Synthesis

Thijs Janzen

Introduction

The large popularity of nature documentaries (there are two TV-stations in the Netherlands alone, solely devoted to nature documentaries[1]), only hints at the tremendous biodiversity and variation in biology we find on our planet. Yet, despite our investigative efforts and copious recordings of animal behavior, our understanding of the underlying processes both generating (evolution) and maintaining (community assembly) biodiversity, remains limited.

One of the first models to focus on community assembly is the mainland-island model, by MacArthur and Wilson (1963). The mainland-island model tries to explain species diversity on newly established islands, where the expected species diversity is assumed to be a dynamic equilibrium between local extinction on the island, and incoming new species from the mainland. Extinction depends on the size of the island, with smaller islands having higher extinction, and immigration depends on the distance to the mainland. The mainland-island theory was tested by removing all terrestrial arthropods on six small mangrove islands in Florida using methyl bromide and establishment of new species confirmed the dynamic equilibrium between local extinction and immigration (Simberloff & Wilson 1969). Building upon mainland-island theory, and incorporating aspects of the neutral theory of molecular evolution (Kimura & Crow 1964), Hubbell proposed a new community assembly theory incorporating migration from the metacommunity (e.g. the mainland) to the local community (e.g. an island), and taking into account extinction and speciation but also birth and death of individuals and coined it "The Unified Neutral Theory of Biodiversity and Biogeography" (UNTB) (Hubbell 2001). Similar to the island-mainland model, UNTB disregards any differences between species and considers birth rates, death rates, speciation rates and extinction rates to be identical between species. This explicitly disregards classic coexistence theory, which explains coexistence between species through the overlap between these species' niches (Diamond 1975; Tilman 1981; Chesson 2000), where a species' niche is defined as the conditions and requirements for a species to survive (Hutchinson 1958). Measuring a species' niche, and the overlap between niches, is troublesome however, and generally the niche concept has difficulties (McInerny & Etienne 2012a; b; c). Instead, a recent development has been to focus on species' traits (HilleRisLambers *et al.* 2012), taking into account that interactions between species depend on their traits, and that species with too similar traits might experience higher competition, and species with traits unadapted to the environment might be selected against (Cornwell & Ackerly 2009).

In chapters 1 and 2 I have focused on trait-based community assembly, and used the trait-based community assembly concept to assess the effect of filtering, limiting similarity and dispersal assembly on community composition in savannah trees in South Africa and cichlid fish in Lake Tanganyika, Zambia. I found that the majority of

[1] National Geographic and Animal Planet (source: upc.nl, ziggo.nl & kpn.com)

community composition was regulated by dispersal assembly, which is invariant to the traits. Building upon that in chapter 3, I incorporated differences in dispersal in the Unified Neutral Model of Biodiversity, and fitted this new model upon a tropical tree dataset from Barro Colorado Island. Including dispersal syndromes turned out to improve the fit of the model and better explain found patterns. It thus appears that means of dispersal and differences in dispersal are an important factor in community assembly, possibly more important than local interactions based on traits.

Although our understanding of community assembly has a rich history, our understanding of evolution goes back even further. Using fossils and insights from breeding programs in pigeons, Darwin developed the concept of natural selection, which weeds out individuals with unfit traits and favors those individuals that possess traits which enable them to have an advantage over others: "survival of the fittest" (Darwin 1859). Variation in such traits could provide the gradient upon which selection could act. How these traits were inherited across generations was not immediately understood, but the work of Mendel (Mendel 1865), provided the first understanding of the inheritance of traits. These notions were combined in the Modern Synthesis (Huxley 1942) and combined with the discovery of the carrier of genetic information, DNA (Watson & Crick 1953), paved the way for modern evolutionary research, which uses molecular similarity to infer evolutionary history (Yang & Rannala 2012).

Using molecular similarity, we can reconstruct phylogenetic trees that represent the ancestry of a group of species. From such a phylogenetic tree we can then in turn estimate diversification rates, using the accumulation of lineages in the tree. In chapter 4 I have tested whether assumptions made during reconstruction of a tree affect estimates of diversification based on that tree. From chapter 4 it turns out that diversification analysis need not be biased by choices made during tree reconstruction, provided that some criteria are met. Many diversification models rely on the absence of such a bias, and any indication of such a bias would mean that many of our estimates of speciation rates would need to be corrected.

As diversification models grow more sophisticated, they also become more mathematically and computationally demanding and in chapter 5 I have looked into a method to deal with that: Approximate Bayesian Computation (ABC). We tested several summary statistics to be used in ABC, and proposed a novel summary statistic that appears to perform much better than previously used statistics: the normalized Lineage-Through-Time statistic. In the last chapter I have used this statistic, relying upon the fact that diversification analysis is unbiased, and inferred past speciation and extinction rates for a clade of cichlid fish: *Lamprologini*. The model used to infer these rates takes into account the effect of environmental changes on speciation (allopatric speciation). I find that the large diversity in the clade of *Lamprologini* was most likely not driven by changes in the environment, but rather either through micro-allopatric events or through sympatric speciation.

The Species Concept

Central to all of my chapters is the concept of the "species". In the first three chapters I use the concept of species as an indicator for shared sets of traits, and shared selection pressures. In the last three chapters, I use the concept of species as a way to distinguish between genetic lineages, and mainly focus on the ability of these lineages to give birth to new lineages (speciation) and their propensity to go extinct. Both these ways of treating species might be a bit different from how people normally conceive or treat species. However, the species concept is itself a complicated concept, often varying between applications, and different between biota (in bacteria for instance, things tend to get complicated). I have generally tried to adhere to the Biological Species Concept: "Species are groups of interbreeding natural populations that are reproductively isolated from other such groups" (Coyne & Orr 2004). The lack of interbreeding is crucial here, as in the trait-based community assembly models it ensures that certain combinations of traits remain together, and for the diversification models it ensures that genetic differences match speciation processes.

Although the trait-based community models I have focused on apply on the species-level, the selection forces they take into account typically act upon on the individual level. Habitat filtering selects against individuals with traits that are maladapted to the local environment and limiting similarity selects against individuals with traits that are too similar to other individuals (not necessarily of other species!). Keeping track of all individuals in the community can be a painstaking exercise however, and current modeling efforts are restricted to the species level, rather than the individual level. All individuals of the same species are assigned the same combination of traits, and instead of assessing the survival of each individual separately, we check the effects of habitat filtering and limiting similarity on a per species basis. However, to obtain species trait values, multiple individuals were measured, and the average measurement was used as the representative trait value of the species. It would be interesting to include data on within-species variation and take this into account to estimate community assembly. Although two species might have different mean trait values, spread in trait values might overlap which could imply very different trait-based effects compared to only using the mean trait values. Within species variation and individual level selection could be combined by generating a metacommunity that consists of individuals with trait values drawn from the actual measurements (instead of the means). Starting from such a metacommunity, individuals can be filtered out due to either habitat effects, limiting similarity effects or dispersal assembly effects. A challenging part of such an approach would be to appropriately define the metacommunity, such that relative frequencies of individuals and the total number of individuals represent our knowledge about the system.

In the second part of my thesis I have used the species concept as a way to keep track of different genetic lineages, and here the species concept reiterates the importance of reproductive isolation. A phylogenetic tree captures the change of genetic

material over time, where branches are lineages that share the same genetic material. When branches diverge into two new branches (a branching point), the genetic material diversifies and we assume two new genetic lineages that do no longer exchange genetic material, e.g. two species. A branching point thus represents a speciation event. There are a number of processes that might obscure this image. First of all, not all genes are tightly linked with speciation events, and some genetic material might be very well preserved across species. Indeed the high degree of shared genetic material across cichlid species makes it often difficult to construct phylogenetic trees (Danley & Kocher 2001; Genner *et al.* 2007; Joyce *et al.* 2011). Furthermore, ongoing hybridization between cichlid species might obscure the links between speciation and genetics (Koblmüller *et al.* 2007), and although hybridization often is considered to reinstate gene flow and remove reproductive barriers (and thus undo speciation) (Seehausen 2004), it can also generate new species, if the hybrids turn out to be fit and later on become reproductively isolated (Salzburger *et al.* 2002a; Selz *et al.* 2014).

Molecular similarity thus provides good insight into past speciation processes, but is not perfect. This is reflected in the uncertainty in branching time estimates, often depicted as grey bars on the branching points (which indicate the posterior distribution of branching points). Current diversification models do not make use of these distributions in estimating diversification rates. Future applications, either through applying a joint inference approach which directly takes into account this variation, or through the implicit accounting of these uncertainties could improve our estimates and remain closer linked to the underlying genetic information.

We have shown here that a two-step approach does not provide substantially different results from the theoretically correct joint inference approach, in which we simultaneously determine the tree topology and the diversification rates. Explorations of different branch-rate models revealed however that these results only apply when using a branch-rate model with estimated or low variance. Surprisingly, when fitting the water levels model on the phylogeny of *Lamprologini*, and allowing all parameters to freely adjust, one of the optima we obtain is exactly the tree prior used to construct the phylogeny of *Lamprologini*. This is the least complex model that can be fitted upon the phylogeny, and as such it might be coincidence that this model is recovered, but alternatively it could be the model used in tree reconstruction has introduced a bias. To estimate the phylogeny of *Lamprologini*, an autocorrelated lognormal branch-rate prior was used, a prior not which was not included in our explorations. The autocorrelated lognormal branch-rate prior has an estimated variance, and we thus do not expect this prior to have introduced any bias, although further analysis is warranted to confirm this. Also, we should take into account that in our comparison between two-step inference and joint inference, we have mainly compared scenarios in which the same model was used both for tree reconstruction and diversification analysis. Estimating the effect of combinations of different models is less straightforward, and runs into problems regarding interpretation: how for instance does one estimate the effect of a time-dependent speciation prior on estimation of the parameters of a diversity-dependent

speciation model? However, considering that most diversification models are an extension of the standard birth-death model, we could exploit this fact in order to shed some light on the effect of prior model choice. The time-dependent speciation model can for instance be reduced to the constant rates birth-death model by takin the limit $\alpha \to 0$, and the diversity-dependent speciation model can be reduced to the constant-rates birth-death model by taking the limit $K \to \infty$. If a constant rates birth-death model is used in tree reconstruction and has an effect on diversification analysis, we expect that diversification analysis using the time-dependent speciation model, on a tree reconstructed using the constant rates birth-death model, is biased towards inferring $\alpha \to 0$, and if diversification analysis is performed with the diversity-dependent model results are biased towards $K \to \infty$. Further tests taking into account more model combinations could further prove or disprove the effect of tree choice in tree reconstruction and explain whether the patterns we find in our estimations using the water layers model are the effect of choices made during reconstruction of the tree, or truly reflect constant rate birth-death dynamics. Alternatively, the implementation of a larger range of diversification models in tree reconstruction software could facilitate the match between tree reconstruction and diversification models, and facilitate the direct estimation of diversification parameters using genetic data, instead of relying on trees constructed with other aims in mind than estimating diversification rates.

Multiple summary statistics in ABC

In chapter 5 I have focused on assessing which summary statistics provide most information and are best used within an ABC framework. Because summary statistics are by definition lower-dimensional than the data itself, some information is lost in the translation from data to summary statistic. Information loss might not be shared between summary statistics and as a result, different summary statistics could provide complementary information about the data. Combining summary statistics within an ABC framework could then maximize the use of information concerning the data, and improve ABC inference. When combining summary statistics there are a few things that need to be taken into account. The most important problem we run into is that it is hard to assess which summary statistic causes most rejections of proposed parameter combinations. To illustrate this, let us consider a model with two parameters, which we are inferring using two summary statistics which both capture a different aspect of the model (e.g. the summary statistics are fully complementary and have no information overlap). For both summary statistics we set a fairly restrictive threshold value, such that approximately 10% of proposed parameter combinations is accepted. As we run our ABC inference, it turns out that the majority of rejections occurs because one of the two summary statistics is rejected, whilst the other is very close to the observed summary statistic. Progress of parameter estimation should proceed along the axis favoured by one summary statistic, but is hindered by the other summary statistic. This

is partially desirable behavior, as we would not want to enter an area of parameter space that generates data which is not favored by one of our summary statistics, but is also undesirable as it might prevent the algorithm from traversing "valleys" of bad summary statistic combinations. Furthermore, with one summary statistic being more restrictive than the other, effectively this summary statistic solely determines acceptance and leaves the other summary statistic obsolete. No clear methods are available yet to counteract this behavior or to appropriately weigh the different summary statistics such that no one summary statistic overpowers the others. Future research could look into how to appropriately combine summary statistics, for instance by weighing the summary statistics or by differently modifying the threshold values.

Model Validation

Before obtaining any estimates for a given model, it is essential to first validate the model and its inference, such that we know the reliability of the obtained parameter estimates, given the model at hand. Model validation can proceed in two ways: either *a priori*, without making use of obtained empirical data; or *a posteriori*, using empirical data to obtain parameter estimates and using these estimates to validate our findings.

A priori model validation assesses the accuracy of the model, given a broad range of circumstances. First, data is simulated using the model for a range parameter values. The simulated data is then used in parameter inference, and the inferred parameters are compared with the parameter values used to simulate the data. If the parameter estimation procedure is accurate, the obtained parameter estimates should closely resemble parameter values chosen to simulate the data. Any structural bias in the estimates can be corrected for, and obtained results provide insight in the overall accuracy of estimates obtained using this model. To validate STEPCAM, we generated datasets for all parameter combinations of filtering, limiting similarity and dispersal assembly, on a resolution of 0.05. For every parameter combination we generated 10 different datasets and used these datasets as the starting point for our parameter inference. We found that the inferred parameter values for these 10 datasets closely matched the parameter values used to generate the datasets, confirming that our model inference method accurately infers parameters. Similarly we tested the water layers model by generating phylogenies for a range of parameter combinations. For each generated phylogeny, we again estimated sympatric speciation, allopatric speciation, extinction and rate of water level change and compared the obtained estimates with the parameter values used to generate the data. Here too estimates were close to used parameters, except for low rates of water level change, that were hard to infer. Similarly, low extinction rates proved to be hard to accurately assess, as extinction rates below 0.01 per million years were indistinguishable from lower rates.

A posteriori model validation uses the posterior distribution of parameters obtained from fitting the model to the empirical data. From this posterior distribution, parameter

combinations are sampled and artificial datasets are simulated for these parameter combinations. For these simulated datasets we can then again perform inference and obtain new posterior likelihood distributions. Comparing the obtained distribution of likelihoods to the likelihood of the empirical data informs then us about the fit of the model. If the likelihood of the empirical data lies well centered within the obtained distribution, we cannot reject that the empirical data has been the result of our modeled processes. If however the likelihood of the empirical data lies outside the obtained distribution, we can reject that the empirical data has been the result of our modeled processes as it is unlikely that the empirical data was the result of our model (see also Etienne (2007)). *A posteriori* testing primarily facilitates a secondary check to assess the applicability of the model, and the uncertainty in model fit. For the guilds model we have used this technique in order to assess the fit of the two-guild model upon the BCI data. Using the parameter estimates obtained with Maximum Likelihood, we generated 100 new datasets, and obtained Maximum Likelihood estimates for these datasets. The relative number of datasets with a higher likelihood than the likelihood obtained for the empirical dataset was used as a measure to assess similarity between the model and empirical data. We found that data generated including guild structure was consistently more similar to the empirical data than data generated without guild structure.

For approaches where the likelihood is not available, instead of the distribution of likelihoods we can assess the distribution of summary statistics. Again we generate a large number of datasets using parameter combinations from the posterior distribution; this time however we do not have to do inference for these datasets. Because the ABC inference would converge towards the summary statistics of the generated data (by design), we can directly use the summary statistics of the generated data. For the water levels model, we generated 1,000,000 trees using the obtained posterior distribution after doing ABC-SMC analysis. After comparing the distribution of obtained summary statistics with the values of the empirical data, we could single out which versions of the model most closely resembled the empirical data.

Ideally, both *a priori* and *a posteriori* model validation techniques are combined to acquire complete knowledge about the models accuracy and fit. If the model has been validated before, and focus is solely on inferring parameters from an obtained dataset, a posteriori validation can be sufficient to obtain the uncertainty in the parameter estimates, and to assess model fit.

A priori and *a posteriori* model validation tests verify the integrity of the model, but only within the framework of the model itself. These tests rely on the ability of the model to simulate data, and only verify that the model can accurately infer itself when confronted with data generated by itself. Although the *a posteriori* test does check to which extent the empirical data conforms to the model, ultimately these tests do not verify that the empirical data at hand really is the result of processes covered by our model. Ideally, we would want to test our model by comparing our model with empirical data for which we exactly know which processes have shaped it. Especially

evolutionary processes do not lend themselves for this way of testing due to the large time-span over which they act (which was the reason in the first place why we chose to use a modeling approach). Nevertheless it would be highly informative to have a phylogeny for which we have independent evidence (apart from molecular evidence) about the timing of speciation events and know the ancestral structure. Such a "known" phylogeny is hard to come by and only a few have so far been constructed, only using organisms with short generation times such as mice (Atchley & Fitch 1991) or using quickly replicating viruses such as bacteriophages (Hillis *et al.* 1992; Poe 1998), the influenza virus (Bush 1999) and the HIV virus (Zhu *et al.* 1998), or by using pedigrees of rDNA (Sanson *et al.* 2002). As a result of the short generation times, and the lack of "known" phylogenies for higher organisms, we are still lacking good empirical evidence to falsify diversification models for many of the vertebrate species groups we focus on.

Alternatively we could make use of the fossil record to infer past levels of diversity and obtain a rough estimate for extinction rates. Fossilization is not a very likely process, and especially rare for larger organisms. For the plankton family of *Foraminifera* we do have a good fossil record, spanning 65 million years, with an average diversity of 30 species (Ezard *et al.* 2011). The excellent state of the fossil record of *Foraminifera* is the result of the calciferous skeleton of this family of plankton, which rains down on the ocean floor. Using the fossil record of *Foraminifera* we can get a good impression of past levels of diversity, and obtain a rough estimate of species turnover. Using the fossil record of *Foraminifera*, Etienne and colleagues fitted a diversity-dependent speciation model and found that past diversity in their model closely matches past diversity in the fossil record. This provides a first step of independently checking diversification models, but further verification using "known" phylogenies or fossil data should be favoured to accurately assess our endeavours in diversification analysis.

Conclusion

Using patterns to study the underlying processes shows to be a powerful approach, and I hope that the past 6 chapters have convinced you that we can reverse-engineer the underlying mechanics and improve our understanding, simply by fitting models to data. Only fitting the model is not sufficient however, and care should be taken in properly validating the models findings, as I have outlined in the previous paragraphs. But even when properly validated, to which extent can we assume a model to be true? True, in the sense that the processes described by the model match processes that have generated the data. Without any additional information this question is hard to answer. Firstly we could try and identify if any comparable models could also generate the observed patterns. If our focal model outperforms all other models, it should be considered the most likely candidate to explain underlying processes. Nevertheless, this does not connect our model yet with the underlying ecology, and to fully test the power of our model we should both test the assumptions the model is based on, and test our

model using different datasets and within different contexts. The more often our model adequately explains observed patterns, the more unlikely alternative models become. Hence it is imperative to not only show that a model provides a suitable fit to the data, but also to provide other researchers with the means to independently apply the model on their own data. We can streamline such a process by providing other researchers with easy to use code, for instance in R-packages. Only by making our models easily applicable to other systems can we expect them to extend their explanatory power beyond the focal study system they have been designed, and can we expect models to improve our understanding of how processes in ecology and evolution have shaped, and still shape, the tremendous biodiversity on our planet.


BARE
BARE