in the world. It spans 4 countries, Tanzania, the Democratic Republic of the Congo, Burundi, and Zambia. It is characterized by a extremely high abundance of cichlid fish species even though this lake is relatively young. This suggests that a huge and fast radiation process must have taken place on a relatively small time scale. This might be due to cyclic water level fluctuations changing connectiveness among different sub-areas of Lake Tanganyika (Janzen T. 2015[13]). In fact this process can create several separated pools allowing different allopatric speciation events to trigger at the same time, leading to an explosive radiation process.

In this context I propose a generalization to the standard birth-death model to include the possibility of multiple births occurring at the same time based on Etienne et al. Q-model [5] as a very general and effective framework to analyze diversity-dependance radiation processes.

However, after developing a multiple birth model it becomes mandatory to find a proper dataset to apply it. Even though I know that the first and natural candidate would be cichlid fish clade from lake Tanganyika, it is also true that current phylogenetic trees are including only one speciation event at each branching time. For this reason my next task will be to reconstruct a different kind of tree out of phylogenetic data.

One of the standard techniques to reconstruct trees out of molecular sequences is by using BEAST (Bayesian Evolutionary Analysis by Sampling Trees, Drummond & Rambaut 2007 [3]). BEAST is a tool that, given molecular alignments, can provide, via a bayesian statistical approach, the maximum clade credibility tree along with uncertainty on marginal likelihoods.

BEAST2 (Bouckaert et al. 2014 [1]) is a re-design of the above-mentioned BEAST that allows third party developers (like me) to write additional functionality that can be directly installed to the main program analysis platform. This will allow me to develop a method to get proper trees with occurrances of multiple births out of molecular data in order to apply my model.

2.2 Does Q equation from Etienne 2012 reduces to Nee et al. 1994 in standard birth-death context?

Here it is possible to look for an analytical solution both for Etienne et al. 2012 and for standard b-d for traits.

2.3 Effects of interaction between diversity dependance and traits on speciation and extinction rates

The main point is to consider both traits and missing species at the same time. This can bring to the table some difficulties because everything should be defined (this is one possible way, not the only one) in terms of some function $Q_{n_0,n_1}^{k_0,k_1}$, referring to Q model's notation, in which 0 and 1 stands for absence or presence of some trait (BiSSE-like). Seems to be quite challenging but it is definitely material to publish some paper.

2.4 Is there a link between standard b-d approach and BiSSE core functions? Can we get more information from this link? (more than provisional, for now only an idea)

It is a very bad defined idea for now. Seems that Lambert & Stadler 2013[16] have already faced this problem finding important results. Something more on this topic is included in Lambert et al. 2015[15] and Etienne et al. 2014[6]. There is a lot of hard stuff to read and it is very mathematical but it seems to include great insights for a research question that I may like.

Chapter 3

Approach

3.1 Q model

In chapter 1 I talked about Q model, or the way Etienne et al.[5] found to provide a likelihood function for a diversity-dependant process. To do that, as I mentioned before, becomes crucial to take into account both branches and nodes contributions.

Given the node at the time t_k in which number of species grows from $k-1$ to k, its contribute can be expressed by the action of a B_k operator:

$$Q_n^{k+1}(t_k) = B_k Q_n^k(t_k) = k\lambda_{n+k} Q_n^k(t_k) \tag{3.1}$$

To evaluate instead branch contribution let's denote with $A_k(t_k - t_{k-1})$ the solver of the dynamical equation1.22 between t_{k-1} and t_k having k fixed species from phylogeny, such that:

$$Q_n^k(t_k) = A_k(t_k - t_{k-1}) Q_n^k(t_{k-1}) \tag{3.2}$$

With this formalism it is possible to evolve $Q_n(t)$ to the present time t_p, from $k = 2$ crown species to q extant ones:

$$Q_n^{k=q}(t_p) = A_q(t_p - t_{q-1}) B_{q-1} A_{q-1}(t_{q-1} - t_{q-2}) \dots B_2 A_2(t_2 - t_1) Q_n^{k=2}(t_1) \tag{3.3}$$

Eventually it is possible to include any number of missing species, even none at all, taking a scalar product of the resulting $Q_n^{k=q}(t_p)$ with a proper vector of missing species v_m, to obtain the probability $P(\overrightarrow{t})$ of the phylogeny with branching times $\overrightarrow{t} = (t_1, t_2, \dots, t_{q-1})$:

$$P_{q,m}(\overrightarrow{t}) = v_m \cdot Q_n^{k=q}(t_p) \tag{3.4}$$

The final likelihood expression is obtained conditioning this probability to the survival of the crown species (as in Nee et al. 1994 [18]) dividing it by the probability $P_c(t_1, t_p)$ (obtained with similar mathematical arguments) that the $c = 2$ crown species at time t_1 have descendants at the present time t_p:

$$L_{q,m} = \frac{P_{q,m}(\overrightarrow{t})}{P_c(t_1, t_p)} \tag{3.5}$$

3.2 Mathematical framework for multiple births

The equation to study the dynamics of the process is:

$$\frac{dP_n(t)}{dt} = \left(\sum_{i=0}^{[n/2]} D_{n-i}(i)\lambda P_{n-i}(t) \right) + (n+1)\mu P_{n+1}(t) - \left(\sum_{i=0}^{n} D_n(i)\lambda P_n(t) \right) - n\mu P_n(t) \tag{3.6}$$

Using normalization property of distribution $D_n(i)$ is possible to simplify the third term on the right-hand side leading to:

$$\frac{dP_n(t)}{dt} = \left(\sum_{i=0}^{[n/2]} D_{n-i}(i)\lambda P_{n-i}(t) \right) + (n+1)\mu P_{n+1}(t) - \lambda P_n(t) - n\mu P_n(t) \tag{3.7}$$

- Here λ gets a different interpretation. In fact it acts as a trigger rate for water level changes and thus for speciation;

- For every infinitesimal time interval dt if a speciation event occurs, it can trigger a number of simultaneous births from 1 to the current number of species. For that reason on the right-hand side of the equation all this speciation modes must be considered and summed up;

- If a speciation event occurs, the number of species generated is described by a binomial probability distribution $D_n(i)$ of having i new species from a starting pool of n:

$$D_n(i) = \binom{n}{i} q^i (1-q)^{n-i} \tag{3.8}$$

- q is the probability for a single speciation event to occur while the binomial factor

$\binom{n}{i}$ takes into account all the possible ways to realize this;

- $D_n(i)$ naturally takes into account the influence of other species on the speciation. It is possible though to further include a carrying capacity K defining $q = q(K)$;

- The total number of species at each time seems to be very important for the whole process, so the tree could be "not breakable" suggesting that Q approach could be the best way to go.

3.3 Likelihood estimations

To fulfill these tasks I am going to use Likelihood models similar to techniques I introduced in chapter 1. So the first step will be to define a set of good functions to describe phenomena (e.g. $E(t)$ and $D(t)$ in Maddison et. al 2007 [17]), to whose give an interpretation useful to link them directly with the likelihood function. Then a set of equations is needed to describe the dynamics of these functions along the whole diversification process. In this way, integrating them over all the branch lines and properly taking into account nodes contribution (e.g. Maddison et al. 2007[17], Etienne et al. 2012[5]) the entire diversification process can be, hopefully, reconstructed despite the complexity of the process.

3.3.1 Likelihood maximization

Several packages are available in R to maximize likelihood function (e.g. subplex [20], nlm [21], mle) in order to estimate best parameter choice for a given model. These will provide me the necessary tools to assess the goodness of the new model, compared to a null model.

3.4 Tests

So far I explained that all needed informations about model and parameters are carried out via the likelihood function. Remarkably, likelihoods can be also used to compare different models to determine which is better and to which extent.

For example is possible to define likelihood ratio test. Let's suppose to have a null model named M_0 (for example a standard birth-death model) and an alternative hypotesis that we want to validate or reject (for example an alternative multiple birth model). Then let's suppose it is possible to carry out two likelihood function L_0 and L_1 for the two models. A likelihood ratio is defined as:

$$\lambda(x) = \frac{sup(L(\underline{\theta}|x) \ : \ \underline{\theta} \in \Theta_0)}{sup(L(\underline{\theta}|x) \ : \ \underline{\theta} \in \Theta_1)} \tag{3.9}$$

where x are data, $\underline{\theta}$ are parameters and Θ_0 and Θ_1 are parameters' sets respectively for M_0 and M_1. Only superior values are considered as likelihoods are considered already maximized. $\lambda(x)$ is always in the interval $[0, 1]$. Small values of $\lambda(x)$ are evidence in favor of the alternative model M_1.

In general it is possible to define a likelihood test distribution $\lambda(X)$, with X being the entire set of possible data. Notably it is known from Wilk's theorem that, if X sample size is enough large, $\Lambda(X) = -2log(\lambda(X))$ tends to distribute as a χ^2 distribution with a number of degrees of freedom equal to $m = dim(\Theta_0) - dim(\Theta_1)$.

Given the distribution $\lambda(X)$, a significance test can be performed arbitrary choosing a value $k \in [0, 1]$ to get a significance level α:

$$P(\lambda(X) \leq k) = \alpha \tag{3.10}$$

3.4.1 Comparing models with AIC

In general, comparing models, one could be also interested in considering the complexity of candidate models as, usually, models involving a minor complexity are preferred. One of the standard techniques to assess quality of models considering the tradeoff between accuracy and complexity is using Akaike's Information Criteria (AIC):

$$AIC = 2k - 2log(L) \tag{3.11}$$

where k is the number of parameters introduced by the model. What it is demanded to a model is to provide the highest possible likelihood with the minimum number of parameters involved, so the model is as good as AIC is low. So, given a null model M_0 and a set of candidate models $\{M_i\}_{i=1}^{R}$ is it possible to evaluate the goodness of each one defining first AIC differences with respect to the null model:

$$\Delta_i = AIC_i - AIC_0 \tag{3.12}$$

and then defining Akaike's weights in the following way:

$$\omega_i = \frac{exp(-\frac{1}{2}\Delta_i)}{\sum_{r=1}^{R} exp(-\frac{1}{2}\Delta_r)} \tag{3.13}$$

Differently from the simple AIC here best models in the set are denoted by a bigger weight.

3.4.2 Check with population dynamics simulations

Given a process (e.g. standard b-d process or its derivatives) it is usually quite simple to run simulations. It is infact only necessary to take into account all possible outcomes (e.g. speciation, extinction events for b-d, but also rate shift for BiSSE-like models and so on) for every infinitesimal time interval to reconstruct one of the possible dynamics for the total number of species in time.

Thus considering a sufficiently great number of different runs it is possible to reconstruct, to some extent, the average behaviour to test versus the result obtained via direct integration of the model.

Given its simplicity, this is actually a pretty standard strategy and it is always worth to use as a preliminary test.

3.4.3 Check with tree simulations

Another standard strategy to assess the validity of a theoretical diversification model is by generating simulated trees. This is widely used in literature (e.g. Etienne et al. 2012[5], Maddison et al. 2007[17], Fitzjohn 2010 [7] and many others) and it is considered the most important test to check consistency of a given model. The idea is very simple: in every case authors can decide beforehand which staple features these trees must respect, such as number of tips, speciation and extinction rates (in case diversity or time dependant), carrying capacity, crown age and so on. Then a family of tree is produced to act as simulated data to test the model, producing a likelihood and estimating back parameters. Varying parameters other family of trees can be produced to perform again the tests. This provides a very robust way to assess the power of the model on estimating parameters when data are completely given.

Chapter 4

Time Planning

$1^{st} year$	$2^{nd} year$	$3^{rd} year$	$4^{th} year$
RQ1			
	RQ2		
	RQ4	RQ4	
		RQ3	RQ3
			Thesis writing

Table 4.1 – Estimated time planning.

Bibliography

[1] Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard, M. A., Rambaut, A. & Drummond, A. J. Beast 2: A software platform for bayesian evolutionary analysis. *PLOS*, 2014.

[2] Dennis, J. E. and Schnabel, R. B. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, NJ, 1983.

[3] Drummond, A. J. & Rambaut, A. Beast: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology*, 2007.

[4] Etienne, R. S. & Haegeman, B. A conceptual and statistical framework for adaptive radiations with a key role for diversity dependence. *The American Naturalist*, 2012.

[5] Etienne, R. S., Haegeman, B., Stadler, T., Aze, T., Pearson, P. N., Purvis, A. & Phillimore, A. B. Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. *Proceedings of the royal society*, 2012.

[6] Etienne, R. S., Morlon, H., Lambert, A. Estimating the duration of speciation from phylogenies. *Evolution*, 2014.

[7] Fitzjohn, R. G. Quantitative traits and diversification. *Systematic Biology*, 2010.

[8] Fitzjohn, R. G. Diversitree: comparative phylogenetic analyses of diversification in r. *Methods in Ecology and Evolution*, 2012.

[9] Fitzjohn, R. G., Maddison, W. P. & Otto, S. P. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Systematic Biology*, 2009.

[10] Goldberg, E. E. & Igic, B. Tempo and mode in plant breeding system evolution. *Evolution*, 2012.

[11] Goldberg, E. E., Lancaster, L. T., Ree, R. H. Phylogenetic inference of reciprocal effects between geographic range evolution and diversification. *Systematic Biology*, 2011.

[12] Haegeman, B. & Etienne, R. S. Entropy maximization and the spatial distribution of species. *The American Naturalist*, 2010.

[13] Janzen, T. *What lies beneath? How patterns in ecology and evolution inform us about underlying processes*. PhD thesis, Rijksuniversiteit Groningen, 2015.

[14] Kendall, D. G. On some modes of population growth leading to r.a. fisher's logarithmic series distribution. *Biometrika*, 1948.

[15] Lamber, A., Morlon, H., Etienne, R. S. The reconstructed tree in the lineage-based model of protracted speciation. *Journal of Mathematical Biology*, 2015.

[16] Lambert, A. & Stadler, T. Birth-death models and coalescent point processes: The shape and probability of reconstructed phylogenies. *Theoretical Population Biology*, 2013.

[17] Maddison, W. P., Midford, P. E. & Otto, S. P. Estimating a binary character's effect on speciation and extinction. *Systematic Biology*, 2007.

[18] Nee, S., May, R. M. & Harvey, P. H. The reconstructed evolutionary process. *Philosophical Transactions: Biological Sciences*, 1994.

[19] Rabosky, D. L. & Lovette, I. J. Density-dependent diversification in north american wood warblers. *Proc. R. Soc. B*, 2008.

[20] Rowan, T. *Functional Stability Analysis of Numerical Algorithms*. PhD thesis, Department of Computer Sciences, University of Texas at Austin, 1990.

[21] Schnabel, R. B., Koontz, J. E. and Weiss, B. E. A modular system of algorithms for unconstrained minimization. *ACM Trans. Math. Software*, 1985.

[22] Stadler, T. Mammalian phylogeny reveals recent diversification rate shifts. *PNAS*, 2011.