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

Robin J. Ryder and Geoff K. Nicholls

University of Oxford, UK

Summary. Nicholls and Gray have described a phylogenetic model for trait data. They used their model to estimate branching times on Indo-European language trees from lexical data. Alekseyenko and co-workers extended the model and gave applications in genetics. We extend the inference to handle data missing at random. When trait data are gathered, traits are thinned in a way that depends on both the trait and the missing data content. Nicholls and Gray treated missing records as absent traits. Hittite has 12% missing trait records. Its age is poorly predicted in their cross-validation. Our prediction is consistent with the historical record. Nicholls and Gray dropped seven languages with too much missing data. We fit all 24 languages in the lexical data of Ringe and co-workers. To model spatiotemporal rate heterogeneity we add a catastrophe process to the model. When a language passes through a catastrophe, many traits change at the same time. We fit the full model in a Bayesian setting, via Markov chain Monte Carlo sampling. We validate our fit by using Bayes factors to test known age constraints. We reject three of 30 historically attested constraints. Our main result is a unimodal posterior distribution for the age of Proto-Indo-European centred at 8400 years before Present with 95% highest posterior density interval equal to 7100–9800 years before Present.

Keywords: Bayesian inference; Dating methods; Markov chain Monte Carlo methods; Missing data; Phylogenetics; Proto-Indo-European; Rate heterogeneity

1. Introduction

The Indo-European languages descend from a common ancestor called Proto-Indo-European. Lexical data show the patterns of relatedness among Indo-European languages. These data are ‘cognacy classes’: a pair of words in the same class descend, through a process of change in sound, from a common ancestor. For example, English sea and German See are cognate to one another, but not to the French mer. Gray and Atkinson (2003) coded data of this kind in a matrix in which rows correspond to languages and columns to distinct cognacy classes, and entries are 0 or 1 as the language possesses or lacks a term in the column class. They analysed these data by using phylogenetic algorithms that are similar to those used for genetic data. Our analysis has the same objectives, but we fit a model that was designed for lexical trait data. We work with data that were compiled by Ringe et al. (2002), recording the distribution of 872 distinct cognacy classes in 24 modern and ancient Indo-European languages. In Section 8, we give estimates for the unknown topology and branching times of the phylogeny of the core vocabulary of these languages.

Address for correspondence: Robin J. Ryder, Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, UK. E-mail: ryder@stats.ox.ac.uk

© 2011 Royal Statistical Society
Nicholls and Gray (2008) analysed the same data by using a closely related stochastic Dollo model for binary trait evolution. However, they could not deal with missing trait records. Missing data arise when we cannot answer the question ‘does language X have a cognate in cognacy class Y?’. Nicholls and Gray (2008) dropped seven languages which had many missing entries, and treated missing trait records in the remainder as absent traits. This is unsatisfactory. However, it is not straightforward to give a model-based integration of missing data for the trait evolution model of Nicholls and Gray (2008). In this paper we integrate the missing trait data, and this technical advance allows us to fit all 24 languages in the original data. The binary trait model of Nicholls and Gray (2008) has been extended by Alekseyenko et al. (2008) to multilbe-level traits and is finding applications in biology. A proper treatment of missing data will be of use in other applications.

We are specifically interested in phylogenetic dating. Because we are working with lexical, and not syntactic, data, it is the age of the branching of the core vocabulary of Proto-Indo-European that we estimate. This is a controversial matter. Workers in historical linguistics have evidence from linguistic palaeontology that the most recent common ancestor of all known Indo-European languages branched no earlier than about 6000–6500 years before the present (BP) (Mallory, 1989). For a recent review of the argument from linguistic palaeontology, and a criticism of phylogenetic dating, see Garrett (2006) and McMahon and McMahon (2005). An alternative hypothesis suggests that the spread began around 8500 BP when the Anatolians mastered farming (Renfrew, 1987) in the early Neolithic. Recent efforts to apply quantitative phylogenetic methods to dating Proto-Indo-European give a time depth around 8000–9500 years BP, supporting the link to farming. For instance, Gray and Atkinson (2003) estimated the age of Proto-Indo-European to lie between 7800 and 9800 years BP; Nicholls and Gray (2008) arrived at similar results. In this work we obtain a unimodal posterior distribution for the age of Proto-Indo-European centred at 8400 years BP with 95% highest posterior density interval equal to 7100–9800 years BP.

One point of view is that the argument from linguistic palaeontology is correct, and the phylogenetic dates are incorrect, owing to rate heterogeneity, or some other model failing. In this respect we advance the search that was started in Nicholls and Gray (2008) for a model misspecification which might explain the 20% difference in the central age estimates from our fitting, and the status quo. The difference is sufficiently large that we are hopeful of finding a single coherent error, if any such error exists. Accounting now for missing data, we can publish a much wider cross-validation test (up from 10 to 30 tested calibrations). Also, we fit a model for explicit rate heterogeneity in time and space. In view of the spatiotemporal homogeneity we find here and in other data, the case for the earlier date seems now fairly strong. The most probable alternative seems to be a step change in the rate of lexical diversification acting in a co-ordinated fashion across the Indo-European territory about 3000–5000 years ago.

In Sections 2 and 3.1 we describe the data and specify a subjective prior for the phylogeny of vocabularies. In Sections 3.2 and 4, we give a generative model for the data and the corresponding likelihood. We include, in Section 4, a recursive algorithm which makes the sum over missing data tractable. In Sections 5 and 6 we give the posterior distribution on tree and parameter space, and briefly describe our Markov chain Monte Carlo (MCMC) sampler. In Section 7, we discuss likely model misspecification scenarios, test for robustness by fitting synthetic data simulated under such conditions and cross-validate our predictions. We fit the model to a data set of Indo-European languages in Section 8. This paper has a supplement (Ryder and Nicholls, 2010) giving an analysis of a second data set, collected by Dyen et al. (1997).

Phylogenetic methods have been used to make inference for tree (Ringe et al., 2002) and network structure (McMahon and McMahon, 2005) in historical linguistics. Warnow et al. (2006)
wrote down a more realistic model for the diversification of vocabulary, accounting for word borrowing between languages (so that their history need not be tree-like) but there has been to date no fitting.

2. Description of the data

The data group words from 328 meaning categories and 24 Indo-European languages into 872 homology classes. The data were collected and coded by Ringe et al. (2002). Meaning categories cover the ‘core’ vocabulary and are assumed relevant to all languages in the study. Two words of closely similar meaning, which are descended from a common ancestor but subject to variation in phonology, are cognate terms. A cognacy class is a homology class of words all belonging to a single meaning category. For example, for the meaning ‘head’, the Italian testa and the French tête belong to the same cognacy class, whereas the English head and the Swedish huvud belong to another cognacy class. An element of a cognacy class is thus a word in a particular language, a sound–meaning pair. These elements are called cognates. The vocabulary of a single language is represented as a set of distinct cognates. In our analysis a cognate has just two properties: its language and its cognacy class. If there are \( N \) distinct cognacy classes in data for \( L \) languages, then the \( a \)th class \( M_a \subseteq \{1, 2, \ldots, L\} \) is a list of the indices of languages which have a cognate in that class. The data are coded as a binary matrix \( D \). A row corresponds to a language and a column to a cognacy class, so \( D_{i,a} = 1 \) if the \( a \)th cognacy class has an instance in the \( i \)th language, and \( D_{i,a} = 0 \) otherwise. See Table 1 for an example. This coding allows a language to have several words for one meaning (such as Old High German stirbit and touwit for ‘he dies’, which are an instance of polymorphism), or no word at all. Missing matrix elements arise because the reconstructed vocabularies of some ancient languages are incomplete. If we cannot answer the question ‘does language \( i \) have a cognate in cognacy class \( a \)?’ then we set \( D_{i,a} = ? \).

| (a)                              | stierþf | stirbit, touwit | miriěte | umiřetũ | morĭtur |
|----------------------------------|---------|----------------|---------|---------|---------|
| Old English                      |         |                |         |         |         |
| Old High German                  |         |                |         |         |         |
| Avestan                          |         |                |         |         |         |
| Old Church Slavonic              |         |                |         |         |         |
| Latin                            |         |                |         |         |         |
| Oscan                            |         |                |         |         |         |

| (b)                              |         |         |         |         |         |
|----------------------------------|---------|---------|---------|---------|---------|
| Old English                      | 1       | 0       |         |         |         |
| Old High German                  | 1       | 1       | 0       |         |         |
| Avestan                          | 0       | 0       | 1       |         |         |
| Old Church Slavonic              | 0       | 0       | 1       |         |         |
| Latin                            | 0       | 0       | 1       |         |         |
| Oscan                            | ?       | ?       | ?       |         |         |

†(a) The word ‘he dies’ in six Indo-European languages and (b) the coding of these data as a binary matrix with ?s for missing data. In the notation of Section 2, the first column in (b) is a cognacy class \( M_1 \subseteq \Omega_1 \) with \( \Omega_1 = \{ \{ \text{Old English, Old High German} \}, \{ \text{Old English, Old High German, Oscan} \} \} \).
We need notation for both matrix and set representations with missing data. Denote by $B_a$ column $a$ of $L \times N$ matrix $B$. For $a = 1, 2, \ldots, N$ let $D_a$ be the set of all column vectors $d^*$ allowed by the data $D_a$ in column $a$ of $D$,

$$D_a = \{d^* \in \{0, 1\}^L : D_{i,a} \in \{0, 1\} \Rightarrow d^*_i = D_{i,a}, i = 1, 2, \ldots, L\}.$$  

For $d^* \in D_a$ let $m(d^*) = \{i : d^*_i = 1\}$. Denote by $\Omega_a$ the set of cognacy classes that are consistent with the data $D_a$, so that

$$\Omega_a = \{\omega \subseteq \{1, 2, \ldots, L\} : \omega = m(d^*), d^* \in D_a\}.$$  

The data $D$ are then equivalently $\Omega = (\Omega_1, \Omega_2, \ldots, \Omega_N)$. The $\Omega_a$-notation generalizes the $M_a$-notation to handle missing data. It is illustrated in Table 1.

Ringe et al. (2002) listed 24 mostly ancient languages. For 11 of these languages (Latin, Modern Latvian, Old Norse...), all the matrix entries are recorded. For the rest, the proportion of missing entries varies between 1% (for Old Irish) and 91% (for Lycian, which is an ancient language of Anatolia). Note that data are usually missing in small blocks corresponding to the cognacy classes for a given meaning category, as in Table 1. We do not model this aspect of the missing data. This is related to the model error that Nicholls and Gray (2008) called ‘the empty field approximation’, under which cognacy classes in the same meaning category are assumed to evolve independently.

For these Indo-European phylogenies, the topology of some subtrees is known from historical records. We have lower and upper bounds for the age of the root node of some subtrees. For example, the Slavic languages are known to form a subtree, and the most recent common ancestor of the Slavic languages in the data is known to be at least 1300 years old. In this way internal nodes of the phylogeny are constrained. We have age bounds for leaf nodes as well, since we are told when the vocabulary of the ancient languages in these data was in use. We combine these calibration constraints with the cognacy class data in Section 3.1. Jumping ahead to our results, Fig. 6 in Section 8 is a sample from the posterior distribution that we find for phylogenies. Calibration constraints are represented by the black bars across nodes in this tree.

3. Models

We specify a subjective prior for phylogenies, representing a state of knowledge that is of interest to us. The model that we give for the diversification of vocabulary in Section 3.2 is, in contrast, generative. We model the diversification of vocabulary as a branching process of sets of cognates, evolving on the phylogeny. Leaves are labelled by languages and a branch represents the ancestral lineage of a vocabulary.

3.1. Prior distribution on trees

The material in this subsection follows Nicholls and Gray (2008). We consider a rooted binary tree $g$ with $2L$ nodes: $L$ leaves, $L - 2$ internal nodes, a root node $r = 2L - 1$ and an Adam node $A = 2L$, which is linked to $r$ by an edge of infinite length. Each node $i = 1, 2, \ldots, 2L$ is assigned an age $t_i$ and $t = (t_1, t_2, \ldots, t_A)$; the units of age are years before the present; for the Adam node, $t_A = \infty$. The edge between parent node $j$ and child node $i$ is a directed branch $(i, j)$ of the phylogeny, with the orderings $i < j$ and $t_i < t_j$. Let $E$ be the set of all edges, including the edge $(r, A)$, let $V$ be the set of all nodes and $V_L = \{1, 2, \ldots, L\}$ the set of all leaf nodes. Our base tree space $\Gamma$ is the set of all rooted directed binary trees $g = (E, V, t)$, with distinguishable leaves and, for $i = 1, 2, \ldots, 2L$, node ages $t_i > 0$ assigned so that the directed path from a leaf $i \in V_L$ to node $A$
passes through nodes of strictly increasing age. With this numbering convention, \((E, V)\) is called
the ordered history of a rooted directed binary tree.

We allow for \(C\) calibration constraints on tree topology and selected node ages. These are
described at the end of Section 2. Each such constraint \(c\) allows trees in just some subspace \(\Gamma^{(c)} \subset \Gamma\) of tree space. We add to these constraints an upper bound on the root time at some age \(T > 0\), and we set \(\Gamma^{(0)} = \{(E, V, t) \in \Gamma : t_r \leq T\}\). The space of calibrated phylogenies with
constraints is then

\[
\Gamma^C = \bigcap_{c=0}^{C} \Gamma^{(c)}. \tag{3.1}
\]

The root age is a sensitive statistic in this inference. A prior distribution on trees, with the
property that the marginal distribution of the root age is uniform over a fixed prior range
\(t_L \leq t_r \leq T\), is of interest. For node \(i \in V\), let \(t^+_i = \sup_{g \in \Gamma^c} (t_i)\) and \(t^-_i = \inf_{g \in \Gamma^c} (t_i)\) be the greatest
and least admissible ages for node \(i\), and let \(S = \{i \in V : t^+_i = T\}\), so that \(S\) is the set of nodes
having ages not bounded above by a calibration (there are 12 such nodes in Fig. 6 in Section 8).
Nicholls and Gray (2008) showed that, before calibration (when \(C = 0\)) and for tree spaces in
which the leaves have fixed equal ages, the prior probability distribution with density

\[
f(g|T) \propto \prod_{i \in S} (t_r - t^-_i)^{-1}
\]

gives a marginal density for \(t_r\) which is exactly uniform in \(t_L < t_r < T\). Nicholls and Gray (2008)
did not comment on the distribution that is determined by \(f\) over tree topologies. The prior
\(f\) has in this case \((C = 0)\) a uniform marginal distribution on ordered histories that is exactly
equal to the corresponding distribution for the Yule model (Yule, 1925). This is not uniform
but favours balanced leaf-labelled topologies, i.e. topologies in which sibling subtrees have the
same number of leaves. For \(C > 0\), and on tree spaces in which leaf ages are constrained by
prior knowledge to lie in an interval only and without the catastrophes that we introduce in the
next section, Nicholls and Gray (2008) argued from simulation studies that this prior gives a
reasonably flat marginal distribution for \(t_r\) if in addition \(T \gg \max_{i \in V \setminus S} (t^+_i)\) (the greatest upper
bound among the calibration constraints is not too close to the upper bound on the root age).
We give a sample from the prior in the supplementary material; the prior does not represent
any reasonable prior belief before about 4500 BP, but this is immaterial as the likelihood rules
these values out (which is an instance of an application of the principle of sufficient reason).

### 3.2. Diversification of cognacy classes

In this subsection we extend the stochastic Dollo model of Nicholls and Gray (2008) to incor-
porate rate heterogeneity in time and space, via a catastrophe process.

Consider the evolution of cognates down the ancestral lineage of the vocabulary of a single
language. An example is given in Fig. 1. A new cognacy class is born when its first cognate is
born. This new word is not cognate with other words in the modelled process. Loan words from
outside the core meaning categories of any language in the study, or from a language outside
the study, may be good for word birth without violating this condition. A cognate dies in a
particular vocabulary when it ceases to be used for the meaning that is common to its cognacy
class: its meaning may change, or it may fall out of use.

Cognates evolving in a single language (i.e. down a single branch of a language phylogeny) are
born independently at rate \(\lambda\), die independently at \(per\ capita\) rate \(\mu\) and are subject to point-like
catastrophes, which they encounter at rate \(\rho\) along a branch. At a catastrophe, each cognate dies
independently with probability \(\kappa\), and a Poisson number of cognates with mean \(\nu\) are born. At
a branching event of the phylogeny, the set of cognates representing the branching vocabulary is copied into each of the daughter languages. See Fig. 1.

The process that we have described is not reversible, and this greatly complicates the analysis. It seems acceptable, from a data modelling perspective, to impose the condition $\nu = \kappa \lambda / \mu$, which is necessary and sufficient for reversibility (see the supplementary material for a proof). Under this condition, adding a catastrophe to an edge is equivalent to lengthening that edge by $T_C(\kappa, \mu) = -\log(1 - \kappa) / \mu$ years: given a branch $\langle i, j \rangle$ of length $t_j - t_i$ with $k_i$ catastrophes, the distribution of the data at $(t_i, i)$ (conditional on the data at $(t_j, j)$) is equal to the distribution that would arise at the end of a branch of length $t_j - t_i + k_i T_C$ but with no catastrophes. This follows because the number of cognates that are generated by the anagenic part of the process in an interval of length $T_C$ is Poisson distributed with mean $(\lambda/\mu) \{1 - \exp(-\mu T_C)\}$ equal to $\kappa \lambda / \mu$, and the probability that a cognate entering an interval of length $T_C$ dies during that interval is $1 - \exp(-\mu T_C)$, which equals $\kappa$.

Because a catastrophe simply extends its edge by a block of virtual time, the likelihood depends only on the number of catastrophes on an edge, and not on their location in time. Let $k_i$ be the number of catastrophes on edge $\langle i, j \rangle$, and $k = (k_1, \ldots, k_{2L-2})$ be the catastrophe state vector. We record no catastrophes on the $\langle R, A \rangle$ edge (its length is already infinite). The tree $g = (V, E, t, k)$ is specified by its topology, node ages and catastrophe state. Calibrated tree space extended for catastrophes is
\[ \Gamma_k^C = \{(V, E, t, k) : (V, E, t) \in \Gamma^C, k \in \mathbb{N}_{0}^{2L-2}\}. \]

We drop the catastrophe process from the calculation in Section 4. It is straightforward to restore it, and we do this in the expression for the posterior distribution in Section 5.

### 3.3. Registration process

Let \( D^* \) denote a notional full random binary data matrix, representing the outcome of the diversification process of Section 3.2. The number of columns in \( D^* \) is random and equal to \( N^* \). For the realization depicted in Fig. 1, \( D^* = D \) with \( D^* \) displayed in Table 2. Note that \( D^* \) has a column for each cognacy class present at the root node, or born below it, whether or not the cognacy class has any cognates that are represented in any leaf languages.

We call the random mapping of the unknown full representation \( D^* \) to the registered data \( D \), which is a random matrix with \( N \) columns, the registration process. There are two stages to this process: the masking of matrix elements of the fully realized data \( D^* \) with ?'s to form an intermediate data matrix \( \tilde{D} \), and the selection of columns from \( \tilde{D} \) to determine the realized data \( D \).

Let \( I^* \) be a random \( L \times N^* \) indicator matrix of independent Bernoulli random variables for observed elements. The 0s of \( I^* \) mark matrix entries in \( D^* \) which will be unobservable. For \( i = 1, 2, \ldots, L \) and given \( a = 1, 2, \ldots, N^* \), let \( \xi_i = \Pr(I^*_i = 1) \). The probability \( \xi_i \) that we can answer the question ‘does language \( i \) have a word in the \( a \)th cognacy class?’ is assumed to be a function of the language index \( i \) only. If we obtain an answer, it is assumed correct. Let \( \xi = (\xi_1, \ldots, \xi_L) \) and denote by \( I^* \) a realization of \( I^* \). Denote by \( \tilde{D} = \tilde{D}(D^*, I^*) \) the masked version of the full random data matrix: if \( I^*_i, a = 1 \) then \( \tilde{D}_{i, a} = D_{i, a} \) and if \( I^*_i, a = 0 \) then \( \tilde{D}_{i, a} = ? \).

Matrix columns may be missing also, so that \( N \leq N^* \). We have missing columns even when there are no missing data. The matrix \( \tilde{D} \) in Table 2 includes some columns with only 0s and ?, corresponding to cognacy classes which existed in the past but are observed in none of the leaf languages from which the data were compiled. These cognacy classes are not included in the registered data. Denote by \( R \) the registration rule \( D = R(D) \) mapping the full data to registered data.

Rule \( R \) may thin additional column types. Let \( Y \) and \( Q \) be functions of the columns of \( D^* \) and \( I^* \) counting the visible 1s and ?'s respectively,

### Table 2. Registration of the vocabulary realized in Fig. 1 supposing a masking matrix \( I^* \)

| \( (D^*) \) | \( (I^*) \) | \( (D^-) \) | \( (I) \) | \( (D) \) |
|---|---|---|---|---|
| 1000000000000 | 1111111111111 | 1000000000000 | 1 | 111 | 1 | 11 | 1 | 000 | 0 | 00 |
| 10100000000001 | 0101111111111 | ?000000000001 | 0 | 111 | 1 | 11 | 0 | 000 | 0 | 01 |
| 00000000011000 | 1011001111110 | ?0?00?00?1?0 | 1 | 100 | 1 | 10 | 0 | 0? | 1 | 0? |
| 0000010000000 | 1111111111111 | 0000100000000 | 1 | 111 | 1 | 11 | 0 | 100 | 0 | 00 |
| 0000001000000 | 1110111111111 | 0000700000000 | 1 | 011 | 1 | 11 | 0 | 00 | 0 | 00 |
| 1000011000010 | 1011111111111 | ?000110000010 | 1 | 111 | 1 | 11 | 1 | 011 | 0 | 10 |
| 1000010000010 | 1111111111111 | 1000010000010 | 1 | 111 | 1 | 11 | 1 | 010 | 0 | 10 |
| 1000000000010 | 1111111111100 | 1000000000000 | 1 | 111 | 1 | 00 | 0 | 100 | 0 | ?? |

\( \dagger \) \( D^* \) is the unobserved full data with a column for each cognacy class and a row for each language of Fig. 1 (thus cognate 1 is present in the first, second, sixth, seventh and eighth rows); 0s in the masking matrix \( I^* \) indicate missing matrix elements. Some cognacy classes are then thinned (in this example, the registration rule keeps cognacy classes with instances displayed in one or more language) to give the registered data \( D \).
Given \( a = 1, 2, \ldots, N^* \), let \( Y_a = Y(D^*_a, I^*_a) \) and \( Q_a = Q(I^*_a) \).

In Appendix A we give an efficient algorithm for computing the likelihood for rules that is formed by compounding the following elementary thinning operations:

(a) \( R_1(\tilde{D}) = (\tilde{D}_a : Y_a > 0) \) (discard classes with no instances at the leaves);
(b) \( R_2(\tilde{D}) = (\tilde{D}_a : Y_a > 1) \) (discard classes—singletons—observed at a single leaf);
(c) \( R_3(\tilde{D}) = (\tilde{D}_a : Y_a < L) \) (discard classes which are observed at all leaves);
(d) \( R_4(\tilde{D}) = (\tilde{D}_a : Y_a < L - 1) \) (discard classes which are observed at all leaves or at all leaves but one);
(e) \( R_5(\tilde{D}) = (\tilde{D}_a : Y_a + Q_a < L) \) (discard classes which are potentially present at all leaves);
(f) \( R_6(\tilde{D}) = (\tilde{D}_a : Y_a + Q_a < L - 1) \) (discard classes which are potentially present at all leaves or at all leaves but one).

The registration rule is chosen in advance by the linguists collecting the data. We assume that the chosen rule includes condition (a). The rule \( D = R(\tilde{D}) \) with

\[
R(\tilde{D}) = R_4 \circ R_2(\tilde{D})
\]

collects ‘parsimony informative’ cognacy classes. Ronquist et al. (2005) gave the likelihood for the finite sites trait evolution model of Lewis (2001) for registration rules like (a)–(f). In the example in Table 2, and in Sections 7.2 and 8, we fit data registered with \( R(\tilde{D}) = R_1(\tilde{D}) \).

The selection of columns is something that we have in general no control over: the column material for the Dyen selection rule simply describes what happened at registration. Results in the supplementary material for the Dyen et al. (1997) data use the rule \( R(\tilde{D}) = R_2(\tilde{D}) \), since singleton columns were not included in those data. However, certain column types may include data which are difficult to model well, and so we may choose to make further thinning by using the other rules.

Recursions for the other rules are given in Appendix A.

Column indices \( a = 1, 2, \ldots, N^* \) are exchangeable. It is convenient to renumber the columns of \( D^*, I^* \) and \( \tilde{D} \) after registration, so that \( \tilde{D}_a = D_a \) and \( I^*_a = I_a \) for \( a = 1, 2, \ldots, N \). The information that is needed to evaluate \( Y_a \) and \( Q_a \) is available in the column \( D_a \) and set \( \Omega_a \) representations. We write \( Y(D_a) = Y(\Omega_a) = Y_a \) and \( Q(D_a) = Q(\Omega_a) = Q_a \).

### 3.4. Point process of births for registered cognacy classes

Fix a catastrophe-free phylogeny \( g \in \Gamma_K \), with \( k = (0, 0, \ldots, 0) \), and let an edge \( \langle i, j \rangle \) and a time \( \tau \in [t_i, t_j) \) be given. Denote by \( [g] \) the set of all points \( (\tau, i) \) on the phylogeny, including points \( (\tau, r) \) with \( \tau \geq t_r \) in the edge \( \langle r, A \rangle \). The locations \( z_D = \{z_1, z_2, \ldots, z_N\} \) of the birth events of the \( N \) registered cognacy classes are a realization of an inhomogeneous Poisson point process \( Z_D \) on \([g]\). Let \( Z \in [g] \) be the birth location of a generic (and possibly unregistered) cognacy class \( M \subseteq \{1, 2, \ldots, L\} \), corresponding to a column of \( \tilde{D} \) with \( Y \) observed 1s and \( Q \) ?s, and let \( \tilde{E}_Z \) be the event that this class generates a column of the registered data. For the \( a \)th cognacy class \( M_a \), born at \( Z_a \), this event is \( \tilde{E}_{Z_a} = \{D_a = R(\tilde{D}_a)\} \) since \( R(\tilde{D}_a) \) is empty for \( \tilde{D}_a \) a column dropped at registration.

Cognacy classes are born at constant rate on the branches of \( g \) but are thinned by registration. However, conditional on the birth location \( Z_a = z_a \), our modelling assumes that the outcome
\( \mathcal{E}_Z \) for the \( a \)th cognacy class is decided independently of all events in all other cognacy classes. The point process \( Z_D \) of birth locations of registered cognacy classes has intensity
\[
\tilde{\lambda}(z) = \lambda \Pr(\mathcal{E}_Z | g, \mu, \lambda, \xi, Z = z)
\]
at \( z \in [g] \) and probability density
\[
f_{Z_D}(z_D) = \frac{1}{N!} \exp\{-\Lambda([g])\} \prod_{a=1}^{N} \tilde{\lambda}(z_a)
\]
with respect to the element of volume \( dz_D = dz_1 dz_2 \ldots dz_N \) in \([g]^N\), where
\[
\Lambda([g]) = \int_{[g]} \tilde{\lambda}(z) \, dz
= \sum_{(i,j) \in E} \int_{t_i}^{t_j} \tilde{\lambda}\{(\tau, i)\} \, d\tau.
\]
The number \( N \) of registered cognacy classes is \( N \sim \text{Poisson}\{\Lambda([g])\}\).

4. Likelihood calculations

We give the likelihood for \( g, \mu, \lambda, \kappa, \rho \) and \( \xi \) given the data, along with an efficient algorithm to compute the sum over all missing data. The catastrophe process is left out, and reincorporated in the next section.

Since we only ever see registered data, the likelihood for \( g, \lambda, \mu \) and \( \xi \) is the probability
\[
P\{D = D | g, \mu, \lambda, \xi, D = R(\tilde{D})\} = \int f_{Z_D}(z_D) \, P\{D = D | g, \mu, \lambda, \xi, Z_D = z_D, D = R(\tilde{D})\} \, dz_D
\]
\[
= \frac{\exp\{-\Lambda([g])\}}{N!} \prod_{a=1}^{N} \lambda \int_{[g]} P(\mathcal{E}_Z | g, \mu, \lambda, \xi, Z_a = z_a) \, dz_a
\times P(D_a = D_a | g, \mu, \lambda, \xi, Z_a = z_a, \mathcal{E}_Z) \, dz_a
= \frac{\exp\{-\Lambda([g])\}}{N!} \prod_{a=1}^{N} \lambda \int_{[g]} P(D_a = D_a | g, \mu, \lambda, Z_a = z_a) \, dz_a.
\]
The last line follows because \( P(\mathcal{E}_Z | D_a = D_a, \ldots, Z_a = z_a) = 1 \) for \( D_a \) a column of registered data: if the outcome of the birth at \( z_a \) was the registerable data \( D_a \) then the event \( \mathcal{E}_Z \) certainly occurs. The likelihood depends on the awkward condition \( D = R(\tilde{D}) \) through the mean number \( \lambda([g]) \) of registered cognacy classes only, whereas \( P(D_a = D_a | g, \mu, \lambda, Z_a = z_a) \) is the probability to realize the data vector \( D_a \) in the unconditioned diversification–missing element process. The calculation has so far extended Nicholls and Gray (2008) to give the likelihood for a greater variety of column thinning rules. We now add the missing element component of the registration process.

We sum over possible values of the missing matrix elements in the registered data. Since \( P(D_a = D_a | g, \mu, \lambda, \xi, Z_a = z_a) \) is not conditioned on the requirement that the column \( D_a \) becomes registered, the entries of the corresponding column \( I_a \) are determined by the unconditioned Bernoulli process, and we have
\[
P(D_a = D_a | g, \mu, \xi, Z_a = z_a) = \sum_{d^* \in D_a} P(I^*_a, D_a^* = d^* | g, \mu, \xi, Z_a = z_a)
= \prod_{i=1}^{L} \xi_i^{h_i} (1 - \xi_i)^{1-h_i} \sum_{d^* \in D_a} P(D_a^* = d^* | g, \mu, \xi, Z_a = z_a).
\]

The likelihood is
\[
P(D = | g, \mu, \lambda, \xi, D = R(\hat{D}))
= \frac{\exp(-\Lambda([g]))}{N!} \prod_{a=1}^{N} \left\{ \prod_{i=1}^{L} \xi_i^{h_i} (1 - \xi_i)^{1-h_i} \right\} \lambda \int_{[g]} \sum_{\omega \in \Omega_a} P(M = \omega | g, \mu, \xi, Z = z_a) \, dz_a,
\]
where we have switched now from summing \(d^* \in D_a\) to the equivalent set representation \(\omega \in \Omega_a\).

For the two integrated quantities in equation (1) we have tractable recursive formulae. We are using a pruning procedure akin to Felsenstein (1981). We begin with \(\Lambda([g])\).

We assume that the registration rule includes at least condition (a). It follows that a cognacy class that is born at \(Z = (\tau, i)\) in \([g]\) must survive down to the node below, at \(Z = (t_i, i)\), to be registered, and so
\[
P(\mathcal{E}_Z | Z = (\tau, i), g, \mu, \xi) = P(\mathcal{E}_Z | Z = (t_i, i), g, \mu, \xi) \exp\{-\mu(\tau - t_i)\}.
\]

We can substitute this into the expression for \(\Lambda([g])\), and integrate, to obtain
\[
\Lambda([g]) = \mu \sum_{(i,j) \in E} P(\mathcal{E}_Z | Z = (t_i, i), g, \mu, \xi) \{1 - \exp\{-\mu(t_j - t_i)\}\}.
\]

Given a node \(i\), let \(V_L^{(i)}\) be the set of leaf nodes descended from \(i\), including \(i\) itself if \(i\) is a leaf.

Let \(s_i = \text{card}(V_L^{(i)})\). Denote by \(u_i^{(n)} = P\{Y = n | Z = (t_i, i), g, \mu, \xi\}\) and \(v_i^{(n)} = P\{Y + Q = n | Z = (t_i, i), g, \mu, \xi\}\). We can compute \(\Lambda\) for rules made up of combinations of condition (a) with any combination of conditions (b)–(f), from \(u_i^{(0)}, u_i^{(1)}, u_i^{(s_i)}, u_i^{(s_i-1)}, v_i^{(s_i)}, v_i^{(s_i-1)}\) and \(v_i^{(s_i)}\). For example,
\[
P(\mathcal{E}_Z | Z = (t_i, i), g, \mu, \xi) = \begin{cases} 1 - u_i^{(0)} & \text{if } R = R_1, \\ 1 - u_i^{(0)} - u_i^{(1)} & \text{if } R = R_2, \\ 1 - u_i^{(0)} - u_i^{(1)} - u_i^{(L-1)} - u_i^{(L)} & \text{if } R = R_3 \circ R_2. \end{cases}
\]

\(u_i^{(n)} = 0\) unless \(s_i \geq n\), so for example \(u_i^{(L)}\) is non-zero at \(i\) the root node only. Since our main application is for data registered under condition (a), we give, in the body of this paper, recursions for \(u_i^{(0)}\) and \(v_i^{(0)}\) only. See Appendix A for the recursions that are needed to evaluate the likelihood under rules involving conditions (b)–(f).

For nodes \(i\) and \(j\), let \(\delta_{i,j} = \exp\{-\mu(t_j - t_i)\}\). Consider a pair of edges \((c_1, i), (c_2, i)\) in \(E\).
\[
u_i^{(0)} = (1 - \delta_{i,c_1} + \delta_{i,c_1} u_{c_1}^{(0)})(1 - \delta_{i,c_2} + \delta_{i,c_2} u_{c_2}^{(0)}),
\]
\[
u_i^{(0)} = \left\{ \delta_{i,c_1} v_{c_1}^{(0)} + (1 - \delta_{i,c_1}) \prod_{j \in V_{c_1}^i} \xi_j \right\} \left\{ \delta_{i,c_2} v_{c_2}^{(0)} + (1 - \delta_{i,c_2}) \prod_{j \in V_{c_2}^i} \xi_j \right\}.
\]

The recursion is evaluated from the leaves \(i \in V_L\), at which
\[
u_i^{(0)} = 1 - \xi_i,
\]
\[
u_i^{(0)} = 0.
\]

We now give the equivalent recursions for \(\lambda \int_{[g]} \sum_{\omega \in \Omega_a} P(M = \omega_a | Z = z_a, g, \mu) \, dz_a\). Consider the set \(m_a = \cap_{\omega \in \Omega_a} \omega\) of leaves that are known to have a cognate in the \(a\)th registered cognacy.
class. Let $E_a$ be the set of branches on the path from the most recent common ancestor of the leaves in $m_a$ up to the Adam node $A$ above the root. Cognacy class $a$ must have been born on an edge in $E_a$. For $a = 1, 2, \ldots, N$, class $M_a$ is non-empty and we can shift the birth location to the node below and convert the integral to a sum,

$$
\lambda \int_{[g]} \sum_{\omega \in \Omega_a} P(M = \omega | Z = z_a, g, \mu) \, dz_a = \lambda \sum_{\mu} \sum_{(i, j) \in E_a} \sum_{\omega \in \Omega_a} P\{M = \omega | Z = (t_i, i), g, \mu\} \{1 - \delta_{i, j}\).
$$

For each $a = 1, 2, \ldots, N$ and $\omega \in \Omega_a$, let $\omega^{(i)} = \omega \cap V_L^{(i)}$ and

$$
\Omega_a^{(i)} = \{\omega^{(i)} : \omega^{(i)} = \omega \cap V_L^{(i)}, \omega \in \Omega_a\}
$$

denote the set of all subsets $\omega^{(i)}$ of the leaves $V_L^{(i)}$ which are cognacy classes consistent with the data that are available for those leaves. Consider two child branches $\langle c_1, i \rangle$ and $\langle c_2, i \rangle$ at node $i$. Since $\Omega_a = \Omega_a^{(c_1)} \times \Omega_a^{(c_2)}$, and events are independent along the two branches,

$$
\sum_{\omega \in \Omega_a} P\{M = \omega | Z = (t_i, i), g, \mu, \xi\} = \sum_{\omega^{(c_1)} \in \Omega_a^{(c_1)}} P\{M = \omega^{(c_1)} | Z = (t_i, c_1), g, \mu\} \\
\times \sum_{\omega^{(c_2)} \in \Omega_a^{(c_2)}} P\{M = \omega^{(c_2)} | Z = (t_i, c_2), g, \mu\}.
$$

Having moved the birth event at $(t_i, i)$ to $(t_i, c_1)$ and $(t_i, c_2)$ (off the node and onto its child edges) we now move the birth event at $(t_i, c)$ to $(t_c, c)$ (down an edge) as follows:

$$
\sum_{\omega \in \Omega_a^{(c)}} P\{M = \omega | Z = (t_i, c), g, \mu\} \begin{cases} \\
\delta_{i,c} \sum_{\omega \in \Omega_a^{(c)}} P\{M = \omega | Z = (t_c, c), g, \mu\} & \text{if } Y(\Omega_a^{(c)}) \geq 1, \\
1 - \delta_{i,c} + \delta_{i,c} \sum_{\omega \in \Omega_a^{(c)}} P\{M = \omega | Z = (t_c, c), g, \mu\} & \text{if } Y(\Omega_a^{(c)}) = 0 \text{ and } Q(\Omega_a^{(c)}) \geq 1, \\
1 - \delta_{i,c} + \delta_{i,c} \delta_{i,c}^{(0)} & \text{if } Y(\Omega_a^{(c)}) + Q(\Omega_a^{(c)}) = 0 \text{ (i.e. } \Omega_a^{(c)} = \emptyset). \end{cases}
$$

The recursion is evaluated from the leaves. If $c$ is a leaf, then

$$
\sum_{\omega \in \Omega_a^{(c)}} P\{M = \omega | Z = (t_c, c), g, \mu\} = \begin{cases} \\
1 & \text{if } \Omega_a^{(c)} = \{\{c\}, \emptyset\} \text{ or } \{\{c\}\} \text{ (i.e. } D_{c,a} \in \{2, 1\}), \\
0 & \text{if } \Omega_a^{(c)} = \emptyset \text{ (i.e. } D_{c,a} = 0). \end{cases}
$$

To restore catastrophes to this calculation, and given $g \in \Gamma_k$, with $k_i$ catastrophes on edge $(i, j) \in E$, replace $t_j - t_i$ with $t_j - t_i + k_i T_C(\kappa, \mu)$ and replace all conditions on $\mu$ with conditions on $\mu$ and $\kappa$ throughout.

5. Posterior distribution

Our prior on the birth rate $\lambda$, death rate $\mu$ and catastrophe rate $\rho$ is $p(\lambda, \mu, \rho) \propto 1/\lambda \mu \rho$ and we take a uniform prior over $[0, 1]$ for the probability of death at a catastrophe $\kappa$ and each missing data parameter $\xi_i$.

Substituting using equations (2) and (3) into equation (1) and multiplying by the prior $f_G(g|T) p(\lambda, \mu, \rho)$, we obtain the posterior distribution
\[ p(g, \mu, \lambda, \kappa, \rho, \xi | D = D) = \frac{1}{N!} \left( \frac{\lambda}{\mu} \right)^N \exp \left( -\frac{\lambda}{\mu} \sum_{(i,j) \in E} P \{ E | Z = (t_i, i), g, \mu, \kappa, \xi \} [1 - \exp \{ -\mu(t_j - t_i + k_i T_C) \}] \right) \times \prod_{a=1}^{N} \left( \sum_{(i,j) \in E_a, \omega \in \Omega_a} P \{ M = \omega | Z = (t_i, i), g, \mu, \kappa \} [1 - \exp \{ -\mu(t_j - t_i + k_i T_C) \}] \right) \times \frac{1}{\mu \lambda \rho} f_G(g | T) \exp \left( -\rho | g | (\rho | g |)^{k_T} \right) \sum_{t=1}^{L} (1 - \xi_t) \alpha_{t_i} \xi_{i}^{N - Q_t} \]

for parameters \( \mu, \lambda, \kappa, \rho > 0, \ 0 < \xi_t < 1 \) and trees \( g \in \Gamma^C_k \).

The posterior is improper without bounds on \( \rho \) since \( k_T = 0 \) is allowed. We place very conservative bounds on \( \rho \) and \( \mu \). Results are not sensitive to this choice; see the supplementary material for details.

### 6. Markov chain Monte Carlo sampling

We use MCMC methods to sample the posterior distribution and to estimate summary statistics. If the prior on the cognacy class birth rate parameter \( \lambda \) has the conjugate form \( \lambda^u \exp(-v \lambda) \) then the conditional distribution of \( \lambda \) in the posterior distribution above has the form \( \lambda^{N+u} \exp\{ -\lambda (X + v) \} \). We took the improper prior \( u = -1 \) and \( v = 0 \) for \( \lambda \) and integrated. The MCMC state is then \( x = (E, V, t, k, \mu, \kappa, \rho, \xi) \) and the target distribution \( p(x | D) \) is the density that is obtained by integrating the density in equation (8) over \( \lambda \).

The MCMC sampler that was described in Nicholls and Gray (2008) has state \( x = (E, V, t, k, \mu) \). We add to the \( (E, V, t) \) - and \( \mu \)-updates of Nicholls and Gray (2008) further MCMC updates acting on the catastrophe vector \( k = (k_1, \ldots, k_{L-2}) \), on the catastrophe parameters \( \kappa \) and \( \rho \) and on \( \xi \), the probability parameter for observable data matrix elements. The catastrophe rate parameter \( \rho \) is added to time scaling updates in which subsets of parameters are simultaneously scaled by a common random factor \( s \): if \( \theta \) is a parameter in the scaled subset having units time\(^u\), then \( \theta \rightarrow s^u \theta \). The probability \( 0 < \xi_t < 1 \) for an element of the registered data matrix to be observable is, for many leaf languages, close to 1, so we update those parameters by scaling \( 1 - \xi_t \).

We incorporate updates adding and deleting catastrophes (the filled dots that are marked on the branches of Fig. 1) plus an update which moves a catastrophe from an edge to a parent, child or sibling edge. For the addition and deletion of catastrophes, we do not need to use reversible jump MCMC methods, as the state vector specifies the numbers, and not the locations, of catastrophes on edges.

We omit the details of these moves but give, as an example, the update that moves a catastrophe from an edge to a parent, child or sibling edge. Let \( k_T = \sum_{i=1}^{L-2} k_i \) give the total number of catastrophes. Given a state \( x = (g, \mu, \kappa, \rho, \xi) \) with \( g = (E, V, t, k) \), we pick edge \( (i, j) \in E \) with probability \( k_i / k_T \). Let \( E_{(i,j)} \) be the set of edges neighbouring edge \( (i, j) \) (child, sibling and parent edges, but excluding the edge \( (R, A) \)) and let \( q_i = \text{card}(E_{(i,j)}) \). We have in general \( q_i = 4 \). However, for \( i \) the index of a leaf node, \( q_i = 2 \) (one parent; one sibling; no children). If \( j \) is the root and \( i \) is non-leaf, then \( q_i = 3 \) (one sibling; two children) and if \( j \) is the root and \( i \) is a leaf we have \( q_i = 1 \) (a sibling edge). Choose a neighbouring edge \( (\tilde{i}, \tilde{j}) \) uniformly at random from \( E_{(i,j)} \) and move one catastrophe from \( (i, j) \) to \( (\tilde{i}, \tilde{j}) \). The candidate state is \( x' = ((E, V, t, k'), \mu, \kappa, \rho) \), with \( k'_i = k_i - 1 \) and \( k'_j = k_j + 1 \) and \( k'_j = k_j \) for \( j \neq i, \tilde{i} \). This move is accepted with probability

\[
\alpha(x' | x) = \min \left\{ 1, \frac{q_i k_j}{q_j k_i} \frac{p(x' | D)}{p(x | D)} \right\}.
\]
We assessed convergence with the asymptotic behaviour of the auto-correlation for the parameters $\mu$, $\kappa$, $\rho$ and $t_R$, as suggested by Geyer (1992). This method indicated that we could use runs of about 10 million samples; we also let the MCMC algorithm run for 100 million samples and checked that the computed statistics did not vary.

7. Validation

We made a number of tests using synthetic data. Fitting the model to synthetic data simulated according to the likelihood $P\{D = \mathbf{D}|g, \lambda, \mu, \kappa, \rho, \xi, \mathbf{D} = R(D)\}$ (in-model data) shows us just how informative the data are for catastrophe placement, as well as making a debug check on our implementation.

We simulated data under our model for various values of $\lambda$, $\mu$, $\kappa$ and $\rho$, to explore the various possible scenarios of rate heterogeneity, such as few large catastrophes or many small catastrophes. We synthesized data for 20 languages, with on average $\lambda/\mu = 40$ traits per language and five clade constraints. We set $\mu = 2 \times 10^{-5}$ deaths per year, since the estimates for that parameter in Nicholls and Gray (2008) were around that value. We studied small ($\kappa = 0.1$), medium ($\kappa = 0.2$) and large ($\kappa = 0.5$) catastrophes (corresponding to catastrophes equivalent to 520 years, 1160 years and 3280 years of change respectively). We show typical results in Fig. 2. The true topology was almost perfectly reconstructed in every case; the position and size of catastrophes were perfectly reconstructed; the posterior evaluations of the parameters were close to the true values. This remains true for a wide variety of parameter values.

When catastrophes are present in the true tree, the signal for them is strong. We wished to check the influence of catastrophes on our reconstructions, so we tried to fit a model without catastrophes to data which did, in fact, evolve with catastrophes. Fig. 3 shows that, if we do not include catastrophes in the model, our reconstructed parameters (root age and death rate $\mu$) are very far from the true values. The reconstructed topology is also much further from the truth than when we fit the model with catastrophes.

We fitted out-of-model data also. These are synthetic data that are simulated under likely model violation scenarios and are used to identify sources of systematic bias. We summarize results for synthetic data simulated by using the parameter values that we estimate in Section 8 on the data. For in-model data we correctly reconstruct topology, root age and the number and position of catastrophes. Further details are given in the supplementary material.

7.1. Model misspecification

Nicholls and Gray (2008) used out-of-model data representing unmodelled loan words (which is called borrowing), rate heterogeneity in time and space, rate heterogeneity across cognacy classes and the empty field approximation. They discussed also model misspecification due to missing data and the incorrect identification of cognacy classes (in particular, a hazard for deeply rooted classes to be split). Rate heterogeneity in time and space, and missing data are now part of the in-model analysis.

We synthesized out-of-model data with the borrowing model of Nicholls and Gray (2008), to see whether unmodelled borrowing biased our results. For what Nicholls and Gray (2008) called low-to-moderate levels of borrowing, we could reconstruct true parameter values well. See the supplementary material for details. With high levels of borrowing, we underestimate the root age and overestimate rate parameters. However, unidentified loan words in the registered data do not generate model misspecification unless they are copies of cognates which occur in the observed meaning categories of languages that are ancestral to the leaf languages. It is not plausible that unidentified borrowing of this kind is present at high levels.
Fig. 2. Simulations of synthetic data showing robustness of the model: (a), (c) true trees; (b), (d) reconstructed consensus trees (consensus trees are very close to the true trees and the reconstructed catastrophes are on the correct branches); samples from the posterior for (e) the root age and (f) death rate $\mu$ are also close to the true value ($\lambda$); other parameters behave similarly to the death rate.
Fig. 3. Importance of including the catastrophes: given data synthesized under a true tree without catastrophes (a), which was well reconstructed by a model with catastrophes (b), we tried to fit a model without catastrophes; the topology (c), root age (d) and death rate $\mu$ (e) were all badly reconstructed.
We have not repeated the Nicholls and Gray (2008) out-of-model analysis of rate heterogeneity across cognacy classes.

In a real vocabulary, distinct cognacy classes that share a meaning category would not evolve independently. Also, a real language might be expected to have a word in each of the core meaning categories. This constrains the number of cognacy classes in each meaning category to be non-zero. Our model allows empty meaning categories. Nicholls and Gray (2008) found no substantial bias in a fit to out-of-model data respecting the constraint.

Our treatment of missing data introduces a new misspecification. We modelled matrix elements as missing independently. However, we obtain missing data when we do not know the word that is used in a given language to cover a given meaning category. Matrix elements are as a consequence typically missing in blocks corresponding to all the cognacy classes for the given meaning. Because the ages of poorly reconstructed languages are well predicted in the cross-validation study below, we have not looked further at this issue.

7.2. Prediction tests for calibration

In the next section we estimate the age of a tree node (the root). In this section we test to see whether the uncertainties that we estimate are reliable.

The calibration data that were described in Section 2 fix the topologies and root ages of the subtrees that are marked with bars in Fig. 6 in Section 8. We remove each calibration constraint in turn and use the data and the remaining constraints to estimate the age of the constrained node, and the probability for the constrained subtree topology. The topological constraints were all perfectly reconstructed. In 25 of the 30 tests the constrained age interval overlaps the 95% highest posterior density interval, as shown in the bottom half of Fig. 4.

How good or bad are each of these predictions? Looking at the Hittite prediction in Fig. 4, the large prediction uncertainty only just allows the calibration interval. Is this bad? We quantify the goodness of fit, for each calibration, by using Bayes factors to replace $p$-values as indices of misfit. For each constraint $c = 1, 2, \ldots, C$ we compute a Bayes factor measuring the support for the fully constrained model compared with a model with just the $c$th constraint removed.

For $c = 1, 2, \ldots, C$ let

$$\Gamma^{C-c} = \bigcap_{c' \neq c} \Gamma^{c'}$$

 denote the enlarged tree space with the $c$th constraint removed, and let $\Gamma^{C-c}_K$ be the enlarged tree space, extended to include catastrophes (as in Section 3.2). For each constraint $c = 1, 2, \ldots, C$ we make a model comparison between a common null model with all the constraints, $H_0 : g \in \Gamma^{C}_K$, and an alternative model $H_1 : g \in \Gamma^{C-c}_K$ with the constraint removed. The Bayes factor $B_{C,C-c}$ for the model comparison is the ratio of the posterior probabilities for these models with model prior $P(H_0) = P(H_1)$,

$$B_{c-c} = \frac{P(D|g \in \Gamma^{C}_K)}{P(D|g \in \Gamma^{C-c}_K)} = \frac{P(D|g \in \Gamma^{C}_K \cap \Gamma^{C-c}_K)}{P(D|g \in \Gamma^{C-c}_K)} = \frac{P(g \in \Gamma^{C}_K|D, g \in \Gamma^{C-c}_K)}{P(g \in \Gamma^{C}_K|g \in \Gamma^{C-c}_K)},$$
Fig. 4. Reconstruction of known node ages (at the top are logarithms of Bayes factors $\log(B_{c,c-1})$ for $c = 1, 2, \ldots, C$; age constraints for different nodes; $\pm$ 95% posterior highest posterior density interval for the reconstructed dates): HI, Hittite; TA, Tocharian A; TB, Tocharian B; LU, Luvian; LY, Lycian; OI, Old Irish; UM, Umbrian; OS, Oscan; LA, Latin; GK, Greek; AR, Armenian; GO, Gothic; ON, Old Norse; OE, Old English; OG, Old High German; OS, Old Church Slavonic; PR, Old Prussian; AV, Avestan; PE, Old Persian; VE, Vedic; CE, Celtic group; IT, Italic group; GE, Germanic group; WG, West Germanic group; NW, Northwest Germanic group; BS, Balto-Slav group; BA, Baltic group; IR, Iranian group; II, Indo-Iranian group; TG, Tocharian group
where the second line follows since $\Gamma_K^c \subset \Gamma_K^{c-c}$ and the third from the definition of the conditional probabilities. The numerator $P(g \in \Gamma_K^c | D, g \in \Gamma_K^{c-c})$ is the posterior probability for the $c$th constraint to be satisfied given the data and the other constraints. The denominator $P(g \in \Gamma_K^c | g \in \Gamma_K^{c-c})$ is the prior probability for the $c$th constraint to be satisfied given the other constraints. We estimate these probabilities by using simulation of the posterior and prior distributions with constraint $c$ removed. The Bayes factors are estimated with negligible uncertainty and $2 \log(B_{c,C^{c-c}})$ is plotted for $c = 1, 2, \ldots, C$ in the top half of Fig. 4.

Strong evidence against a calibration is failure at prediction. Taking a Bayes factor exceeding 12 (i.e. $2 \log(B_{c,C^{c-c}}) \lesssim -5$ in Fig. 4) as strong evidence against the constraint, following Raftery (1996), we have conflict for three of the 30 constraints: the ages of Old Irish and Avestan, and for the age of the Balto-Slav clade. As our analysis in Section 8 shows, there is a high posterior probability that a catastrophe event occurred on the branch between Old Irish and Welsh, and another between Old Persian and Avestan. The evidence for rate heterogeneity in the rest of the tree is so slight that when we try to predict these calibrations we are predicting atypical events.

Our missing data analysis has helped here. The calibration interval for the Hittite vocabulary in these data is 3200–3700 BP. A reconstruction of the age of Hittite which ignores missing data predicts 60–2010 BP, which is well outside the constraints. The 95% highest posterior density interval for the age of Hittite in our model is 430–3250 BP, which just overlaps the constraint. The Bayes factor gives odds that are less than 3:1 against, so the evidence against the constraint is ‘hardly worth mentioning’.

8. Results

In this section we present results for our MCMC simulation of the full posterior: equation (8). An upper limit $T = 16000$ was used in the tree prior of Section 3.1. Any value for $T$ exceeding around $T = 10000$ would lead to the same results. We show a consensus tree in Fig. 5. In a tree, an edge corresponds to a split partitioning the leaves into two sets. A consensus tree displays just those splits that are present in at least 50% of the posterior sample. Splits which receive less than 95% support are labelled. Where no split is present in 50% of the posterior sample, the consensus tree is multifurcating. The date that is shown for a node is the average posterior date given the existence of the split; similarly, the number of catastrophes that is shown on an edge is the average posterior number of catastrophes on that edge given the existence of the split, rounded to the nearest integer. The 95% highest posterior probability density intervals for our estimates for the parameters are as follows: $\mu = 1.86 \times 10^{-4} \pm 1.47 \times 10^{-5}$ deaths per year; $\kappa = 0.361 \pm 0.055$; $\rho = 6.4 \times 10^{-5} \pm 3.3 \times 10^{-5}$ catastrophes per year (corresponding to large but rare catastrophes: about one catastrophe every 15000 years, or an average of 3.4 on the tree, with each catastrophe corresponding to 2400 years of change).

We display in Fig. 6 the calibration constraints on a tree sampled from the posterior. The constraints cannot be shown on a consensus tree, as slices across a consensus tree are not isochronous.

The analysis reconstructs some well-known features of the Indo-European tree. The Germanic, Celtic and Italic families are grouped together, but no particular configuration of their relative positions is favoured. The Indo-Persian group can fall outside the Balto-Slav group but the relative position of these two is uncertain. The deep topology of the tree is left quite uncertain by these data, especially the position of Albanian. We find evidence for catastrophic rate heterogeneity in three positions: on the edges leading to Old Irish and Old Persian, and in the Umbrian–Oscan clade.
Fig. 5. Consensus tree for the data set of Ringe et al. (2002): *, catastrophes supported with probability above \( \frac{1}{2} \).
Fig. 6. Typical sample from the posterior: all the constraints on the node ages are shown, except those for the Italic, Indo-Iranian and Iranian groups, for which we do not have an upper bound, and the root, which has an upper bound at $T = 16000$ years BP
Nicholls and Gray (2008) analysed a subset of the Ringe et al. (2002) data (they discarded languages with too much missing data). Where the comparison is possible, the topology that they reconstructed is similar to the consensus tree that is presented in Fig. 5; the main differences are that we have more uncertainty on the relative positions of the Germanic, Celtic and Italic families and on the relative positions of the Balto-Slav and Indo-Iranian families, and that we find somewhat more support for a subtree grouping Hittite, Luvian, Lycian and the Tocharian languages.

The 95% highest posterior probability density interval for our estimate for the root age of the Indo-European family is 7110–9750 years BP. The distribution of this key statistic is close to normal. This is in line with previous quantitative analyses of Indo-European lexical data (Gray and Atkinson, 2003; Nicholls and Gray, 2008).

9. Conclusions
Our results give a root age for the most recent common ancestor of the Indo-European family of language vocabularies in agreement with earlier phylogenetic studies. Our results are in agreement with models which put this date around 8500 BP, and in conflict with models which require it to be less than 6500 years BP. Our studies of synthetic out-of-model data and reconstruction tests for known historical data support our view that this main result is robust to model error. It would not be robust to a step change in the rate of lexical diversification acting in a co-ordinated fashion across the Indo-European languages extant about 3000–5000 years ago.

The methods that are outlined here for handling missing data and rate heterogeneity in the diversification of languages, as seen through lexical data, will find applications to generic trait data.

Appendix A: Recursions for other registration processes
This section complements Section 4: we give iterations for $u_i^{(0)}$, $u_i^{(1)}$, $u_i^{(n-1)}$, $v_i^{(0)}$, $v_i^{(n-1)}$ and $v_i^{(n)}$. These are the quantities that are needed (as in equation (3)) to evaluate the sum in equation (2), for registration rules which use condition (a) in combination with other conditions from Section 3.3. Consider a pair of edges $(c_1, i)$, $(c_2, i)$ in $E$. In the notation of the text,

\[
 u_i^{(0)} = (1 - \delta_{i,c_1} + \delta_{i,c_1} u_i^{(0)})(1 - \delta_{i,c_2} + \delta_{i,c_2} u_i^{(0)}), \\
 u_i^{(1)} = \delta_{i,c_1}(1 - \delta_{i,c_2} u_i^{(0)} + \delta_{i,c_2} (1 - \delta_{i,c_1}) u_i^{(1)}, \\
 u_i^{(n)} = \delta_{i,c_1} u_i^{(0)} + \delta_{i,c_2} (1 - \delta_{i,c_1}) u_i^{(1)}, \\
 v_i^{(0)} = \left\{ \delta_{i,c_1} v_i^{(0)} + (1 - \delta_{i,c_1}) \prod_{j \in V_L} \xi_j \right\} \left\{ \delta_{i,c_2} v_i^{(0)} + (1 - \delta_{i,c_2}) \prod_{j \in V_L} \xi_j \right\}, \\
 v_i^{(n)} = \left\{ \delta_{i,c_1} v_i^{(0)} + (1 - \delta_{i,c_1}) \prod_{j \in V_L} \xi_j \right\} \left\{ \delta_{i,c_2} v_i^{(0)} + (1 - \delta_{i,c_2}) \prod_{j \in V_L} \xi_j \right\}, \\
 v_i^{(n-1)} = \left\{ \delta_{i,c_1} v_i^{(0)} + (1 - \delta_{i,c_1}) \xi_{c_1} (1 - \xi) v_i^{(n-1)} \right\} \left\{ \delta_{i,c_2} v_i^{(0)} + (1 - \delta_{i,c_2}) (1 - \xi) v_i^{(n-1)} \right\} + \left\{ \delta_{i,c_1} v_i^{(0)} + (1 - \delta_{i,c_1}) (1 - \xi) v_i^{(n-1)} \right\} \left\{ \delta_{i,c_2} v_i^{(0)} + (1 - \delta_{i,c_2}) (1 - \xi) v_i^{(n-1)} \right\}.
\]

The recursion is evaluated from the leaves $i \in V_L$, at which

\[
 u_i^{(0)} = u_i^{(n-1)} = 1 \xi_i ; \\
 u_i^{(1)} = u_i^{(n)} = \xi_i.
\]
\[ v_i^{(0)} = v_i^{(n-1)} = 0, \quad v_i^{(n)} = 1. \]

References

Alekseyenko, A., Lee, C. and Suchard, M. (2008) Wagner and Dollo: a stochastic duet by composing two parsimonious solos. *System. Biol.*, 57, 772–784.

Dyen, I., Kruskal, J. and Black, P. (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/RussellsData.htm.)

Felsenstein, J. (1981) *Inferring Phylogenies*. Sunderland: Sinauer Associates.

Garrett, A. (2006) Convergence in the formation of Indo-European subgroups: phylogeny and chronology. In *Phylogenetic Methods and the Prehistory of Languages* (eds P. Forster and C. Renfrew). Cambridge: McDonald Institute for Archaeological Research.

Geyer, C. (1992) Practical Markov Chain Monte Carlo. *Statist. Sci.*, 7, 473–483.

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

Lewis, P. (2001) A likelihood approach to estimating phylogeny from discrete morphological character data. *System. Biol.*, 50, 913–925.

Mallory, J. (1989) *In Search of the Indo-Europeans: Language Archaeology and Myth*. Thames and Hudson.

McMahon, A. and McMahon, R. (2005) *Language Classification by Numbers*. Oxford: Oxford University Press.

Nicholls, G. K. and Gray, R. D. (2008) Dated ancestral trees from binary trait data and their application to the diversification of languages. *J. R. Statist. Soc. B*, 70, 545–566.

Raftery, A. (1996) Hypothesis testing and model selection. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter). London: Chapman and Hall.

Renfrew, C. (1987) Archaeology and language: the puzzle of Indo-European origins. *Curr. Anthrop.*, 29, 437–441.

Ringe, D., Warnow, T. and Taylor, A. (2002) Indo-European and computational cladistics. *Trans. Philol. Soc.*, 100, 59–129.

Ronquist, F., Huelsenbeck, J. and van der Mark, P. (2005) MrBayes 3.1 manual. School of Computational Science, Florida State University, Tallahassee.

Ryder, R. J. and Nicholls, G. K. (2010) Supplement to Missing data in a stochastic Dollo model for binary traits, and its application to the dating of Proto-Indo-European. (Available from www.ceremade.dauphine.fr/~ryder/research.html.)

Warnow, T., Evans, S., Ringe, D. and Nakhleh, L. (2006) A Stochastic model of language evolution that incorporates homoplasy and borrowing. In *Phylogenetic Methods and the Prehistory of Languages* (eds P. Forster and C. Renfrew). Cambridge: McDonald Institute for Archaeological Research.

Yule, G. (1925) A mathematical theory of evolution based on the conclusions of Dr. JC Willis, FRS. *Phil. Trans. R. Soc. Lond.* B, 213, 21–87.