Calculating the Likelihood

To summarize the calculations, presence/absence information for each cognate set is mapped onto the leaves of a tree. Next, ancestral character states are hypothesized at all internal nodes of the tree. The likelihood of this ancestral-state combination, which depends on substitution rate parameters, is calculated. The likelihood of each tree, for each cognate set, is the sum of the probabilities of all the possible ancestral-state combinations. The overall likelihood of each tree for all of the data can be calculated by taking the product of all the individual cognate-set likelihoods. If a tree has a relatively high likelihood score this means that, given the tree and the model of evolution, the data is a relatively likely outcome. The maximum likelihood (ML) tree is that tree or trees making the data most likely. It is natural to present the age of the Most Recent Common Ancestor (MRCA) of all leaf-languages in this ML tree as the "result" of an analysis. However, there are usually many trees with likelihood scores which are close to the likelihood of the ML tree.

Bayesian inference, MCMC and Estimating Uncertainty

It is easy to show that even when the observation model is an accurate description of cognate evolution there is a very high probability that the true tree will not coincide with the ML tree. For this reason, where feasible, it is preferable to report a confidence interval for the age of the MRCA which takes into account uncertainty in the reconstructed tree. We compute the total support for each possible MRCA age-value by summing the likelihood over all trees with that MRCA age value. For this process to produce a meaningful confidence interval we must weight the trees correctly: we may posses a model of the process which gives rise to branching events in the tree - this model would extend the observation model to include both tree-generation and character evolution; or we may use our subjective knowledge of more and less plausible trees to re-weight the sum of tree-likelihoods. The former case is essentially all observation-model, and is closely related to likelihood-based or frequentist inference; the latter is a variety of Bayesian inference and the weighting function, which represents a prior preference for some trees over others, is called the prior. Where no plausible model of the tree-formation process is available, we are obliged to make Bayesian inference, taking care to consider a range of priors, and choosing priors which represent various degrees of ignorance. Bayes' theorem expresses the posterior probability of a tree (the probability of the tree given the data and subjective prior knowledge) as the product of its likelihood score (the probability of the data given the tree) and its prior probability (a reflection of any prior knowledge about tree topology that is to be included in the analysis). The aforementioned weighted sum of likelihoods of trees of fixed MRCA-age, is therefore just the total posterior probability for that age. However, when we compute this sum we are confronted with an explosion in the number of possible tree topologies. For seven taxa there are 945 possible unrooted trees, for 10 taxa there are over 2 million trees, and for 20 taxa there are over 2 x 10^{20} trees.

Evaluating the posterior probability analytically is almost always impractical. However, we can use Markov Chain Monte Carlo (MCMC) algorithms (Metropolis et al., 1953) to generate a sample of trees in which the frequency distribution of the sample is an approximation of the posterior probability distribution of the trees (Huelsenbeck et al. 2001). In other words, the more likely the tree, the more likely it is to appear in the sample distribution. This sample is used to estimate the sums we need. The MCMC algorithm is typically started from a random phylogeny. The algorithm works by proposing changes to the "current" tree and model parameters and accepting these changes with a probability that depends on the ratio of the proposed and current posterior probabilities. Remarkably, after a "burn-in" period, this algorithm generates trees in proportion to their posterior probability. So, a tree that is twice as probable given the data and the prior information will be sampled twice as often. The limitations of MCMC are well known. It is necessary to check that samples are representative of the posterior distribution. No convenient sufficient condition is available. We make multiple runs from randomly chosen initial conditions, and apply the checks discussed in Geyer (1992). For the synthetic data analyses, more cursory checks were applied (visual inspection of output traces) as the true tree is known, so convergence is not usually in doubt.

Method 1 – Finite-sites Model

The character evolution model

The model used in Gray and Atkinson (2003; and Atkinson and Gray, in press[a, b]) is based on a restriction-site model of binary character evolution implemented in the programme *MrBayes* (Huelsenbeck and Ronquist, 2001). As described above, the binary characters in the data matrix represent the presence (1) or absence (0) of a particular cognate in a particular language. We can model the process of cognate gain and loss using a rate matrix representing the relative probabilities of all possible character-state changes. The rate matrix in Table 1 shows a simplified version of the Gray and Atkinson model. Each cell represents the relative probability of gaining or losing a cognate in any given time period. The model parameters are μ, the mean substitution rate, and π0 and π1, which represent the relative frequencies of 1's (π_1; cognate present) and 0's (π_0; cognate absent). The substitution rate is a parameter estimated in the MCMC analysis and the equilibrium frequency of 1's and 0's can be estimated from the frequency of 1's and 0's in the data.

Table 1 – Rate matrix used for modelling lexical replacement in language evolution. This time-reversible model allows for unequal equilibrium frequencies of 1's and 0's (cognate presence and absence). The model parameters are μ (the mean substitution rate), and π_0 and π_1 (which represent the relative frequencies of 1's and 0's in the data matrix).

	1	0
1	$-\mu\pi_0$	$\mu\pi_0$
0	$\mu\pi_1$	$-\mu\pi_1$

Because there are just two states, 0 and 1, this model is trivially time-reversible – if we follow a particular cognate over a long time evolving within a single language, the number of times we see the state of the cognate change from 1 to 0 and 0 to 1 will be equal . We cannot tell the direction in which the cognate evolved from its history in a single language. This model allows a single cognate to appear in and disappear from a single language more than once over the course of time, allowing the model to mimic the effect of word-borrowing. The direction of time is not determined in a reversible model. As a consequence, we cannot determine the root of the tree from the data – we need to provide an outgroup as a root. For all the method 1 analyses reported here, trees were rooted with Hittite, consistent with independent linguistic analyses (Gamkrelidze and Ivanov, 1995; Rexova, Frynta and Zrzavy, 2003). The method of Nicholls & Gray (in press) predicts a Hittite outgroup. Moreover, Atkinson and Gray (in press[a, b]) found that the root point did not in any case affect age estimates significantly. Note that the lifetime 1/ $\mu\pi_1$ of the “absent” or 0-state (controlled by the rate $u\pi_1$ at which a 0 becomes a 1) need not equal the lifetime 1/ $\mu\pi_0$ of the “present” or 1-state (controlled by the rate $\mu\pi_0$ at which a 1 becomes a 0) for a cognate in a time-reversible model.

A Gamma shape parameter was also added to allow for rate variation between words. The Gamma distribution provides a range of rate categories for the model to choose from when assigning rates to each cognate set. The distribution of these rates is determined by the Gamma shape parameter (α). α can range from 0 to ∞. For small values of α, most cognate sets evolve slowly, but a few can evolve at higher rates. As α increases, the distribution becomes more peaked and symmetrical around a rate of 1 – i.e. rates become more equal (Swofford et al., 1996). As with the overall rate parameter, α was estimated from the data. An α value of 5 was observed, indicating moderate rate variation.

Tree-building

Method 1 uses *MrBayes* (Huelsenbeck and Ronquist, 2001) to perform Bayesian inference of phylogeny. *MrBayes* uses MCMC algorithms to sample trees distributed according to the posterior computed from the Method 1 observation model. After an initial ‘burn-in’ period, trees are sampled in proportion to their likelihood given the data. Each analysis generated 1.3 million trees from a random starting phylogeny. On the basis of an autocorrelation analysis only every 10,000th tree was sampled to ensure that consecutive samples were reasonably independent. A burn-in period of 300,000 trees for each run was used to avoid sampling trees before the run had reached convergence. Log-likelihood plots and an examination of the post burn-in tree topologies using the TreeSet Visualization module (Klinger, 2002) for Mesquite

(Maddison & Maddison, 2002) demonstrated that the runs had indeed reached convergence by this time. Most analyses were repeated 10 times from different random starting trees to produce a total of 1000 trees in each sample, all rooted with Hittite. The branch between Hittite and the rest of the tree was split at the root such that half its length was assigned to the Hittite branch and half to the remainder of the tree - divergence time estimates were found to be robust to threefold alterations of this allocation.

Estimating Dates

A likelihood approach allows us to account for rate variation between words using a Gamma distribution. We can also account for rate variation between lineages and through time by relaxing the assumption of a strict glottoclock. Rate-smoothing algorithms from biology attempt to model rate variation across a phylogeny and thus estimate divergence times without assuming constant rates. One such approach is the "penalized-likelihood" model (Sanderson 2002a) of rate smoothing, which allows for rate variation between lineages while incorporating a "roughness penalty" that costs the model more if rates vary excessively from branch to branch. In a biological context, Sanderson (2002a) has shown that the penalized-likelihood optimization procedure performs significantly better under conditions of rate heterogeneity than procedures that assume a constant rate of evolution.

We can apply the same methods to linguistic data. Using *MrBayes* we produced a distribution of trees with branch-lengths proportional to the inferred amount of evolutionary change. Known divergence times based on historically attested dates were then used to calibrate rates of change across each tree. For example, we know from historical information that the Anglo-Saxons began to settle in Britain in A.D. 449. This would suggest that the English lineage split from the other West Germanic languages at some point during the fifth century A.D. We can constrain the age of this node on the tree accordingly. Similarly, we can constrain the age of extinct languages based on dates associated with the various source texts. For example, we know that Hittite was spoken between 3,200 and 3,700 years ago and we can constrain the age of this node on each tree.

The 87 languages in the modified Dyen et al. (1997) data set allowed for 11 internal clade constraints (see appendix I). Terminal nodes representing contemporary languages were set to 0 years whilst 3 extinct languages (Hittite and Tocharian A & B) were constrained in accordance with estimated ages of the source texts. For the 24 languages in the Ringe et al. (2002) data, 12 internal node constraints were available, whilst 20 extinct languages were constrained in accordance with estimated ages of the source texts (see appendix II). Sanderson's (2002a) penalized-likelihood algorithm, as implemented in *r8s* (Sanderson, 2002b), was then used to smooth rates of evolution across each tree and to calculate divergence times. This procedure was repeated on all of the trees in the MCMC Bayesian sample distribution. Interestingly, high smoothing factors were found to fit the data best, suggesting that the process of evolution is in fact relatively tightly constrained. The result is a distribution of age estimates for various Indo-European language divergence events. The distribution of divergence times at the root can be used to create a confidence interval for the age of Indo-European.

Method 2 – Stochastic-Dollo Model

Dollo's Law states that traits can evolve only once (Farris, 1977). In this context, we treat cognates as traits and assume that the same cognate cannot be independently created in different languages (through time or space). This assumption is equivalent to asserting that the cognate data is homoplasy free (c.f. Ringe et al., 2002). Based on this assumption, we outline a stochastic model of language change appropriate to the cognate data described in section 3.

The model allows language change to occur in three different ways: words (and corresponding cognates sets) are created, words are lost, and words reproduce (when languages split, forming two child copies of a parent language). We assume that words are created in any given language at rate λ. When a word is created, it falls into a new cognate class, so word creation and cognate class creation are synonymous. If there are k languages extant at time t, new cognates are created at total rate $k\lambda$.

Each word is lost from a given language independently at rate μ. If at time t, there are k languages and language i contains l_i words, word death occurs at a total rate of $\mu(l_1+l_2+\ldots+l_k)$.

Each language splits at rate θ. When a language splits, two child copies of the language are made and the parent language dies. At the time of splitting, the child languages are indistinguishable from the parent language and thereafter evolve in exactly the same way as the parent language did. If there are k languages at time t, language splitting occurs at total rate $k\theta$.

We assume that the times between all events causing language change are exponentially distributed and that all rates – the cognate birth rate, λ, the cognate loss rate, μ, and the language splitting rate, θ – are constant across time and space. We assume also that all languages and cognates evolve independently.

The data described in section 2 is collected in such a way that cognates which are present in no languages or only one language at the time of collection are not recorded. Thus the observed cognate birth rate λ^* is different from the actual cognate birth rate λ since words must be born and survive into at least two languages in order to be observed. This data thinning process may result in the birth times of cognates in the data being skewed heavily towards the leaves of the tree. This effect is accounted for in the likelihood calculation for a given tree, the details of which are given in Nicholls and Gray (in press).
Inference for the Stochastic-Dollo model is made within a Bayesian framework and the data is analysed using a MCMC algorithm implemented in Matlab by two of the authors (GN and DW). The relevant software, called TraitLab, can be downloaded from (aitken.math.auckland.ac.nz/~nicholls/TraitLab/).

Age constraint Tables

Tabel 2 Dyen et al. data age constraints - from Gray and Atkinson (2003) - Age constraints for the Dyen et al. (1997) data set, used to calibrate the divergence time calculations on the basis of known historical information. Terminal node constraints representing ancient languages are shown in italics.

Calibration	Age constraint
Iberian-French	450AD-800AD
Italic-Romanian	150AD-300AD
North/West Germanic	50AD-250AD
Welsh/Breton	400AD-550AD
Irish/Welsh	before 300AD
Indic	before 200BC
Iranian	before 500BC
Indo-Iranian	before 1,000BC
Slavic	before 700AD
Balto-Slavic	1,400BC-100AD
Greek split	before 1,500BC
Tocharic	140BC-350AD
Tocharian A & B	*500AD-750AD*
Hittite	*1,800BC-1,300BC*

Table 3 - Ringe et al data age constraints - Age constraints for the Ringe et al. (2002) data set, used to calibrate the divergence time calculations on the basis of known historical information. Terminal node constraints representing ancient languages are shown in italics.

Calibration	Age constraint
Italic	before 800BC
Germanic	750BC-250BC
North-West Germanic	50AD-250AD
West Germanic	400AD-500AD
Celtic	650BC-300AD
Indic	before 200BC
Iranian	before 500BC
Indo-Iranian	before 1,000BC
Baltic	600AD-700AD
Balto-Slavic	1,400BC-400BC
Greek split	before 1,500BC
Tocharic	140BC-350AD
Vedic	*1,500BC-800BC*
Old Persian	*600BC-300BC*
Avestan	*600BC-400BC*
Old Prussian	*1,250AD-1,600AD*
Old Chruch Slavonic	*900AD-1,100AD*
Old High German	*850AD-1,050AD*
Old English	*900AD-1,100AD*
Old Norse	*1,150AD-1,350AD*
Gothic	*300AD-400AD*
Armenian	*400AD-800AD*
Greek	*500BC-300BC*
Latin	*200BC-100AD*
Oscan	*400BC-50BC*
Umbrian	*300BC-50BC*
Old Irish	*600AD-900AD*
Tocharian A & B	*500AD-750AD*

Lycian	*500BC-200BC*
Luvian	*1,700BC-1,200BC*
Hittite	*1,700BC-1,200BC*

Ringe et al (2002) Consensus Network

One problem with using consensus trees, is that they cannot display the strength of evidence for conflicting clades. For example, we may be able to show that the Germanic languages group with the Celtic languages 42% of the time, but we cannot simultaneously show that 38% of the time the Germanic languages group with the Italic or Balto-Slavic languages, even though this may be very interesting. One way of summarizing a distribution of trees without losing this information is to display conflicting clades or 'splits' simultaneously. We can do this using consensus networks (Holland and Moulton, 2003). Figure 1 shows the RF1 distribution of trees summarized as a consensus network displaying all those clades with greater than 10% support. Each edge or 'split' separating one set of languages from another corresponds to a clade. This clearly shows the lack of resolution at the base of the tree – the box like structures in the centre of the figure indicate incompatible clades in the sample distribution of trees. This picture is consistent with acknowledged uncertainties, such as the position of Albanian. It is this uncertainty in the branching structure that we can integrate out by estimating divergence times across the sample distribution of trees.

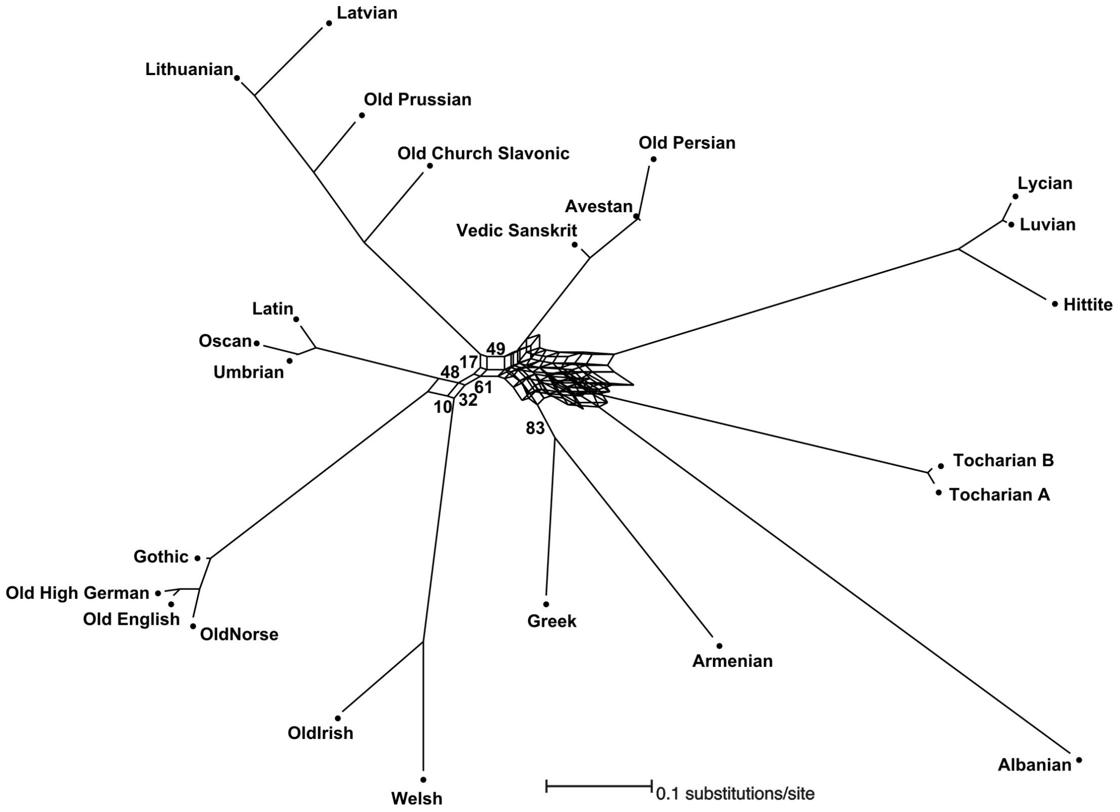

Figure 1 - Consensus network from the initial Bayesian MCMC sample of 1,000 trees based on the Ringe et al (2002) data, constructed using SplitsTree (Huson, 1998). Values express percentage support for some of the splits. A threshold of 10% was used to draw this splits graph – i.e. only those splits occurring in at least 10% of the observed trees are shown in the graph. Branch lengths represent the median number of reconstructed substitutions per site across the sample distribution.

References

Atkinson, Q. D. and R. D. Gray. in press[a]. Are accurate dates an intractable problem for historical linguistics? In *Mapping our Ancestry: Phylogenetic Methods in Anthropology and Prehistory.* (eds.) C. Lipo, M. O'Brien, S. Shennan & M. Collard. Chicago: Aldine.

Atkinson, Q. D. and Gray, R. D. (in press[b]). How old is the Indo-European language family? Illumination or more moths to the flame? In *Phylogenetic methods and the prehistory of languages* Eds. J. Clackson, P. Forster and C. Renfrew. MacDonald Institute (Cambridge).

Dyen, I., J. B. Kruskal, and P. Black. 1992. *An Indoeuropean Classification: A Lexicostatistical Experiment.* American Philosophical Society, Transactions 82(5). Philadelphia.

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

Farris, J. S. 1977. Phylogentic analysis under Dollo's Law. Systematic Zoology, 26:77-88.

Gamkrelidze, T. V., and V. V. Ivanov. 1995. Indo-European and the Indo-Europeans: A Reconstruction and Historical Analysis of a Proto-Language and Proto-Culture. Mouton de Gruyter, Berlin.

Geyer, C.J. 1992. Practical Markov chain Monte Carlo. *Statistical Science,* 7, 473-511

Gray, R. D. and Q. D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426, 435-9.

Holland, B., and V. Moulton. 2003. Consensus networks: a method for visualising incompatibilities in collections of trees. Pp. 165–176 *in* G. Benson and R. Page, eds. Algorithms in bioinformatics, WABI 2003. Springer-Verlag, Berlin, Germany.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology. *Science* 294:2310–2314.

Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian Inference of Phylogeny. *Bioinformatics* 17:754–755.

Huson, D. H. 1998. SplitsTree: Analyzing and visualizing evolutionary data. Bioinformatics 1.4:68–73.

Klingner, J. (2002). *Tree Set Visualization*, version 2.0. http://www.cs.utexas.edu/users/phylo/

Maddison, W. P. & Maddison, D.R. (2002). *Mesquite: a modular system for*

Evolutionary analysis. Version 0.992. http://mesquiteproject.org

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics* 21:1087–1091.

Nicholls, G., and R. Gray. in press. Quantifying uncertainty in a stochastic dollo model of vocabulary evolution. *In* Phylogenetic methods and the prehistory of languages. (J. Clackson, P. Forster and C. Renfrew. ed.). McDonald Institute for Archaeological Research, Cambridge.

Rexová, K., D. Frynta, and J. Zrzavy. 2003. Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics* 19, 120–127.

Ringe, D., T. Warnow, and A. Taylor. 2002. Indo-European and Computational Cladistics. *Philological Society, Transactions* 100:59–129.

Sanderson, M. 2002a. Estimating absolute rates of evolution and divergence times: A penalized likelihood approach. *Molecular Biology and Evolution* 19, 101–109.

Sanderson, M. 2002b. R8s, Analysis of Rates of Evolution, version 1.50. http://ginger.ucdavis.edu/r8s/

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic Inference. In *Molecular Systematics*, second ed., edited by D. M. Hillis, C. Moritz, and B. K. Marble, pp. 407-514. Sinauer, Sunderland, Mass.