Stochastic network modelling of the evolutionary tree

Francisco Richter and Rampal Etienne and Ernst C. Wit

Abstract The mechanisms that control the diversification of species are poorly understood. Sophisticated diversification models have been developed, but they have been developed on a case-by-case basis and no general method to study the combined effect of ecological factors exists. Such a general method has remained elusive for several reasons. Firstly, evolutionary processes have extremely complex dynamics. Secondly, decay and fossilization degrade crucial evidence useful for phylogenetic analyses. Thirdly, diversification processes have many potential explanatory variables, which increases the dimensionality of the models enormously. To overcome these issues, we propose a general diversification model expressing the evolutionary species diversification dynamics as a combination of two generalized linear models. The fact that we typically only have data on currently existing species can be described as a missing data problem and we developed an MCEM-type algorithm for it. We show that our method performs well for cases where an exact solution is available, and discuss potential future usage of our approach.

Abstract *I meccanismi che controllano la diversificazione delle specie sono capiti male. Sono stati sviluppati sofisticati modelli di diversificazione, ma sono stati sviluppati caso per caso e senza un metodo generale. Un tale metodo generale rimasto elusivo per diverse ragioni. In primo luogo, i processi evolutivi hanno dinamiche estremamente complesse. In secondo luogo, il decadimento e la fossilizzazione hanno degradato le prove cruciali utili per le analisi filogenetiche. In terzo luogo, i processi di diversificazione hanno molte potenziali variabili esplicative, che aumentano enormemente la dimensionalit dei modelli. Per superare questi problemi, proponiamo un modello generale di diversificazione che esprima le dinamiche*

Francisco Richter
JBI, University of Groningen, NL, e-mail: f.richter@rug.nl

Rampal Etienne
GELIFES, University of Groningen, NL e-mail: r.s.etienne@rug.nl

Ernst C. Wit
JBI, University of Groningen (NL), ICS, Universitá Svizzera Italiana (CH) e-mail: e.c.wit@rug.nl, ernst.wit@usi.ch

di diversificazione evolutiva delle specie come una combinazione di due modelli lineari generalizzati. Il fatto che di solito abbiamo dati sulle specie esistenti presenti pu essere descritto come algoritmo di tipo MCEM per questo. Dimostriamo che il nostro metodo funziona bene per i casi in cui disponibile una soluzione esatta e discutiamo il potenziale utilizzo futuro del nostro approccio.

Key words: species assemblages, Fokker-Planck, non-homogeneous Poisson process, generalized linear models, EM

1 Introduction and motivation

Biodiversity, the wide variety of species on Earth, is declining at enormous rates [2]. That compromises ecosystem stability and productivity, which negatively impacts the ecosystem services on which human communities depend [3]. To conserve biodiversity, we must understand the mechanisms how it comes about and how it is maintained, in assemblages of species, so-called ecological communities.

It is our aim to incorporate the sources of high-dimensional ecological and biological data under a unified statistical framework in order to overcome the main challenges that evolutionary biologists currently face. Particularly, the lack of information of extinct species and the huge complexity of current stochastic differential diversification models are bottlenecks for a proper inference on a general scenario. In this report we propose a general method with the potential to provide practical solutions for a large number of open questions in evolutionary biology and ecology.

2 Non-homogeneous Poisson process as driver of species diversification

Birth-death processes has been systematically used to explain the evolutionary dynamics described on a phylogenetic tree [12],[13]. On that kind of process we assume that each lineage has their speciation rate λ and extinction rate μ which can be influenced by individual attributes, ecological or environmental factors, composition of the local biodiversity, etc. In the literature we find models where diversifications are assumed to be constant [11], change through time [15], depends on diversity [4], individual traits [14] and many other factors [10]. In order to achieve a more realistic model, we are interested on flexible rates able to change dynamically through all those factors simultaneously and taking into account the ecological nature of species assemblages. In that sense the speciation rate of species j at moment t_i could also depends on other species traits and local interactions, as well as any ecological influences described above.

We assume that the evolutionary process of diversification is driven by a Markov process, i.e, the diversification event i is only influenced by the state on the previous

event $i-1$

$$P(S(t_n)=T_n|S(t_0)=T_0,...,S(t_{n-1})=T_{n-1})=P(S(t_n)=T_n|S(t_{n-1})=T_{n-1})$$

Thus, the distribution of the waiting times Δt_j for a species j to have a diversification event will be exponential [1] with rate $\lambda_i+\mu_i$, and if a diversification event occurs it will be an speciation with probability $\frac{\lambda_j}{\lambda_j+\mu_j}$ or an extinction with probability $\frac{\mu_j}{\lambda_j+\mu_j}$ [16].

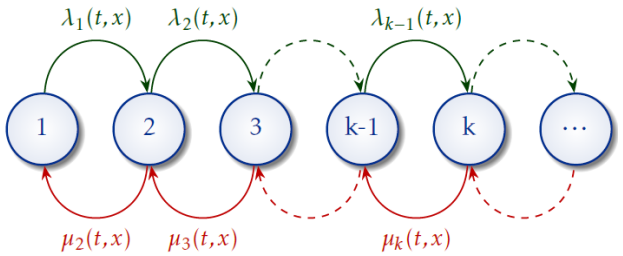

Fig. 1 Representation of a general birth-death process when birth and death rates depends on several multidimensional covariates.

Similarly, looking at the whole phylogenetic process, given a previous diversification time t_{i-1}, we know that the next waiting time

$$\Delta t_i=\min\{\Delta t_{i,1},...,\Delta t_{i,n_i}\}$$

would be again exponential with rate $\sum\lambda_{i,j}+\mu_{i,j}$ and it will be an speciation of species j with probability $\frac{\lambda_{i,j}}{\sum\lambda_{i,j}+\mu_{i,j}}$ or an extinction of species j with probability $\frac{\mu_{i,j}}{\sum\lambda_{i,j}+\mu_{i,j}}$. We will define

$$\rho_i=\begin{matrix}\lambda_{i,j}\text{ speciation}\\ \mu_{i,j}\text{ extinction}\end{matrix}$$

In that sense, if we focus on the number of species (diversity) per time $N(t)$ we find that is driven by a non-homogenous Poisson process with rate function $\sigma(t)$, defined as the sum of all individual speciation and extinction rates at moment t ,

$$\sigma(t)=\sigma_{i,\lambda}(t,\mathbf{x_i})+\sigma_{i,\mu}(t,\mathbf{x_i}) \tag{1}$$

where

$$\sigma_{i,\lambda}(t,\mathbf{x_i})=\sum_{j=1}^{n_i}\lambda_{i,j}(t,x_{i,j}),$$

$$\sigma_{i,\mu}(t,\mathbf{x_i})=\sum_{j=1}^{n_i}\mu_{i,j}(t,x_{i,j})$$

Note that equation 1 represents the connection between the continuous (time) and discrete (species) spaces. In this analysis **we assume that $\sigma(t)$ is step-wise function of t.** The possible extinctions to the continuous case will be discussed. That rate is potentially dependent on many ecological variables, including diversity itself.

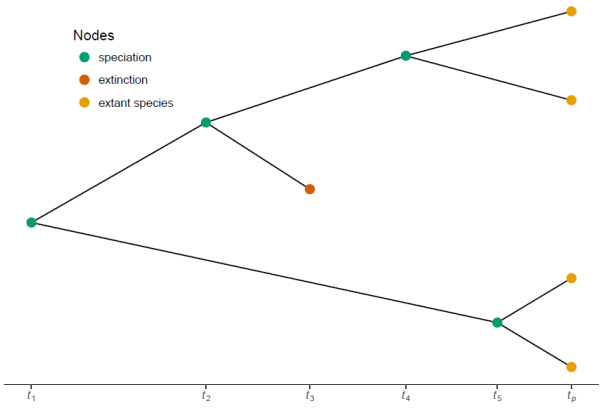

Fig. 2 Evolutionary process with 4 extant species at the present time t_p and one extinction.

Considering that we can write the likelihood function of the Markov chain representing the diversification process of species and performs statistical inference winch would give us biological insight helpful to grow a better understanding of it.

3 Inference of species assemblages process

As described on the previous section, the Markov nature of the process means that the likelihood is exactly the product of the conditional densities, in other words, the likelihood of the tree is then described as a multiplication of an exponential distribution and a multinomial distribution

$$\mathscr{L}(\theta;\tau,t) = \prod P_E(t = t_i;\theta) P_M(\tau = \tau_i | t = t_i;\theta) \tag{2}$$

or

$$\mathscr{L}(\theta;\tau,t) = \prod_{i=1}^{M} e^{-\sigma_i t_t} \rho_i \tag{3}$$

which, on the logarithm scale would be

$$l(\phi;\tau,t) = \sum_{i=1}^{M} -\sigma_i t_t + log(\rho_i) \tag{4}$$

corresponding to the log-likelihood function. We say that the tree $T = (t, \tau)$

$$C = \cup_{m=1}^{\infty} C_m$$

where

$$C_m(T) = \{(\tau, t) \in R^m \times Z_1^m | t_1 < t_2 < ... < t_n, \sum t_i = 15\}$$

3.1 Observing the full phylogenetic tree

If we observe the full phylogenetic tree, we can use directly the MLE to find the most likely values for different parameters associated with ecological covariates which are potentially involved on the diversification process of species. In that sense we search for the solutions of the equations

$$\frac{\partial l(\phi; \tau_i, t_i)}{\partial \phi} = \sum_{i=1}^{M} -\frac{\partial \sigma_i}{\partial \phi} t_t + \frac{1}{log(\rho_i)} \frac{\partial \rho_i}{\partial \phi} = 0$$

which for simple cases could lead to an analytical solution [16], but in most cases these expressions are too complex and a numerical minimization method is needed [14]. The numerical optimization is normally straightforward though since we assume this function is unimodal, however, as stated on this section the calculation of those values requires complete information on the phylogeny, which as discussed previously, in most cases is not realistic. For the case when we do not have information about fossil record, we would need to implement an EM algorithm to deal with the missing data.

3.2 Observing the extant phylogenetic tree

Mathematically, we denote $x \in \mathscr{X}$ as a random variable in the time-tree space and probability distribution given by equation (2). We define the random variable $y = \mathscr{Y}$ as the *observed tree*, which lives in the space of ultrametric trees [5] and has probability distribution given by

$$g(y|\phi) = \int_{\mathscr{X}(y)} f(x|\phi) dx \tag{5}$$

considering $\mathscr{X}(y)$ as the subset of $\mathscr{X}$ in agreement with the observed tree y, that is, $\mathscr{X}(y) = \{x \in \mathscr{X} | \mathscr{Y}(x) = y\}$.

In order to infer meaningful information about evolutionary dynamics we would like to find the parameters ϕ which maximizes $g(y|\phi)$ given the observed tree y. However, given the complexity of the space $\mathscr{X}(y)$, a close form for equation (5) is not available [6]. A standard way to sort out the difficulties driven by missing in-

formation is the EM algorithm. Thus, we describe its implementation on the species diversification modeling context. One natural approach to this setup would be the application of the EM algorithm to the time-tree space. We define the EM iteration $\phi^* \to \phi$ as

E. Compute $Q(\phi|\phi^*) = E_{\phi^*}(\log f(x|\phi)|y)$,
M.Choose ϕ to be the value of $\phi \in \Omega$ which maximizes $Q(\phi|\phi^*)$

Note that

$$E_{\phi^*}(\log f(x;\phi)|y) = \int_{\mathscr{X}(y)} \log f_X(x;\phi) f_{X|Y}(x|y,\phi^*)dx$$

As shown previously in the case of equation (5), the calculation of $Q(\phi|\phi^*)$ has not a close form due to the huge complexity of the space $\mathscr{X}(y)$, so numerical calculations are needed. One way to perform this task is considering a Monte-Carlo sampling [17], where, given a set of sampled trees $x_1,...,x_p$ from $f_{X|Y}(x|y,\phi^*)$, we approximate $Q(\phi|\phi^*)$ by

$$E_{\phi^*}(\log f(x;\phi)|y) \approx \frac{1}{p}\sum_{i=1}^{p} \log f_{X;\phi^*}(x_i;\phi) \tag{6}$$

However, sampling complete trees given the observed ones from $f_{X|Y}(x|y,\phi^*)$ is computationally infeasible. Thus, we use an approximation algorithm, inspired by the Gillespie algorithm [9], correcting the error of the approximation via importance sampling [7].

The idea of importance sampling techniques is to give more weight in the likelihood estimation to those sampled values that seems to be in more agreement with the true sampling distribution $f_{X|Y}(x|y,\phi^*)$. In that sense we re-write equation 6 including weights such that

$$E_{\phi^*}(\log f(x;\phi)|y) \approx \frac{1}{p}\sum_{i=1}^{l} \log f_{X;\phi^*}(x;\phi|y) w_i$$

where

$$w_i = \frac{f_{X|Y}(x|y,\phi^*)}{g_{X|Y}(x|y,\phi^*)} = \frac{f_{X,Y;\phi^*}(x,y)}{g_{X,Y;\phi^*}(x,y)}$$

Finally we perform the EM routine iteratively using the MC sampling with an importance sampling correction.

4 Conclusions and discussion

In this paper, we looked at networks from a slightly different point of view. We considered the species diversification process as a natural process that generates a special type of network, namely a species tree. The speciation process is subject to many random influences, having partly to do with the topology of the tree and

partly on external influences. In this paper we define a random network/tree process by means of a stochastic differential equation. The aim is to infer the parameters underlying the process from the information we can obtain from the extant species at the tips of the tree. This may seem an impossible process, but fortunately genetics is helping out, typically allowing us to infer the – incomplete – evolutionary tree of these species. In this paper we show how a MCEM algorithm can be used to infer the kinetic parameters of the speciation process.

Finally, we like to raise one additional point. We consider a phylogenetic tree, mathematically expressed as a set $Y = (\mathscr{T}, \Upsilon)$, where $\mathscr{T}$ represent the set of branching times, and Υ has the information of the topology of the tree. P_E is the probability for t to be the minimum waiting time given by an exponential distribution. And P_M is the probability of, given the waiting time, that the topology event corresponds either to extinction or speciation of one of the extant species at time t. That probability is given by a multinomial distribution. However, equation (2) does not consider the restriction

$$\sum t_i < C, \tag{7}$$

i.e., we implicitly assume that that likelihood is 0 when $\sum t_i > C$, where C defined as the *crown time*.

References

1. Allen, L.J.S.: An introduction to stochastic processes with applications to biology. CRC Press (2010)
2. Ceballos, G. et al: Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science advances*, 1(5) (2015)
3. Chivian, E., Bernstein, A.: Sustaining life: how human health depends on biodiversity, Oxford University Press (2008)
4. 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. Proc. R. Soc. B, rspb20111439 (2011)
5. Gavryushkin, A., Drummond, A.J.:The space of ultrametric phylogenetic trees. Journal of theoretical biology, **403**, 197–208, (2016)
6. Gavryushkin, A., Whidden, C., Matsen, F.: The combinatorics of discrete time-trees: theory and open problems. bioRxiv, 063362 (2016)
7. Glynn, P.W., Iglehart, D.L.: Importance sampling for stochastic simulations. Management Science, **35**(11), 1367–1392 (1989)
8. Morlon, H., Parsons, T.L., Plotkin, J.B.: Reconciling molecular phylogenies with the fossil record. Proceedings of the National Academy of Sciences, **108**(39), 16327–16332 (2011)
9. Meng, T.C., Somani, S., Dhar, P.: Modeling and simulation of biological systems with stochasticity. In silico biology, **4**(3), 293–309 (2004)
10. Morlon, H.: Phylogenetic approaches for studying diversification. Ecology letters, **17**(4), 508–525 (2014)
11. Nee, S., May, R.M., Harvey, P.H.: The reconstructed evolutionary process. Philosophical Transactions of the Royal Society of London B: Biological Sciences, **344**(1309), 305–311 (1994)
12. Nee, S.: Birth-death models in macroevolution. Annu. Rev. Ecol. Evol. Syst., **37**, 1–17 (2006)
13. Novozhilov, A.S., Karev, G.P., Koonin, E.V.: Biological applications of the theory of birth-and-death processes, Briefings in bioinformatics, **7**(1), 70–85 (2006)

14. Paradis, E.: Statistical analysis of diversification with species traits. Evolution, **59**(1), 1–12 (2005)
15. Rabosky, D.L., Lovette, I.J.: Explosive evolutionary radiations: decreasing speciation or increasing extinction through time? Evolution, **62**(8), 1866–1875 (2008)
16. Reynolds, J.F.: On estimating the parameters of a birth-death process. Australian & New Zealand Journal of Statistics, **15**(1), 35–43 (1973)
17. Wei, G.C.G., Tanner, M.A.: A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. Journal of the American statistical Association, **85**(411), 699–704 (1990)