1 *Austronesian language phylogenies: myths and misconceptions about Bayesian computational methods*

SIMON J. GREENHILL and RUSSELL D. GRAY [1]

Historical linguistics has never been particularly intimate with computers. The first wave of computational historical linguistics—lexicostatistics—was developed in the 1950s (Swadesh 1952; Lees 1953) and quickly applied to language groups around the world from Indo-European to Austronesian (Lees 1953; Hymes 1960; Embleton 1986). However, critics were quick to point out the problems caused by assuming a single constant rate of lexical replacement and repeatedly noted the erroneous results that this produced (Hoijer 1956; Bergsland and Vogt 1962; Blust 1981; McMahon and McMahon 2006). As a consequence of these critiques lexicostatistics has been widely rejected by mainstream historical linguists (Campbell 2004).

The last few years have seen a second wave of computational approaches entering historical linguistics: phylogenetic methods. These techniques, drawn from evolutionary biology, have been used to investigate some provocative and controversial claims about human prehistory. For example, we have applied phylogenetic methods to lexical data compiled by Bob Blust to test hypotheses about the settlement of the Pacific (Gray and Jordan 2000; Greenhill and Gray 2005; Gray et al. 2009). Our results reflected a settlement pattern through Island South-East Asia, New Guinea and then into Oceania, consistent with the 'Out of Taiwan' scenario (e.g. Blust 1999; Pawley 2002; Diamond and Bellwood 2003). We have also used these methods to investigate the origins of the Indo-European (Gray and Atkinson 2003) and Bantu languages (Holden 2002; Holden and Gray 2006). Other groups have applied phylogenetic methods to investigate the internal subgrouping of these families (Ringe et al. 1998; Rexovà et al. 2003, 2006). The application of computational phylogenetic methods has not been restricted to just lexical data. Phylogenetic analyses of structural features have revealed historical signals in Papuan

[1] We would like to thank Andreea Calude, Andy Pawley, and Malcolm Ross for comments on this paper. We would like to note that some of the analyses reported in this paper have been superseded by those conducted using a better fitting model of cognate evolution reported in Gray et al. (2009).

Alexander Adelaar and Andrew Pawley, eds.
Austronesian historical linguistics and culture history: a festschrift for Robert Blust, ???–.
Pacific Linguistics, 2009.

languages that may stretch back around 10,000 years (Dunn et al. 2005). Nor have phylogenetic methods been restricted to just building trees. Phylogenetic network methods have been used to investigate conflicting signals in Indo-European (Bryant et al. 2005), Bantu (Holden and Gray 2002), Chinese dialects (Hamed and Wang 2004), and Polynesian (Bryant 2006; Gray 2007). Finally, phylogenetic methods have recently been used to investigate general claims about the factors that affect the rate of language change. Pagel et al. (2007) used phylogenetic methods to estimate the rates of lexical replacement in Indo-European languages and showed an almost hundred-fold difference between the rates of rapidly evolving words (e.g. 'dirty') and the slowly evolving words (e.g. 'tongue'). They then calculated the frequency at which these words were currently used in four large language corpora. Their results showed a strong correlation between the frequency with which words are used today and their stability over time: the more a word is used, the slower it evolves. This striking result suggests that over the 9000 years of Indo-European language history, there have been consistent underlying mechanisms controlling lexical replacement. A second study (Atkinson et al. 2008) used phylogenetic methods to test claims that speakers often use their language as a social tool for increasing group cohesion and demarcating groups (Labov 1994). The results showed a strong relationship between the total amount of lexical change and the number of language splitting events along the tree: between 10% to 33% of the total lexical change in the Bantu, Indo-European, and Austronesian languages occurred as a rapid burst of change shortly after languages diverged. This punctuational change (e.g. Bowern 2006) is consistent with rapid language change in small founder populations and differentiation as a cultural marker.

Given the combination of strong claims, new techniques, and the high-profile reporting of results, it is not surprising that these studies are often controversial. Responses have ranged from the positive: 'Computational methodologies of this kind can only be helpful for historical linguistics' (April McMahon in Balter 2003:1491), to the skeptical: 'There is no reason whatsoever to assume that vocabulary would behave the same way that organisms do.' (Alexander Lehrman in Balter 2004:1326), to the negative: '… have ignored the fatal shortcomings of glottochronology …' (Eska and Ringe 2004:569), and the painfully incorrect; 'sledg(ing) the dead horse of the Swadesh algorithm' (Holm 2007:201).

Sadly many of these criticisms are mired in misunderstanding. Computational phylogenetic methods are not just lexicostatistics redux, but a powerful supplement to the comparative method used in historical linguistics. On several occasions Bob Blust has challenged us to specify exactly how phylogenetic methods differ from lexicostatistics and explain why they are superior. Here we respond to his challenge. To do this, we will focus on one of the great battlegrounds between lexicostatistics and the traditional comparative method: the Austronesian language family. First, we will describe how Bayesian phylogenetic methods work, and then give a step-by-step explanation of an analysis of a large lexical dataset for 400 Austronesian languages (Gray et al. 2009; Greenhill et al. 2008).

1 The Austronesian language family

The Austronesian language family is one of the two largest in the world, containing around 1000 to 1200 languages (Gordon 2005). Before Columbus, these Austronesian languages were also the most widely dispersed with speakers in Mainland and Island South-East Asia, Madagascar, Micronesia, Melanesia, and Polynesia (Bellwood et al. 1995). The groundwork that identified this family began in the 16th and 17th centuries as European scholars began to compare word lists that trickled back from early explorers and

missionaries (e.g. Houtman 1603; Reland 1708; Forster 1778; Brandes 1884; Kern 1886). Dempwolff (1934, 1938) systematically reconstructed early Austronesian phonology and lexicon, and identified a large subgroup, Oceanic, to which he assigned the languages of Melanesia, Polynesia and (most of) Micronesia (Dempwolff 1937). The evidence that all these Oceanic languages formed a subgroup of Austronesian implied that they stem from a single Austronesian settlement of this region from the west (Grace 1961, 1964a; Pawley and Green 1973; Pawley and Ross 1995).

A major challenge to this hypothesis came from Dyen's lexicostatistical analyses of vocabulary from 352 Austronesian languages (1962, 1965). Lexicostatistics had previously been applied to subgroups within Austronesian (an early paper by Elbert (1953) explored Polynesia), but Dyen's was by far the largest in scale. At the time, Dyen's analysis was an impressive computational feat; his program compared 7,000,000 pairs of words. The lexicostatistical results suggested a tree with 40 first-order branches, no fewer than 30 of which were located in Melanesia. Dyen took this to indicate that the most probable area of origin of the Austronesian languages was in Melanesia, possibly in the Bismarck Archipelago north of New Guinea, with subsequent expansions east into Polynesia, and west into Indonesia then to the Philippines and Taiwan. This study was hailed by Murdock (1964:117) as '… a significant work—one which may conceivably be as revolutionary for Oceanic linguistics and culture history as was the work of Greenberg (1949–54) for the interpretation of African languages and cultures'.

This enthusiasm was short-lived. Grace (1964b, 1966) was quick to to suggest that the difference between the lexicostatistical view of Austronesian relationships and that of the traditional view may be a consequence of faster rates of lexical replacement in Melanesia. Blust (1981, 2000) quantitatively demonstrated that the Austronesian languages varied markedly in their retention rates across a 200-item basic vocabulary word-list. Retention rates in Malayo-Polynesian languages ranged from 5% to 60% in the interval between Proto Malayo-Polynesian and the present, a time period of around 4000 years. Moreover, Blust (2000) argued that the inability of lexicostatistics to discriminate between shared retentions and innovations—a distinction that had been critical in historical linguistics since Brugmann (1884)—exacerbated the effect of different rates. These differences in retention rates, especially in regions such as Melanesia where there have been high levels of language contact and borrowing (Ross 1996) rendered the lexicostatistical conclusions invalid.

In contrast to a Melanesian origin for Austronesian languages suggested by lexicostatistics, the comparative method has provided strong evidence that all languages outside Taiwan belong to a single sub-group (Dahl 1973; Blust 1977), which Blust (1977) named Malayo-Polynesian. In a series of publications Blust (e.g. 1977, 1978, 1982, 1999) marshalled a large array of evidence for the claim that the Proto Austronesian (PAn) homeland lay in Formosa (Taiwan). First, Blust (1999) concluded there are at least nine primary subgroups of Austronesian within Taiwan, whereas all Austronesian languages spoken outside of Taiwan fall into a single first order subgroup. There are a number of phonological and morphological innovations that are shared by the Malayo-Polynesian subgroup but are not found in the Formosan languages. If we assume that the region with the most primary subgroups is likely to be the primary dispersal centre Taiwan is thus strongly favoured as the Austronesian homeland. Blust (1982) also used the distribution of flora and fauna lexicon to delimit the range of possible Austronesian homelands. The distribution of cognate words for placental and marsupial mammals in Austronesian languages suggests that ancestral Austronesian society was located in the Asiatic faunal

zone to the west of the Wallace line. Archaeological evidence indicates that the spread of Neolithic cultures from Taiwan parallels the directions and dates of the Austronesian linguistic expansion. This conjunction of different lines of evidence has convinced most specialists in Austronesian historical linguistics that the Austronesian-speaking people were present in Taiwan around 5500 years ago, before spreading into the Philippines, Indonesia and through the Pacific (e.g. Shutler and Marck 1975; Bellwood 1997; Blust 1995; Kirch 2000; Kirch and Green 2001; Pawley 2002).

The failure of lexicostatistics to get Austronesian 'right' is not surprising—computing Austronesian language relationships is a very difficult problem. First, the rapid expansion of the Austronesian family means that it is likely to be difficult to resolve the fine branching structure of the Austronesian language tree as there is little time for the internal branches on the tree to develop numerous shared innovations (Pawley 1999). Second, as these languages moved across the Pacific they encountered new environments and the consequent need for new terminology may have increased the rates of language replacement. This acceleration in rates is likely to be exacerbated by the effects of language contact—particularly within Near Oceania (Ross 1996). Additionally, many Austronesian languages have small speech communities, which are also likely to speed up the rates of language evolution (Nettle 1999). The effects of these factors can be seen in the substantial variation in cognate retention rates in Austronesian languages (Blust 1981, 2000; Pawley this volume). Finally, the sheer scale of the Austronesian language family is daunting—with around 1000 to 1200 languages there are more than 10^{2864} possible rooted family trees. In the following section we will outline a Bayesian phylogenetic analysis on the Austronesian languages.

2 A phylogenetic approach

Much of biology and linguistics is historical. That is, to understand these systems properly we need to know their history. Where did particular languages or species come from? When did they arise and diverge? What sequence of changes took place? Are two characteristics similar because they share common ancestry or are they similar because they've evolved to fill the same function? To investigate these questions biologists have developed a large collection of tools collectively known as phylogenetics. Biologists initially constructed phylogenetic trees with clustering algorithms such as UPGMA ('Unweighted Pair-Group Method using Arithmetic averages', Sneath and Sokal 1963), that analysed pairwise similarity matrices (just like the lexicostatistical percentage shared cognacy matrices). Not surprisingly, this approach also produced inaccurate results when there were substantial differences in the rates of genetic change between lineages (Felsenstein 1978). However, rather than abandon a computational approach when confronted with this difficulty, biologists improved the computational methods. In the last few decades phylogenetic methods have revolutionised biology and have become the dominant way of testing historical evolutionary hypotheses (Huelsenbeck and Rannala 1997; Pagel 1999). Currently, the Bayesian phylogenetic approach is seen as the most powerful and robust approach available (Lewis 2001; Huelsenbeck et al. 2001, 2002). In the section below we will outline the major components of Bayesian phylogenetic analysis: dataset construction, maximum likelihood modeling, and the search for the most probable evolutionary trees.

2.1 Data

For successful phylogenetic analysis we need a large amount of well-sampled data with sufficient historical information to resolve the aspects of the phylogeny we are interested in. The comparative method commonly used in historical linguistics takes a sample of lexicon and proceeds to reconstruct systematic sound correspondences between the languages in order to uncover historically related 'cognate' forms (Durie and Ross 1996). This information about cognate sets can easily be coded as binary characters. An example of this is shown in Table 1. The data, in this case the words meaning 'bone' in a number of Austronesian languages (Column A) are divided into cognate sets on the basis of systematic sound correspondences (Column B). Once the cognate sets have been determined and any known loan words removed, then the data can be coded into a binary matrix showing the presence or absence of each cognate set for every language (Column C). In the 400-language dataset used in this paper, the cognate sets in a 210-item word-list produced 34,440 binary characters.

It is worth emphasising that whilst most recent work computing language phylogenies has primarily been based on cognate datasets (e.g. Atkinson et al. 2008; Gray and Atkinson 2003; Gray and Jordan 2000; Greenhill and Gray 2005; Gray et al. 2009; Holden 2002, Holden and Gray 2006; Pagel et al. 2007; Rexovà et al. 2003), other linguistic characters could also be used as long as there is sufficient data and an appropriate way of modeling the changes in these characters. Indeed, some studies have used combinations of lexical and grammatical data (Rexovà et al. 2006) and typological information (Dunn et al. 2005).

Table 1: Cognate data coding from original lexical data (A), to cognate set information (B), to binary characters (C)

Language	(A) Item	(B) Cognacy	(C) Binary Coding
Paiwan	tsuqela	1	1000
Itbayaten	tuqgan	1	1000
Bare'e	wuku	2	0100
Mangarrai	toko	2	0100
Numfor	kor	3	0010
Motu	turia	4	0001
Fijian (Bau)	sui-na	4	0001
Tongan	hui	4	0001
Samoan	ivi	4	0001
Maori	iwi	4	0001

2.2 Maximum likelihood models

The next step is to analyze the data. Bayesian phylogenetic inference builds on an older tradition of Maximum Likelihood methods (Fisher 1922; Edwards 1964; Felsenstein 1981; Pagel 1999). In this framework the data is treated as a fixed and given observation, and the analysis aims to find the values of model parameters that explain this data well (Pagel 1999; Steel and Penny 2000). To do this we need a stochastic model of language evolution that specifies how the changes between the character states should be counted. In modeling

language evolution in this way we make simplifying assumptions about relevant processes and explicitly build these into the model. For example, a very simple model of lexical evolution would require one parameter—the rate of change between the absence of a specific cognate and the presence of that cognate. In this simplest model, this rate would be symmetrical in the sense that the rate at which any cognate was gained would be equal to the rate at which a cognate was lost. Obviously, this is not very realistic. Once a cognate set has arisen it is much more likely to be lost than for another language to independently derive it. A more realistic model would accommodate the differential ease of losing a cognate over gaining it by adding a second parameter, so there is now one rate for cognate gain and one rate for cognate loss (we will refer to this as the two-parameter mode below).

What other important parameters could be added? One of the major problems with lexicostatistics is that it assumed a constant rate of cognate loss of around 19% every thousand years in the 200-item Swadesh list (Lees 1953). This fixed rate did not allow for differences in rates of change between cognate sets, or for differences in rates of change between languages. Both of these types of rate variation are common in Austronesian languages (Blust 1981, 2000). Site-specific rate heterogeneity (different sites in DNA sequences evolving at different rates) was also a problem for early phylogenetic methods (Posada and Crandall 2001). More recent approaches, however, have solved this by enabling a distribution of rates instead of a single rate. One common method is to estimate a gamma distribution of rate changes from the data (Yang 1994). This method gives each character an *inherent* rate of change so that some cognates are gained or lost rapidly, whilst others are more resistant to change. Modeling lexical change in this way allows for the differences between highly persistent characters like reflexes of 'hand—Proto Austronesian **(qa)-lima* (Blust 1999)—and highly unstable characters such as words meaning 'dirty'.

The full model with two rate parameters and gamma-distributed rate heterogeneity can then be used to calculate a numeric value known as the likelihood. The likelihood measures how well the data are explained by the tree under this model. Our aim is to find the set of trees that explain the data well, or in other words, find those trees with the maximum likelihood. The general approach to finding trees here is to take a tree and then permute it in some fashion (e.g. by changing the tree shape, or the amount of change along a branch, or model parameters, etc.) to give a second tree. The likelihood of both those trees under the given model of language evolution can then be compared to find the better tree. Here the search algorithm (usually a Markov Chain Monte Carlo approach—described below), preferentially selects the tree with the better likelihood, and iterates over this procedure many times, to find a set of good trees.

Critics of early language studies using Bayesian phylogenetic methods claimed that the models were 'inappropriate' as they had been designed for biological analyses rather than linguistic change (Eska and Ringe 2005; Naklekh et al. 2005). This criticism demonstrates a misunderstanding of the rationale behind model-based inference. Whilst it is true that language change is complex, and the model employed here and elsewhere (e.g. Gray and Atkinson 2003; Gray et al. 2009) is simple, this simplicity does not necessarily discredit or invalidate the methodology. Developing a model is a trade-off between over- and under-fitting model parameters (Burnham and Anderson 1998). Typically the fit will improve as parameters are added to the model, especially if the new parameters capture an important aspect of the process. More complex models are not uncommon in biology; one of the most popular models used for genetic data is the General Time-Reversible model (Yang et al.

1994). This model has six parameters: one for each of the rates of change between each combination of the four bases found in DNA. This is often coupled with gamma distributed rate heterogeneity, and an allowance for invariant sites, giving a total of eight parameters.

However, as parameters are added the sampling error also increases and therefore it becomes difficult to reliably estimate the model parameters (Swofford et al. 1996). Therefore, the goal of modeling language evolution is not to build a complex model that captures every aspect of language change, but rather to construct the simplest model that provides reliable estimates of the parameters with finite amounts of data. Choosing the most appropriate model is not an issue for armchair speculation. We can evaluate the performance of the model by analysing the data with a range of models, and then selecting the best model with a standard model comparison test such as the Likelihood Ratio Test (Goldman 1993), or Bayes Factor Comparison (Suchard et al. 2001).

2.3 An example of a likelihood calculation

To clarify the way in which likelihood scores are calculated we have outlined a simple example in Figure 1 (adapted from Swofford et al. 1996, and Atkinson and Gray 2006). This figure shows the basic procedure on a set of data coded in a binary matrix as described above (1A). We will follow the process of likelihood calculation for one of these characters: character ‘a’. Character ‘a’ represents a cognate set found in the Oceanic languages Motu, Fijian, Samoan and Maori, and absent from the other languages in our example dataset. To show how the likelihood can measure how well a topology describes the data we will compare two different trees (1B). The tree on the left represents the accepted linguistic history of these languages, whilst the tree on the right does not. First, character ‘a’ is mapped onto both the trees, and all the possible ancestral states of this character are enumerated. The likelihood of this distribution of character state change on the tree is then calculated using the chosen model of cognate evolution that specifies the probabilities of transitions between cognate presence and absence (1C). The likelihood of the distribution of character ‘a’ on the tree is the product of all possible ancestral state reconstructions for this character (1D). Finally, the overall likelihood of each tree can be calculated by repeating this process for all the characters in the data, giving rise to a single score for each tree. Note that, in contrast to lexicostatistics, actual character state changes are inferred on the tree. This means that the distinction between retentions and innovations is part of the analysis. The overall likelihood score, generally reported as the log of the likelihood (lnL), represents how well the data is explained by the tree given the model. Better trees are characterized by less negative log likelihoods (1E). In figure 1, the tree on the left has a log likelihood of -4178, whilst the second tree scores -4627. Thus, the former tree is a better explanation of the data.

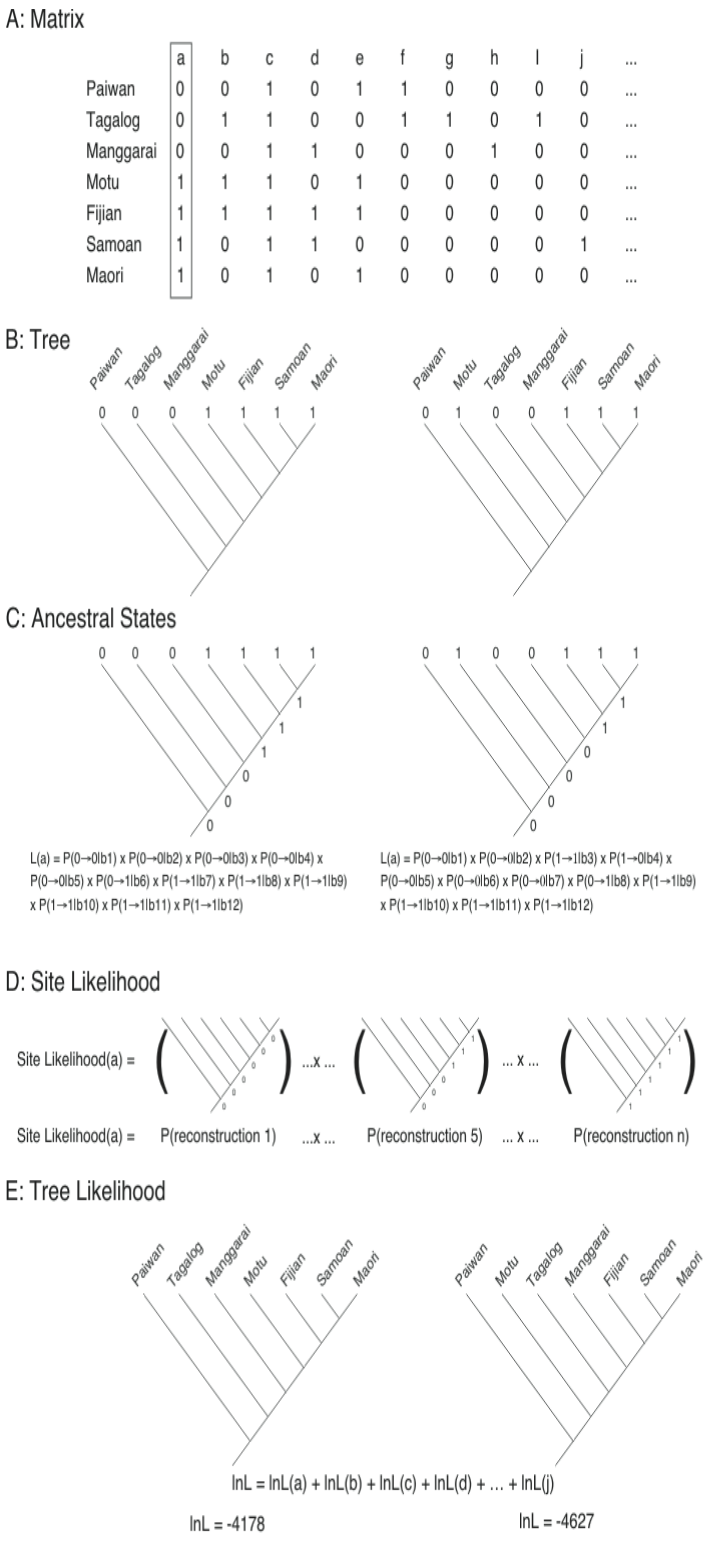

Figure 1: The calculation of the log likelihood of a tree. (A) A hypothetical cognate presence/absence matrix for seven Austronesian languages. (B) Two different trees for these languages with character 'a' mapped onto them. (C) An example of one possible ancestral state reconstruction of character 'a' on these topologies. (D) The site likelihood for character 'a' on the tree is calculated as the product of the probability of all possible ancestral state combinations for that character. (E) The overall tree likelihood is calculated as the sum of

the likelihoods of each site, where the tree with the lower (less negative) log likelihood fits the data better. Here, under a two-parameter model of cognate gain/loss (with no gamma distribution), the tree on the left is a better fit to the data with a log likelihood of -4178, whilst the other tree fits the data less well with a likelihood of -4627.

2.4 Finding the most probable trees

Once we've chosen an appropriate model we then have an explicit *optimality criterion* with which to measure how good a tree is. This means that we can search through the range of possible trees until we find the one(s) with the highest likelihood under this optimality criterion. However, as the number of languages analysed increases so does the number of possible trees. If a tree is strictly bifurcating (i.e. each node can only have two daughter languages), then the number of trees can be calculated as shown in (1) where *n* is the number of languages (Graham and Folds 1972).

(1) $$\frac{(2n-3)!}{2^{n-1}(n-1)!}$$

Thus, when there are four languages there are 15 possible trees. Adding one more language increases the number of possible trees to 105. When the data contains more than 50 languages, there are more possible trees than there are atoms in the universe. If Austronesian has around 1000 languages, then there are an intimidating 3.8 x 10^{2864} possible combinations. This unfortunately means that it is not possible to search through all the trees in any non-trivial dataset. A systematic technique for finding a subset of the good trees from this huge space of possible trees is therefore required. Moreover, as with any statistical estimate, we need some way of evaluating how robust our inferences are.

To do this we use a Bayesian inferential approach that combines the likelihood with our prior knowledge of the trees to give the *posterior probability distribution* of trees. This can be calculated using Bayes's theorem as in (2) (Huelsenbeck et al. 2001).

(2) $$\mathrm{P[Tree \mid Data]} = \frac{\mathrm{P[Data \mid Tree]} \times \mathrm{P[Tree]}}{\mathrm{P[Data]}}$$

The posterior distribution contains the trees that have high likelihoods and fit the data well, given the data and the priors. Priors are the initial values of the model parameters. Often the prior distribution of the parameters is 'flat'; that is all values are considered equally probable. However, if there is strong external evidence supporting some hypothesis, then this can be taken into account explicitly (Lewis 2001). For example, if one wanted to assume that new languages were born at a constant rate across the tree, then a 'Yule' prior on branching rate could be implemented. The ability to incorporate extra information using priors is very powerful—but must be justified. Calculating the posterior probability distribution is hard as it involves the integration of all model parameters, across all branch length combinations, over every single tree (Huelsenbeck et al. 2001). However, using Markov Chain Monte Carlo methods, (MCMC, Metropolis et al. 1953; Huelsenbeck et al. 2001), we can sample from the posterior probability distribution. The phrase 'Monte Carlo' refers to a random sampling method, and a 'Markov Chain' is a process which draws each sample from the probability distribution of the previous state (Larget 2005). To find trees

this method starts with a tree (usually randomly generated) and permutes it in some fashion (e.g. changing the topology, branch lengths or model parameters)—this is the Markov Chain process. The chain preferentially samples trees from this distribution according to their likelihood scores—the Monto Carlo process. If run long enough the chain provides a representative sample of the most probable trees. There are two further considerations in the use of Bayesian MCMC methods. First, the initial trees sampled are heavily contingent on the model's starting parameters (i.e. the priors). To avoid this early samples in an MCMC run are usually discarded as 'burn-in'. Second, each successive tree in an MCMC run is a permutation of the previous one due to the nature of the Markov Chain process (i.e. tree 2 is tree 1 with a branch moved or a change in branch length, etc). This means that each tree is highly correlated with its neighbors. To avoid this auto-correlation, and thus make each sample statistically independent, it is common to only keep every 1,000th or 10,000th tree from the post-burn-in set of trees.

3 Using phylogenetic trees

Using this procedure we will be left with a collection of trees sampled from the posterior probability distribution that should explain the data well. The results we present here are drawn from an analysis using the two-parameter model of cognate gain/loss and gamma-distributed rate variation (Pagel and Meade 2004). This was run for 100,000,000 generations on a cluster of over 150 processors (over 21 years of computer time). The trees were sampled every 10,000 generations after a burn-in of 20,000,000 generations. This gave us a final sample of 8000 trees. However, the endpoint of a phylogenetic analysis is not finding the trees: trees by themselves are boring. Instead, the rationale is to use them to test hypotheses and to investigate the process of evolution. There are many things one can do with trees (Gray et al. 2007). Here we will describe how this set of 8000 most probable trees from the MCMC run can be used to test hypotheses about subgrouping, to date events on the trees, and to trace character change.

3.1 Subgrouping

In historical linguistics it is common to use a family tree to depict the groupings (families, groups, clades, etc) once the groups have been identified using the comparative method. However, there is no formal way of quantifying the support for subgroups. The phylogenetic trees provide a statistical *estimate* of the sub-groupings in the data, and provide a measure of the uncertainty in this estimate. A common way of doing this is to use a Majority Rule 'consensus' tree. This combines the groupings present in all trees in the posterior tree sample. The percentage of trees containing a certain group can be taken as a measure of the support for that grouping in the data. Figure 2 shows an example majority rule consensus tree from our Austronesian data. Subgroups with posterior probability values close to 1.0 are well-supported. For example, the grouping of the Philippine languages is strongly supported by the data (0.99). More surprisingly, the branch grouping the languages of Vanuatu and New Caledonia is also well-supported (0.98). These values mean that 99% and 98% of the 8000 trees in the posterior tree distribution contain those respective groupings. In contrast, other regions of the tree are more poorly supported (e.g. the branch placing the Admiralties languages inside Oceanic after the New North Guinea/Papuan Tip languages has only 0.58 support). Groups with very weak support (<0.50) are not shown. Weakly supported groups could either be the

consequence of little signal in the data due to rapid population expansions, or conflicting stronger signals (perhaps produced by borrowing), or non-tree-like descent processes such as dialect chains and linkages.

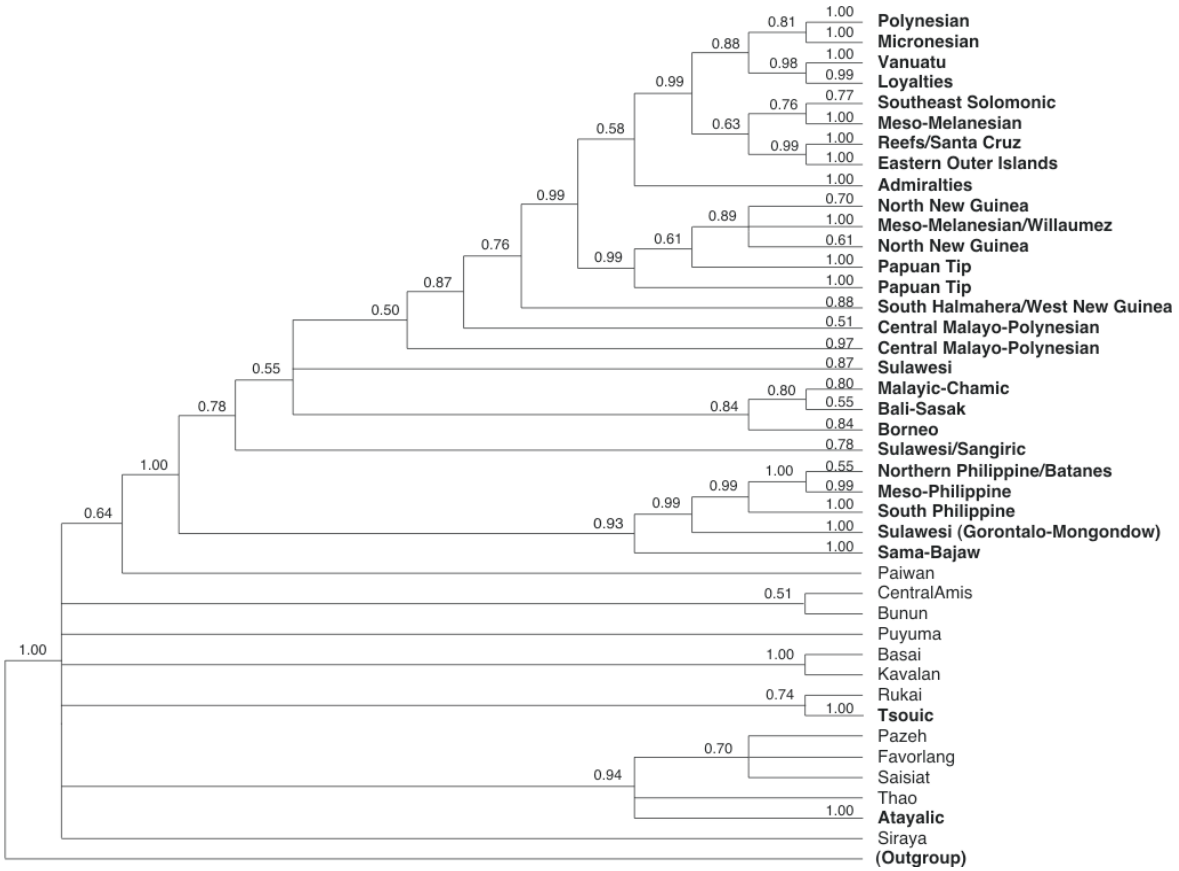

Figure 2: The majority-rule consensus tree of all post burn-in Austronesian trees. Labels in bold represent subgroups of languages, normally-weighted labels denote languages. Where subgroups appear twice in the tree this indicates that they are not monophyletic (e.g. Central Malayo-Polynesian). The numbers on the branches denote the posterior probability of each node. For example, the split between the Northern- and Meso-Philippine languages is strongly supported (1.00). Posterior probability values below 0.50 are considered weak and are not included.

Recall that the lexicostatistical analyses incorrectly 'rooted' the Austronesian languages in East New Guinea. Our phylogenetic analyses, however, support Blust's (1999) rooting of the Austronesian languages in Taiwan. The Formosan languages are placed at the base of the tree after the outgroup languages. There is no unified Formosan subgroup but at least seven higher-order branches of Formosan derived from Proto Austronesian. Moreover, whilst the Tsouic and Atayalic subgroups of Formosan languages are robust, there is little support for other higher-order sub-groupings within Formosan. These results are all concordant with Blust (1999).

Not only do the phylogenetic trees support a Formosan origin of the Austronesian languages, the sequence of the higher-order subgroups closely conforms to the 'Out of Taiwan' scenario of Austronesian settlement. Moving down the tree, after Formosan languages we find the languages of Island South-East Asia, with strong support for the

Philippine and Malayic-Chamic language groups. This is followed by two weakly supported groups of Central Malayo-Polynesian languages, and then the well-supported South Halmahera/West New Guinea group. Finally, there is a well-supported Oceanic subgroup, with strong support for the recognized subgroups within Oceanic (Polynesian, Micronesian, Southeast Solomonic, Eastern Outer Islands, Admiralties). Our results split Oceanic into two major groups, both strongly supported (0.99). The first of these Oceanic subgroups is comprised of the Papuan Tip, North New Guinea and Meso-Melanesian languages. This represents the Western Oceanic group identified by Ross (1988). However, only the Willaumez languages of Meso-Melanesian are in this subgroup, the remainder is located in our second Oceanic grouping. This second Oceanic group contains the Remote Oceanic language subgroups and the majority of the Meso-Melanesian languages. Interestingly, we show strong support (0.99) for the recently identified subgroup Temotu containing the languages from the Eastern Outer Islands and the Reefs–Santa Cruz region (Ross and Næss 2007). In contrast to Blust (1998), the Admiralties subgroup is not at the base of the entire Oceanic subgroup, but is situated—albeit very weakly (0.58)—between Western and Remote Oceanic. Some of the higher-order nodes within our two Oceanic groupings are only weakly supported, such as the cluster grouping Temotu to Southeast Solomonic (0.63). These low values may reflect the rapid dispersal of languages through this region (Pawley 1999), or the large amounts of contact induced change in large-scale dialect networks found in this region (Ross 1996).

3.2 Dating

One of the great attractions of lexicostatistics was its apparent ability to calculate absolute dates of language divergence times through a method known as *glottochronology* (Lees 1953). This technique calculated absolute ages by assuming that as languages split they lost vocabulary at a constant rate. Accordingly, a simple decay curve of cognate loss could be used to calculate divergence times by solving the equation in (3) where C is the percentage of shared cognates between the two languages, r is the retention rate, and t is the estimated time depth.

(3) $$t = \frac{\log C}{2 \log r}$$

Over 1000 years the retention rate r was often assumed to be 81% for the 200 item Swadesh list (Lees 1953). Therefore, if two languages shared 90% of their basic vocabulary, they should have diverged 250 years ago, whilst languages that were 75% similar should have diverged around 680 years ago. However, these glottochronological calculations magnified all the shortcomings of lexicostatistics. Languages vary substantially in their retention rates, and this rate variation produced some obviously inaccurate dates (Bergsland and Vogt 1962; Blust 2000). For example, Icelandic shares over 95% of its core vocabulary with Old Norse. According to glottochronology Old Norse and Icelandic would have diverged less than 200 years ago. This is incorrect—Old Norse was spoken around 1000 years ago (Bergsland and Vogt 1962). Problems such as this led to such a strong rejection of glottochronology that over fifty years later we are still being cautioned about its inaccuracy (McMahon and McMahon 2006).

The age of the Indo-European language family has been a topic of considerable interest and much debate. There are two main theories. The first proposes that Proto Indo-European broke up 5000–6000 years ago when Indo-European languages spread with the expansion of

the archaeological culture known as Kurgan (Gimbutas 1973). The main alternative account that Indo-European spread with the advent of farming technology around 8000–9000 years ago (Renfrew 1987). Naturally, one of the first uses we put phylogenetic methods to was dating the divergence of particular branches of Indo-European (Gray and Atkinson 2003, Atkinson and Gray 2006). Our results showed strong support for an initial breakup of the Indo-European family around 8000–9000 years ago, with a subsequent breakup of 'Nuclear Indo-European' (Indo-European minus Anatolian and Tocharian) around 6000 years ago. The results were robust to different calibrations, cognate coding, and likelihood models (Atkinson et al. 2005). However, we were promptly criticized for merely, 'reintroducing glottochronology by the back door' (Gamble et al. 2005:208), and 'ignor(ing) the fatal shortcomings of glottochronology' (Eska and Ringe 2004:569). These are unfortunate misunderstandings. Phylogenetic dating methods, such as the Penalized Likelihood rate smoothing approach (Sanderson 1997, 2002) used by Gray and Atkinson (2003), as well as newer methods which can 'relax the clock' (Drummond et al. 2006), do not have the fatal shortcomings of glottochronology. These approaches need not assume that there is a single 'clock-like' rate of lexical change (Atkinson and Gray 2006).

To demonstrate how divergence date estimation can be obtained without a strict 'glottoclock' we will estimate the age of Proto Austronesian on the (expanded) tree from Figure 2. The branches on the trees in our posterior sample are proportional to the amount of change along that lineage. This is usually expressed as the rate of substitutions (in this case the gain or loss of cognates in a language). These branch lengths can be converted to time by adding historically attested calibration points. For example, the Eastern Polynesian subgroup can be constrained to around 1200 to 1300 years ago on the basis of initial settlement times (Green and Weisler 2002). Similarly, the Chamic subgroup can also be calibrated based on the fact that Chamic speakers were mentioned in Chinese records around 1800 years ago, and probably entered Vietnam around 2600 years ago (Thurgood 1999). This calibration of nodes on the tree within a historical time range allows the method to estimate how fast the changes measured by the branch lengths are occurring. The Penalized Likelihood rate-smoothing approach can then convert branch lengths into time estimates by smoothing the rates of change across the tree. Instead of assuming a constant retention rate, this allows certain parts of the tree to change faster or slower than others. We applied this approach to one tree from the posterior distribution of trees for our analysis. The resulting dated tree (Figure 3) shows an age of around 5310 years for Proto Austronesian, and an age of 4240 years for Proto Malayo-Polynesian. We must emphasise at this point that the date estimates should be done on all trees in the posterior sample and not just a single one. Calculating divergence dates on all the trees would produce a distribution of the most probable age of Proto Austronesian. This distribution can then be used to provide a confidence interval on any date estimate. As our aim in this paper is to illustrate the overall approach rather than to test specific hypotheses, we have just dated one tree for illustrative purposes. However, dates from this tree support the emergence of Proto Austronesian in Taiwan around 5500 years ago (e.g. Blust 1995; Pawley 2002). Note also the presence of pauses and rapid pulses of expansion as has been argued by Blust (1999), Green (1999) and Pawley (1999, 2002). In this tree, we see a pause of around 1000 years before Proto Malayo-Polynesian arises, and a subsequent rapid pulse of expansion through to Proto Oceanic. Another pause then expansion pulse occurs after the initial settlement of the Central Pacific region.

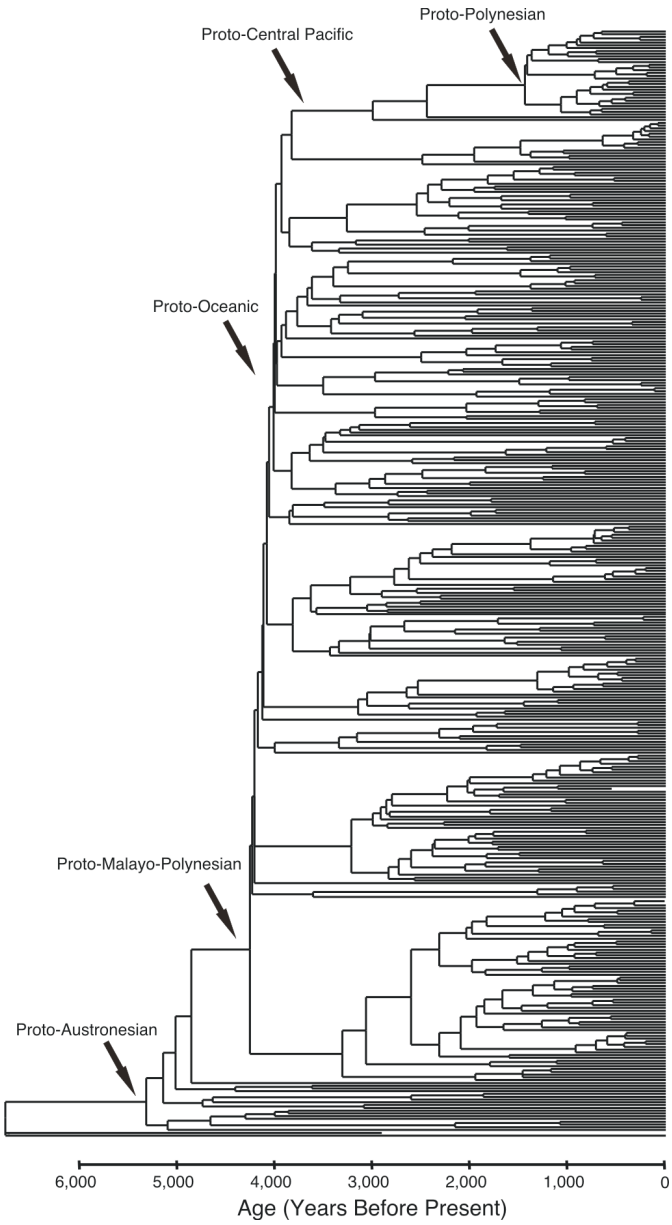

Figure 3: A dated tree of 400 Austronesian languages showing the age of a number of protolanguages estimated using Penalized Likelihood rate smoothing. On this tree Proto Austronesian is estimated to be 5310 years old and Proto Malayo-Polynesian 4240 years.

3.3 Tracing character history

Much of historical linguistics is concerned with the reconstruction of protoforms. This is done both as a means to subgrouping and as a way of making inferences about society and culture of ancestral speech communities. Biologists have also developed phylogenetic methods to reconstruct ancestral states. These methods have been used to tackle problems such as identifying the origin of ancestral genes in the eukaryote genome (Lester et al. 2006). One common phylogenetic approach essentially 'maps' a character of interest onto the posterior tree sample using a continuous-time Markov model of trait evolution (Pagel et al. 2004). Under this model a character can change between a finite number of states over infinitesimally small time periods. The rates of change between these states along the

branches can be estimated directly from the posterior tree sample. These model parameters can then be used to calculate the probability of a certain state at any given node. For example, one might want to evaluate how the words for 'earth, soil' had evolved in the Polynesian languages, and infer what variant was spoken by Proto Polynesian. Figure 4 shows three cognate sets for words meaning 'earth/soil' mapped onto a tree of the Central Pacific subfamily (the expanded form of Figure 2). Cognate set A (colored white) reflects forms like Tongan *kelekele,* Samoan *'ele'ele* and Fijian (Bau) *qele*. Cognate set B (colored gray) reflects forms like the Tahitian *repo* and Hawaiian *lepo*. Cognate set C (colored black) reflects forms like Vaeakau-Taumako's *pela*. Using the Bayesian ancestral state reconstruction method (Pagel et al. 2004) we can estimate that, on this tree, the probability that Proto East Polynesian and Proto Tahitic had cognate set B was 0.99. This is concordant with the comparative method, where the reconstructed Proto East Polynesian form is **repo* (Biggs and Clark 2000). Deeper in the tree, the Proto Polynesian and Proto Central Pacific nodes reflect cognate set A with a probability very close to 1. Again, this matches the reconstructed Proto Central Pacific form **g(w)ele* (Ross et al. 1998). Cognate set C presumably reflects Prot Oceanic **pela* 'muddy' (Biggs and Clark 2000) with semantic change. We emphasise again that ideally this estimation should be integrated over the set of trees in the posterior sample, not just a single tree.

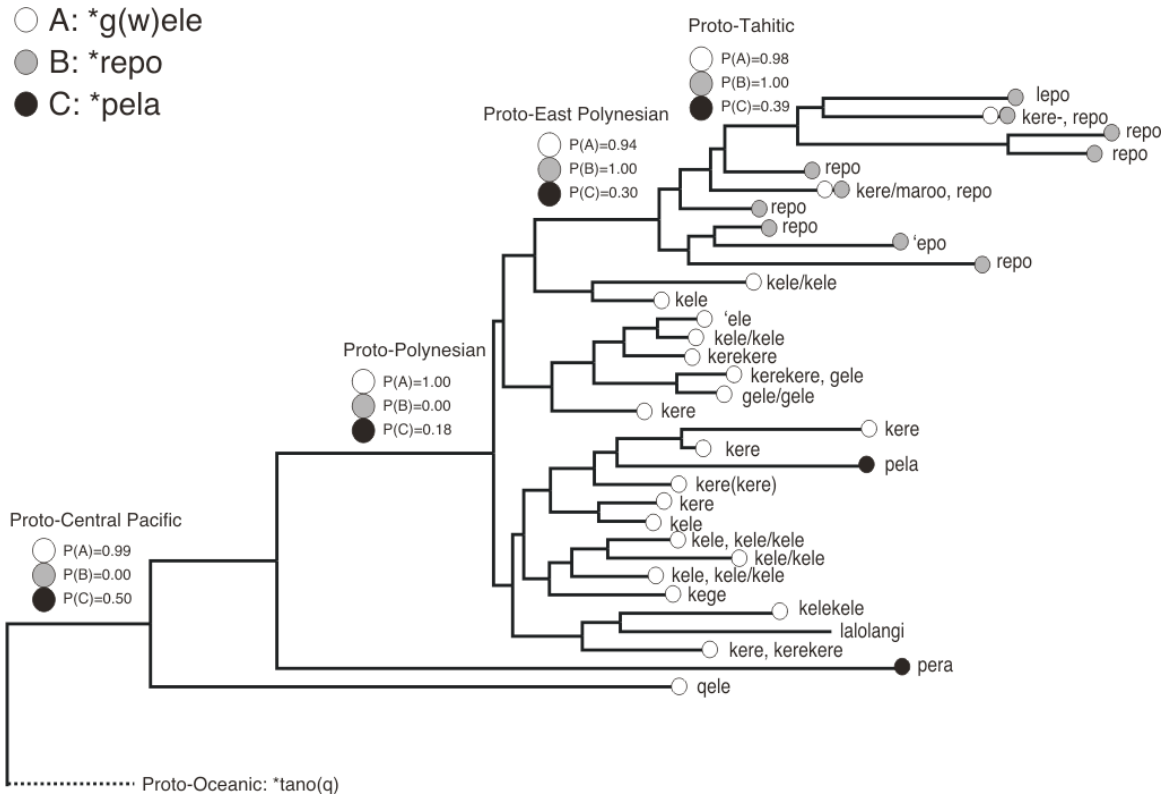

Figure 4: Tree of Central Pacific languages showing the distribution of three cognate sets A (in white), B (in gray), and C (in black) containing words for 'earth/soil'. Branch lengths are proportional to amount of change along the lineage. The probability of the ancestral states are marked for a number of protolanguages. The probability of Proto Tahitic and Proto East Polynesian belonging to cognate set B (*repo) is >0.99. Proto Polynesian and Proto Central Pacific instead contain cognate set A (*g(w)ele) also with a probability of >0.99.

4 Conclusion

We hope that this chapter has corrected most of the persistent myths and misconceptions about the application of computational phylogenetic methods to historical linguistics. Let us be very clear. Phylogenetic methods do not make the flawed assumptions of lexicostatistics or glottochronology. They do not count cognates to calculate pairwise similarity measures. Instead, the likelihood calculations are based on each cognate set and how it fits onto the tree. Phylogenetic methods do not require a single 'one size fits all' rate of lexical replacement. These methods can allow for different rates of change both between cognate sets and between different lineages. Moreover, this framework can explicitly take into account external evidence such as archaeological dates and known historical events to make robust inferences about divergence dates. In marked contrast to lexicostatistics, the phylogenetic methods we have detailed here perform exceptionally well on the very difficult problem of the Austronesian subgrouping and dating. First, the trees are rooted in Taiwan, in line with the results of the comparative method. Second, the sequence and subgrouping of these phylogenetic trees strongly reflect the structure of the family tree suggested by the comparative method, at least in those cases where there is a consensus among comparative linguists. Third, the timing of events on these trees again corresponds extremely well to the 'Out-of-Taiwan' scenario.

We also hope to have laid to rest a final vexing misconception about phylogenetic linguistics: 'this method is not giving anything new' (Jasanoff in Wade 2004:1). Not only do phylogenetic methods work well and outperform lexicostatistics, they also provide a range of new tools that can be of great benefit to linguistics. First, phylogenetic methods provide an explicit optimality criterion for evaluating how well different trees (i.e. historical scenarios) are supported by the data. Second, they provide an empirical way of assessing the statistical robustness of any subgroup in those trees. We have shown here a number of Austronesian examples where the support values on our trees coincide well with linguistic intuitions about the strength of support for these groupings. Third, despite the failure of glottochronology to provide robust date estimates, the attraction of absolute dating is strong. Dates are critically important for inferences about human prehistory. They provide a powerful way of linking linguistic, archaeological, cultural, and genetic evidence. It is not uncommon to still see glottochronological age estimates cited in publications, along with the standard disclaimer that this method cannot be trusted (e.g. Campbell 1997; Comrie 2002; Pawley 2002). Phylogenetic dating methods can, when used carefully and appropriately, help integrate our inferences about human prehistory without these glaring disclaimers. Fourth, these methods enable us to investigate how linguistic and cultural traits have evolved in families by tracing their history. These tools can infer ancestral states and can even be used to infer functional dependency between linguistic characters (Gray et al. 2007). Far from being lexicostatistics-redux, Bayesian phylogenetic methods provide exciting new tools for historical linguistics.

References

Atkinson, Q.D., G. Nicholls, D. Welch and R.D. Gray. 2005. From words to dates: water into wine, mathemagic or phylogenetic inference? *Transactions of the Philological Society* 103:193–219.

Atkinson, Q.D. and R.D. Gray. 2006. Are accurate dates an intractable problem for historical linguistics? In C.P. Lipo, M.J. O'Brien, M. Collard and S.J. Shennan, eds

Mapping our Ancestors: phylogenetic approaches in Anthropology and Prehistory, 269–298. New Brunswick: Aldine.

Atkinson, Q.D., A. Meade, C. Venditti, S.J. Greenhill and M. Pagel. 2008. Languages evolve in punctuational bursts. *Science* 319:588.

Balter, M. 2003. Early date for the birth of Indo-European languages. *Science* 302:1490–1491.

Balter, M. 2004. Search for the Indo-Europeans. *Science* 303:1323–1326.

Bellwood, P. 1997. *Prehistory of the Indo-Malaysian Archipelago*. Honolulu: University of Hawai'i Press.

Bellwood, P., J.F. Fox and D. Tryon. 1995. *The Austronesians: historical and comparative Perspectives*. Canberra: Research School of Pacific and Asian Studies, The Australian National University.

Bergsland, K. and H. Vogt. 1962. On the validity of glottochronology. *Current Anthropology* 3:115–153.

Blust, R.A. 1977. The proto-Austronesian pronouns and Austronesian subgrouping: a preliminary report. *University of Hawai'i Working Papers in Linguistics* 9:1–15.

—— 1981. *Variation in retention rate among Austronesian languages*. Talk given to the Third International Conference on Austronesian Linguistics. Bali.

—— 1982. The linguistic value of the Wallace line. *Bijdragen tot de taal-, land- en. volkenkunde* 138:231–250.

—— 1995. The prehistory of the Austronesian-speaking peoples: the view from language. *Journal of World Prehistory* 9:453–510.

—— 1998. A note on higher-order subgroups in Oceanic. *Oceanic Linguistics* 37:182–188.

—— 1999. Subgrouping, circularity and extinction: some issues in Austronesian comparative linguistics. In E. Zeitoun and P. Jen-kuei Li, eds *Selected papers from the Eighth International Conference on Austronesian Linguistics* vol. 1, 31–94. Taipei, Taiwan: Symposium Series of the Institute of Linguistics, Academia Sinica.

—— 2000. Why lexicostatistics doesn't work: the 'universal' constant hypothesis and the Austronesian languages. In C. Renfrew, A. McMahon and L. Trask, eds *Time depth in historical lLinguistics*, 311–331. Cambridge: The McDonald Institute for Archaeological Research.

Bowern, C. 2006. Punctuated equilibrium and language change. In K. Brown, ed. *Encyclopedia of Language and Linguistics*, 286–289. Oxford: Elsevier.

Brandes, J.L.A. 1884. *Bijdrage tot de vergelijkende klankleer der westersche afdeeling van de Maleisch-Polynesische taalfamilie*. Utrecht.

Brugmann, K. 1884. Zur Frage nach den Verwandtschaftsverhältnissen der Indogermanischen Sprachen. *Internationale Zeitschrift für allgemeine Sprachewissenschaft* 1:226–256.

Bryant, D. 2006. Radiation and Network Breaking in Polynesian Language Evolution. In P. Forster and C. Renfrew, eds Phylogenetic Methods and the Prehistory of Languages, 111–118. Cambridge: McDonald Institute Press, University of Cambridge.

Bryant, D., F. Filimon and R.D. Gray. 2005. Untangling our past: languages, trees, splits and networks. In R. Mace, C.J. Holden and S. Shennan, eds *The evolution of cultural diversity: phylogenetic approaches*, 67–84. London: UCL Press.

Burnham, K.P. and D.R. Anderson. 1998. *Model selection and inference — a practical information-theoretic approach*. New York: Springer.

Campbell, L. 1997. *American Indian languages: the historical linguistics of Native America*. Oxford: Oxford University Press.

—— 2004. *Historical linguistics: an introduction.* 2nd edition. Edinburgh: Edinburgh University Press.

Comrie, B. 2002. Farming dispersal in Europe and the spread of the Indo-European language family. In P. Bellwood and C. Renfrew, eds *Examining the farming/ language dispersal hypothesis*, 409–419. Cambridge: The McDonald Institute for Archaeological Research.

Dahl, O.C. 1973. *Proto-Austronesian.* Scandinavian Institute of Asian Studies Monograph Series No.15. Studentlitteratur: Lund, Sweden.

Dempwolff, O. 1934. *Vergleichende lautlehre des austronesischen Wortschatzes. Zeitschrift für Eingeborenen-Sprachen*. 1 Induktiver Aufbau einer indonesischen Ursprache, 15, Berlin: Deitrich Reimer.

—— 1937. *Vergleichende lautlehre des austronesischen Wortschatzes. Zeitschrift für Eingeborenen-Sprachen*. 2, Deductive Anwendung des urindonesischen auf austronesische Einzelsprachen, 17, Berlin: Deitrich Reimer.

—— 1938. *Vergleichende lautlehre des austronesischen Wortschatzes. Zeitschrift für Eingeborenen-Sprachen*. 3 Austronesisches Wörterverzeichnis, 19, Berlin: Deitrich Reimer.

Diamond, J. and P. Bellwood. 2003. Farmers and their languages: the first expansions. *Science* 300:597–603.

Dixon, R.M.W. 1997. *The rise and fall of languages*. Cambridge: Cambridge University Press.

Dunn, M., A. Terrill, G. Reesink, R.A. Foley and S.C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science* 309:2072–2075.

Durie, M. and M. Ross. 1996. *The comparative method reviewed: regularity and irregularity in language change*. Oxford: Oxford University Press.

Drummond, A.J., S.Y.W. Ho, M.J. Phillips and A. Rambaut. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology* 4:e88.

Dyen, I. 1962. The lexicostatistical classification of the Malayopolynesian languages. *Language*, 38:38–46.

—— 1965. A lexicostatistical classification of the Austronesian languages. In *Indiana University Publications in Anthropology and Linguistics: Memoir 19 of the International Journal of American linguistics*. Indiana: Indiana University.

Edwards, A.W.F. and L.L. Cavalli-Sforza. 1964. Reconstruction of evolutionary trees. In J. McNeill, ed. *Phenetic and phylogenetic classification*, 67–76. Cambridge: Cambridge University Press.

Elbert, S. H. 1953. Internal relationships of Polynesian languages and dialects. *Southwest Journal of Anthropology* 9:147–173.

Eska, J.F. and D. Ringe. 2004. Recent work in computational linguistic phylogeny. *Language* 80:569–582.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* 27:401–410.

—— 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* 17:368–376.

Fisher, R.A. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London* Series A, 222:309–368.

Forster, J.R. 1778. *Observations made during a voyage round the world.* London.

Gamble, C., W. Davies, P. Pettitt, L. Hazelwood and M. Richards. 2005. The archaeological and genetic foundations of the European population during the Late Glacial: implications for 'agricultural thinking'. *Cambridge Archaeological Journal* 15:193–223.

Gimbutas, M. 1973. Old Europe c. 7000–3500 B.C., the earliest European cultures before the infiltration of the Indo-European peoples. *Journal of Indo-European Studies* 1:1–20.

Goldman, N. 1993. Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* 36:182–198.

Gordon, Raymond G., Jr. 2005. *Ethnologue: languages of the World*, Fifteenth edition. Dallas, Texas: SIL International. Online version: http://www. ethnologue. com/.

Grace, G.W. 1961. Austronesian linguistics and culture history. *American Anthropologist* 57:359–368.

—— 1964a. Movement of the Malayo-Polynesians 1500 BC to AD 500: the linguistic evidence. *Current Anthropology* 5:361–368.

—— 1964b. The linguistic evidence. *Current Anthropology*, 5:361–368.

—— 1966. Austronesian lexicostatistical classification: a review article. *Oceanic Linguistics* 5:13–31.

Graham, R.L. and L.R. Foulds. 1982. Unlikelihood that minimal phylogenies for a realistic biological study can be constructed in reasonable computational time. *Mathematical Biosciences* 60:133–142.

Gray, R.D. 2007. *Tangled trees: what do phylogenetic networks reveal about Oceania linguistic history?* Talk given to the Seventh International Conference on Oceanic Linguistics. Noumea, New Caledonia.

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

Gray, R.D. and F.M. Jordan. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature* 405:1052–1055.

Gray, R.D., A.J. Drummond and S.J. Greenhill. 2009. Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement. *Science*, 323:479–483.

Green, R.C. 1999. Integrating historical linguistics with Archaeology: insights from research in remote Oceania. *Indo-Pacific Prehistory Association Bulletin* 18 (Melaka Papers), vol. 2:3–16.

Green, R.C. and M.I. Weisler. 2002. The Mangarevan Sequence and dating of the geographic expansion into Southeast Polynesia. *Asian Perspectives* 41:213–241.

Greenberg, J.H. 1949–54. Studies in African linguistic classification. *South-western Journal of Anthropology* 5:79–100, 190–198, 309–317; 6:47–63, 143–160, 223–237, 388–398; 10:405–415.

Greenhill, S.J. and R.D. Gray. 2005. Testing population dispersal hypotheses: Pacific settlement, phylogenetic trees and Austronesian languages. In R. Mace, C.J. Holden and S. Shennan, eds *The evolution of cultural diversity: phylogenetic approaches*, 31–52. London: UCL Press.

Greenhill, S.J., R. Blust and R.D. Gray. 2008. *The Austronesian basic vocabulary database*: From Bioinformatics to Lexomics. *Evolutionary Bioinformatics,* 4:271–283. http://language.psy.auckland.ac.nz

Hamed, M.B. and F. Wang. 2006. Stuck in the forest: Trees, networks and Chinese dialects. *Diachronica* 23:29–60.

Hoijer, H. 1956. Lexicostatistics: a critique. *Language* 32:49–60.

Holden, C.J. 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony approach. *Proceedings of the Royal Society of London, B Biological Sciences* 269:793–799.

Holden, C.J. and R.D. Gray. 2006. Rapid radiation, borrowing and dialect continua in the Bantu languages. In P. Forster and C. Renfrew, eds *Phylogenetic methods and the prehistory of languages*, 19–31. Cambridge: McDonald Institute for Archaeological Research.

Holm, H.J. 2007. The new arboretum of Indo-European 'trees'. Can new algorithms reveal the phylogeny and even prehistory of Indo-European? *Journal of Quantitative Linguistics* 14:167–214.

Houtman, F. de. 1603. *Spraeck ende Woord-Boeck, Inde Maleysche ende Madagaskarsche Taln met vele Arabische ende Turcsche Woorden*. Amsterdam.

Huelsenbeck, J.P. and B. Rannala. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* 276:227–232.

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., B. Larget, R.E. Miller and F. Ronquist. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. *Systematic Biology* 51:673–688.

Hymes, D. H. 1960. Lexicostatistics so far. *Current Anthropology* 1:3–44.

Kern, H. 1886. *De Fidji-taal vergeleken met hare verwanten in Indonesië en Polynesië*. Verhandelingen der Koninklijk Akademie van Wetenschappen, 16, Amsterdam.

Kirch, P.V. 2000. *On the Road of the Winds: an archaeological history of the Pacific Islands before European contact*. Berkeley: University of California Press.

Kirch, P. and R. Green. 2001. *Hawaiki, Ancestral Polynesia: an essay in Historical Anthropology.* Cambridge: Cambridge University Press.

Labov, W. 1994. *Principles of linguistic change: internal factors*. Oxford: Blackwell.

Larget, B. 2005. Introduction to Markov Chain Monte Carlo methods in molecular evolution. In R. Nielsen, ed. *Statistical Methods in Molecular Evolution*, 45–62. New York: Springer.

Lees, R.B. 1953. The basis of Glottochronology. *Language* 29:113–127.

Lester, L., A. Meade and M. Pagel. 2006. The slow road to the eukaryotic genome. *BioEssays* 28:57–64.

Lewis, P.O. 2001. Phylogenetic systematics turns over a new leaf. *Trends in Ecology and Evolution* 16:30–37.

Lynch, J., M.D. Ross and T. Crowley. 2002. *The Oceanic languages*. Richmond: Curzon Press.

McMahon, A. and R. McMahon. 2006. Why linguists don't do dates: evidence from Indo-European and Australian languages. In P. Forster and C. Renfrew, eds *Phylogenetic methods and the prehistory of languages*, 153–160. Cambridge: McDonald Institute for Archaeological Research.

Murdock, G.P. 1964. Genetic classification of the Austronesian languages: a key to Oceanic culture history. *Ethnology* 3:117–126.

Nakhleh, L., T. Warnow, D. Ringe and S.N. Evans. 2005. A comparison of phylogenetic reconstruction methods on an Indo-European dataset. *Transactions of the Philological Society* 103:171–192.

Nettle, D. 1999. Is the rate of linguistic change constant? *Lingua* 108:119–136.

Nichols, J. 1997. Modeling ancient population structures and movement in linguistics. *Annual Review of Anthropology* 26:359–384.

Pagel, M. 1999. Inferring the historical patterns of biological evolution. *Nature* 401:877–884.

Pagel, M. and A. Meade. 2004 A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology* 53:571–581.

Pagel, M., Q.D. Atkinson and A. Meade. 2007. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449:717–720.

Pawley, A. 1999. Chasing Rainbows: implications of the rapid dispersal of Austronesian languages for subgrouping and reconstruction. In E. Zeitoun and P. Jen-kuei Li, eds *Selected papers from the Eighth International Conference on Austronesian Linguistics* vol. 1, 95–138. Taipei, Taiwan: Symposium Series of the Institute of Linguistics, Academia Sinica.

—— 2002. The Austronesian dispersal: languages, technologies and people. In P. Bellwood and C. Renfrew, eds *Examining the farming/language dispersal hypothesis*, 251–274. Cambridge: McDonald Institute for Archaeological Research.

Pawley, A. and R.C. Green. 1973. Dating the dispersal of the Oceanic languages. *Oceanic Linguistics* 121–2:1–67.

Pawley, A. and M. Ross. 1995. The prehistory of the Oceanic languages: a current view. In P. Bellwood, J.J. Fox and D.T. Tryon, eds *The Austronesians: historical and*

comparative perspectives, 39–74. Canberra: Research School of Pacific and Asian Studies, The Australian National University.

Penny, D., B.J. McComish, M.A. Charleston and M.D. Hendy. 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. *Journal of Molecular Evolution* 53:711–723.

Posada, D. and K.A. Crandall. 2001. Selecting the best-fit model of nucleotide substitution. *Systematic Biology* 50:580–601.

Renfrew, C. 1987. *Archaeology and language: the puzzle of Indo-European Origins*. London: Cape.

Reland, H. 1708. *Dissertatio de linguis insularum quarundem orientalium*. Trajecti ad Rhenum.

Rexovà, K., Y. Bastin and D. Frynta. 2006. Cladistic analysis of Bantu languages: a new tree based on combined lexical and grammatical data. *Naturwissenschaften* 93:189–194.

Rexovà, 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. *Transactions of the Philological Society* 100:59–129.

Ross, M. 1988. *Proto Oceanic and the Austronesian languages of Western Melanesia*. Canberra: Pacific Linguistics.

—— 1996. Contact-induced change and the comparative method: cases from Papua New Guinea. In M. Durie and M.D. Ross, eds *The comparative method reviewed: regularity and irregularity in language change*, 180–217. New York: Oxford University Press.

—— 1997. Social networks and kinds of speech-community events. In R. Blench and M. Spriggs, eds *Archaeology and Language*, vol. 1, 209–261 London: Routledge.

Ross, M. and A. Næss. 2007. An Oceanic origin for Äiwoo, the language of the reef islands? *Oceanic Linguistics* 46:456–498.

Sanderson, M.J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution* 14:1218–1231.

—— 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Molecular Biology and Evolution* 19:101–109.

Shutler, R. and J.C. Marck. 1975. On the dispersal of the Austronesian horticulturalists. *Archaeology and Physical Anthropology in Oceania* 10:81–113.

Sokal, R.R. and P.H.A. Sneath. 1963. *Principles of numerical taxonomy*. San Francisco: W.H. Freeman.

Steel, M. and D. Penny. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. *Molecular Biology and Evolution* 17:839–850.

Suchard, M.A., R.E. Weiss and J.S. Sinsheimer. 2001. Bayesian selection of continuous-time Markov Chain evolutionary models. *Molecular Biology and Evolution* 18:1001–1013.

Swadesh, M. 1952. Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96:453–463.

Swofford, D.L., G.J. Olsen, P.J. Waddell and D.M. Hillis. 1996. Phylogenetic inference. In D.M. Hillis, C. Moritz and B.K. Mable, eds *Molecular Systematics*, 407–514. Sinauer Associates, Sunderland, MA.

Thurgood, G. 1999. *From ancient Cham to modern dialects: two thousand years of language contact and change.* Hawaii: University of Hawaii Press.

Wade, N. 2004, 16 March. A biological dig for the roots of language. *The New York Times*.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* 39:306–314.

Yang, Z., N. Goldman and A.E. Friday. 1994. Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. *Molecular Biology and Evolution* 11:316–324.