1 Introduction

It is becoming increasingly common to use the joint analysis of fossil temporal data and morphological and molecular data to estimate species phylogenies. The total-evidence approach initially introduced by Pyron (2011) has been recently improved by using more appropriate models for speciation-fossilisation process (Stadler, 2010; Heath et al., 2014; Gavryushkina et al., 2014; Zhang et al., 2016; Gavryushkina et al., 2017). The advantage of these new models is in direct modelling of the fossil and extant species sampling processes: in trees produced by the birth-death process, any lineage can fossilise with Poisson rate ψ and extant species are included into an analysis with probability ρ.

However, the applicaitons of these models violated the sampling process assumptions because several fossil specimens of the same species (from different localities and different ages) were treated as a single fossil occurrence. The standard FBD model assumes that every fossil specimen ever discovered should be included in an analysis as a distinct sample. However, such an analysis is often not possible for two main reasons. Firstly, paleontological datasets usually only contain morphological data combined from several well preserved specimens identified as belonging to the same species. Thus, a specimen by specimen morphological data is not available. Secondly, even if such data were avialable, the number of specimens could be too large making an analysis infeasible.

The available fossil data then contains of a morphological character sequence for every fossil species and the ages of the first and the last occurrences of fossil specimens of that species. To accomodate this data in a joint dating analysis, Stadler et al. (2018) developed stratigraphic range birth-death model (SBD), — a birth-death model with fossil sampling that allows to assign multiple fossil samples to species.

Here we implemented the SBD model as an addon to BEAST2 (Bouckaert et al., 2014) phylogenetic package. We tested the properties of the model in simulation studies and analysed two empirical datasets.

2 Methods

2.1 Stratigraphic range model

Here we implemented an extension of the fossilised birth-death model that allows assigning fossils to species Stadler et al. (2018). We only considered the case of pure budding speciation (that is, $\beta = \lambda_a = 0$ in

2.2 Bayesian inference

2.3 MCMC kernel

2.4 Simulation studies

2.5 Empirical data analysis

3 Results

3.1 Simulation studies

3.2 Total-evidence analysis of penguins

3.3 Total-evidence analysis of North American canids

4 Discussion

5 Supplementary

6 Operators

Table 1: The table that describes the change in the dimension of the tree for different types of the move.

pruning from/ attaching to	SA	branch	range branch
SA	0	+1	0
branch	-1	0	-1
range branch	0	+1	0

Table 2: The table that describes the multiplier in the Hastings ratio that reflects the orientation.

pruning from/ attaching to	SA	branch	range branch
SA	1	2	1
branch	$\frac{1}{2}$	1	$\frac{1}{2}$
range branch	1	2	1

1. pruning from the root SA (normal case)
2. pruning from the root SA (special case)
3. pruning from an internal SA (normal case)
4. pruning from an internal SA (special case)
5. pruning from a range branch (normal case)
6. pruning from a range branch (special case)
7. pruning from a non-range branch
    - pruning from an internal branch (normal case)
    - pruning from an internal branch (special case)
    - pruning from the root branch (normal case)
    - pruning from the root branch (special case)
8. attaching to a leaf (normal case)
9. attaching to a leaf (special case)
10. attaching to a non-range branch
    - attaching to an internal branch (normal case)
    - attaching to an internal branch (special case)
    - attaching to the root branch (normal case)
    - attaching to the root branch (special case)
11. attaching to a range branch (normal case)
12. attaching to a range branch (special case)

Normal case:
Pruning:

1. when pruning from a non-root SA or a non-root branch then keep the orientation of CiP the same as iP

Attaching:

1. when attaching to a leaf make iP the same as j, also make i the same as j used to be
2. when attaching to a non-root (range or non-range) branch make iP the same as j
3. when attaching to a range (which is always not the root) branch make j left and i right
4. when attaching to a non-range branch assign left and right randomly to j and i

Special case 1:
Pruning:

1. when pruning from a non-root SA or a non-root branch then keep the orientation of CiP the same as iP (DON'T DO THIS)

Attaching:

1. when attaching to a leaf make iP the same as j, also make i the same as j used to be (IS NEVER THE CASE)
2. when attaching to a non-root (range or non-range) branch make iP the same as j (DON'T NEED TO DO THIS iP=j)
3. when attaching to a range (which is always not the root) branch make j left and i right (MAKE CiP left and i right)
4. when attaching to a non-range branch assign left and right randomly to j and i (ASSIGN left and right randomly to CiP and i)

Special case 2:
Pruning:

1. when pruning from a non-root SA or a non-root branch then keep the orientation of CiP the same as iP (DON'T DO THIS)

Attaching:

1. when attaching to a leaf make iP the same as j, also make i the same as j used to be (DO THIS but don't touch iP just make i the same as iP)
2. when attaching to a non-root (range or non-range) branch make iP the same as j (IS NEVER THE CASE)
3. when attaching to a range (which is always not the root) branch make j left and i right (IS NEVER THE CASE)
4. when attaching to a non-range branch assign left and right randomly to j and i (IS NEVER THE CASE)

7 Simulations

We performed simulations under two Scenarios.

7.1 Simple prior

We fixed the time of origin to $x_0 = 100$. Then we draw a set of parameters η composed of $d = \lambda - \mu$, $\nu = \frac{\mu}{\lambda}$, $s = \frac{\psi}{\mu+\psi}$ and ρ from distribution:

$$\begin{array}{lcl} d & \sim & \text{Uniform(0.03, 0.04)} \\ \nu & \sim & \text{Uniform(0.4, 0.7)} \\ s & \sim & \text{Uniform(0.2, 0.5)} \\ \rho & \sim & \text{Uniform(0.5,1.0)} \end{array} \tag{1}$$

For each set of parameters η_i we simulate a tree under the stratigraphic range birth-death model assuming only budding speciation for $x_0 = 100$ time units. We simulate conditioning on sampling at least five individuals at present, that is, if the process dies out or four or less individuals are sampled at present we repeat the simulations with the same set of parameters.

We repeat the above procedure 100 times to obtain 100 trees (with more than five present taxa each). We then remove all fossil occurrences between the first and the last sample of the same species and simulate DNA sequences along the trees. We keep a sequence of the last occurrence of every species. Then we either estimate the parameters, (x_0, η), from each tree or we estimate the tree, (x_0, η), and the molecular evolution model parameters.

In all analyses, we use a Uniform(0,200) prior distribution for the time of origin (x) and distributions (1) for η. We use the default prior for the substitution and clock model parameters.

7.2 Compound prior

Let l' be the number of sampled extinct species. Let m be the number of extant species samples. For the simulations studies we have chosen parameters λ, μ, ψ, ρ in such a way that l and m are large enough to be informative about the rates but small enough for a fast analysis: $l, m \in \{10, 100\}$. We chose $t_0 = 100$ for numerical reasons (too large or too small values can lead to numerical problems) but we note that as we infer the rates the results can be scaled to any t_0.

Note that

$$E(m) = \rho e^{(\lambda-\mu)t_0}$$

and we approximate the number of all species that existed between time t_0 and present time with

$$N = \int\limits_0^{t_0} \lambda e^{(\lambda-\mu)(t_0-t)} dt = \frac{\lambda}{\lambda-\mu}(e^{(\lambda-\mu)t_0} - 1).$$

Then $\rho e^{(\lambda-\mu)t_0}$ approximate the number of ρ-sampled species, m, and $\frac{\psi}{\psi+\mu}(N - e^{(\lambda-\mu)t_0})$ approximate the number of ψ-sampled extinct species, l'.

$$m = \rho e^{(\lambda-\mu)t_0}$$

$$l' = \frac{\psi}{\psi+\mu}(N - e^{(\lambda-\mu)t_0}) = \frac{\psi}{\psi+\mu}(\frac{\lambda}{\lambda-\mu}(e^{(\lambda-\mu)t_0} - 1) - e^{(\lambda-\mu)t_0}) =$$

$$\frac{\psi}{\psi+\mu}(\frac{\lambda}{\lambda-\mu}(\frac{\mu}{\lambda}e^{(\lambda-\mu)t_0} - 1)) = \frac{\psi(\mu e^{(\lambda-\mu)t_0} - \lambda)}{(\psi+\mu)(\lambda-\mu)}$$

$$10 \leq \rho e^{(\lambda-\mu)t_0} \leq 100$$

$$10 \leq \frac{\psi(\mu e^{(\lambda-\mu)t_0} - \lambda)}{(\psi+\mu)(\lambda-\mu)} \leq 50$$

$$\frac{10}{s} \leq \frac{(\nu e^{dt_0} - 1)}{1-\nu} \leq \frac{50}{s} \tag{2}$$

Out of all extant species the number of species that were not ρ sampled but were ψ sampled is:

$$\frac{\psi}{\lambda+\psi}(1-\rho)e^{(\lambda-\mu)t_0}$$

then

t_0	100
ρ	[0.1, 1]
d	$[\frac{log(\frac{10}{\rho})}{100}, \frac{log(\frac{100}{\rho})}{100}]$
s	[0.01,0.5]
ν	[0,1] so that inequality (2) holds

7.3 Conditioning on sampling at least N extant individuals

Let $\hat{p}_n(t|t_0, \lambda, \mu, \psi, \rho)$ be the probability that an individual alive at time t before today has n sampled extant descendants and an arbitrary number of sampled extinct individuals, then according to Stadler et al. (2011) with $\rho > 0$:

$$\hat{p}_0(t) = 1 - \frac{\rho(\lambda-\mu)}{\rho\lambda + (\lambda(1-\rho)-\mu)e^{-(\lambda-\mu)t}}$$

$$\hat{p}_1(t) = \frac{\rho(\lambda-\mu)^2 e^{-(\lambda-\mu)t}}{(\rho\lambda + (\lambda(1-\rho)-\mu)e^{-(\lambda-\mu)t})^2}$$

$$\hat{p}_n(t) = \hat{p}_1(t)\Big(\frac{\rho\lambda(1-e^{-(\lambda-\mu)t}}{\rho\lambda + (\lambda(1-\rho)-\mu)e^{-(\lambda-\mu)t}}\Big)^{n-1} \text{ for } n > 1$$

Let $a = \rho\lambda(1-e^{-(\lambda-\mu)t}$, $b = \rho\lambda + (\lambda(1-\rho)-\mu)e^{-(\lambda-\mu)t}$ and $\zeta = \frac{a}{b}$

Let $N \geq 1$ then the probability that an individual alive at time t before today has at least N sampled extant descendants and an arbitrary number of sampled extinct individuals with $\rho > 0$ is:

$$\hat{p}_N(t) = 1 - \sum_{n=0}^{N} \hat{p}_n(t) = 1 - \hat{p}_0(t) - \hat{p}_1(t) \sum_{n=0}^{N-1} \zeta^n = 1 - \hat{p}_0(t) - \hat{p}_1(t) \frac{1-\zeta^N}{1-\zeta} =$$

$$\frac{\rho(\lambda-\mu)}{b} - \frac{\rho(\lambda-\mu)^2 e^{-(\lambda-\mu)t}}{b^2} \frac{b^N - a^N}{b^{N-1}(b-a)} =$$

note that $b - a = \rho\lambda + (\lambda(1-\rho)-\mu)e^{-(\lambda-\mu)t} - \rho\lambda(1-e^{-(\lambda-\mu)t}) = (\lambda-\mu)e^{-(\lambda-\mu)t}$

$$\rho(\lambda-\mu)\Big(\frac{1}{b} - \frac{b^N - a^N}{b^{N+1}}\Big) = \rho(\lambda-\mu)\frac{a^N}{b^{N+1}} = \frac{\rho(\lambda-\mu)(\rho\lambda(1-e^{-(\lambda-\mu)t})^N}{(\rho\lambda + (\lambda(1-\rho)-\mu)e^{-(\lambda-\mu)t})^{N+1}}$$

8 File format

The first and the last fossil (or present sample) in a stratigraphic range are distinguished by suffixes `_first` and `_last` in the sampled node labels. : `taxon_name_first`, `taxon_name_last`, for example. If a stratigraphic range is represented by a single occurrence or only a present sample then the taxon either does not have the suffix or it has suffix `_last`.

9 Stratigraphic range birth-death tree likelihood tests

$$\begin{aligned}
p(t) &= 1 + \frac{-(\lambda-\mu-\psi) + c_1 \frac{e^{-c_1 t}(1-c_2)-(1+c_2)}{e^{-c_1 t}(1-c_2)+(1+c_2)}}{2\lambda}, && (3)\\
\widetilde{q}_{asym}(t) &:= \sqrt{e^{-t(\lambda+\mu+\psi)} q(t)}, && (4)\\
q(t) &= \frac{4e^{-c_1 t}}{(e^{-c_1 t}(1-c_2) + (1+c_2))^2} && (5)\\
c_1 &= |\ \sqrt{(\lambda-\mu-\psi)^2 + 4\lambda\psi}\ |, \\
c_2 &= -\frac{\lambda-\mu-2\lambda\rho-\psi}{c_1}. && (6)
\end{aligned}$$

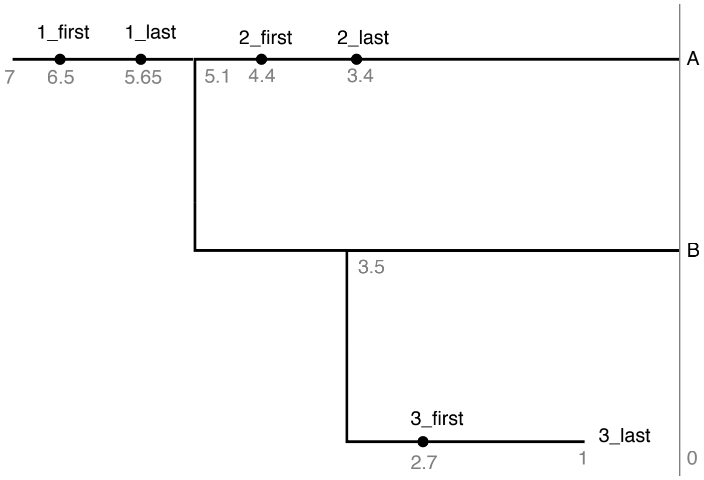

Figure 1: Example tree

$$f[\mathcal{T}_s^o \mid \lambda, \mu, \psi, \rho, x_0] = \psi^k \rho^l \lambda^{n-j-1} \prod_{i=1}^{3n-j-1} \hat{q}_{asym}(B_i) e^{\psi L_s}$$

$$f[\mathcal{T}_s^o \mid \lambda, \mu, \psi, \rho, x_0] = \psi^6 \rho^2 \lambda^2 q(x_0) q(x_1) q(x_2) \frac{r(o_1)}{r(y_1)} \frac{r(o_2)}{r(y_2)} \frac{r(o_3)}{r_{tip}(y_3)} e^{\psi L_s}$$

where

$$\begin{aligned} L_s &= o_1 - y_1 + o_2 - y_2 + o_3 - y_s \\ r(x) &= \frac{\tilde{q}_{asym}(x)}{q(x)} = \sqrt{\frac{e^{-t(\lambda+\mu+\psi)}}{q(x)}} \\ r_{tip}(x) &= \frac{\tilde{q}_{asym}(x)}{p(x)} \end{aligned}$$

For the tree in Figure 1, $x_0 = 7.0$, $x_1 = 5.1$, $x_2 = 3.5$, $o_1 = 6.5$, $y_1 = 5.65$, $o_2 = 4.4$, $y_2 = 3.4$, $o_3 = 2.7$, and $y_3 = 1$.

Let

$$\begin{aligned}
p(t) &= 1 + \frac{-(\lambda - \mu - \psi) + c_1 \frac{e^{-c_1 t}(1-c_2)-(1+c_2)}{e^{-c_1 t}(1-c_2)+(1+c_2)}}{2\lambda}, \\
\widetilde{q}_{asym}(t) &:= \sqrt{e^{-t(\lambda+\mu+\psi)} q(t)}, \\
q(t) &= \frac{4e^{-c_1 t}}{(e^{-c_1 t}(1-c_2) + (1+c_2))^2} \\
c_1 &= |\ \sqrt{(\lambda - \mu - \psi)^2 + 4\lambda\psi}\ |, \\
c_2 &= -\frac{\lambda - \mu - 2\lambda\rho - \psi}{c_1}.
\end{aligned}$$

Let κ' be the total number of sampled fossils that represent the start and end times of a stratigraphic range. If a stratigraphic range is represented by a single fossil then this fossil only contributes one towards κ'. Let the sum of all stratigraphic range lengths be $L_s = \sum_{i=1}^{n} o_i - y_i$.

References

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., and Drummond, A. J. (2014). "BEAST 2: a software platform for Bayesian evolutionary analysis." *PLoS Comput Biol*, 10(4): e1003 537.

Gavryushkina, A., Heath, T. A., Ksepka, D. T., Stadler, T., Welch, D., and Drummond, A. J. (2017). "Bayesian Total-Evidence Dating Reveals the Recent Crown Radiation of Penguins." *Systematic Biology*, 66(1): 57.

Gavryushkina, A., Welch, D., Stadler, T., and Drummond, A. J. (2014). "Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration." *PLoS Computational Biology*, 10(12): e1003 919.

Heath, T. A., Huelsenbeck, J. P., and Stadler, T. (2014). "The fossilized birth–death process for coherent calibration of divergence-time estimates." *Proceedings of the National Academy of Sciences*, 111(29): E2957–E2966.

Pyron, R. A. (2011). "Divergence Time Estimation Using Fossils as Terminal Taxa and the Origins of Lissamphibia." *Systematic Biology*, 60: 466–81.

Stadler, T. (2010). "Sampling-through-time in birth-death trees." *Journal of Theoretical Biology*, 267(3): 396–404.

Stadler, T., Gavryushkina, A., Warnock, R. C., Drummond, A. J., and Heath, T. A. (2018). "The fossilized birth-death model for the analysis of stratigraphic range data under different speciation modes." *Journal of theoretical biology*, 447: 41–55.

Stadler, T., Kouyos, R., von Wyl, V., Yerly, S., Böni, J., Bürgisser, P., Klimkait, T., Joos, B., Rieder, P., Xie, D., Günthard, H. F., Drummond, A., Bonhoeffer, S., and the Swiss HIV Cohort Study (2011). "Estimating the basic reproductive number from viral sequence data." *Mol Biol Evol*, 29: 347–357.

Zhang, C., Stadler, T., Klopfstein, S., Heath, T. A., and Ronquist, F. (2016). "Total-evidence dating under the fossilized birth–death process." *Systematic Biology*, 65(2): 228–249.