# QUANTIFYING UNCERTAINTY IN A STOCHASTIC DOLLO MODEL OF VOCABULARY EVOLUTION

G.K. NICHOLLS AND R.D. GRAY

ABSTRACT. We specify a stochastic model of binary-valued trait-character evolution. Although the model is a simple and natural model for cognate data, no model of this type has been considered in the past. Because of the technical difficulty of fitting models, authors have tended to avoid ab intitio model specification and fitting. The model is based on a point process-representation of the birth and death of cognates. It is essentially a stochastic variant of Dollo parsimony in which traits correspond to the presence or absence of words. A branching process is used to model the development of a language tree.

We fit the model to Indo-European cognate data using our own implementation of our own Metropolis-Hastings Markov chain Monte Carlo algorithms. The historical processes and modern observation practices which determine the data are complex. In an exploratory analysis, we identify some features of the data which raise difficulties for quantitative analysis. In particular, words generating cognate classes in the data are those words with descendants in two or more observed languages. We show that at least some of the difficulties may be overcome. We tend to support the results of Gray & Atkinson, (2003).

## 1. INTRODUCTION

What is the data, and how was it gathered? Dyen et al., (1997) group the words of a number of human languages into equivalence classes called cognate classes. Words in a cognate class have two important things in common. First, they have (in their respective languages) roughly the same meaning. Secondly, words in a cognate class are descended from a common ancestor, so that they constitute homologous language traits. Cognate classes are generated by starting with two hundred meaning categories, and collecting together all words in all languages under study which fall in one of the meaning categories. These categories are then further subdivided into homologous groupings. In Gray & Atkinson, (2003) each cognate class (labels $j = 1, 2 \dots N$) generates a binary valued trait ($D_{i,j}$ say) for each language (labels $i = 1, 2 \dots L$). Language $i$ either posseses ($D_{i,j} = 1$) or lacks ($D_{i,j} = 0$) a word in cognate class $j$. The data analysed by Gray & Atkinson, (2003) and here contain $N = 2398$ cognate classes observed in $L = 87$ Indo-European languages and $K = 200$ meaning categories.

Cognate classes present in the data are of course a subset of the cognate classes which existed in the history of the languages represented in the data. Cognate classes which existed in the past, but are present in no data-language are of course absent from the data. However, in the course of building the Gray & Atkinson,

*Key words and phrases.* Phylogenetics, human language, binary trait presence-absence, Bayesian inference, Markov chain Monte Carlo.

The authors acknowledge advice and assistance from Quentin Atkinson and David Welch of the University of Auckland.

(2003) data from the Dyen et al., (1997) corpus, cognate classes represented by a single word in a single data language were dropped from the data. It follows that words generating cognate classes in the data are those words with descendants in two or more of the observed languages. Cognate classes which existed in the past, and are present in just one data language are absent from the data. We refer to cognate classes present in the data as data-cognates.

A great deal is known about the recent history of the languages in our study. We use the information summarized in the addendum to Gray & Atkinson, (2003). Those authors imposed 16 language clades, and corresponding constraints on clade-root times. The clade information adds little to the cognate data which is already very informative on that score. However the clade root age constraints are important, in particular the few upper bounds which are available for clade root times. It is this historical knowledge which sets the clock throughout the tree. For example, the Brythonic clade is made up of two Welsh and three Breton languages and we impose $1450 \leq t_{\texttt{Brythonic}} \leq 1600$ for the age of the root of this clade. The Italic clade contains Romanian, Vlach, Italian, Ladin, Provencal, French, Walloon, two French creole languages, three Sardinian languages, Spanish, Portuguese, Brazilian and Catalan, and $1700 \leq t_{\texttt{Italic}} \leq 1850$.

How do we interpret the data? We assume that language 'speciation' is fundamentally tree-like. We specify a model of cognate birth and death in which any given cognate class is born exactly once in the language phylogeny. It has two parameters, representing the birth rate for cognate classes and the per capita death rate for each word in a cognate class. We do not impose the constraint that all meaning categories of all ancient languages are filled at all times by words from data-cognate classes, since those meaning categories may have been filled by words belonging to cognate classes which are not represented in the data. Our model predicts that there are about two lost cognate classes (below the root) for each cognate class present in the data. Words from these classes filled meaning categories not occupied by data-cognates. We extend our model to account for lateral cognate transfer (word 'borrowing'). This second model has an additional parameter, the per capita borrowing rate (which it is natural to express relative to the per capita death rate).

The models we describe are of a simple but novel kind, and there is therefore some value in our model specification itself. They are the simplest instances of a new class of models for trait-based phylogenetic inference. However, it very easy to come up with new models of language evolution. It is typically much harder to fit models to data, that is, it is difficult to find model parameters which make the data a relatively likely outcome of the model. We have set up analytical tools and software for model fitting for our cognate birth-death model. We simulate the cognate birth-borrowing-death model as part of our model mis-specification analysis.

Gray & Atkinson, (2003) adapt finite-sites Markov mutation models of the kind proposed by Felsenstein, (1981) to fit these data. In the class of models Gray & Atkinson, (2003) consider, the number of sites is fixed, and corresponds to the number of distinct cognate classes in the data. In their model, a cognate class can come into existence independently in more than one language. It might be observed that this allows the model to fit lateral word transfer events. However, two parameters are doing the work of three. The two rate parameters for cognate

birth and death must be chosen to fit the borrowing rate, the equilibrium probability to find a cognate at any given site, and the cognate loss rate. We accept that some model mis-specification is inevitable for a large and complex data set of the kind Gray & Atkinson, (2003) treat. However, we would like to begin with a model which is at least a mathematical description of idealized cognate evolution.

## 2. Data exploration

There is a representation of the data as a list of sets. Suppose language $i$ has words in $N_i$ data-cognate classes, so $N_i = \sum_{j=1}^{N} D_{i,j}$. Suppose words from data-cognate class $c$ are found in $M_c$ languages. For cognate $c = 1, 2 \ldots N$, let the set $\mathcal{C}_c$ contain a list of the $M_c$ language-labels in which cognate $c$ was present. Let $\mathcal{C} = (\mathcal{C}_1, \mathcal{C}_2 \ldots \mathcal{C}_N)$. $\mathcal{C}$ and $D$ are equivalent representations of the data.

We begin by plotting histograms for the distributions of $M_c$ ('languages per cognate') and $N_i$ ('cognates per language'). The top graph in Fig. 1 shows the distribution of $N_i$ in the data. The distribution is centred around 200, the number of items in the Swadesh, (1952) word list. It has a tail skewed toward smaller values. Hittite has relatively few cognates. This need not indicate lost vocabulary. Hittite would share relatively few words with other IE languages (and therefore possess less data-cognates) if it was the first to branch. The bottom graph in Fig. 1 shows the $M_i$-distribution in the data. The number of cognates present in one language and no others is zero. This thinning could explain the direction of the skew in the $N_i$-distribution at top. Most cognates are present in just a few languages. However there is a shoulder at around 20 languages and a flat tail running the full extent of the graph. Three cognates are present in all 87 languages. The shoulder is caused by language selection. The dense sampling of the Germanic and Italic clades gives two large clades of relatively closely related languages. The long tail suggests rate heterogeneity among words. A small fraction of data-cognates are evolving at a distinctly slower pace.

## 3. A cognate-birth and word-death model

We define basic terms in the context of a simple model. A language is summarized by a set of words. Each word carries a label indicating its cognate class. Each language evolves according to a set-branching process to form a language tree. At a branching, the language-set entering the branch point is copied to yield two identical sets of words. Each word in the parent language is copied to give two sibling words, one in each of the offspring languages. Sibling words belong to the same cognate class. Set elements evolve according to a constant rate birth-death process corresponding to the birth and death of individual words. We suppose a 'raw' cognate-class-creation process generates words at a rate $\lambda$ constant in all languages at all times (as we explain below it is necessary to modify this raw process - also, the process acts at constant rate, but the number of word births is a Poisson distributed random variable). When a word is born in this way a new cognate class is created. Each word in each language dies independently at constant per capita rate $\mu$. The number of words in a language is randomly variable over time. A cognate class is born exactly once. The first word in a cognate class reproduces as the language-set branching process generates new language sets. Its descendants may die in the distinct languages into which they are copied. The full history of a word is a subtree of the language-set tree.

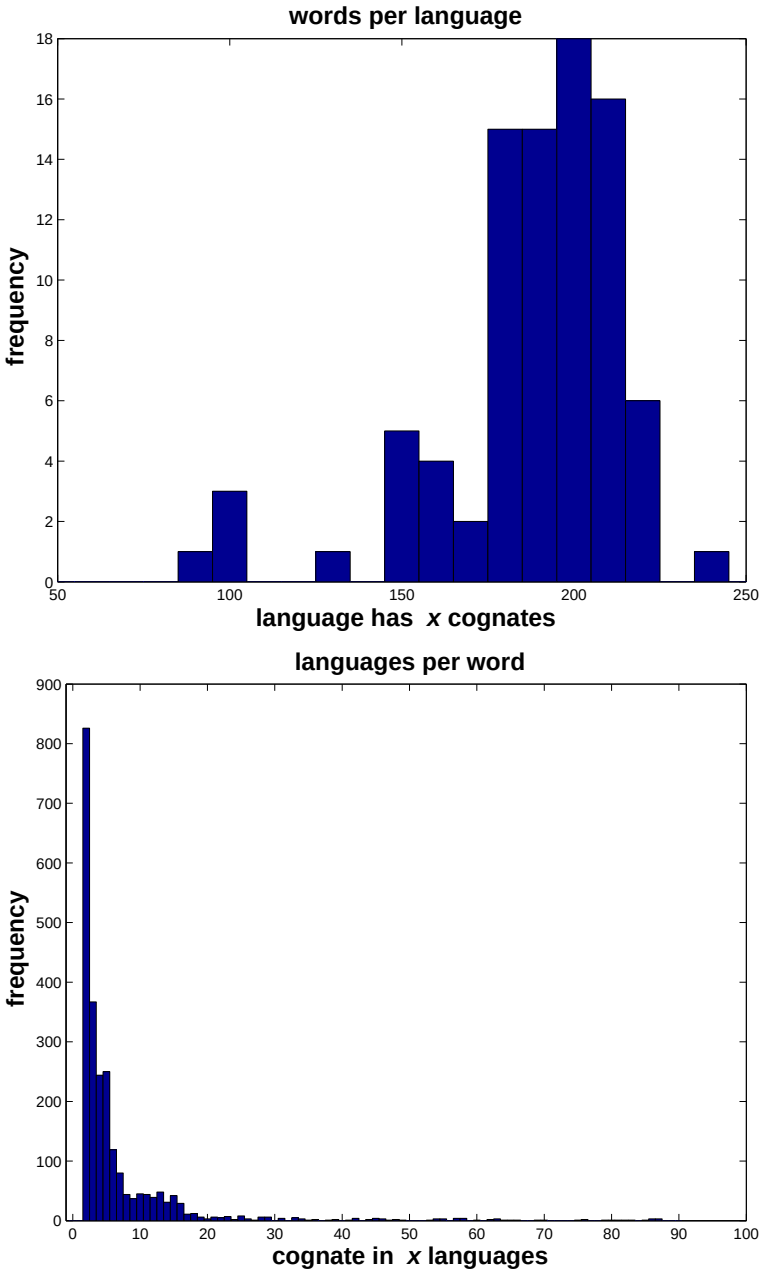


FIGURE 1. (top) the distribution of the number of cognates per language. Each language has around 200 cognates. Hittite has 94 cognates, the least in the data. Nepali has 236, the most. (bottom) the distribution of the number of languages to which cognates belong. Cognates are typically shared by just a few languages. However, no data-cognates are in just one language, while 11 are in 80 or more of the 87 languages.

Languages extant in some fixed time slice of the evolution are observed. Homologous words are grouped into cognate classes. In the data-gathering process, cognate classes with just one extant modern word are dropped from the data. While every word birth is associated with the birth of a cognate class, just a subset of all word births are associated with the birth of one of the cognate classes observed at the leaves of the tree. As a consequence of this heavy word thinning, a language-set, ancient or modern, may at any given time posses no data-cognates in any given meaning class. Because data-cognates are selected by the observation process on the basis that they survive to the leaves, the birth times of data-cognates are biased towards the leaves. The birth rate for data-cognates declines towards the root, and goes to zero on the edge leading away from the root into the depths of the past.

3.1. **Model details.** *This section may be omitted at a first reading.* Let $\mathbf{g}$ be a binary tree graph with $L+1$ nodes of edge-degree one and $L-1$ nodes of degree three. Let $\mathcal{V}$ denote the set of language-tree node labels. Let $t_i$ denote the time associated with node $i \in \mathcal{V}$. Time is measured in years before the present. Node labels are ordered by increasing age, so that $j > i \;\Rightarrow\; t_j \geq t_i$. Let $t = (t_1, t_2 \dots t_{2L})$. Of the degree one nodes, $L$ are leaf nodes, corresponding to data-languages, and one is ancestral to the root node, labelled $A \in \mathcal{V}$ with $A = 2L$ and assigned age $t_A = \infty$. One degree three node is identified as the root node and labelled $R \in \mathcal{V}$ with $R = 2L-1$. Suppose $i, j \in \mathcal{V}$ are two nodes of $\mathbf{g}$ connected by an edge. Let $\langle i, j\rangle$ denote the corresponding edge, ordered so that $i < j$, and $E$ the set of all edges of $\mathbf{g}$. Let $E^- = E \setminus \{\langle R, A\rangle\}$. We define $\mathbf{g} = (E, t)$.

If $\theta$ denotes the branching rate for language-sets, and

$$\ell(\mathbf{g}) = \sum_{\langle i,j\rangle \in E^-} t_j - t_i$$

gives the tree length $\ell(\mathbf{g})$ excluding edge $\langle R, A\rangle$, then $\mathbf{g} \sim f_G(\mathbf{g}|\theta)$ with

$$f_G(\mathbf{g}|\theta) = \theta^{L-1} \exp(\theta \ell(\mathbf{g})).$$

Consider data-cognate $c = 1, 2 \dots N$ born on $\mathbf{g}$ in edge $\langle i, j\rangle \in E$ at time $\tau$. The edge is identified by the node $i$ at its base. Let $x_c = (\tau, i)$ denote a point on $\mathbf{g}$. The set of birth points for all data-cognates is $x = \{x_1, \dots x_N\}$. Denote by $[\mathbf{g}]$ the set of all points $(\tau, i)$ in $\mathbf{g}$, by $\Omega_{\mathbf{g}}^{(n)}$ the space of all regular subsets of $[\mathbf{g}]$ containing $n$ distinct points, and let

$$\Omega_{\mathbf{g}} = \bigcup_{n=0}^{\infty} \Omega_{\mathbf{g}}^{(n)}$$

so that $x \in \Omega_{\mathbf{g}}$. Let $Y \in \Omega_{\mathbf{g}}$ be the point process of 'raw' tree-cognate births, including those which fail to survive into two or more leaves. Denote by $\lambda$ the raw tree-cognate birth rate and suppose that each copy of each cognate dies out with instantaneous per capita death rate $\mu$ per year.

Let $f_X(x|\mathbf{g}, \mu, \lambda)$ denote the point-process density for data-cognate birth times $x$ on a given tree, for known birth and death rates. The measure $dx = dx_1 \dots dx_N$ of the point process distribution $F(dx|\mathbf{g}, \mu, \lambda)$ is a product measure of sets in $\Omega_{\mathbf{g}}$ defined in terms of the element of length $d\tau$ on the random set $[\mathbf{g}]$. Counting measure is taken over the edges of $\mathbf{g}$, including $\langle R, A\rangle$. The point process $X$ generating data-cognates is derived from the 'raw' point process $Y$ of tree-cognate births by thinning. A birth at $y \in [\mathbf{g}]$ is included in the data-cognate process $X$ if it generates descendants in two or more languages. The intensity $\lambda(y)$ of the thinned process is

a function of position on the tree. The distribution function for the inhomogeneous point-process of data-cognate births is

$$F(dx|\mathbf{g},\mu,\lambda) = f_X(x|\mathbf{g},\mu,\lambda)dx = \exp\left(-\int_{[\mathbf{g}]}\lambda(y)dy\right)\prod_{c=1}^{N}\lambda(x_c)dx_c. \tag{3.1}$$

The number of leaf languages, $M(y)$ say, in which a generic tree-cognate born at $y$ appears is a discrete random variable. If $\Pr(M(y)>1|y,\mathbf{g},\mu)$ is the probability that $M$ exceeds one then

$$\lambda(y) = \lambda \times \Pr(M(y)>1|y,\mathbf{g},\mu)$$

is the instantaneous birth rate for data-cognates at $y \in [\mathbf{g}]$. The total rate is

$$\int_{[\mathbf{g}]}\lambda(y)dy \equiv \sum_{\langle i,j\rangle\in E}\int_{t_i}^{t_j}\lambda\Pr(M(y)>1|y=(\tau,i),\mathbf{g},\mu)d\tau.$$

This is evaluated by observing that

$$\Pr(M(y)>1|y=(\tau,i),\mathbf{g},\mu) = \Pr(M(y)>1|y=(t_i,i),\mathbf{g},\mu)e^{-\mu(\tau-t_i)}$$

so that

$$\int_{[\mathbf{g}]}\lambda(y)dy = \frac{\lambda}{\mu}\sum_{\langle i,j\rangle\in E(g)}\Pr(M(y)>1|(t_i,i),\mathbf{g},\mu)(1-e^{-\mu(t_j-t_i)}). \tag{3.2}$$

Notice that the contribution to Equation (3.2) from $\langle R,A\rangle$ is finite, as births in the distant past are killed off by the thinning inherent in the data cognate-collection process.

In order to compute $\int_{[\mathbf{g}]}\lambda(y)dy$ in Equation (3.2), we need only evaluate $\Pr(M(y)>1|y,\mathbf{g},\mu)$ for $y$ at nodes of $\mathbf{g}$, and this can be done using a pruning recursion of the kind described in Felsenstein, (1981). Let $u_i^{(0)} = \Pr(M(y)=0|y=(t_i,i),\mathbf{g},\mu)$ denote the probability for zero offspring at the leaves descended from node $i$, and let $u_i^{(1)} = \Pr(M(y)=1|y=(t_i,i),\mathbf{g},\mu)$ denote the corresponding probability for one offspring. Notice that $u_i^{(0)}=0$ and $u_i^{(1)}=1$ if $i$ is a leaf. Since $\Pr(M(y)>1|y=(t_i,i),\mathbf{g},\mu) = 1-u_i^{(0)}-u_i^{(1)}$, the quantity of interest can be computed once we have $u_i^{(0)}$ and $u_i^{(1)}$ at each $i\in\mathcal{V}$. Suppose $\langle j,i\rangle$ and $\langle k,i\rangle$ are edges of $\mathbf{g}$, so that $j$ and $k$ are child nodes of $i$, and let $p_{i,j}=\exp(-\mu(t_i-t_j))$. The tree recursions

$$\begin{aligned} u_i^{(0)} &= (1-p_{i,j}(1-u_j^{(0)}))\times(1-p_{i,k}(1-u_k^{(0)})),\\ u_i^{(1)} &= u_i^{(1)} = (1-p_{i,j}(1-u_j^{(0)}))p_{i,k}u_k^{(1)} + (1-p_{i,k}(1-u_k^{(0)}))p_{i,j}u_j^{(1)}, \end{aligned}$$

are easily evaluated.

The model we have described reminds us of the infinite sites model introduced in Watterson, (1975). Mutations generate new segregating sites at constant rate. However cognate death corresponds to back-mutation at a segregating site. The absence of back mutation in the infinite sites model follows from its definition as a limit model of the finite sites model, and is fundamental to the model. In the infinite alleles model of Kimura & Crow, (1964) the setup is again similar. However new allele types are generated by mutation of existing alleles so the total number of alleles in each individual is constant. The model we have described does

not appear to be related to existing models by conditioning or taking limits. We welcome comment on this point.

3.2. **Borrowing model.** In the model above, cognates generated by the 'raw' tree-cognate process evolve subject to an instantaneous per capita death rate $\mu$ (applied to each copy of the cognate independently at each instant in each language). It is straightforward to allow each copy of the cognate to be subject to instantaneous per capita borrowing at rate $b\mu$, where $b \geq 0$ and we expect values between $b = 0$ and $b = 1$ are of practical interest. When a borrowing event occurs to an instance of a cognate, a target language-set is chosen uniformly at random from the languages existing at the time of the event, and a copy of the cognate is dropped into that set. If the language already possess the cognate in question, there is no change. Otherwise, the target language adds a cognate in a new class.

3.3. **Inferential framework.** *This section may be omitted at a first reading.* The aim of the inference is to reconstruct the unknown true tree $\mathbf{g}$, and the birth, death and branching rates $\lambda$, $\mu$ and $\theta$ from the cognate data $D$ (or equivalently $\mathcal{C}$). Let $h(x, \mathbf{g}, \mu, \lambda, \theta|D)$ denote the joint posterior probability density for the unknown data-cognate birth times, $x$, and $\mathbf{g}$, $\lambda$, $\mu$ and $\theta$. The $x$ are to some extent nuisance parameters. We may not be interested in reconstructing these birth times. However, the observation model cannot be defined without them.

The likelihood $P(D|\mathbf{g}, \mu, \lambda)$ is given in terms of the point process density $f_X$ for data-cognates and $P(D|x, \mathbf{g}, \mu, M(x_c) > 1)$ (the probability to realize data $D$ given the tree, the data-cognate birth times, the death rate, and the requirement that each realized point $x_c$ generates two or more data-cognates) by

$$P(D|\mathbf{g}, \mu, \lambda) = \int_{\Omega_\mathbf{g}} P(D|x, \mathbf{g}, \mu, M(x_c) > 1) f_X(x|\mathbf{g}, \mu, \lambda) dx.$$

The evolution of cognate $c$ into the languages $\mathcal{C}_c$ is conditionally independent of the evolution of all other cognates given $x_c, \mathbf{g}$ and $\mu$, hence

$$P(D|x, \mathbf{g}, \mu, M(x_c) > 1) = \prod_{c=1}^{N} P(\mathcal{C}_c|x_c, \mathbf{g}, \mu, M(x_c) > 1).$$

It is not feasible to carry out Monte-carlo inference for all 2398 data-cognate birth times $x_c, c = 1, 2 \ldots N$. However it is feasible to integrate $x$ explicitly, by a combination of hand integration and a recursion like the one used to evaluate Equation (3.2). We have

$$\begin{aligned} h(\mathbf{g}, \mu, \lambda, \theta|D) &\propto f_G(\mathbf{g}|\theta) p(\mu, \lambda, \theta) \int_{\Omega_\mathbf{g}} P(D|x, \mathbf{g}, \mu, M(x_c) > 1) f_X(x|\mathbf{g}, \mu, \lambda) dx \\ &= \frac{\lambda^N}{N!} \exp\left( - \int_{[\mathbf{g}]} \lambda(y) dy \right) f_G(\mathbf{g}|\theta) p(\mu, \lambda, \theta) \quad \times \\ &\qquad \prod_{c=1}^{N} \int_{[\mathbf{g}]} P(\mathcal{C}_c|x_c, \mathbf{g}, \mu, M(x_c) > 1) \Pr(M(x_c) > 1|x_c, \mathbf{g}, \mu) dx_c. \end{aligned} \tag{3.3}$$

When we carry out MCMC for $h$, we must evaluate the right hand side of Equation (3.3) thousands or even millions of times. The difficulty is that we must compute the probability $P(\mathcal{C}_c|x_c, \mathbf{g}, \mu, M(x_c) > 1)$ to generate the cognate distribution pattern $\mathcal{C}_c$ given cognate $c$ was born at $x_c \in [\mathbf{g}]$ and given (this is the difficult part)

that the cognate survived into two or more languages. However the inconvenient condition cancels out of the expression because

$$P(\mathcal{C}_c|x_c,\mathbf{g},\mu) = P(\mathcal{C}_c|x_c,\mathbf{g},\mu,M(x_c)>1)\Pr(M(x_s)>1|x_c,\mathbf{g},\mu),$$

(using the fact that the outcome $\{\mathcal{C}_c, M(x_c) > 1\}$ is trivially equivalent to the outcome $\{\mathcal{C}_c\}$ whenever $\mathcal{C}_c$ has at least two members, which it has, precisely because each data-cognate is present in at least two languages). We have only to calculate the probability to realize $\mathcal{C}_c$ given a birth in the unconditioned 'raw' tree-cognate process $Y$, at $x_c$, rather than a birth in $X$, the data-cognate process.

We have now to evaluate $\int_{[\mathbf{g}]} P(\mathcal{C}_c|x_c,\mathbf{g},\mu)dx_c$ for each $c = 1, 2 \ldots N$. For any two vertices $i, k \in \mathcal{V}$ define the path $\mathrm{Path}(i,k)$ connecting $i$ and $k$ on $\mathbf{g}$ as the sequence of $p$ edges

$$\mathrm{Path}(i,k) = (\langle i, j_1\rangle, \langle j_1, j_2\rangle, \ldots \langle j_{p-2}, j_{p-1}\rangle\langle j_{p-1}, k\rangle)$$

running from $i$ to $k$ in $\mathbf{g}$ without backtracking. Recall that $A \in \mathcal{V}$ is the node above the root, located in the distant past. For cognate $c = 1, 2 \ldots N$ let

$$E^{(c)} = \bigcap_{i\in\mathcal{C}_c} \mathrm{Path}(i, A)$$

denote the "covering" edge set for cognate $c$, that is, $E^{(c)}$ is the set of edges ancestral to every language-leaf-node containing the $c$'th cognate. Clearly $P(\mathcal{C}_c|x_c,\mathbf{g},\mu) = 0$ whenever $x_c$ is not on an edge in $E^{(c)}$ so

$$\int_{[\mathbf{g}]} P(\mathcal{C}_c|x_c,\mathbf{g},\mu)dx_c = \sum_{\langle i,j\rangle\in E^{(c)}} \int_{t_i}^{t_j} P(\mathcal{C}_c|x_c,\mathbf{g},\mu)dx_c.$$

The integral $\int_{t_i}^{t_j} \Pr(\mathcal{C}_c|x_c,\mathbf{g},\mu)dx_c$ is evaluated in much the same way as the total rate (Equation (3.2)). We omit the details. Explicitly, for $h$,

$$\begin{aligned} h(\mathbf{g},\mu,\lambda,\theta|D) \quad \propto \quad & \left(\frac{\lambda}{\mu}\right)^N f_G(\mathbf{g}|\theta)p(\mu,\lambda,\theta) \quad \times \\ & \exp\left(-\frac{\lambda}{\mu}\sum_{\langle i,j\rangle\in E}\Pr(M(y)>1|(t_i,i),\mathbf{g},\mu)\left(1-e^{-\mu(t_j-t_i)}\right)\right) \\ & \times \quad \prod_{c=1}^{N}\left(\sum_{\langle i,j\rangle\in E^{(c)}} P(\mathcal{C}_c|(t_i,i),\mathbf{g},\mu)\left(1-e^{-\mu(t_j-t_i)}\right)\right) \end{aligned} \tag{3.4}$$

with $\Pr(M(y) > 1|(t_i, i), \mathbf{g}, \mu)$ and $P(\mathcal{C}_c|(t_i, i), \mathbf{g}, \mu)$ given by tree recursions.

3.4. **Priors.** The space $\Gamma$ of trees $\mathbf{g}$ is restricted to those trees which satisfy the clade constraints discussed in Section 1 and detailed in Gray & Atkinson, (2003). The timescale would be unresolved without this data. This is apparent in the expression for $h$ we have written above. The rate parameters appear only in dimensionless multiples $\lambda/\mu$, $\theta t_i$ and $\mu t_i$ for $t_i$ a tree node (language branching) time.

For simple $\theta$ and $\lambda$ priors, we can integrate $\lambda$ and $\theta$ out of the formulae above. We are left with a posterior density $h = h(\mathbf{g}, \mu|D)$. We imposed Jefferys priors $p(\mu, \lambda, \theta) = (\mu\lambda\theta)^{-1}$, a choice which is non-informative with respect to the time scale of $\mathbf{g}$. The bulk of our simulations were carried out with a tree prior which was

not the exponential branching process prior $f_G(\mathbf{g}|\theta)$ described above, but a distribution in which the root time is uniformly distributed between 0 and a conservative maximum $T_{\max}$, and all trees with a given root time have equal prior probability density. The expression is

$$f_G(\mathbf{g}|T_{\max}) \propto \mathbb{I}_{0\leq t_R\leq T_{\max}} t_R^{-L+2}(T_{\max}-t_R)^{-1},$$

with $\mathbb{I}_{0\leq t_R\leq T_{\max}}$ the indicator function for $0 \leq t_R \leq T_{\max}$. We used $T_{\max} = 16,000$ years before the present, which is intended to be uncontroversial. This prior is uninformative with respect the root time, which is the sensitive statistic. In the presence of clade constraints it is necessary to revise the formula for $f_G$ above to get uniform marginal prior $t_R$, and carry out prior simulations to check imposed priors are non-informative with respect to the hypotheses of scientific interest.

3.5. **Markov chain Monte Carlo.** *This section may be omitted at a first reading.* We designed our own Markov chain Monte Carlo simulation algorithms and implemented them ourselves in the MatLab programming language. We summarize $h(\mathbf{g}, \mu|D)$ using samples

$$(\mathbf{g}^{(s)}, \mu^{(s)}) \sim h, s = 1, 2 \ldots S.$$

Using these samples we can quantify support from the data and prior for any particular hypothesis. For example the probability that the unknown true root time (branching of Indo-European) was smaller than some time $T$ is estimated directly from the proportion of sampled $\mathbf{g}^{(s)}$ in which $t_R^{(s)} < T$. These samples are drawn via MCMC. The software used to carry out these simulations is available from the authors. A user friendly interface has been implemented by David Welch of Auckland University and a user manual, Nicholls & Welch, (2004), is in preparation. MCMC for simulation over a space of trees is now fairly standard. Our MCMC updates are for the most part identical to those of Drummond et al., (2002). The main novelty in our implementation is in the rapid evaluation of $h$ itself, using Equation (3.4) and careful implementation of the defining recursions. Our code has been tested on a number of control problems. One non-trivial test, for the likelihood of a point process density, is the requirement that the likelihood function should sum to one over the data. We fix a tree, and fix birth and death rates, then enumerate all possible data sets with $N = 0, 1, 2 \ldots$ cognates from the empty data set with $N = 0$ up. For each data set we compute $P(D|\mathbf{g}, \mu, \lambda)$. When $\lambda/\mu$ is small (around one or smaller) and the number of leaves is small (up to five), the likelihood for data sets at $N > 7$ is tiny, and we find that $\sum_D P(D|\mathbf{g}, \mu, \lambda)$ converges to one rapidly with $N$ increasing. We checked we could reconstruct synthetic data also. The MCMC convergence analysis depended principally on the visual inspection of traces.

## 4. Sample results

We begin by presenting results from a straightforward analysis of the data, assuming constant death-rate at all times and in all languages and no borrowing. A tree sampled from the posterior is presented in Fig. 2. It must be understood that no one tree can summarize the information available in such a complex data set, in particular the uncertainty. The figure is included simply for illustration, to give readers a feeling for the kind of analysis which was undertaken. We would like to move people away from summarizing a complex data set like Dyen et al., (1997)

| Clade | $\hat{\mu}$ | $(\hat{\sigma}_{\hat{\mu}})$ |
|---|---|---|
| Italic | 0.000268 | (11) |
| Iberian-French | 0.000224 | (21) |
| Germanic | 0.000246 | (13) |
| Balto-Slav | 0.000260 | (36) |
| Brythonic | 0.000176 | (16) |
| Celtic | 0.000299* | (24) |

TABLE 1. The per capita death rate (deaths per year per word), estimated from each of the age constrained clades in turn. There is rate constancy, with the exception of the Brythonic clade. The convention 'Italic 0.000268(11)' reports an Italic word death rate estimated at $\hat{\mu} = 0.000268$ with standard error $\hat{\sigma}_{\hat{\mu}} = 0.000011$. *We are grateful for input at the meeting enabling us to fix an extra rate. We have not otherwise incorporated that information in the analysis presented in this paper.

with a consensus tree with bootstrap confidence labels. The confidence interval for the root time was $[6880, 8170]$ years BP at $3\sigma$, (in a normal approximation to the distribution, which was reasonable).

We would like to know how important model mis-specification is for these data. We ask the following questions. To what extent is rate heterogeneity from one language group to another important? How about rate heterogeneity from one word to another? Is word borrowing important? In order to answer the first questions we estimate the death rate parameter $\mu$ from each of the constrained clades independently. Results are presented in Table (1).

Tocharian has been omitted from this analysis, as it is a two-taxon clade. It is infeasible to estimate rates from two-taxon clades when cognates present in just one language are dropped. Root time estimates are insensitive to mild rate heterogeneity (30 percent in one of 6 clades) of the kind displayed in Table (1). It is possible that more evidence for rate heterogeneity would appear if constraints were available for the Indo-Persian languages. These languages are sampled at a lower density than the western European languages. If there is cladagenic loss (spiking cognate loss at branching events associated for example with the founder effect) in this data then more rapidly branching languages should show more rapid evolution.

There is some evidence for rate heterogeneity from one word to another. In Fig. 3 we simulate synthetic data from the model on a typical tree drawn from the posterior, and plot a histogram showing the distribution of $M_c$, the number of languages containing a given synthetic data-cognate. Compare this graph with the lower graph in Fig. 1. The synthetic data does not typically include values of $M_c$ close to $L$. In other respects the distribution mimics that of the real data reasonably well. Most cognates are in just a few languages, the shoulder at 20 cognates is visible and there is a tail of cognates present in many languages. However, the synthetic data has no cognates present in more than about half the 87 languages, in contrast to the real data.

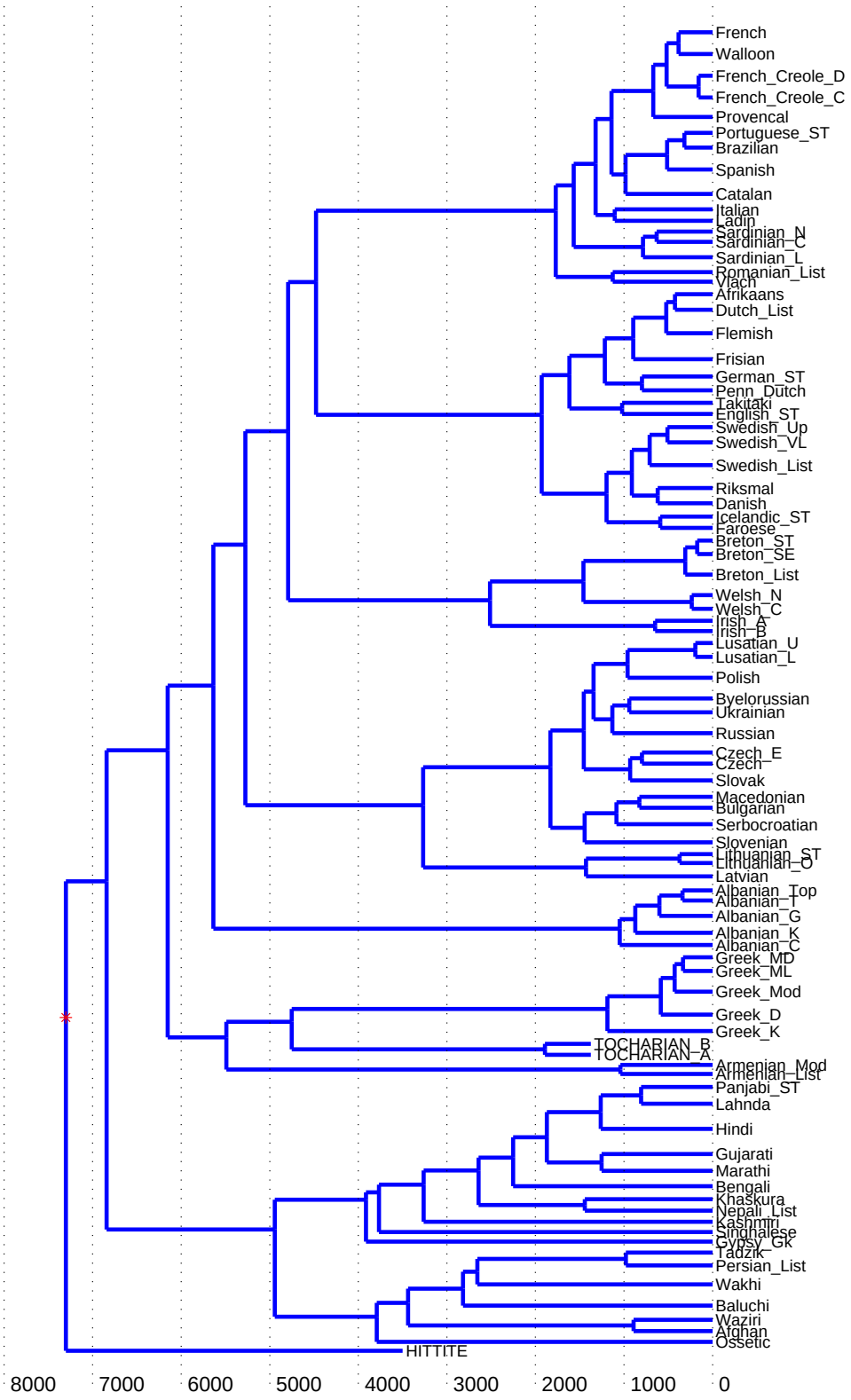


FIGURE 2. It must be understood that no one tree can summarize the information available in the Dyen et al., (1997) data. The tree above does not constitute our 'result'. Above is one Indo-European language tree sampled from the posterior distribution of the data using a prior with a uniform marginal tree root time, and otherwise uniform over trees.

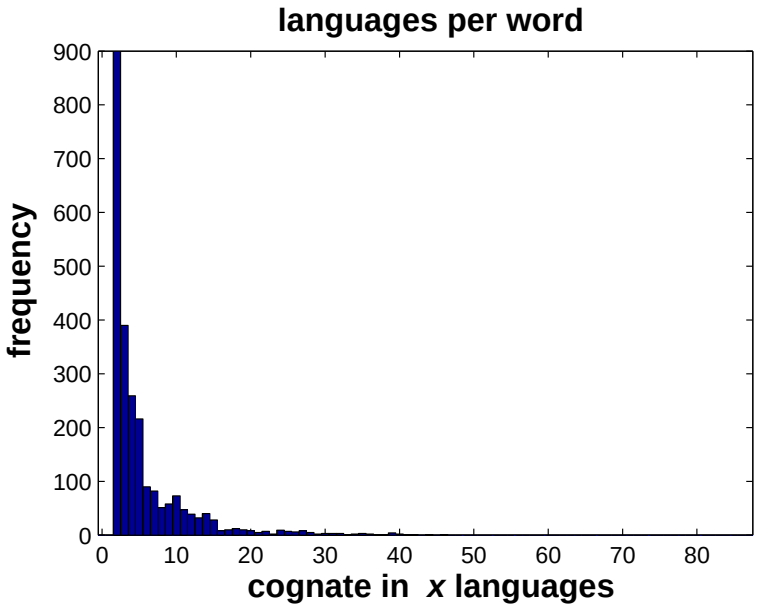


FIGURE 3. The distribution of the number of languages $M_c$ which possess a given synthetic data-cognate. Data simulated on tree sampled from posterior. Compare Fig. 1.

| $b$ | $\hat{t}_R$ | $\hat{\sigma}_{t_R}$ |
|---|---|---|
| 0 | 6720 | 250 |
| 0.1 | 6925 | 225 |
| 0.2 | 6282 | 184 |

TABLE 2. Effect of borrowing on reliability of root time estimates using synthetic data simulated on a tree like that in Fig. 2 with true root time 6900 years BP.

In order to study the effect of borrowing, we simulate synthetic data from our model with borrowing, the birth-borrow-death model, and then ask, what do we estimate when we fit the model without borrowing, the birth-death model? We find that borrowing of the kind described in our model causes us to underestimate the root age. We simulated data with no borrowing, ($b = 0$) and mild borrowing ($b = 0.1$ and $b = 0.2$) on a tree drawn from the posterior distribution (like the tree in Fig. 2) which happened to have true root age 6900 years BP. We wish to explain what we mean by mild borrowing. In the birth-borrow-death model, $b$ is the mean number of times an instance of any particular cognate evolving in a given language, is copied into other languages before it dies in the given language. For small $b$ this is roughly the probability a cognate is borrowed before it dies. Results are shown in Table (2). When there is no borrowing in the synthetic data, and we simulate and fit the same model, the posterior distribution of the root time covers the true

root time. The same applies if there is a little ($b = 0.1$) borrowing. At $b = 0.2$ the borrowing is sufficiently strong to distort the estimate. The estimated root age is too low. The direction of the bias may depend on the details of the borrowing model, see Section 3.2 and the comments associated with Fig. 4.

Estimates are robust to some model mis-specification. Rates are estimated from the same data we are extrapolating. If the nature of the model mis-specification is the same in both the later and earlier time intervals then rate estimates from the later interval can correctly predict time intervals in the early period. This observation applies in particular to mild rate heterogeneity from one cognate to another, if this difference is constant across time and language, and to borrowing. Rate estimates distorted to fit the language evolution where it is given can be good rate estimates for prediction. Mild model mis-specification does not typically cause catastrophic loss of predictive power. Evidence for this is seen in analysis of subsets of languages. Analyzed languages are selected so that the clade time constraints remain appropriate. When we carry out this exercise on synthetic data we find, as we expect, that we recover similar results to those obtained on the full set of languages, but with greater uncertainty. When we make the same study for the real Indo-European dataset we see a similar picture. A tree sampled from the posterior of such a reduced run is shown in Fig. 4. Of the 87 languages in the full data set, 31 are selected for this analysis. The new range is $[6970, 9140]$ at $3\sigma$, double the interval obtained on the full data ($[6880, 8170]$) but in good agreement. However the deathrate parameter $\mu$ is in the range $[0.00017, 0.00024]$ for the reduced data but $[0.00024, 0.00028]$ for the full data, a conflict. The impact of borrowing is weakened by sub-selection of languages (borrowing from words which are no longer in the analysis is not a model violation, it is subsumed in the birth rate $\lambda$). However the estimated root time is insensitive to the reduced model mis-specification as we move from the full to the small data set, because the rate parameters must adjust to accomodate the new picture in the part of the tree where the language evolution is given.

## 5. Discussion

We have described a stochastic generalization of Dollo parsimony. Although the model is a simple and natural model for cognate data, no model of this type has been considered in the past. Because of the technical difficulty of fitting models, authors have tended to avoid ab intitio model specification and fitting.

We have taken into account in our analysis the fact that cognates present in zero or just one of the languages in the data are dropped from the analysis. We have found, in simulation studies which we do not report, that substantial biases can result if this property of the observation process is ignored.

We have sumarized our model mis-specifcation analysis. We find evidence for mild model violation. We have presented evidence that our estimates are robust to violations of the type and degree detected. Our somewhat simplified analysis does tend to support the results presented in Gray & Atkinson, (2003).

## References

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., & Solomon, W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequnce data. *Genetics* 161, 1307–1320.

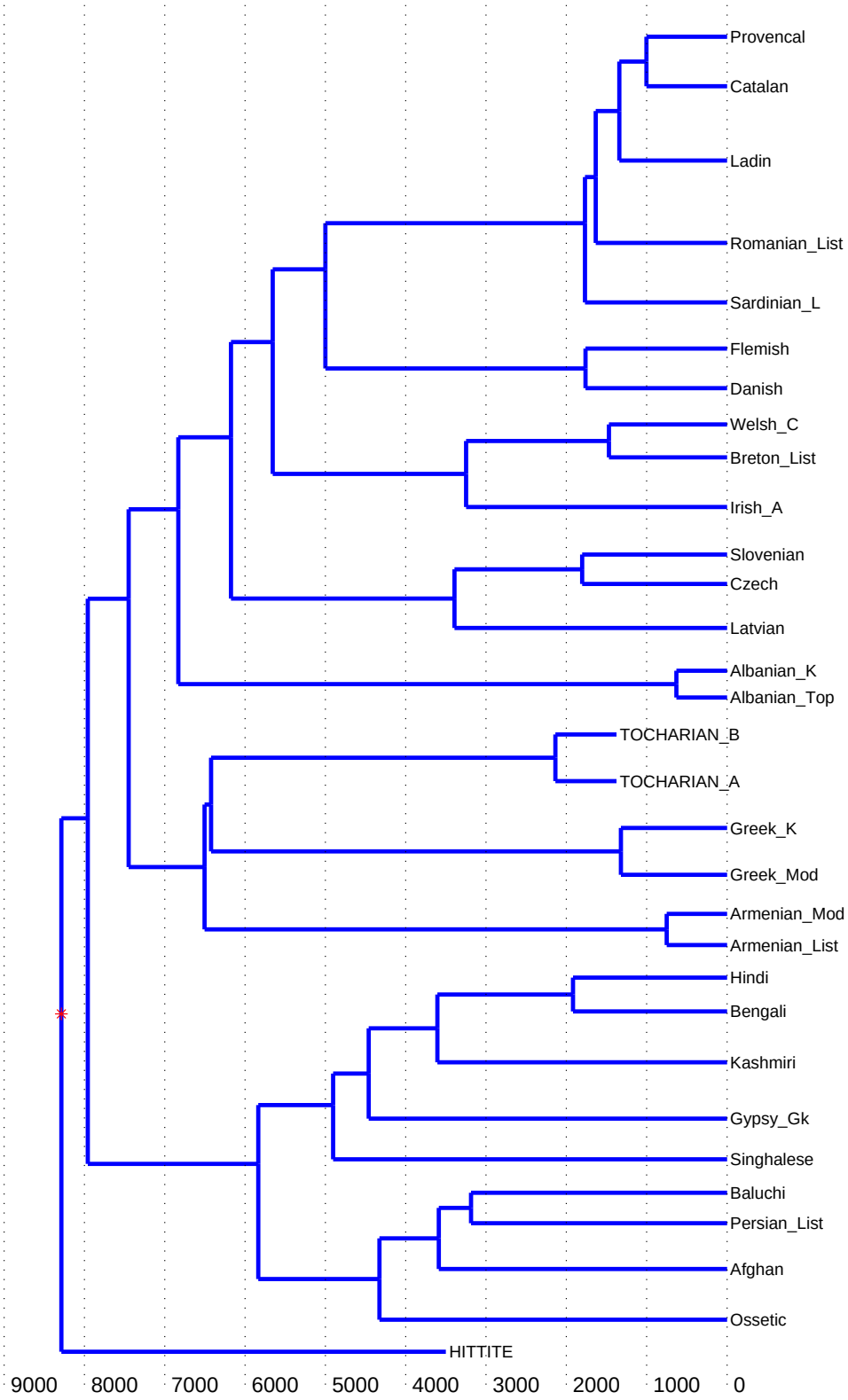


FIGURE 4. It must be understood that no one tree can summarize the information available in the Dyen et al., (1997) data. The tree above does not constitute our 'result'. One language tree sampled from the posterior distribution of a subsample of 31 languages from the 87 in the data. The analysis of a subset of languages gave results consistent with those on the full dataset.

Dyen, I., Kruskal, J., & Black, P. 1997. FILE IE-DATA1, Available at http://www.ntu.edu.au/education/langs/ielex/IE-DATA1.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17, 368–376.

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

Kimura, M. & Crow, J. 1964. The number of alleles that can be maintained in a finite population. *Genetics* 49, 725–738.

Nicholls, G. & Welch, D. 2004. Software manual for MCMC fitting a stochastic Dollo-model to binary character data. In preparation.

Swadesh, M. 1952. Lexico-statistic dating of prehistoric ethnic contacts. *Proc. Am. Phil. Soc.* 96, 453–463.

Watterson, G. 1975. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* 7, 256–276.

Department of Mathematics, The University of Auckland, Auckland, New Zealand
*E-mail address*: nicholls@math.auckland.ac.nz

Department of Psychology, The University of Auckland, Auckland, New Zealand
*E-mail address*: rd.gray@auckland.ac.nz