Outline

- We looked at evolution with an individual — short time scales
- Today, an MCMC/ABC problem from paleontology — long time scales
  - Trees and phylogenetics
  - Primate fossil record
  - Introduction to inference problem
  - Gibbs-ABC for inference
- Some concluding remarks

Charles Darwin (1809 - 1882)

Ernst Haeckel (1834-1919)

Haeckel tree of life (detail)
August Schleicher (1821-1868)

Different Types of Tree

- Across species
  - Phylogenetics
    - Species tree/gene tree
- Within species
  - Population genetics
    - Random trees (coalescents)

Questions about Phylogenetic Trees

- Topology of tree (the branching order)
- Branch lengths (time of common ancestor)
  - Calibrated from fossil record?
- How are trees found?
  - Morphological data
  - Molecular data (GHF Nuttall, 1904)
  - Genome sequences

Building Phylogenetic Trees

- Active research area
  - Challenging statistical inference problems
  - Bayesian methods — MCMC
  - . . . even if aim is maximum likelihood
- Dealing with genome-wide data

**Primate Evolution**

- 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

Reconciling molecular and fossil records?
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

Primate Data

<table>
<thead>
<tr>
<th>Epoch</th>
<th>k</th>
<th>T</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>37.6</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>
Model for speciation

A non-homogeneous Markov branching process:

- There are two species at time 0
- Species have exponential lifetimes with mean $1/\lambda$, time being measured in millions of years
- A species that goes extinct at time $u$ is replaced by an average of $m(u)$ new species
- $Z_t$ = number of species alive at time $t$

Stratigraphic intervals

- Divide time into $k$ stratigraphic intervals, beginning from the present and proceeding into the past
- The base of the first stratigraphic interval is at $T_1$ my and the base of the $k^{th}$ is at $T_k$ my
- The founding species originate at time $T = T_k + \tau$ my

Aim: estimate the parameter $\tau$

Model for Fossil Finds

- Given the number of distinct species $N_j$ in the $j$th stratigraphic interval ($j = 1, 2, \ldots, k + 1$), assume that the number of species $D_j$ actually found in the fossil record in this interval is
  \[
  \text{Binomial}(N_j, \alpha_j), \quad j = 1, 2, \ldots, k
  \]
- The parameter $\alpha_j$ gives the probability of sampling a species in the $j$th stratigraphic interval

Under this sampling model, the expected number of species found in the $j$th interval is
\[
ED_j = \alpha_j EN_j, \quad j = 1, \ldots, k + 1.
\]

Variance of the number of species sampled in the $j$th stratigraphic interval is
\[
\text{Var}D_j = \alpha_j(1 - \alpha_j)EN_j + \alpha_j^2\text{Var}N_j \approx \alpha_jEN_j,
\]
assuming that $\alpha_j$ is small

A Minimum $\chi^2$ Estimator

Statistic is
\[
X^2 = X^2(\theta) = \sum_{j=1}^{k+1} \frac{(D_j - \alpha_jEN_j)^2}{\alpha_jEN_j}
\]

The estimator is
\[
\hat{\theta} = \arg\min_{\theta} X^2(\theta)
\]

Some details

- We chose
  \[
  \alpha_j = \alpha p_j, \quad j = 1, 2, \ldots, k + 1,
  \]
  where the $p_j$ are known sampling intensities, and the scale parameter $\alpha$ is to be estimated
- The mean diversity curve was taken to be logistic:
  \[
  EZ_u = \frac{2}{\gamma + (1 - \gamma)e^{-\rho u}}
  \]
To fix some of the parameters, assume

\[ N_0 = E(Z_T) \]

• 90% of existing diversity by the Middle Eocene 49 mya
• Average species lifetime of 2.5 my

Statistical Approach

• Approximate confidence intervals for \( \tau \) and \( \alpha \) using bootstrap:
  – We simulate \( b \) realizations of the process \( Z_t \) using the estimated values \( \tau_0 \) and \( \alpha_0 \) in each run
  – For each run, \( i \), we re-estimate \( \tau \) and \( \alpha \), getting values \( \tau_i, \alpha_i, i = 1, \ldots, b \)
  – Run \( i \) also gives value \( X_i^2 \) of the statistic \( X^2 \)

What Happened?

and got . . .

• average sampling fraction of 5.7%
  – upper 95% limit 7.4%
• estimated divergence time 81.5 mya
  – 95% CI (72.0, 89.6) mya


Front Page News!

An Unusual Statistical Problem . . .

• There is no obvious sense in which these estimators should be well-behaved
• Typically not consistent
• Does this version of bootstrap work?

• Other approaches?
  – Implicit likelihood (O. Will)
  – Uses backward simulation of birth-death process

---

We simulate \( b \) realizations of the process \( Z_t \) using the estimated values \( \tau_0 \) and \( \alpha_0 \) in each run

\[ X_i^2 \] of the statistic \( X^2 \)
ABC Approach

Data can be thought of in two parts:

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

A suitable metric might be

$$\left| \frac{F}{F_{obs}} - 1 \right| + \frac{1}{2} \sum_{j=1}^{k+1} |p_j - p_{j,obs}|$$

We took $\epsilon = 0.1$ and used rejection approach

Note: no data summaries here

Results

We assumed logistic diversity curve

Priors:

$\rho \sim U(0.1, 0.5)$, $\gamma \sim U(0.005, 0.01)$, $\tau \sim U(0, 100)$

- Median sampling fraction was 6.0%
  - upper 95% credible limit 16.7%
- Median divergence time 82.1 mya (MLE $\approx$ 76.7 mya)
  - 95% credible interval (65.8, 118.3) mya

Results agree with standard MCMC: V. Plagnol

Posterior Density of Temporal Gap

Goodness of Fit:

Posterior for Number of Extant Primates

Dealing with Sampling Fractions

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

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

Sensitivity: Exploring Other Models

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

- Choice of $d$
- 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
Gibbs-ABC Example

Richard Wilkinson (Cambridge)

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

Sample from the two conditional distributions

- \( f(\alpha \mid D, \lambda, \tau, N) \)
- \( f(\tau, \lambda, N \mid D, \alpha) \)

\[ f(\alpha \mid D, \lambda, \tau, N) \propto f(\alpha, \lambda, \tau, N \mid D) \]
\[ \propto P(N \mid \tau, \lambda) f(\tau) f(\lambda) f(\alpha) P(D \mid \tau, \lambda, N, \alpha) \]
\[ \propto \Pi_{i=1}^{14} \alpha_i^d_i (1 - \alpha_i)^{N_i - d_i} \alpha_i^{a-1} (1 - \alpha_i)^{b-1} \]
\[ \propto \Pi f_\beta(\alpha_i ; d_i + a, N_i - d_i + b) \]

Posterior mean of \( \alpha_i = \frac{a + d_i}{a + b + N_i} \approx \frac{d_i}{N_i} \)

Conditional distribution of \((\tau, \lambda, N)\)

\[ f(\tau, \lambda, N \mid D, \alpha) \propto f(\lambda, \tau, N, \alpha \mid D) \]
\[ \propto P(D \mid \lambda, \alpha, N, \alpha) P(N \mid \tau, \lambda) f(\tau) f(\lambda) \]

Simulate from this using ABC: accept \((\lambda, \tau, N)\) if \( \rho(D, D') < \epsilon \), where \( D' \) 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

\[ \rho(D, D') = \sum_{i=1}^{k} \left| \frac{D_i - D'_i}{D_i + D'_i} \right| + \left| \frac{D'_i}{D_i} - 1 \right| + \left| \frac{N'_0}{N_0} - 1 \right| \]

- Penalises trees with \( N_0 \) values far from 235
Results: $\epsilon = 0.3$

<table>
<thead>
<tr>
<th>$N_0$</th>
<th>min</th>
<th>LQ</th>
<th>Median</th>
<th>mean</th>
<th>UQ</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>184</td>
<td>212</td>
<td>224</td>
<td>226</td>
<td>238</td>
<td>279</td>
<td></td>
</tr>
</tbody>
</table>

| $\tau$ | 0.0  | 8.0  | 18.6   | 26.3 | 36.8| 99.5|

Old World/New World Split

<table>
<thead>
<tr>
<th>Epoch</th>
<th>$k$</th>
<th>$T_k$</th>
<th>Hap/Strep number of species ($D_k$)</th>
<th>Plat/Cat number of species ($D^*_k$)</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>1.6</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>26.5</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>66</td>
<td>0</td>
</tr>
<tr>
<td>Pre-Eocene</td>
<td>14</td>
<td></td>
<td>14</td>
<td>0</td>
</tr>
</tbody>
</table>

Dating Two Splits

Details

- $N_0 = 235$ species for the Strep/Hap,
- $\epsilon = 0.4$ for both metrics

<table>
<thead>
<tr>
<th>$N_0$</th>
<th>min</th>
<th>LQ</th>
<th>Median</th>
<th>mean</th>
<th>UQ</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>159</td>
<td>212</td>
<td>234</td>
<td>233</td>
<td>254</td>
<td>303</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\tau$</th>
<th>0.9</th>
<th>12.1</th>
<th>17.6</th>
<th>20.1</th>
<th>25.3</th>
<th>94.5</th>
</tr>
</thead>
<tbody>
<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 ($\times 100$)

<table>
<thead>
<tr>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_4$</th>
<th>$\alpha_5$</th>
<th>$\alpha_6$</th>
<th>$\alpha_8$</th>
<th>$\alpha_9$</th>
<th>$\alpha_{10}$</th>
<th>$\alpha_{11}$</th>
<th>$\alpha_{12}$</th>
<th>$\alpha_{13}$</th>
<th>$\alpha_{14}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>10</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
</tr>
</tbody>
</table>

Posterior

Joint posterior of $\tau$ and $\tau^*$
Conclusions

- Primates originated while dinosaurs roamed the earth (revise textbooks)
  - Median date is 73 mya
- Results suggest that New World monkeys have an early origin
  - Median is 55 mya
- Times in primate tree need revising: chimps and humans diverged \( \approx 8 \) mya

Implications

- Increasing Rate of Fossil Primate Discovery
- Pangea 50 mya
- Mirabile dictu!
Collaborators

Richard Wilkinson  [Univ of Cambridge]
Bob Martin       [Field Museum, Chicago]
Christophe Soligo [Natural History Museum, London; UCL]
Chris Scotese    [Paleomap Project]
Oliver Will      [Univ of Canterbury, NZ]