The Impact of the Representation of Fossil Calibrations on Bayesian Estimation of Species Divergence Times

Jun Inoue, Philip C.J. Donoghue and Ziheng Yang
Introduction

The Bayesian method \rightarrow Integrating information about divergence times based on the fossil record.

Rates
Times <- Fossils
Likelihood <- Sequence
\rightarrow MCMC
\rightarrow Posterior distributions

Which of them have the greatest impact on time estimates?

Ex. MCMCTREE produced about 1.4 times older estimates than MULTIDIVTIME

(Rannala and Yang 2007)
Objective

Assess the impact of the prior and other factors on Bayesian estimation of divergence times.
### Difference between MCMCTREE (PAML) and MULTIDIVTIME

<table>
<thead>
<tr>
<th></th>
<th>MULTIDIVTIME</th>
<th>MCMCTREE in PAML</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Use of constraints on the root age</strong></td>
<td>Requires gamma prior.</td>
<td>Requires the root age to be constrained loosely from above, if no upper bound exists on the root.</td>
</tr>
<tr>
<td><strong>Outgroup</strong></td>
<td>Requires outgroup.</td>
<td>Rooted ingroup tree only.</td>
</tr>
<tr>
<td><strong>Minimum and maximum bounds</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Specification of the prior on times.</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Specification of the prior on times in MCMCTREE and MULTIDIVTIME

**MULTIDIVTIME**

- Root age based on the gamma prior.
- Specify the distribution of other node ages.
- Minimum and maximum bounds are applied by truncating the joint density of times in the MCMC.

**MCMCTREE in PAML**

- The joint distribution of the ages for the **calibration nodes** is generated by multiplying independent densities.
- The age distribution of the **non-calibration nodes** given the ages of the calibration nodes is specified.
- Truncation in the MCMC.
The most probable time should be older than the fossil minimum since the acquisition of the apomorphies will postdate the actual divergence (e.g., Steiper and Young, 2008).

The probability density should increase with increasing age from the minimum, peak at the most probable time, and decrease the strength of evidence.

We used a **truncated Cauchy distribution** to represent the minimum bound.
Cauchy distribution for lower bound in MCMCTREE ‘L(t_L, p, c)’

Example:

- Mode: \( t_L(1+p) \)
- Scale parameter: \( ct_L \)

- \( c \) has a greater impact than \( p \) on posterior time estimation.
- The user is advised to use different values of \( p \) and \( c \) for each minimum bound.
Summary of the differences among three programs

<table>
<thead>
<tr>
<th></th>
<th>Specification of the prior on times</th>
<th>Implementation of minimum bound</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MCMCTREE4.1</strong></td>
<td>Joint distribution of the calibration node ages.</td>
<td>Uniform distribution</td>
</tr>
<tr>
<td><strong>MCMCTREE4.2</strong></td>
<td>MCMC sampling.</td>
<td>Cauchy distribution</td>
</tr>
<tr>
<td><strong>MULTIDIVTIME</strong></td>
<td>Root age with gamma prior.</td>
<td>Uniform distribution</td>
</tr>
</tbody>
</table>

- **MCMCTREE4.1**: Joint distribution of the calibration node ages.
- **MCMCTREE4.2**: MCMC sampling.
- **MULTIDIVTIME**: Root age with gamma prior.
Settings for the three programs
- as similar as possible -

<table>
<thead>
<tr>
<th></th>
<th>MULTIDIVTIME</th>
<th>MCMCTREE4.1, 4.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood</td>
<td>Normal approximation of the MLEs</td>
<td>= Normal approximation of the MLEs</td>
</tr>
<tr>
<td>Substitution model</td>
<td>F84+ G⁵</td>
<td>= F84+ G⁵</td>
</tr>
<tr>
<td>Gamma priors of the rate-drift parameter</td>
<td>mean, standard deviation</td>
<td>= shape (α) and scale (β) parameters</td>
</tr>
<tr>
<td>Rate correlation model</td>
<td>Auto-correlated</td>
<td>= Auto-correlated</td>
</tr>
<tr>
<td>Gamma prior on the root age</td>
<td>Three datasets</td>
<td>≈ Frog dataset only</td>
</tr>
</tbody>
</table>
Three datasets

Frog (Amphibia)  
San mauro et al. (2005)

RAG1 gene: 1 gene  
52 spp.  
1368 bp.  
2 max. and 8 min. bounds

mt Genome: 36 genes  
28 spp.  
10,327 bp.  
2 max. and 12 min. bounds

Cat (Felidae)  
Johnson et al. (2006)

Nuclear genes: 30 genes  
38 spp.  
19,984 bp.  
1 max. and 11 min. bounds

Fish (Actinopterygii)  
Inoue et al. (2005)

300 bp.  
19,984 bp.  
1 max. and 11 min. bounds

Three datasets

Frog (Amphibia)  
San mauro et al. (2005)

RAG1 gene: 1 gene  
52 spp.  
1368 bp.  
2 max. and 8 min. bounds

mt Genome: 36 genes  
28 spp.  
10,327 bp.  
2 max. and 12 min. bounds

Cat (Felidae)  
Johnson et al. (2006)

Nuclear genes: 30 genes  
38 spp.  
19,984 bp.  
1 max. and 11 min. bounds

Fish (Actinopterygii)  
Inoue et al. (2005)

Three datasets

Frog (Amphibia)  
San mauro et al. (2005)

RAG1 gene: 1 gene  
52 spp.  
1368 bp.  
2 max. and 8 min. bounds

mt Genome: 36 genes  
28 spp.  
10,327 bp.  
2 max. and 12 min. bounds

Cat (Felidae)  
Johnson et al. (2006)

Nuclear genes: 30 genes  
38 spp.  
19,984 bp.  
1 max. and 11 min. bounds

Fish (Actinopterygii)  
Inoue et al. (2005)

Three datasets

Frog (Amphibia)  
San mauro et al. (2005)

RAG1 gene: 1 gene  
52 spp.  
1368 bp.  
2 max. and 8 min. bounds

mt Genome: 36 genes  
28 spp.  
10,327 bp.  
2 max. and 12 min. bounds

Cat (Felidae)  
Johnson et al. (2006)

Nuclear genes: 30 genes  
38 spp.  
19,984 bp.  
1 max. and 11 min. bounds

Fish (Actinopterygii)  
Inoue et al. (2005)

Three datasets

Frog (Amphibia)  
San mauro et al. (2005)

RAG1 gene: 1 gene  
52 spp.  
1368 bp.  
2 max. and 8 min. bounds

mt Genome: 36 genes  
28 spp.  
10,327 bp.  
2 max. and 12 min. bounds

Cat (Felidae)  
Johnson et al. (2006)

Nuclear genes: 30 genes  
38 spp.  
19,984 bp.  
1 max. and 11 min. bounds

Fish (Actinopterygii)  
Inoue et al. (2005)

Three datasets

Frog (Amphibia)  
San mauro et al. (2005)

RAG1 gene: 1 gene  
52 spp.  
1368 bp.  
2 max. and 8 min. bounds

mt Genome: 36 genes  
28 spp.  
10,327 bp.  
2 max. and 12 min. bounds

Cat (Felidae)  
Johnson et al. (2006)

Nuclear genes: 30 genes  
38 spp.  
19,984 bp.  
1 max. and 11 min. bounds

Fish (Actinopterygii)  
Inoue et al. (2005)

Three datasets

Frog (Amphibia)  
San mauro et al. (2005)

RAG1 gene: 1 gene  
52 spp.  
1368 bp.  
2 max. and 8 min. bounds

mt Genome: 36 genes  
28 spp.  
10,327 bp.  
2 max. and 12 min. bounds

Cat (Felidae)  
Johnson et al. (2006)

Nuclear genes: 30 genes  
38 spp.  
19,984 bp.  
1 max. and 11 min. bounds

Fish (Actinopterygii)  
Inoue et al. (2005)

Three datasets

Frog (Amphibia)  
San mauro et al. (2005)

RAG1 gene: 1 gene  
52 spp.  
1368 bp.  
2 max. and 8 min. bounds

mt Genome: 36 genes  
28 spp.  
10,327 bp.  
2 max. and 12 min. bounds

Cat (Felidae)  
Johnson et al. (2006)

Nuclear genes: 30 genes  
38 spp.  
19,984 bp.  
1 max. and 11 min. bounds

Fish (Actinopterygii)  
Inoue et al. (2005)

Three datasets

Frog (Amphibia)  
San mauro et al. (2005)

RAG1 gene: 1 gene  
52 spp.  
1368 bp.  
2 max. and 8 min. bounds

mt Genome: 36 genes  
28 spp.  
10,327 bp.  
2 max. and 12 min. bounds

Cat (Felidae)  
Johnson et al. (2006)

Nuclear genes: 30 genes  
38 spp.  
19,984 bp.  
1 max. and 11 min. bounds

Fish (Actinopterygii)  
Inoue et al. (2005)
The impact of parameters $p$ and $c$ in the minimal-bound density for the prior mean node ages (MCMC TREE 4.2)

Increasing either $p$ or $c$ made node ages older, with $c$ having a larger effect.
The impact of parameters $p$ and $c$ in the minimal-bound density for the **posterior** mean node ages (MCMCTREE4.2)

Importantly the effect was present in the posterior as well as in the prior. Based on these results, we used $p = 0.1$ and $c = 1.0$ as default values (Different values of $p$ and $c$ should be used for each minimum-bound).
Sensitivity analysis for substitution models and rate-drift parameters

Substitution model had minimal impact. The prior on the rate-drift parameter $\sigma^2$ was more important.
The prior and posterior means of node ages using MCMCTREE and MULTIDIVTIME

The posterior time estimates were sensitive to different strategies for representing minimum and maximum fossil bounds.
The uncertainties in the posterior and the impact of the prior will not disappear with the addition of sequence data. The amount of sequence data approaches infinity, and the posterior means of times and the 95% CIs for different nodes fall on a straight line.

Correlation ($R^2$):
- Weak
  - Not enough
- Stronger
  - Better
- Nearly perfect
  - Nearly reached saturation

Slope ($w$):
- Every 1 Myr of div. time adds 0.44 MY of uncert.
- Smaller
  - More informative
- Larger
  - Most imprecise

Constraints:
- The uncertainties in the posterior and the impact of the prior will not disappear with the addition of sequence data.
**Summary**

**Sensitivity analysis:** The prior on time (with fossil calibrations) had the greatest impact on posterior time estimation.

**Implementation of minimum bound calibration:** The posterior time estimates were sensitive to different strategies for representing minimum and maximum fossil bounds.

**Infinite-sites theory:** The uncertainties in the posterior and the impact of the prior will not disappear with the addition of sequence data.

- Critical importance of fossil calibrations to molecular dating.
- Urgent need for research into ways of summarizing the fossil data to provide useful calibrations.