What fruits can we get from this tree?

Laudanno, Giovanni

DOI:
10.33612/diss.155031292

Chapter 7

Summary

This thesis provides new tools for extracting information on the process of diversification from a phylogenetic tree. To do so the standard approach is to employ likelihood functions in order to estimate the best parameters for the diversification models via likelihood maximization. The parameters for this kind of models usually represent the rates at which the various evolutionary events (e.g., speciations or extinctions) can take place in the process. As presented in chapter 1 many diversification models are already available. They can answer different questions and therefore the parameters that are estimated in each case reflect several biological aspects. Many of these methods rely on the solution of the so-called P-equation, used in standard Birth-Death models (BD), presented in 1.3.2 and are applied according to the Nee et al. framework (sometimes referred to as P-framework in this thesis) using eqs. 1.3.3. Later in the same chapter we also presented the so-called Q-framework 1.3.6. Such framework, originally developed to deal with diversity-dependent diversification allows to keep track of species number along the process. As a consequence this approach can accommodate a broader set of problems, if expanded, as in Valente, Phillimore, and Etienne (2015).

In chapter 2 we studied the Q-framework in all cases in which it is also possible to apply the P-framework (n- and ρ- sampling schemes, constant rates and time-dependent rates). In such cases, we analytically proved that the Q-framework yield solutions that are equivalent to those provided by models already available in the literature. Such a proof was needed, as the paper where the Q-framework

was originally introduced only provided an heuristic justification obtained using numerical methods. This is far from optimal and can only be considered true for a finite number of parameter settings. Since the framework had the potential to be expanded to new cases we needed stronger foundations. After providing additional analytical support for it, we used it in two different other chapters of this thesis.

In chapter 3 we established the core equations for the Multiple-Births and Death (MBD) model using the Q-framework. This model allows for explosive bursts of simultaneous speciation events of which the intensity depends on the current number of species. This is not a general model, but it is tailored specifically to deal with crowded phylogenies. The focus is to study how the effect of (cyclic) environmental changes can influence the phylogenetic history of a clade. We showed that we can reliably retrieve the parameters with maximum likelihood for a large range of simulated phylogenies (with known parameters). We also studied whether the BD likelihood is adequate for capturing the characteristics of an MBD process. We found that this is not trivial and developed a new metric, the DNBT metric, that can distinguish between BD and MBD trees.

In chapter 4 the goal was not to build an entirely novel model. Instead we focussed on improving some of the models already available in literature that account for the influence of a single lineage shift on phylogenetic likelihoods. This is the case, for example, when a single lineage in the clade develops a key innovation. If this occurs, such a lineage can escape from competition with other species in the clade (as in Etienne and Haegeman, 2012), diversifying at a higher pace with respect to the background. In this chapter we first identified critical aspects of current models, then we presented the correct analytical expressions for the likelihood in the case of a phylogeny featuring: (1) one observable lineage shift with constant rates; (2) one observable rate shift with diversity-dependent rates; (3) one unobservable lineage shift with constant rates; (4) multiple observable rate shifts with constant rates; (5) multiple observable rate shifts with diversity-dependent rates. We also proved that, when one rate shift is present, it is possible to retrieve the original Nee et al. formula by combining the likelihoods for cases of unobservable and observable shifts. This shows the consistency of our approach.

In chapter 5 we built a method to assess if the development of a new likelihood model is necessary or if, instead, currently available inference models are good enough. To do so we developed an R package called `pirouette`. Then, from every phylogeny in the simulated distribution, `pirouette` will generate a posterior distribution obtained using BEAST2 standard inference models. Finally, the user can select an error statistic to estimate the error made by BEAST2 when

trying to reconstruct a phylogeny with currently available tools. However, the so obtained error distribution results from several other sources as well, of which a major one is the huge stochasticity naturally involved in the process. To account for that `pirouette` also runs a second, parallel, pipeline, which is almost identical to the original pipeline. The only difference is that this pipeline takes as input a phylogeny created under a standard diversification model. We regard the output of this parallel pipeline as the baseline error. If the two error distributions are similar then the development of a new likelihood model (and its subsequent implementation as tree prior in BEAST2) is not required. If they are very dissimilar, one would need to develop a module for the corresponding tree prior.