A short course on trait evolution modeling

Marc Manceau

December 5, 2017

Abstract

This document is a summary of an introductory course given to Biology students in their first year of Master's degree at the École Normale Supérieure in Paris. It is part of a special day organized by Hélène Morlon, focusing on *macroevolution*, the study of evolutionary processes over long time-scales.

Contents

1 Introduction

1.1 A huge variety of traits and questions

The main sources of raw data available to biologists to study life evolution over long time-scales are the various traits caracterizing distinct organisms. These can be molecular traits, like amino-acid or nucleotidic sequences. It could also be morphological traits, like body mass, presence/absence of a given organ, eye color. Or it could be a life history trait, like the life-span, or some reproductive behavior.

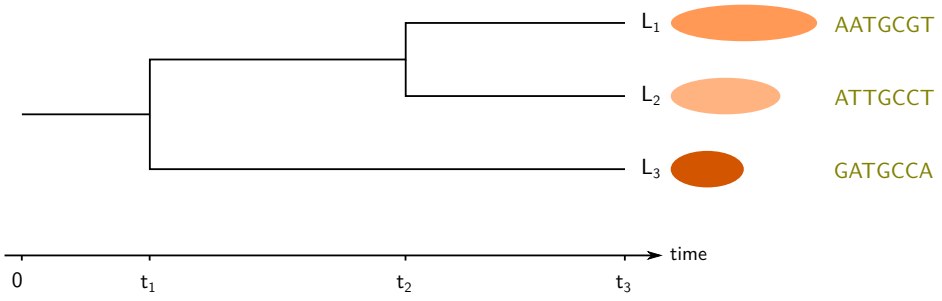

Figure 1: Traits (color, size, genetic sequences...) of the organisms at present time t_3 are the product of some evolutionary processes unfolding on the phylogenetic tree.

Whatever the traits, we consider that they are the product of some evolutionary processes unfolding on the phylogenetic tree linking our organisms. Their observation allows us to address many questions, for example :

- How could we use our traits to reconstruct the most likely ancestral relationships between individuals ?
- What did ancestral organisms look like ?
- At which speed do distinct traits evolve ?
- Can we see the signature of selective constraints leading the joint evolution of distinct traits ?
- Which are the main processes influencing trait evolution ? Is there a signal for biotic versus abiotic drivers ?

1.2 One main method

All these questions are studied using the same type of probabilistic models. The goal of this course is to get an idea of what these are, how they work, and which questions they allow to address. Everything relies on Markov processes. Here, we only present the basics that we need to get an intuition of how these models work, but we also ask you to admit many results. In order to get a more in-depth understanding of these processes, which are really accessible to you with your math background, we strongly recommend to follow Amaury Lambert's courses.

2 Trait evolution in a discrete state space

2.1 Markov chains in discrete time

Let's call, $\forall n \in \mathbb{N}$, X_n the random variable which value gives us the state of the nucleotide at step n.

Definition : We say that (X_n) is a Markov chain if $\forall n \in \mathbb{N}^*$, we have,

$$\mathbb{P}(X_n = x_n | X_0 = x_0, X_1 = x_1, ..., X_{n-1} = x_{n-1}) = \mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1})$$

This intuitively means that the state of the nucleotide at any time step only depends on the past through the value it had just before. We also sometimes say that the process has no memory.

Markov chains in discrete time are thus usually described with the law of the initial state, here X_0, and a transition matrix P, which gives the probability, for any two states, to go from the first one to the second one. In our example, the state space is $\{A, T, G, C\}$. We could assume that the initial state X_0 of our nucleotide is uniformely chosen on $\{A, T, G, C\}$, and we could further define a transition matrix such as the one shown here :

$$\pi_0 = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$$

$$P = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \end{pmatrix}$$

This matrix means that, at each time-step, the nucleotide state changes and takes a new value uniformly chosen among the three other values. Figure 2 illustrates the state space and gives an example of the Markov chain trajectory.

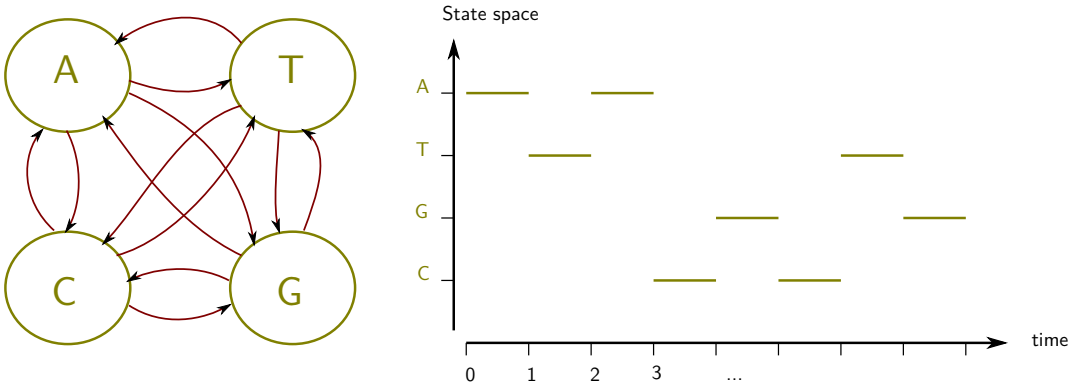

Figure 2: The state space $\{A, T, G, C\}$ of our Markov chain, together with an example of trajectory.

Markov chains in discrete time are very useful to model any trajectory where it makes sense to discretize the time-steps. However, in our case, the time corresponds to the real continuous time. Fortunately, there is an extension of Markov chains to continuous time, which is known as Markov processes.

2.2 Markov processes in continuous time

In order to get a continuous-in-time process we can consider infinitesimal timesteps. Over each of these, the process might jump from one state to the other, at a given rate. We also somehow need the simplifying assumption that at any time in the trajectory, the future does not depend on the past, but only depends on the present value. More formally, we want to consider the following definition :

Definition : A Markov process $(X_t,\ t \in [0, +\infty[)$ is a stepwise right-continuous process satisfying the following property, known as the Markov property, which is analoguous to the one presented before for Markov chains : $\forall 0 < t_1 < ... < t_{n-1} < t_n$,

$$\mathbb{P}(X_{t_n} = x_n | X_{t_{n-1}} = x_{n-1}, ..., X_{t_1} = x_1, X_0 = x_0) = \mathbb{P}(X_{t_n} = x_n | X_{t_{n-1}} = x_{n-1})$$

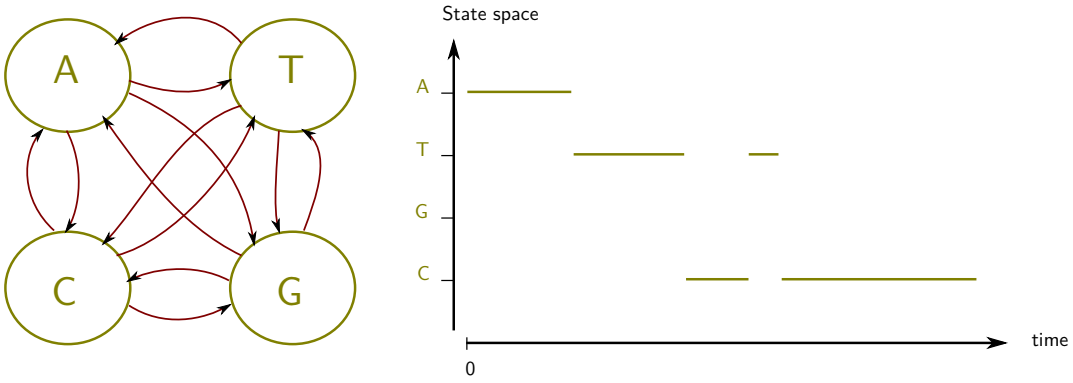

Figure 3: The state space $\{A, T, G, C\}$ of our Markov chain, together with an example of trajectory.

We get a memory-less process in continuous time by relying on the exponential distribution to describe the waiting time before the next jump.

A Markov process with the same state space as before $\{A, T, G, C\}$, but continuous in time, is described by its starting distribution π_0 and the instantaneous matrix of rates $Q = (q_{ij})$, where $\forall i \neq j,\ q_{ij}$ is the rate at which a process in state i goes to state j, and q_{ii} is fixed by convention at $q_{ii} = -\sum_{i \neq j} q_{ij}$.

Intuitive description of the process

1. It starts first in a state drawn in the π_0 distribution.
2. When it is in state i, the waiting time before the next jump follows $\mathcal{E}(-q_{ii})$.
3. At a jumping time, the next state is j with probability $\frac{q_{ij}}{-q_{ii}}$.

One good way to get the intuition behind Markov process is to provide a simple example. Let's consider again the state space $\{A, T, G, C\}$. We now define a Markov process, well-known as the Jukes-Cantor model, parametrized by the substitution rate m :

$$\pi_0 = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$$

$$Q = \begin{pmatrix} -3m & m & m & m \\ m & -3m & m & m \\ m & m & -3m & m \\ m & m & m & -3m \end{pmatrix}$$

Figure 3 illustrates one realization (simulation) of the process through time, on a single interval. Here, the initial state is A. We then simulate the next time at which something happen : it follows an exponential distribution with parameter $3m$. At jumping time, the new state is uniformly chosen on $\{T, G, C\}$. And we iterate through time.

Interestingly, the standard theory on Markov processes also provides us a way to compute the transition matrix $P(t) = (p_{ij}(t))_{i,j}$ where $p_{ij}(t)$ is the probability, starting in state i at time 0, to be in state j after a time t. We again recommend to follow Amaury Lambert's course to derive and understand everything about this.

2.3 Considering a process running on a phylogeny

We now consider the same process, but we no longer make it run on a single interval. Instead, we would like the process to evolve on a dated phylogenetic tree. As before, we consider that the process starts at the root of the tree with a given law π_0. The process then runs on the root branch, between $t = 0$ and t_1 like previously on a simple interval. At the branching time t_1, we get one more lineage. We consider that the process splits into two independent processes, starting with initial value X_{t_1} and having the same law as before.

And so on recursively. Between any two branching points, the process evolves like presented on a simple interval. It then splits into independent processes having the same law in the daughter branches. Until reaching the leaves,

where it is killed. Figure 4 presents the whole trajectory of the process unfolding on a fixed tree.

In empirical problems, we never have access to the full (past) trajectory. We will see that our probabilistic approach nevertheless allows us to compute the probability of the present-day observation easily.

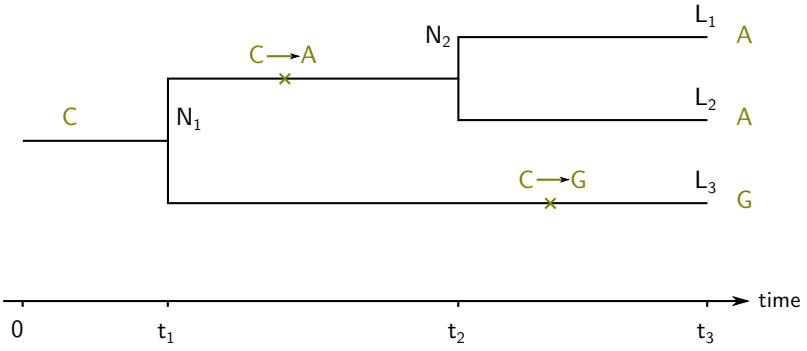

Figure 4: The nucleotide state evolves through time on the phylogeny. It corresponds to the example exposed in the main text.

For any two states, we know the probability to go from one state to the other in any amount of time. We consider the small tree in Figure 4, with the leaf observations $(X_{L_1}, X_{L_2}, X_{L_3}) = (A, A, G)$. The total probability formula allows us to write the probability, summing on all possible states for the internal nodes. It leads to,

$$\begin{aligned}
&\mathbb{P}(X_{L_1} = A, X_{L_2} = A, X_{L_3} = G) \\
&= \sum_{(i,j)\in\{A,T,G,C\}^2} \mathbb{P}(X_{L_1} = A, X_{L_2} = A, X_{L_3} = G, X_{N_1} = i, X_{N_2} = j) \\
&= \sum_{(i,j)\in\{A,T,G,C\}^2} \mathbb{P}(X_{N_1} = i)\mathbb{P}(X_{N_2} = j|X_{N_1} = i)\mathbb{P}(X_{L_1} = A|X_{N_2} = j)\mathbb{P}(X_{L_2} = A|X_{N_2} = j)\mathbb{P}(X_{L_3} = G|X_{N_1} = i)
\end{aligned}$$

This would quickly become very intensive on bigger trees, because of this sum over all internal states. Fortunately, there is a useful algorithm to simplify this calculus, which is called *tree pruning*.

Consider the simple rearrangement of the sum :

$$\begin{aligned}
&\mathbb{P}(X_{L_1} = A, X_{L_2} = A, X_{L_3} = G) \\
&= \sum_{i\in\{A,T,G,C\}} \mathbb{P}(X_{N_1} = i)\mathbb{P}(X_{L_3} = G|X_{N_1} = i) \left(\sum_{j\in\{A,T,G,C\}} \mathbb{P}(X_{L_1} = A, X_{L_2} = A|X_{N_2} = j)\mathbb{P}(X_{N_2} = j|X_{N_1} = i) \right)
\end{aligned}$$

With this rearrangement of the terms, we now remark that it is sufficient to recursively compute, at each node w, the probability to get the observations at the leaves X_{w_1} and X_{w_2} knowing the trait values X_w, for all four possible state. This simple observation simplifies very much the computation of our probability.

If we call L_w the 4 elements vector $(\mathbb{P}(\text{ observing the trait values subtended by node } w \mid X_w = i))_{i\in\{A,T,G,C\}}$, here is a simple recursive algorithm to compute it :

1. if w is a leaf, $L_w = (\mathbb{1}_{\text{ observing } A}, \mathbb{1}_{\text{ observing } T}, \mathbb{1}_{\text{ observing } G}, \mathbb{1}_{\text{ observing } C})$.
2. otherwise, $L_w = (P(t_1)L_{w_1}) \cdot (P(t_2)L_{w_2})$ (where the branch leading to w_1 has length t_1 and the branch leading to w_2 has length t_2).

where $\cdot$ stands for the entrywise product.
Finally, the probability of our observations is obtained by multiplying the conditional probability at the root ρ with the law of its state :

$$\mathbb{P}(\text{ nucleotide state observations at the leaves }) = \pi_0 L_\rho$$

This allows us to compute the probability of our observation at one nucleotide site. Fortunately, there is always much more than only one nucleotide in available biological sequences. If our raw data is a multiple alignment of 10000 nucleotides, people usually consider that nucleotides are independent and identically distributed, and each follows the same Markovian process. We thus compute the probability of the whole alignment by multiplying the probability to observe present-day state at each nucleotide. Because this fairly quickly becomes a very low number, we consider for computational reasons the log-probability,

$$\ln \mathbb{P}(\text{ observing the whole alignement }) = \sum_{n=1}^{10000} \ln \mathbb{P}(\text{ nucleotide states at site } n)$$

2.4 Application to biological questions

The principle of parameter estimation by likelihood maximization comes up now and allows us to make statistics with our data. Its verbal formulation is the following : we estimate the parameters of our models by maximizing the probability to observe the data we have under our model. More formally,

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\text{argmax}}\ \mathbb{P}(\text{ our data observations };\ \theta)$$

In our example of nucleotide evolution, many parameters may be of interest.

Molecular evolution parameters We consider a model slightly more general than the Jukes-Cantor model that we were considering before. We suppose now that there is a difference between the transitions ($A \leftrightarrow G$, $T \leftrightarrow C$) and transversions (all other substitutions). The Markov process we are interested in is attributed to Kimura and is parametrized with :

$$\pi_0 = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$$

$$Q = \begin{pmatrix} -(2\beta+\alpha) & \beta & \beta & \alpha \\ \beta & -(2\beta+\alpha) & \alpha & \beta \\ \beta & \alpha & -(2\beta+\alpha) & \beta \\ \alpha & \beta & \beta & -(2\beta+\alpha) \end{pmatrix}$$

If we consider that we know the tree linking our species, we could consider optimizing the likelihood to fit the transition rate and transversion rate :

$$(\hat{\alpha}_{\text{MLE}}, \hat{\beta}_{\text{MLE}}) = \underset{\alpha,\beta}{\text{argmax}}\ \mathbb{P}(\text{ our data observations };\ \alpha, \beta)$$

Dating a phylogeny We consider now that we know the topology of the phylogenetic tree. We also know the substitution rate, and we would like to use it to date the tree. The parameters here are the $(t_i)_{i=1}^3$ values, and we are looking for :

$$(\hat{t_1}, \hat{t_2}, \hat{t_3}) = \underset{t_1,t_2,t_3}{\text{argmax}}\ \mathbb{P}(\text{ our data observations };\ t_1, t_2, t_3)$$

Reconstructing the phylogeny If we were interested in reconstructing the phylogeny of a set of species, we could use such an approach, and optimize the likelihood directly over the parameter T, which is a dated, binary, tree. The maximum likelihood reconstruction of our phylogeny would thus be :

$$\hat{T} = \underset{T}{\text{argmax}}\ \mathbb{P}(\text{ our data observations };\ T)$$

It might be more complicated to explore the space of all dated binary trees and a whole litterature exists on this subject. We can just remember that this simple model, depending on what is fixed and what is to be inferred, can be useful to answer a huge number of questions.

Transition Many other discrete traits might be modeled in the same way. It could be appropriate to model in the same way life history traits like the migratory versus sedentary behavior among birds. Some morphological traits might also be well described with a discrete state space, for example, the number of digits among tetrapods. But still, many other traits might be best described using a continuous state space : for example the size, mass, color, longevity, or some shape descriptors of our organisms...

3 Trait evolution in continuous state space

3.1 A word on the Brownian motion

The Brownian motion is the basic process which modelers use to model continuous traits. Here is a quick reminder of the intuition we need.

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

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

This random walk is represented in Figure 5i). By simply changing the scale at which we look this random walk, we want to get a non-trivial continuous in time process. Let's try something simple first : what is the state of the walk if we zoom out in time and state space by a factor n ?
Remark that $\mathbb{E}(\xi_i) = 0$. Applying the law of large numbers, we get, $\forall t > 0$:

$$\frac{1}{n} S_{[nt]} = \frac{[nt]}{n} \frac{1}{[nt]} \sum_{i=1}^{[nt]} \xi_i \ \longrightarrow\ 0$$

This is really not the most interesting scale change we can propose, because it leads at the limit to a trivial process (Figure 5ii)). Let's try something a bit different, by zooming out a bit less on the state space, with an order $\sqrt{n}$ instead of n, as in Figure 5iii). The central limit theorem then gives us the convergence in law :

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

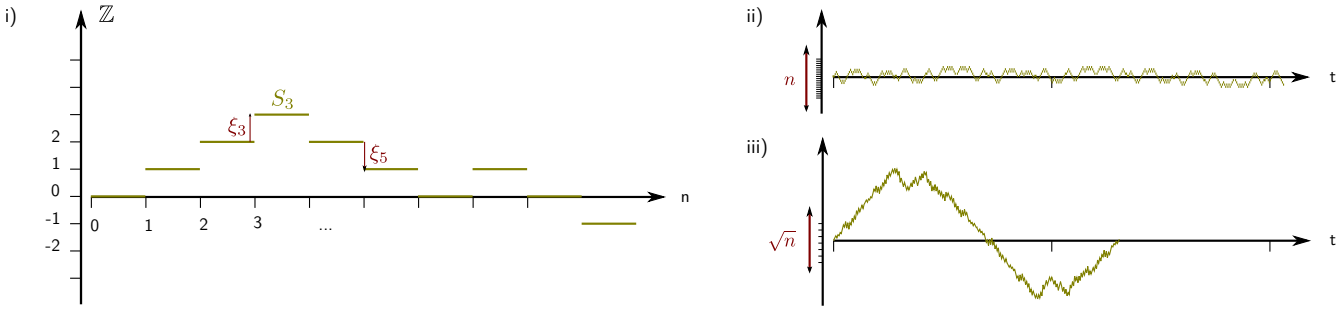

Figure 5: From a random walk on $\mathbb{Z}$ to the Brownian motion. i) S_n is a random walk on $\mathbb{Z}$, with iid increments (ξ_i). ii) Changing the scale by zooming out by a factor n, then letting $n \to \infty$, the process becomes trivially zero. iii) Zooming out a bit less in the state space leads to a non-trivial process, with Gaussian, independent, increments.

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

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

1. $t \mapsto B_t$ is continuous
2. $\forall s, t$, the law of $(B_{t+s} - B_t)$ does not depend on t
3. $(B_s)_{s\leq t}$ and $(B_{t+s} - B_t)_{s\geq 0}$ are independent
4. $B_0 = 0$
5. $B_1 \sim \mathcal{N}(0,1)$.

We need only to remember that it's a continuous process, with random, Gaussian increments between any two time points.

3.2 Considering a Brownian motion running on a phylogeny

As previously with our Markov processes in a discrete space, we make our Brownian motion run on a phylogeny, as shown in Figure 6. We also consider a non-standard Brownian motion, accelerated with a parameter σ, meaning that $X_1 = \sigma B_1 \sim \mathcal{N}(0, \sigma^2)$, and we call *evolutionary rate* of our trait this parameter. The process starts at the root of the tree at a given value X_0. At each branching point, it splits into two independent Brownian motions, starting at the same value. As a result, sister taxa tend to have closer trait values than distantely related species.

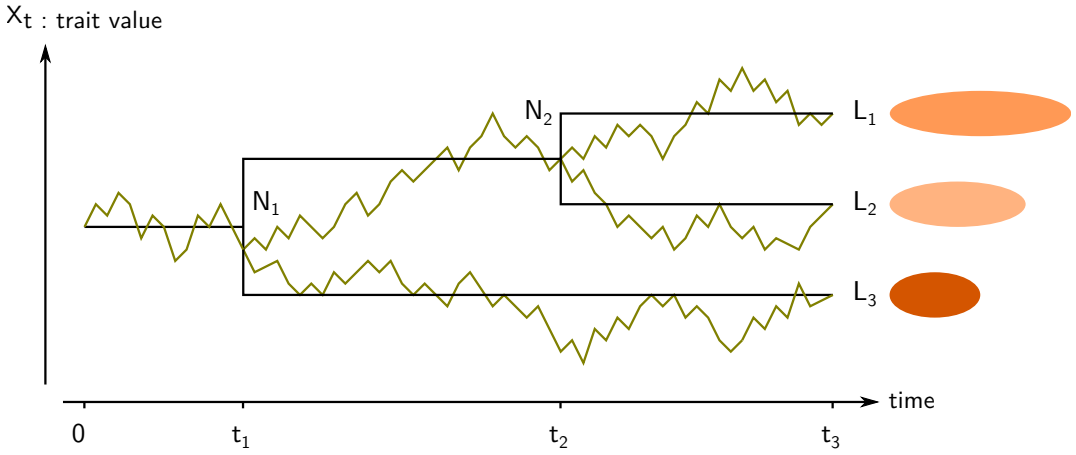

Figure 6: Evolution of a continuous trait on a phylogeny. At each branching point, the process splits into two independent processes starting at the same value.

The description above is useful to understand the forward-in-time simulation process illustrated in Figure 6. However, we would also like to get the law of the trait values at present in each of our species. This is key to our ability to make inferences with empirical data. Because we are dealing with continuous traits, we look for the probability density of a vector of traits at present $(X_{L_1}, X_{L_2}, X_{L_3})$.

Taken individually, each trait value is Gaussian, with mean 0 and variance $\sigma^2 t_3$. But this does not tell us much about the joint law of $(X_{L_1}, X_{L_2}, X_{L_3})$. We can do a bit more, by breaking up the process into branch pieces which are independent by construction. We see that :

$$\begin{aligned}
\mathrm{Cov}(X_{L_1}, X_{L_2}) &= \mathrm{Cov}\left(X_{N_2} + (X_{L_1} - X_{N_2}), X_{N_2} + (X_{L_2} - X_{N_2})\right) \\
&= \mathrm{Cov}(X_{N_2}, X_{N_2}) \\
&= \sigma^2 t_2 \\
\mathrm{Cov}(X_{L_1}, X_{L_3}) &= \mathrm{Cov}\left(X_{N_1} + (X_{L_1} - X_{N_1}), X_{N_1} + (X_{L_3} - X_{N_1})\right) \\
&= \mathrm{Cov}(X_{N_1}, X_{N_1}) \\
&= \sigma^2 t_1 \\
\mathrm{Cov}(X_{L_2}, X_{L_3}) &= \mathrm{Cov}\left(X_{N_1} + (X_{L_2} - X_{N_1}), X_{N_1} + (X_{L_3} - X_{N_1})\right) \\
&= \mathrm{Cov}(X_{N_1}, X_{N_1}) \\
&= \sigma^2 t_1
\end{aligned}$$

We call $\Sigma = (\mathrm{Cov}(X_i, X_j))_{i,j}$ the covariance matrix between any two leaf traits, and $m = (\mathbb{E}(X_i))_i = (0, 0, 0)$ the expectation vector. The following result will not be surprising because it seems very similar to the Gaussian density that we are used to in $\mathbb{R}$. But for simplicity we will admit it here : the law of $(X_{L_1}, X_{L_2}, X_{L_3})$ is called *multivariate normal* and is simply parametrized by the mean vector m and the covariance matrix Σ. The density of the vector is given by the following function :

$$\forall x \in \mathbb{R}^3, \quad p(x) = \frac{1}{\sqrt{(2\pi)^3 \det \Sigma}} e^{-\frac{1}{2}(x-m)^T \Sigma^{-1} (x-m)}$$

3.3 Application to biological questions

Having this law of tip data in hand, we can start making inferences. We use the maximum likelihood principle to infer knowledge on the processes that have generated our observed data. Instead of maximizing the probability to observe our data, we maximize the density of the observations. More formally, we look for :

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\text{argmax}}\ p(\text{ our data observations }\ ;\ \theta)$$

Typically, on the example above, we can assume that we are given the phylogeny. We used only one other parameter, σ. This is the rate at which the trait evolves, and its maximum likelihood estimator is :

$$\hat{\sigma}_{\text{MLE}} = \underset{\sigma}{\text{argmax}}\ p(\text{ our data observations }\ ;\ \sigma)$$

Most of the time, this type of likelihood optimization is achieved on a computer. Yet, in this particular example, we could derive the estimator formula analytically. (Exercise : look at the optima of the log-likelihood function by deriving it with respect to σ.)

3.4 Introduction to more complex models of continuous trait evolution

More complex models might be considered using the same kind of framework. I provide here a rapid overview of recent questions in this research area.

First, we could be interested to look for periods of time that are likely to have distinct evolutionary rates. For example, competition release after a mass extinction might lead to an accelerated rate of trait evolution. Or some climatic event might have had the opposite effect, and so on. Considering two periods of time, we could write a model with two evolutionary rates, one in each of our two periods, and use likelihood optimization to fit these parameters with our data. Because we can compute the likelihood under two distinct scenarios, we could design a statistical test aimed at testing the equality or difference of the two evolutionary rates.

In another study, with another trait dataset, we could be interested to look for some branches, which could show distinct evolutionary rates. For example, a key innovation might lead to competition release in some part of the phylogeny, in which many traits would suddenly evolve faster. Again, we can fit a model with distinct evolutionary rates in distinct subclades, and we could design a test aimed at inferring wheter or not this scenario is likely to have happened.

Many more studies have tried to look for factors influencing evolutionary rates. Models have been designed to make them depend on the temperature, or on the number of species in a clade. The parametrization might become quite different, and there might be some technical difficulties, but they rely on the same idea.

Other studies are interested in the correlation between distinct traits. However, there could be a correlation between two different traits for the only reason that they evolved on the same tree. In order to distinguish between a scenario where two traits evolve independently on the phylogeny, and a scenario where one trait is tightly linked to the other one for selective reason, maximum likelihood optimization is, again, a good option.

Finally, some other stochastic processes, a bit different from the Brownian motion, are being used to model distinct evolutionary modes, for example for traits evolving under strong stabilizing selection, or for traits attracting or repulsing each other. They, however, always rely on the same kind of framework, considering stochastic processes running on a tree.

4 Conclusion

4.1 Back to the main idea behind these models

Two take-home messages :

- Whatever trait we are interested in, we could consider building a Markov process in the appropriate space, and make it run on a tree, with a split into two new independent processes at each branching time.

- Macroevolutionary inferences rely on likelihood optimization. It consists in maximizing the probability to observe the present-day trait values, upon some unknown parameter. Quite similarly, model selection consists in comparing the maximum likelihood of distinct modeled scenarios.

4.2 Getting further : studying jointly trait evolution and diversification

These types of models are routinely used by the macroevolutionary research community. There is also a current tendency to develop more complex models, linking for example diversification and trait evolution. They rely on the idea that some traits might lead to more or less rapid diversification. The model called BiSSE (for *Binary State Speciation and Extinction*, in Maddison et al. (2007)) makes a connection between this course and the previous one.

Consider that we follow lineages, that have one trait either of type 0, or of type 1. We start a branching process with one lineage, say, of type 0. It might :

- speciate, and give birth to two type-0 lineages at rate λ_0
- die, at rate μ_0
- become of type 1, at rate $q_{0\to1}$.

Similarly, type-1 lineages can speciate at rate λ_1, die at rate μ_1 or become of type 0, at rate $q_{1\to0}$. The process runs from the root to the present.

This model may be fitted to empirical data to get an idea of the influence of some binary trait on the speciation-extinction dynamics. For example, it has been applied to the study of migratory/sedentary behaviour in birds. In this case, we must compute the probability to observe the bird phylogeny, together with the trait at present. For more information, and to get the biological results, see Rolland et al. (2014).

References

Maddison, W. P., P. E. Midford, and S. P. Otto. 2007. Estimating a binary character's effect on speciation and extinction. Syst. Biol. .

Rolland, J., F. Jiguet, K. A. Jønsson, F. L. Condamine, and H. Morlon. 2014. Settling down of seasonal migrants promotes bird diversification. Proc. R. Soc. B 281:20140473–20140483.