Estimating the divergence time of primates

An introduction to ABC lecture

June 11 2019
Statistical inference on trees: timescales

- Introduction
- Primate fossil record
- Dating splits by ABC
- Today’s posterior is tomorrow’s prior: molecular data
- Conclusions
Charles Darwin (1809 - 1882)
Reconciling molecular and fossil records?

- Extant primates are strepsirrhines (lemurs and lorises) and haplorhines (tarsiers and anthropoids)
- Molecular estimate of time of divergence is approximately 90 mya
- Fossil record suggests 60-65 mya
- Fossil record is patchy

Problem: Use the fossil record to estimate the age of the last common ancestor of extant primates
Table 1. Data and relative sampling intensities for the primate fossil record, taking a total of 235 modern species. References for the data can be found in the Supplementary Information.

<table>
<thead>
<tr>
<th>Epoch</th>
<th>$k$</th>
<th>$T_k$</th>
<th>Observed number of species ($D_k$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Late Pleistocene</td>
<td>1</td>
<td>0.15</td>
<td>19</td>
</tr>
<tr>
<td>Middle Pleistocene</td>
<td>2</td>
<td>0.9</td>
<td>28</td>
</tr>
<tr>
<td>Early Pleistocene</td>
<td>3</td>
<td>1.8</td>
<td>22</td>
</tr>
<tr>
<td>Late Pliocene</td>
<td>4</td>
<td>3.6</td>
<td>47</td>
</tr>
<tr>
<td>Early Pliocene</td>
<td>5</td>
<td>5.3</td>
<td>11</td>
</tr>
<tr>
<td>Late Miocene</td>
<td>6</td>
<td>11.2</td>
<td>38</td>
</tr>
<tr>
<td>Middle Miocene</td>
<td>7</td>
<td>16.4</td>
<td>46</td>
</tr>
<tr>
<td>Early Miocene</td>
<td>8</td>
<td>23.8</td>
<td>36</td>
</tr>
<tr>
<td>Late Oligocene</td>
<td>9</td>
<td>28.5</td>
<td>4</td>
</tr>
<tr>
<td>Early Oligocene</td>
<td>10</td>
<td>33.7</td>
<td>20</td>
</tr>
<tr>
<td>Late Eocene</td>
<td>11</td>
<td>37.0</td>
<td>32</td>
</tr>
<tr>
<td>Middle Eocene</td>
<td>12</td>
<td>49.0</td>
<td>103</td>
</tr>
<tr>
<td>Early Eocene</td>
<td>13</td>
<td>54.8</td>
<td>68</td>
</tr>
<tr>
<td>Pre-Eocene</td>
<td>14</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
The evolutionary process
What happened?

- Average sampling fraction of 5.7%
  - upper 95% limit 7.4%
- Estimated divergence time 81.5 mya
  - 95% CI (72.0, 89.6) mya


- Pravda, Times, BBC, . . . , assorted religious fanatics, . . .
Primate Evolution
Why more?

- Bayesian approach more natural
- Allows us to incorporate prior information
- Sampling fractions
  - probability of finding a fossil in bin \( i \) is \( \alpha_i \)
  - \( \alpha = \alpha p, p \) known
  - reasonable?
- Other models for finds?
- Allowing for dinosaur extinction at K/T boundary?
Fossil record: ABC approach

Data can be thought of in two parts:

(a) the observed number of fossils $F_{\text{obs}}$ found
(b) the proportions $p_{j,\text{obs}}$ found in $j$th bin

A suitable metric might be

$$\left| \frac{F}{F_{\text{obs}}} - 1 \right| + \sum_{j=1}^{k+1} |p_j - p_{j,\text{obs}}|$$
Results $\varepsilon = 0.1$
Some ABC technicalities

Hybrid ABC schemes
Sensitivity: Exploring Other Models

One advantage of ABC – it is easy to change the input . . .

- Choice of $\rho$
- Demography
- **Sampling fractions**
- K/T crash 65 mya
  - the time of origin of primates is even further back in the Cretaceous
- Poisson sampling scheme: length in bin matters
- **Dating other split points**
Hybrid ABC schemes: ABC-Gibbs

J1 If currently at $\mathbf{\theta} = (\theta_1, \theta_2)$, draw $\theta'_1$ from $\pi(\theta_1 | \mathcal{D}, \theta_2)$ and set $\mathbf{\theta} = (\theta'_1, \theta_2)$.

J2 Draw $\theta'_2$ from $\pi(\theta_2)$ and simulate data $\mathcal{D}'$ using parameter $\mathbf{\theta} = (\theta'_1, \theta'_2)$.

J3 If $\mathcal{D} = \mathcal{D}'$, set $\mathbf{\theta} = (\theta'_1, \theta'_2)$ and return to step J1. Otherwise stay at $\mathbf{\theta} = (\theta'_1, \theta_2)$ and return to step J2.

Steps J2 and J3 above are the mechanical version of the rejection algorithm which gives samples from $\pi(\theta_2 | \mathcal{D}, \theta_1)$. 

20
By replacing step J3 with

\[ J3' \text{ If } \rho(D, D') \leq \epsilon, \text{ set } \theta = (\theta'_1, \theta'_2) \text{ and return to step J1. Otherwise stay at } \theta = (\theta_1', \theta_2) \text{ and return to step J2.} \]

we can generate approximate draws from \( \pi(\theta_2|D, \theta_1) \).

- Could also use Approximate Metropolis-within-Gibbs and other variants
Dealing with Sampling Fractions

\[ f(\lambda, \tau, \mathcal{N}, \alpha | D) \propto P(D | \alpha, \lambda, \tau, \mathcal{N}) P(\mathcal{N} | \tau, \lambda) f(\tau) f(\lambda) f(\alpha) \]

where

- \( \lambda = (\lambda, \gamma, \rho) \) growth parameters,
- \( \alpha = (\alpha_1, \ldots, \alpha_{14}) \) sampling fractions
- \( \mathcal{N} \) is the underlying tree structure

Give sampling fractions independent Beta\((a, b)\) priors
Gibbs-ABC Example

Split the random variable into two parts: \( \alpha \) and \( (\lambda, \tau, N) \)

Sample from the two conditional distributions

- \( f(\alpha \mid D, \lambda, \tau, N) \)
- \( f(\tau, \lambda, N \mid D, \alpha) \)
Conditional distribution of $\alpha$

\[
f(\alpha \mid D, \lambda, \tau, N) \\
\propto f(\alpha, \lambda, \tau, N \mid D) \\
\propto \mathbb{P}(N \mid \tau, \lambda) f(\tau) f(\lambda) f(\alpha) \mathbb{P}(D \mid \tau, \lambda, N, \alpha) \\
\propto f(\alpha) \mathbb{P}(D \mid N, \alpha) \\
\propto \Pi_{i=1}^{I} \alpha_i^{d_i} (1 - \alpha_i)^{N_i - d_i} \alpha_i^{a_i - 1} (1 - \alpha_i)^{b_i - 1} \\
\propto \Pi_{i=1}^{I} f_{\beta}(\alpha_i ; d_i + a, N_i - d_i + b)
\]

Posterior mean of $\alpha_i = \frac{a + d_i}{N_i + a + b} \approx \frac{d_i}{N_i}$
Conditional distribution of \((\tau, \lambda, \mathcal{N})\)

\[
f(\tau, \lambda, \mathcal{N}|\mathcal{D}, \alpha) \propto f(\lambda, \tau, \mathcal{N}|\mathcal{D}) \\
\propto \mathbb{P}(\mathcal{D}|\lambda, \alpha, \mathcal{N}, \alpha) \mathbb{P}(\mathcal{N}|\tau, \lambda) f(\tau) f(\lambda)
\]

Simulate from this using ABC: accept \((\lambda, \tau, \mathcal{N})\) if

\(\rho(\mathcal{D}, \mathcal{D}') < \epsilon\), where \(\mathcal{D}'\) represents the simulated data.
Metric and Priors

\( \tau \sim U[0, 100] \)
\( \alpha \sim U[0, 0.6] \)
\( \rho \sim U[0, 0.8] \)
\( \gamma \sim U[0.005, 0.015] \)
\( 1/\lambda \sim U[2, 3] \)
\( a = 0.1 \)
\( b = 1 \)
\( \epsilon = 0.2 \)

Same metric as before
No free lunches
Tweak metric

• The observed $N_0$ values are too small
  
  – require $N_0 > 235$
  
  – change the metric

\[
\rho(D, D') = \sum_{i=1}^{k} \left| \frac{D_i}{D} - \frac{D_i'}{D'} \right| + \left| \frac{D_i'}{D'} - 1 \right| + \left| \frac{N_0'}{N_0} - 1 \right|
\]

• Penalises trees with $N_0$ values far from 235
Results: $\epsilon = 0.3$

<table>
<thead>
<tr>
<th></th>
<th>min</th>
<th>LQ</th>
<th>Median</th>
<th>mean</th>
<th>UQ</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_0$</td>
<td>184</td>
<td>212</td>
<td>224</td>
<td>226</td>
<td>238</td>
<td>279</td>
</tr>
<tr>
<td>$\tau$</td>
<td>0.0</td>
<td>8.0</td>
<td>18.6</td>
<td>26.3</td>
<td>36.8</td>
<td>99.5</td>
</tr>
</tbody>
</table>

![Histogram of accepted $N_0$ values](image1)

![Histogram of accepted $\tau$ values](image2)

![Histogram of accepted $\rho$ values](image3)

![Histogram of accepted $\gamma$ values, Strep](image4)

![Histogram of accepted $1/\lambda$ values](image5)
One advantage of ABC – it is easy to change the input . . .

- Choice of $\rho$
- Demography
- Sampling fractions
- K/T crash 65 mya
  - the time of origin of primates is even further back in the Cretaceous
- Poisson sampling scheme: length in bin matters
- Dating other split points
<table>
<thead>
<tr>
<th>Epoch</th>
<th>k</th>
<th>$T_k$</th>
<th>Hap/Strep number of species ($D_k^h$)</th>
<th>Plat/Cat number of species ($D_k^p$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Late Pleistocene</td>
<td>1</td>
<td>0.15</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>Middle Pleistocene</td>
<td>2</td>
<td>0.9</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td>Early Pleistocene</td>
<td>3</td>
<td>1.8</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>Late Pliocene</td>
<td>4</td>
<td>3.6</td>
<td>47</td>
<td>44</td>
</tr>
<tr>
<td>Early Pliocene</td>
<td>5</td>
<td>5.3</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>Late Miocene</td>
<td>6</td>
<td>11.2</td>
<td>38</td>
<td>33</td>
</tr>
<tr>
<td>Middle Miocene</td>
<td>7</td>
<td>16.4</td>
<td>46</td>
<td>43</td>
</tr>
<tr>
<td>Early Miocene</td>
<td>8</td>
<td>23.8</td>
<td>36</td>
<td>30</td>
</tr>
<tr>
<td>Late Oligocene</td>
<td>9</td>
<td>28.5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Early Oligocene</td>
<td>10</td>
<td>33.7</td>
<td>20</td>
<td>6</td>
</tr>
<tr>
<td>Late Eocene</td>
<td>11</td>
<td>37.0</td>
<td>32</td>
<td>2</td>
</tr>
<tr>
<td>Middle Eocene</td>
<td>12</td>
<td>49.0</td>
<td>103</td>
<td>0</td>
</tr>
<tr>
<td>Early Eocene</td>
<td>13</td>
<td>54.8</td>
<td>68</td>
<td>0</td>
</tr>
<tr>
<td>Pre-Eocene</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Dating Two Splits

32
• \( N_0 = 235 \) species for the Strep/Hap,

• \( \epsilon = 0.4 \) for both metrics

<table>
<thead>
<tr>
<th></th>
<th>min</th>
<th>LQ</th>
<th>Median</th>
<th>mean</th>
<th>UQ</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N_0 )</td>
<td>159</td>
<td>212</td>
<td>234</td>
<td>233</td>
<td>254</td>
<td>303</td>
</tr>
<tr>
<td>( \tau )</td>
<td>0.9</td>
<td>12.1</td>
<td>17.6</td>
<td>20.1</td>
<td>25.3</td>
<td>94.5</td>
</tr>
<tr>
<td>( \tau^* )</td>
<td>1.6</td>
<td>14.5</td>
<td>18.2</td>
<td>19.6</td>
<td>23.5</td>
<td>82.9</td>
</tr>
</tbody>
</table>

The median posterior sampling fractions (\( \times 100 \))
Dating Two Splits, revisited
The structure of branching processes

• Our approach to inferring multiple split points is heuristic

• What other approaches might work?

• Consider conditioning the process on a split at a fixed time
  – leads to a size-biased GW process
  – For ABC, need to be able to simulate the process
  – Can use rejection ...
Another fishbone process
Which metric?

\[ \rho(D, X) = \sum_{i=1}^{14} \left| \frac{D_i}{D_+} - \frac{X_i}{X_+} \right| + \left| \frac{X_+}{D_+} - 1 \right| + \left| \frac{X_0}{N_0} - 1 \right|. \]

Match up:

- Proportions of fossils observed in each bin
- Total number of fossils observed
- Number of extant species
What happened?
Combining fossil record with molecular data

Yesterday’s posterior is tomorrow’s prior . . .

- Estimate posterior for two primate divergence times
- Use as prior for dating nodes from molecular data
  \textit{(memetree)}

- Data are updated from earlier analysis
The posteriors
The phylogeny of the species (Poisson model)
The molecular data
References

