# Supplement to Missing data in a stochastic Dollo model for binary traits, and its application to the dating of Proto-Indo-European

Robin J. Ryder and Geoff K. Nicholls
*Department of Statistics, University of Oxford, UK*

This supplement to Ryder and Nicholls (2009) gives results for a second data set, by Dyen et al. (1997), as well as details of validations using synthetic data.

## 1. Description of the data

Ryder and Nicholls (2009) analyses a data set of 24 Indo-European languages put together by Ringe et al. (2002). The data by Dyen et al. (1997) comprise of 84 modern Indo-European languages, to which Gray and Atkinson (2003) add 3 ancient languages (Hittite, Tocharian A and Tocharian B). These data contain 2449 cognate classes over 207 meaning categories and 19 constraints on node ages. There are much less missing data points: only 2% of the data are missing (18% are missing in the Ringe et al. (2002) data). The registration process is also different; see Section 2.4.

## 2. Models

### *2.1. Prior distribution on trees*

Our main interest is in estimating the age of the root. The two main competing hypotheses for the diversification of the Indo-European family are the Kurgan hypothesis, which places the root between 6000 and 6500 BP (Before Present), and the Anatolian hypothesis, which places it between around 8500 BP. We were therefore looking for a uniform prior on the root age over these regions. Figure 1 shows a sample from the prior described in Section 2.1 of the main paper, showing that it is approximately uniform over the region of interest.

### *2.2. Diversification of cognacy classes*

### *2.3. Time reversibility*

In this section, we show that the process is time-reversible if and only if $\nu = \kappa\lambda/\mu$, as claimed in section 2 of the main paper. For this, we look at the transition rates

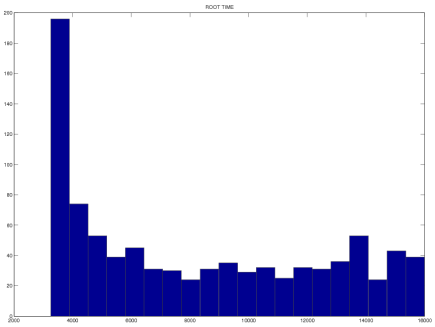


**Fig. 1.** *Sample from the prior distribution on the root age. The prior is approximately flat over the region of interest.*

$R_{i,j}$ from the state with $i$ traits to the state with $j$ traits. These transition rates are the sum of the transition rates for the anagenic process, which allows transitions from $i$ to $i+1$ and $i-1$, and the transition rates for the catastrophe process, which allows transitions from any $i$ to any $j$.

- for $|i-j| \neq 1$, the transition from $i$ to $j$ has to go through a catastrophe (these occur at rate $\rho$). Let $k$ be the number of traits existing after the deaths have occurred, but before the births (if this catastrophe occurs at time $\tau$, this would be the number of traits in existence between the times $\tau - \epsilon$ and $\tau - 2\epsilon$). Only deaths occurred to go from $i$ to $k$, so $k \leq i$, and only births occur to go from $k$ to $j$, so $k \leq j$; $k$ can take any value between 0 and $\min(i,j)$. Summing over all these values, we get:

$$R_{i,j} = \rho \sum_{k=0}^{k=\min(i,j)} Bin(k; i, 1-\kappa) \times Poi(j-k; \nu) \tag{1}$$

(starting with $i$ traits, $k$ traits have to survive the thinning process with survival probability $1-\kappa$; then the remaining $j-k$ traits are born through the $Poi(\nu)$ process). This becomes

$$R_{i,j} = \rho \kappa^i e^{-\nu} \nu^j i! \; {}_2F_0\left(-i, -j; \frac{1-\kappa}{\nu\kappa}\right) \tag{2}$$

where

$${}_2F_0(-i,-j;\theta) = \sum_{k=0}^{k=\min(i,j)} \theta^k \frac{1}{k!(i-k)!(j-k)!}$$

is a generalized hypergeometric function. Note that ${}_2F_0(-i,-j;\theta) = {}_2F_0(-j,-i;\theta)$.

- for $j = i+1$, the transition rates are

$$R_{i,j} = \lambda + \rho e^{-\nu} \frac{\nu^j \kappa^{j-1}}{j!} \, {}_2F_0\left(-i,-j;\frac{1-\kappa}{\nu\kappa}\right) \tag{3}$$

$$R_{j,i} = \mu j + \rho e^{-\nu} \frac{\nu^{j-1} \kappa^{j}}{(j-1)!} \, {}_2F_0\left(-j,-i;\frac{1-\kappa}{\nu\kappa}\right). \tag{4}$$

Now, assume that $\nu = \kappa\lambda/\mu$. We have proven in Section 2.2 of the main paper that $\pi_i = \frac{e^{-\frac{\nu}{\kappa}}\nu^i}{i!\kappa^i}$. If $|j-i| \neq 1$, it is straightforward to check that $\pi_i R_{i,j} = \pi_j R_{j,i}$. If $j = i+1$, then

$$\begin{aligned}
\frac{\pi_j R_{j,i}}{\pi_i R_{i,j}} &= \frac{\lambda}{\mu}\frac{1}{j} \times \frac{\mu j + \rho e^{-\nu} \frac{\nu^{j-1}\kappa^j}{(j-1)!} \, {}_2F_0\left(-j,-i;\frac{1-\kappa}{\nu\kappa}\right)}{\lambda + \rho e^{-\nu} \frac{\nu^j \kappa^{j-1}}{j!} \, {}_2F_0\left(-i,-j;\frac{1-\kappa}{\nu\kappa}\right)} \\
&= \frac{\lambda}{\mu} \times \frac{\mu + \rho e^{-\nu} \frac{\nu^{j-1}\kappa^j}{j!} \, {}_2F_0\left(-j,-i;\frac{1-\kappa}{\nu\kappa}\right)}{\lambda + \rho e^{-\nu} \frac{\nu^j \kappa^{j-1}}{j!} \, {}_2F_0\left(-i,-j;\frac{1-\kappa}{\nu\kappa}\right)} \\
&= \frac{\frac{\lambda}{\mu}\mu + \frac{\nu}{\kappa}\rho e^{-\nu} \frac{\nu^{j-1}\kappa^j}{j!} \, {}_2F_0\left(-j,-i;\frac{1-\kappa}{\nu\kappa}\right)}{\lambda + \rho e^{-\nu} \frac{\nu^j \kappa^{j-1}}{j!} \, {}_2F_0\left(-i,-j;\frac{1-\kappa}{\nu\kappa}\right)} \\
&= 1.
\end{aligned}$$

For all $i$ and $j$, $\pi_i R_{i,j} = \pi_j R_{j,i}$ and so the process is time-reversible.

Suppose conversely that the process is time-reversible. Take $i \geq 2$ and let $j = i+1$. Then

$$\begin{aligned}
\pi_i &= \frac{\pi_0 R_{0,i}}{R_{i,0}} \text{ (by time-reversibility)} \\
&= \frac{\nu^i}{\kappa^i}\frac{\pi_0}{i!} \text{ (by equation (2))} \qquad (5) \\
\pi_j &= \frac{\nu^j}{\kappa^j}\frac{\pi_0}{j!} \qquad (6)
\end{aligned}$$

Hence $\frac{\pi_i}{\pi_j} = (i+1)\frac{\kappa}{\nu}$. Equations (3) and (4) give

$$\frac{R_{j,i}}{R_{i,j}} = \frac{\mu(i+1) + \frac{\rho e^{-\nu}\nu^i\kappa^{i+1}}{i!} \, {}_2F_0\left(-i,-j;\frac{1-\kappa}{\nu\kappa}\right)}{\lambda + \frac{\rho e^{-\nu}\nu^{i+1}\kappa^i}{(i+1)!} \, {}_2F_0\left(-j,-i;\frac{1-\kappa}{\nu\kappa}\right)} \tag{7}$$

Since the process is time reversible, we have $\frac{\pi_i}{\pi_j} = \frac{R_{j,i}}{R_{i,j}}$, *i.e.*

$$(i+1)\frac{\kappa}{\nu} = \frac{\mu(i+1) + \frac{\rho e^{-\nu}\nu^i\kappa^{i+1}}{i!} \, {}_2F_0\left(-i,-j;\frac{1-\kappa}{\nu\kappa}\right)}{\lambda + \frac{\rho e^{-\nu}\nu^{i+1}\kappa^i}{(i+1)!} \, {}_2F_0\left(-j,-i;\frac{1-\kappa}{\nu\kappa}\right)} \tag{8}$$

Dividing the numerator by $(i+1)\kappa$ and the denominator by $\nu$ gives

$$1 = \frac{\frac{\mu}{\kappa} + \frac{\rho e^{-\nu}\nu^i\kappa^i}{(i+1)!}\,{}_2F_0\left(-i,-j;\frac{1-\kappa}{\nu\kappa}\right)}{\frac{\lambda}{\nu} + \frac{\rho e^{-\nu}\nu^i\kappa^i}{(i+1)!}\,{}_2F_0\left(-i,-j;\frac{1-\kappa}{\nu\kappa}\right)}. \tag{9}$$

It follows that $\frac{\mu}{\kappa} = \frac{\lambda}{\nu}$.

This shows that the process is time reversible if and only if $\nu = \lambda\kappa/\mu$.

### *2.4. The registration process*

The Dyen et al. (1997) data have a different registration process from the Ringe et al. (2002) data. Cognacy classes diversify in the same way, but the linguists entering the data chose to exclude any traits which were not informative of the topology, *i.e.* traits which only appear at a single leaf; this corresponds to thinning operation $R_2$ in the main paper.

The mathematical details for this registration process are given in Appendix A of the main paper.

### *2.5. Point process of births for registered cognacy classes*

## 3. Likelihood calculations

## 4. Posterior distribution

In Equation 8, we have shown that the posterior distribution is

$$\begin{aligned} & p(g,\mu,\lambda,\kappa,\rho,\xi|\mathbf{D}=D) \\ &= \frac{1}{N!}\left(\frac{\lambda}{\mu}\right)^N \exp\left(-\frac{\lambda}{\mu}\sum_{\langle i,j\rangle\in E} P[\mathcal{E}_Z|Z=(t_i,i),g,\mu,\kappa,\xi](1-e^{-\mu(t_j-t_i+k_iT_C)})\right) \\ &\quad \times\prod_{a=1}^{N}\left(\sum_{\langle i,j\rangle\in E_a}\sum_{\omega\in\Omega_a} P[M=\omega|Z=(t_i,i),g,\mu,\kappa](1-e^{-\mu(t_j-t_i+k_iT_C)})\right) \\ &\quad \times\frac{1}{\mu\lambda\rho} f_G(g|T)\frac{e^{-\rho|g|}(\rho|g|)^{k_T}}{k_T!}\prod_{i=1}^{L}(1-\xi_i)^{Q_i}\xi_i^{N-Q_i} \end{aligned}$$

For the parameters $\rho$, $\mu$ and $\lambda$, we have used the improper prior $p(\rho,\mu,\lambda)\propto 1/\rho\mu\lambda$ (and $\kappa$ has a uniform prior on $(0,1)$). We have in addition used a Gamma prior $\rho\sim\Gamma(1.5,0.0002)$ (so that the number of catastrophes on a typical tree varies between around 1 and 50 - the posterior for dating, tree structure and catastrophe placement

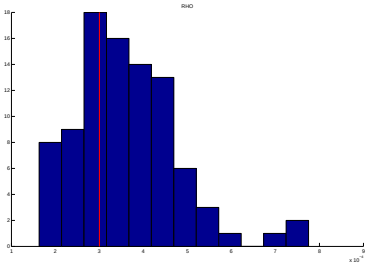


**Fig. 2.** Reconstruction of the catastrophe rate $\rho$ from the same in-model validation analysis as the one used for Figure 3 in the article. Blue: histogram from the posterior sample; red: true value.

was, in this second study, almost identical to the results we obtained for the scale invariant prior $1/\rho$, and are omitted). We integrate out $\lambda$.

We can show that under conditions which are verified by any realistic data set, the posterior is proper when $\mu \to \infty$ at any fixed $\rho$ (see Ryder (2010) for details). However, the posterior is improper when $\rho \to 0$, $\rho \to \infty$ or $\mu \to 0$. We can make the posterior proper by placing very conservative bounds on these parameters. We can put a bound on $\mu$ so that $\mu > 10^{-7}$. For values of $\mu$ less than $10^{-7}$, the probability that not a single cognate death occur on the entire tree under the anagenic process is of the order of 50%. In other words, all deaths would have to occur at catastrophes; we rule out these parameter values (where the catastrophe process mimics the anagenic process) and force catastrophes to model "catastrophic" change, and the birth death process to represent the anagenic evolution. In our analyses, the MCMC never visits values of $\mu$ less than $10^{-5}$ so the bound may be removed to arbitrarily conservative values without altering our results. Similarly, we can impose the constraint $10^{-7} < \rho < 1$; here again, the MCMC always remains very far from these bounds, so results are insensitive to the details of these bounds. This the same in essence as restricting the parameters to a compact set, and observing that the posterior density is finite there. In fact, we can allow no upper bound on $\mu$.

## 5. Markov Chain Monte Carlo

## 6. Validation

The article presents an analysis of in-model synthetic data, which shows that the topology, the root age, and the death rate $\mu$ are well reconstructed. Figure 2 shows that the catastrophe rate $\rho$ is also well reconstructed.

### 6.1. Model mis-specification: Borrowing

Borrowing between languages is a frequent phenomenon, which we do not include in our model. Even though the levels of borrowing for core vocabulary are much lower than for general vocabulary (Embleton, 1986), we searched for potential systematic bias. We simulated data with different levels of borrowing $b$ and under two different models of borrowing (global and local). Our model of borrowing is as follows: borrowing events occur at rate $b\mu$. At a borrowing event, two languages $l_1$ and $l_2$ are chosen uniformly at random; one trait is chosen uniformly at random amongst those present in language $l_1$, and it is copied into $l_2$. Under the global model, borrowing can occur between any two languages; under the local model, it can occur only between languages which split less than $T_b$ years previously (we used $T_b = 1000$ years). We looked at low levels ($b = 0.1$) and high levels ($b = 0.5$) of borrowing. We did not consider "catastrophic" borrowing, in which one language would borrow many words from another language in a short amount of time, as did Warnow et al. (2004).

For $b = 0.1$, the topology was well reconstructed, with only minor differences between the true tree and the output (Figure 3 (a)-(b)). The dates, catastrophes and parameters were also correctly reconstructed. This is typical, so the effect of low levels of borrowing is negligible, under both global and local models of borrowing. For $b = 0.5$, the topology was surprisingly well reconstructed in the examples we looked at, given the amount of noise in the data (Figure 3 (c)-(d)). However, we found that for $b = 0.5$, we systematically underestimated the root age and overestimated the rate parameters by up to 75% (Figure 3 (e)-(f)). This is of little concern to us, since we have reason to believe that no such high levels of borrowing occurred in the data we are analysing (see Nicholls and Gray (2008)); furthermore, even borrowing might cause a small downward bias, but we focus on evidence for upward bias.

### 6.2. Reversibility

In our model, we have imposed the reversibility condition $\nu = \kappa\lambda/\mu$. In order to check for systematic bias arising from this condition, we simulated data with different values of $\nu$ and estimated all parameters under the reversibility condition. Here again, we do not expect to be able to correctly estimate all parameters, but we hope that no systematic bias will be introduced in the estimates of the topology and of the root age. The data we simulated used the parameter values $\mu = 2.23 \cdot 10^{-4}$, $\kappa = 0.2$, $\lambda = 4.46 \cdot 10^{-2}$ and we studied $\nu = 2\kappa\lambda/\mu$ and $\nu = \kappa\lambda/2\mu$. The data were simulated on a tree with 20 languages and 8 internal constraints.

The reconstruction of the topology is not affected: the topology is still almost perfectly reconstructed, as shown in Figure 4. However, the position of catastrophes is much more uncertain: no catastrophes are supported in more than 50% of the posterior. The trees shown are for $\nu = \kappa\lambda/2\mu$; the situation is similar for $\nu = 2\kappa\lambda/\mu$.

The parameters $\mu$ and $\kappa$ are not well reconstructed, as shown in figure 5: the

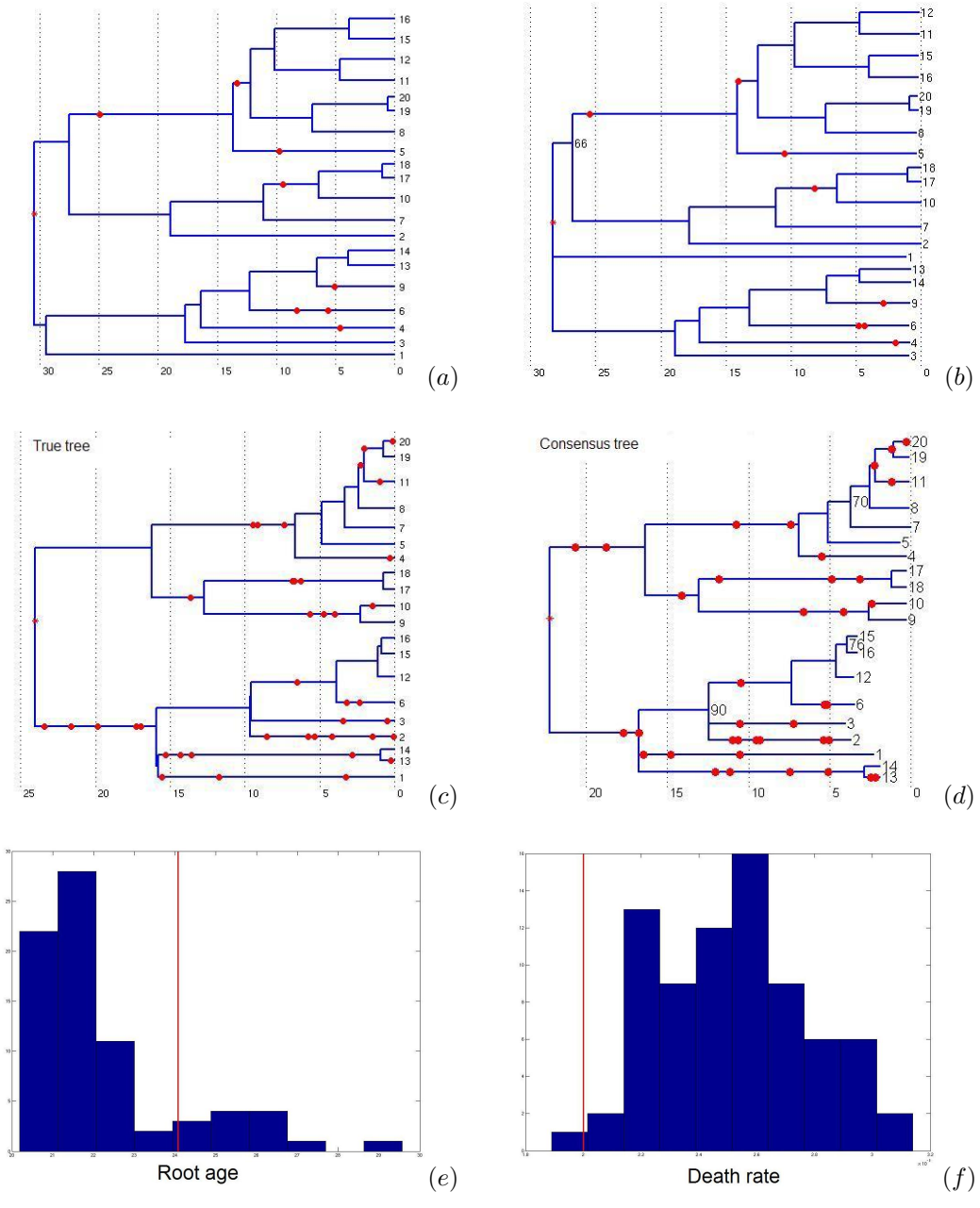


**Fig. 3.** *Influence of borrowing. (a)-(b): low levels of borrowing (b=0.1) have negligible effects on the topology and the parameter estimates. (c)-(d): high levels of borrowing (b=0.5) still allow to reconstruct most of the topology, but the root age and parameter estimates, shown here for* $b = 0.5$*, are biased (e-f).*

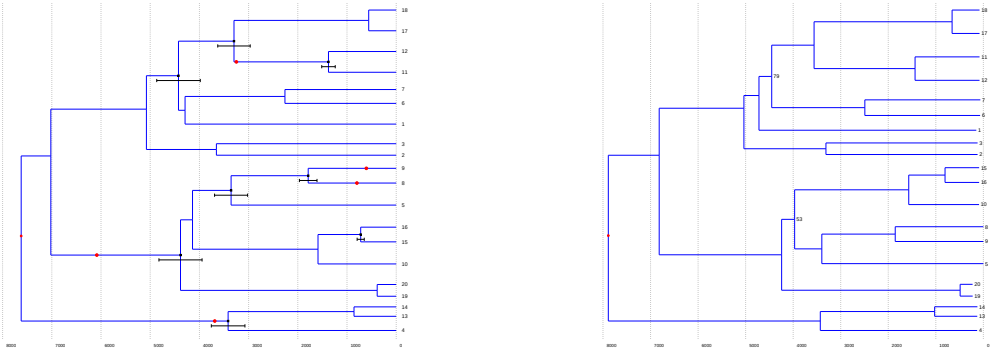

**Fig. 4.** *Influence of the reversibility condition. Left: true tree under which data were simulated, under the condition $\nu = \kappa\lambda/2\mu$; right: reconstructed consensus tree.*

posterior distribution is highly uninformative of $\kappa$ and the death rate $\mu$ is systematically overestimated. However, the root age, which is the parameter of interest, is well reconstructed: for $\nu = \kappa\lambda/2\mu$, the true root age was 7622BP and the 95% HPD is 6932–8584BP; for $\nu = 2\kappa\lambda/\mu$, the true root age is 7664BP and the 95% HPD is 6472–7701BP. In both cases, the 95% HPD covers the true value.

### *6.3. Cross-validation*

## 7. Results

Figure 6 presents a consensus tree for our analysis of the Dyen et al. (1997) data. Nodes with more than 50% support in the posterior are displayed, and nodes with less than 95% support are labeled. Our estimates for the parameters are as follows: $\mu = 2.37 \cdot 10^{-4} \pm 1.08 \cdot 10^{-5}$ deaths/year; $\kappa = 0.121 \pm 0.048$; $\rho = 2.17 \cdot 10^{-4} \pm 5.9 \cdot 10^{-5}$ catastrophes/year (corresponding to smaller but more common catastrophes than for the Ringe et al. (2002) data: about 1 catastrophe every 4600 years, or an average of 31.3 on the tree, with each catastrophe corresponding to 550 years of change).

The analysis of the Dyen et al. (1997) data strongly supports Indo-Iranian as an outgroup. It also supports the Germanic and Italic subfamilies being siblings, with Celtic as the next closest cousin, though the configurations Germanic-Celtic and Italic-Celtic are also present in the posterior (with about 15% posterior probability each). On the other hand, the analysis of the Ringe et al. (2002) data does not support any particular outgroup, and it shows a preference for a Germanic-Celtic subgrouping. Here again, the other configurations also appear in the posterior sample in non-negligible frequencies. In both cases, the position of Albanian is very unclear.

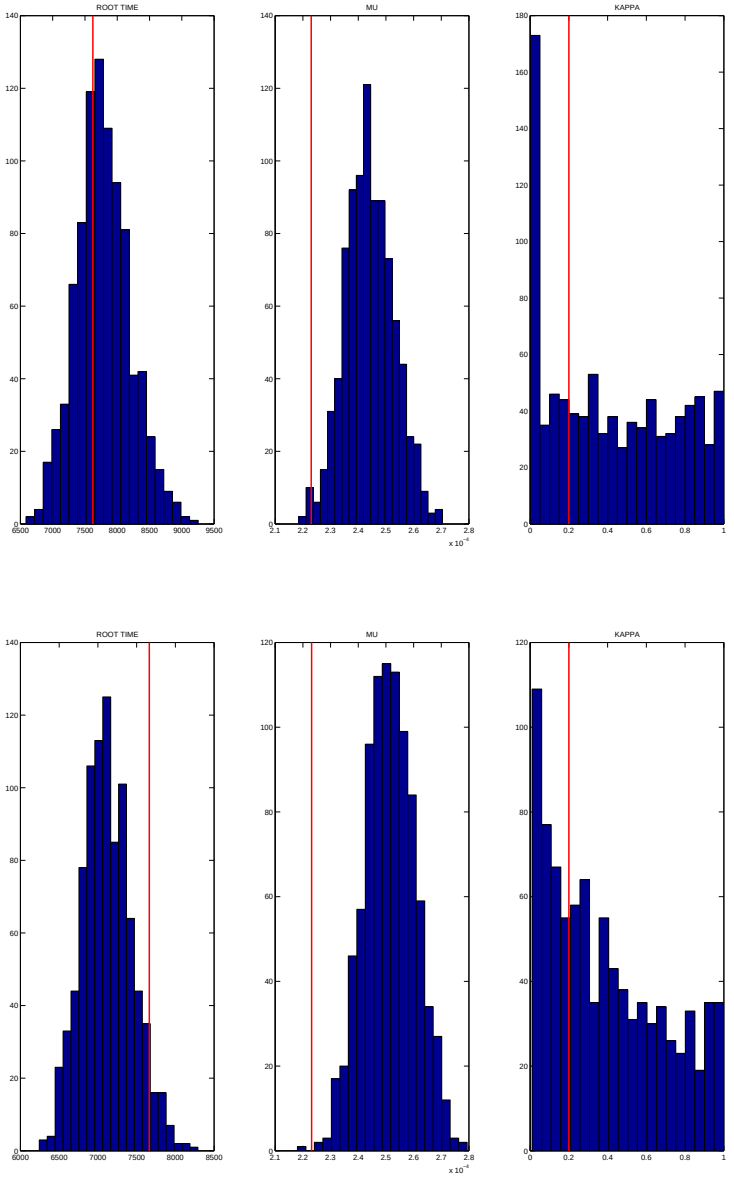


**Fig. 5.** *Influence of the reversibility condition. Top: data simulated under the condition* $\nu = \kappa\lambda/2\mu$*; bottom: data simulated under the condition* $\nu = 2\kappa\lambda/\mu$*. Blue: posterior sample from the reconstruction under the reversibility condition* $\nu = \kappa\lambda/\mu$*; red: true value of the root age, death rate* $\mu$ *and probability of death at a catastrophe* $\kappa$*.*

There is agreement between the analyses for the other topological features; these also correspond to the results linguists have obtained through the comparative methods.

There is rate heterogeneity in a number of positions. Superficially, some of these positions could be expected. For example, French Creoles, Pennsylvania Dutch and the Gypsy language of Greece all went through some rate heterogeneity, which could be linked to the large geographical distance from their parent language. We do not have an explanation for the other catastrophes.

The analysis of the Dyen et al. (1997) data gives a 95% highest posterior density interval for the root age of 7080 – 8350 BP.

## 8. Discussion

## References

Dyen, I., J. Kruskal, and B. Black (1997). FILE IE-DATA1. Raw data available from http://www.ntu.edu.au/education/langs/ielex/IE-DATA1. Binary data available from http://www.psych.auckland.ac.nz/psych/research/RusselsData.htm.

Embleton, S. (1986). *Statistics in historical linguistics*. Brockmeyer.

Gray, R. and Q. Atkinson (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature 426*(6965), 435–439.

Nicholls, G. K. and R. D. Gray (2008). Dated ancestral trees from binary trait data and its application to the diversification of languages. *Journal of the Royal Statistical Society, series B 70*(3), 545–566.

Ringe, D., T. Warnow, and A. Taylor (2002). Indo-European and Computational Cladistics. *Transactions of the Philological Society 100*(1), 59–129.

Ryder, R. (2010). *Phylogenetic Models of Language Diversification*. Ph. D. thesis, University of Oxford.

Ryder, R. and G. Nicholls (2009). Missing data in a stochastic dollo model for binary traits, and its application to the dating of proto-indo-european. *Submitted to the Journal of the Royal Statistical Society, Series C*.

Warnow, T., S. Evans, D. Ringe, and L. Nakhleh (2004). A Stochastic model of language evolution that incorporates homoplasy and borrowing. *Phylogenetic Methods and the Prehistory of Languages*.

**A. Time reversibility**

**B. Recursions for other registration processes**

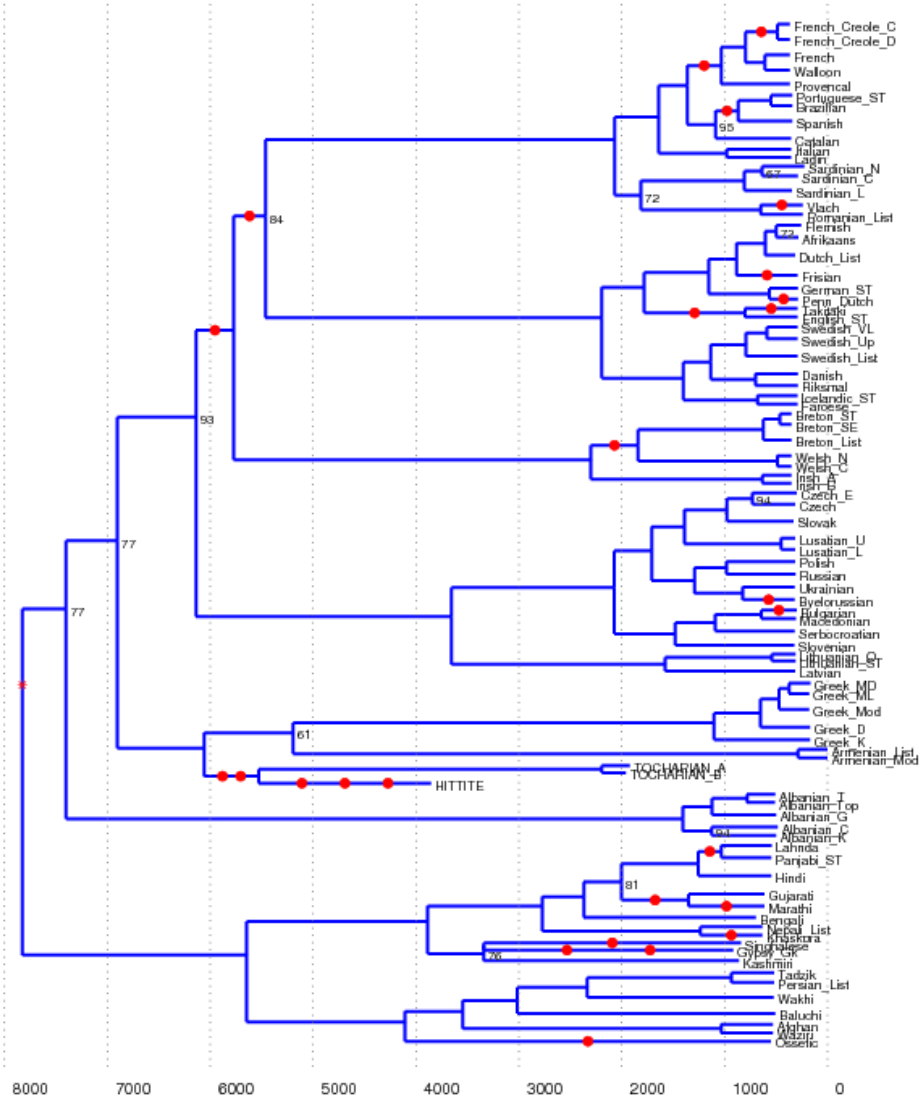


**Fig. 6.** Consensus tree for the Dyen et al. (1997) data set.