Macroevolutionary Analysis of Stratigraphic Range Data

Walker Pett, Rachel C M Warnock, Alexandra Gavryushkina

The FBD skyline model applied to stratigraphic ranges

Stadler [2] introduced the fossilised birth-death process for the phylogenetic analysis of extant and fossil samples, and Gavryushkina et al. [1] and Zhang et al. [5] extended this model to account for rate variation across different time intervals (the FBD skyline model). Stadler et al. [3] later introduced the FBD range model for the analysis of stratigraphic range data, defined as the interval between first and last appearances in the fossil record, in the absence of information about the underlying phylogenetic relationships. Here, we extend this modelling framework to account for temporal rate variation in speciation, extinction and fossil sampling.

Model notation assuming constant rates

First, we define the key model parameters assuming constant rates. The process begins with a single lineage at origin time x_0. Each lineage has instantaneous branching speciation rate λ and extinction rate μ. Here, we assume all speciation occurs via asymmetric (budding) speciation, i.e. speciation events give rise to a single new species and do not result in the extinction of the ancestor (parent) species. Fossil sampling occurs along each branch with rate ψ and extant species are sampled with probability ρ. The process gives rise to a total of n species, with m extinct and $n - m$ extant species. The total number of fossils sampled is k. Given potential incomplete species sampling, species i attaches to the tree at time b_i and goes extinct at time d_i. The origin x_0 is equal to the oldest b_i time. The age of the first and last fossil samples are o_i and y_i, respectively. Note

if species i has been sampled only once, i.e. species i is a singleton, $o_i = y_i$, and if species i is extant $y_i = d_i$. If species i has been sampled only once at the present $o_i = y_i = d_i$. To account for uncertainty in phylogenetic relationships, we can integrate over all possible tree topologies by defining γ_i, which is the number of co-existing lineages at time b_i. Since typically we only sample o_i and y_i we can also marginalise over all possible attachment and extinction times, assumming $b_i > o_i$ and $d_i < y_i$. Stadler et al. [3] derived the probability density for a given set of n stratigraphic ranges, $\mathcal{D} = (k, \{b_i, d_i, o_i\}_{i \in 1 \ldots n})$, given that we have no information about the underlying topology.

Let $Q(t)$ represent the probability of a not-yet sampled lineage evolving from time t according to a process in which one lineage arising from each unobserved speciation represents a new species, and the other lineage represents the continuation of the parent species. However, only one descendant lineage is ultimately sampled. Since the lineage has not yet been sampled, either descendant lineage can potentially represent the continuation of the parent species (i.e. the parent species may be ultimately sampled, or it may go extinct while the descendant species is ultimately sampled). This process describes the evolution of a lineage during the time when it has not yet been sampled, i.e. between its origination point b_i and the time of its first fossil sample o_i. The probability $Q(t)$ is described by the differential equation,

$$\frac{d}{dt}Q(t) = -(\lambda + \mu + \psi)Q(t) + 2\lambda Q(t)p(t) \tag{1}$$

where $p(t)$ is the probability that an individual at time t in the past does not leave any sampled fossils or sampled extant descendants (derived in [3]).

Let $\widetilde{Q}(t)$ represent the probability of a lineage evolving according to a similar process, but assuming at least one sampling event has occurred, so that the lineage arising from each speciation event that is ultimately sampled can always be identified as a continuation of the parent species lineage, and therefore by the same descendant branch

(left or right). This process describes the evolution of a lineage since the time when it was first sampled, i.e. between its first fossil sample o_i and its extinction time d_i. This probability is described by the differential equation,

$$\frac{d}{dt}\widetilde{Q}(t) = -(\lambda + \mu + \psi)\widetilde{Q}(t) + \lambda\widetilde{Q}(t)p(t) \tag{2}$$

Let $q(t)$ and $\widetilde{q}(t)$ represent solutions to equations (1) and (2), respectively. For stratigraphic range i, the initial condition for $Q(t)$ is $Q(o_i) = \widetilde{Q}(o_i)$, and therefore $Q(t) = \frac{q(t)}{q(o_i)}\widetilde{Q}(o_i)$. The initial condition for $\widetilde{Q}(t)$ is $\widetilde{Q}(d_i) = c_i$, where $c_i = \mu$ if $d_i > 0$ and $c_i = \rho$ if $d_i = 0$, and therefore $\widetilde{Q}(o_i) = \frac{\widetilde{q}(o_i)}{\widetilde{q}(d_i)}c_i$. Therefore, the probability of observing range i starting from time b_i is $Q(b_i) = \frac{q(b_i)}{q(o_i)}\frac{\widetilde{q}(o_i)}{\widetilde{q}(d_i)}c_i$.

Thus, the joint probability density for the full set of species ranges and k fossil samples $\mathcal{D} = \{k, (b_i, d_i, o_i)_{i\in 1\ldots n}\}$ is

$$\begin{aligned} f[\mathcal{D} \mid \lambda, \mu, \psi, \rho] &= \psi^k \lambda^{n-1} \prod_{i=1}^{n} \gamma_i Q(b_i) \\ &= \psi^k \lambda^{n-1} \mu^m \rho^{n-m} \prod_{i=1}^{n} \gamma_i \frac{q(b_i)}{q(o_i)} \frac{\widetilde{q}(o_i)}{\widetilde{q}(d_i)} \end{aligned} \tag{3}$$

Analytical expressions for $q(t)$ and $\widetilde{q}(t)$ are derived in [3], and are as follows

$$\begin{aligned} q(t) &= \frac{4e^{-c_1 t}}{(e^{-c_1 t}(1-c_2) + (1+c_2))^2} \\ c_1 &= \mid \sqrt{(\lambda - \mu - \psi)^2 + 4\lambda\psi} \mid \\ c_2 &= -\frac{\lambda - \mu - 2\lambda\rho - \psi}{c_1} \\ \widetilde{q}(t) &= \sqrt{q(t)e^{-(\lambda+\mu+\psi)t}} \end{aligned}$$

Note $\widetilde{q}(t)$ here refers to $\widetilde{q}_{asym}(t)$ in Stadler et al [3].

Conditioning on observing at least one sampled descendant

In some cases, we may want to condition on the event that at least one speciation occurs and that this speciation event leaves at least one sampled descendant. Denote the latter event by $\mathcal{S}$, where $f[\mathcal{S} \mid \lambda, \mu, \psi, \rho] = \lambda(1 - p(x_0))$. Then,

$$f[\mathcal{D} \mid \lambda, \mu, \psi, \rho, \mathcal{S}] = \frac{f[\mathcal{D} \mid \lambda, \mu, \psi, \rho]}{1 - p(x_0)}$$

where

$$p(t) = \frac{\lambda + \mu + \psi - c_1 \frac{(1+c_2)-(1-c_2)e^{-c_1 t}}{(1+c_2)+(1-c_2)e^{-c_1 t}}}{2\lambda} \tag{4}$$

Incorporating piece-wise constant rate variation

Next, we derive the probability density for $\mathcal{D}$ allowing speciation, extinction and sampling to change over time in a piece-wise fashion, following theory already outlined in [4, 1, 5]. We define l time intervals $[t_i, t_{i-1})$ for $i \in \{1, ..., l\}$ with minimum ages $t_1 > t_2 > \ldots > t_l$ between the present $t_l = 0$ and the past $t_0 = \infty$. Within each interval i, constant birth, death and sampling parameters λ_i, μ_i, ψ_i apply. Parameter ρ only applies to the present time $t_l = 0$.

For this model, we must define piece-wise expressions for $Q_i(t)$ and $\widetilde{Q}_i(t)$ in each interval i. In this case, the initial conditions for $\widetilde{Q}_i(t)$ at time $t \geq t_i$ in interval $i < l$ for stratigraphic range j are $\widetilde{Q}_i(t_i) = \widetilde{Q}_{i+1}(t_i)$ if $t_i > d_j$, and $\widetilde{Q}_i(d_i) = c_j$ if $d_j < t_i$. Thus, we have

$$\widetilde{Q}_i(t) = \begin{cases} \widetilde{Q}_{i+1}(t_i)\dfrac{\widetilde{q}_i(t)}{\widetilde{q}_i(t_i)} & t \geq t_i > d_j \\ c_j \dfrac{\widetilde{q}_i(t)}{\widetilde{q}_i(d_j)} & t \geq d_j > t_i \end{cases}$$

and therefore

$$\widetilde{Q}_i(o_j) = c_j \frac{\widetilde{q}_i(o_j)}{\widetilde{q}_{l(d_j)}(d_j)} \prod_{k=i}^{l(d_j)-1} \widetilde{q}_{k+1}(t_k)$$

where $l(t)$ gives the index i such that $t_{i-1} > t \geq t_i$.

Similarly,

$$Q_i(t) = \begin{cases} Q_{i+1}(t_j)\dfrac{q_i(t)}{q_i(t_i)} & t \geq t_i > o_j \\ \widetilde{Q}_i(o_j)\dfrac{q_i(t)}{q_i(o_j)} & t \geq o_j > t_i \end{cases}$$

and therefore,

$$Q_i(b_j) = \widetilde{Q}_{l(o_j)}(o_j)\frac{q_i(b_j)}{q_{l(o_j)}(o_j)} \prod_{k=i}^{l(o_j)-1} q_{k+1}(t_k)$$

Thus, for a given set of n stratigraphic ranges allowing for piece-wise constant rate variation across l intervals, assuming we have fossil counts k_j for each interval j, the joint probability density of the joint set of species ranges and interval-specific fossil counts $\mathcal{D} = \{(b_i, d_i, o_i)_{i\in 1\ldots n}, (k_j)_{j\in 1\ldots l}\}$ is

$$\begin{aligned} f[\mathcal{D} \mid \bar{\lambda}, \bar{\mu}, \bar{\psi}, \bar{t}, \rho] &= \frac{1}{\lambda_{l(x_0)}} \prod_{i=1}^{l} \psi_i^{k_i} \prod_{i=1}^{n} \lambda_{l(b_i)} \prod_{i=1}^{n} \gamma_i Q_{l(b_i)}(b_i) \\ &= \frac{\rho^{n-m}}{\lambda_{l(x_0)}} \prod_{i=1}^{l} \psi_i^{k_i} \prod_{i=1}^{n} \lambda_{l(b_i)} \prod_{i=1}^{m} \mu_{l(d_i)} \\ &\quad \times \prod_{i=1}^{n} \gamma_i \left(\frac{q_{l(b_i)}(b_i)}{q_{l(o_i)}(o_i)} \frac{\widetilde{q}_{l(o_i)}(o_i)}{\widetilde{q}_{l(d_i)}(d_i)} \prod_{j=l(b_i)}^{l(o_i)-1} q_{j+1}(t_j) \prod_{j=l(o_i)}^{l(d_i)-1} \widetilde{q}_{j+1}(t_j) \right) \end{aligned} \quad (5)$$

with the following solutions to the differential equations for $Q_i(t)$ and $\widetilde{Q}_i(t)$

$$\begin{aligned}
q_i(t) &= \frac{4e^{-A_i(t-t_i)}}{\left(e^{-A_i(t-t_i)}(1-B_i)+(1+B_i)\right)^2}, \\
A_i &= \sqrt{(\lambda_i-\mu_i-\psi_i)^2+4\lambda_i\psi_i}, \\
B_i &= \frac{(1-2(1-\rho_i)p_{i+1}(t_i))\lambda_i+\mu_i+\psi_i}{A_i}, \\
p_i(t) &= \frac{\lambda_i+\mu_i+\psi_i-A_i\frac{(1+B_i)-(1-B_i)e^{-A_i(t-t_i)}}{(1+B_i)+(1-B_i)e^{-A_i(t-t_i)}}}{2\lambda_i}, \\
\widetilde{q}_i(t) &= \sqrt{q_i(t)e^{-(\lambda+\mu+\psi)(t-t_i)}}.
\end{aligned}$$

Again we can condition on the event $\mathcal{S}$, of sampling at least one individual, where $f[\mathcal{S} \mid \bar{\lambda}, \bar{\mu}, \bar{\psi}, \bar{t}, \rho] = \lambda_{l(x_0)}(1-p_1(x_0))$

$$f[\mathcal{D} \mid \bar{\lambda}, \bar{\mu}, \bar{\psi}, \bar{t}, \rho, \mathcal{S}] = \frac{f[\mathcal{D} \mid \bar{\lambda}, \bar{\mu}, \bar{\psi}, \bar{t}, \rho]}{1-p_1(x_0)}$$

Marginalising over the number of fossils within a stratigraphic range

Next we extend the skyline model to incorporate alternative sampling scenarios. In some cases we may not know the exact number of samples collected during a given interval. Instead, we may only know the ages of the first and last fossil occurrence for each species (o_i, y_i). Thus, we define κ'_j as the total number of fossil samples representing either first or last appearances in interval j (i.e. within the interval $[t_j, t_{j-1})$). Next, letting $d_j(t) = t - t_j$ we define $L_j = \sum_{i=0}^{n} d_j^+(o_i) - d_j^+(y_i)$ as the total duration of all sampled stratigraphic ranges overlapping interval j. Using the same approach as Theorem 14 in [3], we integrate over the unknown occurrence times of the $k_j - \kappa'_j$ intermediate fossil samples, then sum over k_j to give the probability density for the joint set of species

ranges and first/last occurrence counts $\mathcal{D}_r = \{(b_i, d_i, o_i)_{i\in 1\ldots n}, (\kappa'_j)_{j\in 1\ldots l}\}$

$$f[\mathcal{D}_r \mid \bar{\lambda}, \bar{\mu}, \bar{\psi}, \rho] \propto f[\mathcal{D} \mid \bar{\lambda}, \bar{\mu}, \bar{\psi}, \rho] \prod_{j=1}^{l} e^{\psi_j L_j} \tag{6}$$

where the constant of proportionality is $\prod_{i=1}^{l} \psi_i^{k_i - \kappa'_j}$.

Marginalising over the number of fossils within a stratigraphic interval

Instead of recording the fossil sampling times or the age of first and last appearance times for a set of stratigraphic ranges, we may only know whether a species was sampled during the time interval $[t_i, t_{i-1}]$. We refer to this as presence/absence fossil sampling. We use $S_{i,j}$ to indicate the sub-branch spanned by species lineage i in interval j. Let $k_{S_{i,j}}$ indicate the number of fossil occurrences along sub-branch $S_{i,j}$, then let $\kappa_{S_{i,j}} = \mathbf{1}_{\mathbb{Z}^+}(k_{S_{i,j}})$ indicate the presence a fossil specimen for species i in interval j. Define $L_{S_{i,j}}$ as the duration of sub-branch $S_{i,j}$. If species i spans the entire interval j and $\kappa_{S_{i,j}} = 1$, then $L_{S_{i,j}}$ is simply equal to the total length of the interval $(t_{j-1} - t_j)$.

In order to compute the full likelihood under presence absence sampling, we first consider the likelihood only in the earliest interval for a single species. Let F indicate the set of species with at least one recovered fossil sample. Then, let $\alpha_i = \min\{j : \kappa_{S_{i,j}} = 1\}$ indicate the index of the earliest interval in which species $i \in F$ was sampled. Next, consider the fossil samples appearing in interval $l(o_i)$ other than the first sample at o_i. Using the same approach as Theorem 14 in [3], we integrate over the unknown occurrence times of these $k_{S_{i,\alpha_i}} - 1$ fossils, then sum over $k_{S_{i,\alpha_i}}$ to give the following expression for the joint density (for fixed o_i) of the datum $\kappa_{S_{i,\alpha_i}}$ and parameters b_i, d_i

$$f[\kappa_{i,\alpha_i}, b_i, d_i, o_i = t \mid \bar{\lambda}, \bar{\mu}, \bar{\psi}, \rho] \propto f[\mathcal{D}_i, o_i = t \mid \bar{\lambda}, \bar{\mu}, \bar{\psi}, \rho]\ e^{\psi_{\alpha_i}(t - \delta_i)}$$

where $\delta_i = \max\{d_i, t_{\alpha_i}\}$ and the constant of proportionality is $\psi_{\alpha_i}^{k_{S_{i,\alpha_i}} - 1}$

Next, because we ultimately consider o_i as unknown, we integrate the probability density over o_i, giving the following integrated expression for the joint density of of $\kappa_{S_{i,\alpha_i}}$ and b_i, d_i

$$\begin{aligned} f[\kappa_{S_{i,\alpha_i}}, b_i, d_i \mid \bar{\lambda}, \bar{\mu}, \bar{\psi}, \rho] &= \int_{\delta_i}^{\delta_i + L_{S_{i,\alpha_i}}} f[\kappa_{S_{i,\alpha_i}}, b_i, d_i, o_i = t \mid \bar{\lambda}, \bar{\mu}, \bar{\psi}, \rho] dt \\ &\propto \int_{\delta_i}^{\delta_i + L_{S_{i,\alpha_i}}} e^{\psi_{\alpha_i}(t-\delta_i)} f[\mathcal{D}_i, o_i = t \mid \bar{\lambda}, \bar{\mu}, \bar{\psi}, \rho] dt \\ &\propto f[\mathcal{D}_i \mid \bar{\lambda}, \bar{\mu}, \bar{\psi}, \rho] \int_{\delta_i}^{\delta_i + L_{S_{i,\alpha_i}}} e^{\psi_{\alpha_i}(t-\delta_i)} \frac{\widetilde{q}_{\alpha_i}(t)}{q_{\alpha_i}(t)} dt \\ &\propto f[\mathcal{D}_i \mid \bar{\lambda}, \bar{\mu}, \bar{\psi}, \rho] \; e^{-\psi_{\alpha_i}(\delta_i - t_{\alpha_i})} H_i \end{aligned}$$

where

$$H_i = \left| e^{-\frac{1}{2}(\lambda_{\alpha_i} + \mu_{\alpha_i} - \psi_{\alpha_i} - A_{\alpha_i})(t - t_{\alpha_i})} \left(\frac{1 + B_{\alpha_i}}{A_{\alpha_i} - (\lambda_{\alpha_i} + \mu_{\alpha_i} - \psi_{\alpha_i})} - \frac{(1 - B_{\alpha_i}) e^{-A_{\alpha_i}(t - t_{\alpha_i})}}{A_{\alpha_i} + (\lambda_{\alpha_i} + \mu_{\alpha_i} - \psi_{\alpha_i})} \right) \right|_{\delta_i}^{\delta_i + L_{S_{i,\alpha_i}}}$$

and the constant of proportionality is

$$\frac{\widetilde{q}_{l(o_i)}(o_i)}{q_{l(o_i)}(o_i)} \psi_{\alpha_i}^{k_{S_{i,\alpha_i}} - 1}$$

Then, using the same approach as Theorem 17 in [3], we marginalize the counts and occurrence times of the fossil samples for species i in the remaining intervals, yielding the joint probability density of $(\kappa_{S_{i,j}})_{j \in 1 \ldots l}$ and b_i, d_i

$$f[(\kappa_{S_{i,j}})_{j \in 1 \ldots l}, b_i, d_i \mid \bar{\lambda}, \bar{\mu}, \bar{\psi}, \rho] \propto f[\kappa_{S_{i,\alpha_i}}, b_i, d_i \mid \bar{\lambda}, \bar{\mu}, \bar{\psi}, \rho] \prod_{j > \alpha_i} \left[e^{\psi_j L_{S_{i,j}}} (1 - e^{-\psi_j L_{S_{i,j}}}) \right]^{\kappa_{S_{i,j}}}$$

where the constant of proportionality is

$$\prod_{j > \alpha_i} \psi_j^{k_{S_{i,j}}}$$

Finally, the full joint density of all species ranges and all presence absence data $\mathcal{D}_l = \{(b_i, d_i, (\kappa_{S_{i,j}})_{j\in 1\ldots l})_{i\in 1\ldots n}\}$ is therefore

$$\begin{aligned} f[\mathcal{D}_l \mid \bar{\lambda}, \bar{\mu}, \bar{\psi}, \rho] &= \prod_{i=0}^{n} f[(\kappa_{S_{i,j}})_{j\in 1\ldots l}, b_i, d_i \mid \bar{\lambda}, \bar{\mu}, \bar{\psi}, \rho] \\ &\propto f[\mathcal{D} \mid \bar{\lambda}, \bar{\mu}, \bar{\psi}, \rho] \prod_{i\in F} e^{-\psi_{\alpha_i}(\delta_i - t_{\alpha_i})} H_i \prod_{j>\alpha_i} \left[e^{\psi_j L_{S_{i,j}}}(1 - e^{-\psi_j L_{S_{i,j}}})\right]^{\kappa_{i,j}} \end{aligned} \tag{7}$$

where the constant of proportionality is

$$\prod_{i\in F} \frac{\widetilde{q}_{l(o_i)}(o_i)}{q_{l(o_i)}(o_i)} \psi_{\alpha_i}^{k_{S_{i,j}}-1} \prod_{j>\alpha_i}^{l} \psi_j^{k_{S_{i,j}}}$$

References

[1] Alexandra Gavryushkina, David Welch, Tanja Stadler, and Alexei J Drummond. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS Computational Biology*, 10(12):e1003919, 2014.

[2] Tanja Stadler. Sampling-through-time in birth–death trees. *Journal of theoretical biology*, 267(3):396–404, 2010.

[3] Tanja Stadler, Alexandra Gavryushkina, Rachel C M Warnock, Alexei J Drummond, and Tracy A Heath. The fossilized birth-death model for the analysis of stratigraphic range data under different speciation modes. *Journal of Theoretical Biology*, 447:41–55, 2018.

[4] Tanja Stadler, Denise Kühnert, Sebastian Bonhoeffer, and Alexei J Drummond. Birth–death skyline plot reveals temporal changes of epidemic spread in hiv and hepatitis c virus (hcv). *Proceedings of the National Academy of Sciences*, 110(1):228–233, 2013.

[5] Chi Zhang, Tanja Stadler, Seraina Klopfstein, Tracy A Heath, and Fredrik Ronquist. Total-evidence dating under the fossilized birth–death process. *Systematic Biology*, 65(2):228–249, 2015.