An experimental study comparing linguistic phylogenetic reconstruction methods

François Barbançon,[1] Tandy Warnow,[1] Steven N. Evans,[2] Donald Ringe,[3] and Luay Nakhleh[4]

[1]Department of Computer Sciences, University of Texas at Austin, Austin, TX 78712, USA
[2]Department of Statistics, University of California at Berkeley, Berkeley CA 94720-3860, USA
[3]Department of Linguistics, University of Pennsylvania, Philadelphia, PA 19104, USA
[4]Department of Computer Sciences, Rice University, Houston TX 77005, USA

Abstract

The estimation of linguistic evolution has intrigued many researchers for centuries, and in just the last few years, several new methods for constructing phylogenies from languages have been produced and used to analyze a number of language families. These analyses have led to a great deal of excitement, both within the field of historical linguistics and in related fields such as archaeology and human genetics. They have also been controversial, since the analyses have not always been consistent with each other, and the differences between different reconstructions have been potentially critical to the claims made by the different groups. In this paper, we report on a simulation study we performed in order to help resolve this controversy, which compares some of the main phylogeny reconstruction methods currently being used in linguistic cladistics. Our simulated datasets varied in the number of contact edges, the degree of homoplasy, the deviation from a lexical clock, and the deviation from the rates-across-sites assumption. We find the accuracy of the unweighted methods maximum parsimony, neighbor joining, lexico-statistics, and the method of Gray & Atkinson, to be remarkably consistent across all the model conditions we studied, with maximum parsimony being the best, followed (often closely) by Gray & Atkinson's method, then neighbor joining, and finally lexico-statistics (UPGMA). The accuracy of the two weighted methods (weighted maximum parsimony and weighted maximum compatibility) depends upon the appropriateness of the weighting scheme, and so depends upon the homoplasy levels produced by the model conditions; for low-homoplasy levels, however, the weighted methods generally produce the most accurate results of all methods, while the use of inappropriate weighting schemes can make for poorer results than maximum parsimony and Gray & Atkinson's method under moderate to high homoplasy levels.

1 Introduction

In a phylogenetic analysis, an evolutionary history is proposed for a given set of "taxa"; in biology, the taxa are likely to be biological species or biomolecular sequences, and in historical linguistics, the taxa are languages, or perhaps dialects, which are presumed to have a common ancestor. In both biological and linguistic phylogenetic analyses, a set of characters common to all taxa are considered, and each taxon is represented by its states for these characters. A (linguistic) character is any feature of languages that can take one or more forms; these different forms are called the "states" of the character. Linguistic characters are of three types: lexical, phonological, and morphological. For lexical characters, the different states are cognate classes, so that two languages exhibit the same state for the lexical character if and only if they have cognates for the meaning associated with the lexical character. Phonological characters record the occurrence of sound changes within the (pre)history of the language; thus a typical phonological character has two states, depending on whether or not the sound change (or, more often, constellation of changes) has occurred in the development of each language. Most morphological characters represent inflectional markers; like lexical characters, they are coded by cognation. Thus each character defines

an equivalence relation on the language family, such that two languages are equivalent if they exhibit the same state for the character. Thus, if two languages exhibit the same state for the same character, then the presumption is (generally) that the shared state arose due to common inheritance. However, shared states can also arise due to borrowing, or through random chance, with some linguistic characters being much more likely to evolve by random chance or borrowing than others. Thus, not all linguistic characters provide the same quality of "phylogenetic signal".

Thus, decisions related to character selection – whether to rely only upon lexical characters, or to use morphological and phonological characters as well – have the potential to impact a phylogenetic analysis, and these decisions also raise other issues, such as whether all characters should be treated identically, or whether "weighting schemes" should be used to reflect the assumed reliability of the character. In [30], we examined the impact of character selection on phylogenetic analyses of an Indo-European (IE) dataset compiled by Ringe and Taylor, and showed how phylogenetic analyses using the same method can differ when based upon different sets of characters. For example, phylogenies obtained on the basis of lexical characters can be very different from phylogenies obtained based upon a mixture of the three different types of characters, and phylogenies based upon "screened" datasets (whereby characters are removed if they are considered to be likely to be "homoplastic") can differ from phylogenies based upon unscreened datasets. These differences in some cases can be minor, but in other cases can be significant!

The study in [30] suggests that aspects of character evolution are likely to be significant when evaluating the impact of characters on phylogenetic accuracy. For example, a character's resistance to borrowing could be important, since analyzing characters that have evolved through undetected borrowing could lead to an incorrect estimation of the underlying true tree (known in linguistics as the "genetic tree"). However, incorrect phylogenetic reconstructions arise due to a host of reasons. For example, there can be too little evolution in some particular branch of the true phylogeny for that branch to be correctly reconstructed, resulting typically in an incompletely resolved tree. There can be "rogue taxa", which in this case would be languages which have evolved so quickly from their parents that they can attach fairly arbitrarily throughout the tree without changing the quality of the resultant phylogeny; Albanian is an example of this property, to some extent. There can also simply be inadequate data - just not enough information to resolve the evolutionary history. The degree of deviation from a lexical clock can negatively impact methods, as can the degree of homoplasy (parallel evolution or back-mutation).

All of these issues have the potential to impact all phylogenetic reconstruction methods, and yet it is clear that different methods respond differently to these challenges, with some methods more negatively impacted by some conditions than others.

How, therefore, is an interested researcher to determine whether a particular phylogenetic analysis proposed for a given language family is reliable? Or to determine what phylogenetic reconstruction method to use when given a particular character dataset? Or to determine which characters to use in a new phylogenetic analysis? Or to understand why two phylogenetic analyses might differ? Explicit models of language evolution – especially parametric ones – will greatly enable the exploration of how different conditions impact the accuracy of different phylogeny reconstruction methods, and help us answer these questions.

Models of linguistic character evolution Various stochastic models of linguistic character evolution have been proposed or implicitly suggested in simulation studies and statistical analyses of language evolution [14, 22, 25, 1, 31]. Models of linguistic character evolution differ in several ways: (a) they may assume that all evolution is treelike, so that no borrowing occurs, or they may explicitly model borrowing, (b) they may assume that evolution is clock-like or not, (c) they may assume that characters evolve identically or not, (d) they may assume that different characters evolve independently, or not, and (e) they may allow homoplasy to occur, or not. For example, in [25] we proposed the "perfect phylogenetic network" non-parametric model of language evolution in which every character evolves down a tree contained within a network, but without any homoplasy, and in [39] we proposed a parametric model that allows for borrowing and limited homoplasy.

Parametric stochastic models are those in which the probability distribution of the observed data comes from a given family of possible distributions, with the actual member of the family being determined by a collection of

numerical parameters. For example, a parametric model may assume that all characters evolve independently, but without any homoplasy, and require extra parameters to specify the probability that a given character will change its state (and thus evolve into a new state) on a given edge in the tree. Thus, the model is fully specified by the underlying tree and the parameter values for the substitution mechanism. One of the virtues of a parametric model is that simulation studies can be performed under the model, which allows a researcher to study the accuracy of a reconstruction method under a range of conditions. That is, in a simulation study, the researcher selects (or randomly generates) a model phylogeny, which may either be a tree or a network (a tree with contact edges, which are added to model borrowing). Characters are then evolved down the phylogeny, thus producing states for each character at every leaf in the phylogeny. This resultant character state matrix (where the rows correspond to the languages that occupy the leaves of the phylogeny, and the columns correspond to the characters) can then be given to a collection of phylogeny reconstruction methods. The resultant phylogenies are computed, of course, without knowledge of the model phylogeny, and hence may have errors. However, these estimated phylogenies can each be compared to the model phylogeny, and the degree of error can be quantified. By performing simulation studies, it is possible to determine various aspects of the performance of phylogeny reconstruction methods. For example, it is then possible to characterize the model conditions (i.e., parameter values) under which a method will yield a highly accurate estimation, and those under which the method will be more likely to have errors. It can also be possible to characterize the model conditions under which all methods do well, or all methods do poorly, or - perhaps - one method outperforms another.

To our knowledge, the most complex model of language evolution is the one we provided in [39], which allows characters to evolve with limited (but identifiable) homoplasy, borrowing between lineages, and assumes the characters evolve independently but not identically. Under this complex model, we showed in [39] that with limited borrowing, the evolutionary history is identifiable (which implies that given enough data, the true history can be reconstructed if a good reconstruction method is used), and we provided polynomial time algorithms which are statistically consistent under the model. However, these theoretical results do not provide any direct insight into the relative performance of any methods on finite datasets (as identifiability and statistical consistency are concepts that address what is possible given unbounded amounts of data).

Simulation studies have been used to evaluate the performance of phylogeny reconstruction methods in molecular systematics (i.e., the estimation of phylogenies from DNA, RNA, or amino-acid sequences), and have been able to shed light on how different molecular evolutionary processes impact phylogenetic accuracy. For example, such simulation studies have been used to show how rates of evolution, taxonomic sampling, reticulation events, and deviations from a molecular clock (the biological equivalent of a lexical clock, which asserts that the number of times each site within a molecular sequence changes should be proportional to time) impact absolute accuracy as well as the relative performance of different phylogenetic reconstruction methods (the scientific literature is too large to provide a comprehensive list of papers, but see [6, 15, 19, 24, 26, 27, 28]). These studies tend to lend support to the conjecture that statistical estimation methods, such as maximum likelihood, will produce the most accurate results, provided that the statistical estimation methods are based upon models that are a good fit to the underlying evolutionary processes (which, of course, cannot be known on a real dataset).

Linguistic evolution has many of the same features as biological sequence evolution, but certain issues are of particular relevance because of differences between the two domains. In particular, in linguistic evolution (as compared to biological sequence evolution), there is generally much less homoplasy. That is, in our experience, careful application of the Comparative Method [16] by an experienced and knowledgeable historical linguist can identify many of the borrowings between languages, and thus produce data matrices which have very little homoplasy. In addition, since certain characters are known to evolve through parallel evolution, these can also be identified in advance, and screened out (removed from the analysis). Finally, some phylogenetic reconstructions of language families have been based solely upon lexical data, while others have used morphological and complex phonological characters as well.

In this paper we provide a simulation study in order to address how different features of evolutionary processes impact the accuracy of phylogenies estimated from linguistic character data, using phylogeny reconstruction methods that have been used in recent analyses of linguistic datasets. We describe the experimental design in Section

2, and report on the results of this experiment in Section 3. We then conclude in Section 4 with a discussion about ongoing work in modelling language evolution.

2 The simulation study

2.1 Overview

We performed a simulation study under the model we provided in [39], in order to evaluate the performance of six existing phylogeny reconstruction methods under a wide range of model conditions. All model networks and trees we used had 30 leaves, and ranged from no contact edges (i.e., tree-like evolution) to networks with three contact edges. To capture the characteristics of a real dataset, such as the IE dataset that was analyzed in [30], we evolved from 301 to 361 characters down the trees, the bulk of which (300 or more) were modelled after lexical characters, and the remainder were morphological. We set the parameters of the simulation in order to produce datasets with different homoplasy levels, deviations from a lexical clock, and deviations from the rates-across-sites assumption.

We had two types of characters, lexical and morphological, and we divided lexical characters into three types according to the rate of evolution, obtaining fast lexical, medium lexical, and slow lexical characters. Within each of the four types of characters, the parameters of the evolutionary process were drawn identically and independently from a distribution which we describe below.

Our experiment was designed to help us understand how the conditions of the evolutionary process (e.g., the presence of borrowing between lineages (i.e., reticulations), relaxing the strong molecular clock, relaxing the rates-across-sites assumption, and the degree of homoplasy) impact the accuracy of the different phylogeny reconstruction methods we studied. However, we were also interested in seeing if there were any clear indications of relative performance between different methods, in evaluating the consequences for "screening datasets" to remove likely homoplastic characters, in using weighting schemes to give higher weight to those characters which were considered likely to be more resistant to homoplasy, and in restricting analyses to lexical-only datasets as compared to using lexical and morphological characters together.

2.2 Phylogeny reconstruction methods

The phylogeny reconstruction methods we study in this paper include most of the standard methods used in molecular phylogenetics as well as two newer methods proposed explicitly for reconstructing phylogenies on languages. The methods studied include four character-based methods and two distance-based methods. The four character-based methods each produce several trees, and hence we use a standard consensus method (the "majority consensus") in order to return a single estimate of the evolutionary history. (See [11, 38] for two books providing information on phylogenetic reconstruction methods used in biology, including many of the methods studied here.)

UPGMA The UPGMA (unweighted pair grouping method of agglomeration) algorithm is a distance-based method which is designed to work well when the evolutionary processes obeys the *lexical clock* assumption. This is the same method used in lexicostatistical analyses. As is standard for this method, we use Hamming distances (the number of characters on which a given pair of languages have different states) to define the distance matrix between the set of languages.

Neighbor joining NJ, or *Neighbor Joining* [34], is a particular agglomerative clustering technique used in molecular phylogenetics, which is able to reconstruct accurate phylogenies even when the clock assumption does not hold. Of all distance-based methods, NJ is believed to be one of the best. The corrected distance $D(i,j)$ between two languages i and j is computed by calculating corrected distances for each type of character (i.e., slow lexical (SL), medium lexical (ML), fast lexical (FL) and morphological (Mo)), and then averaging them:

$$D(i,j) = \frac{\text{num}_{SL} D_{SL}(i,j) + \text{num}_{ML} D_{ML}(i,j) + \text{num}_{FL} D_{FL}(i,j) + \text{num}_{Mo} D_{Mo}(i,j)}{\text{num}_{SL} + \text{num}_{ML} + \text{num}_{FL} + \text{num}_{Mo}}$$

where num_X is the number of characters in class X as X ranges over the four classes of characters, $HD_X(i,j)$ is the Hamming Distance between languages i and j computed only on the basis of the characters in the class X, and $D_X(i,j) = -\log(1 - HD_X(i,j)/\mathrm{num}_X)$. Under the model we propose, if we do not allow reticulation, homoplasy or heterotachy (that is, violation of the rates-across-sites assumption), then the $D(i,j)$ will be consistent statistical estimators of genuine tree distances that are concordant with the topology of the underlying genetic tree. That is, when the numbers of replicates num_X are large, the $D(i,j)$ will be close to a collection of leaf-to-leaf distances on a tree with edge lengths whose shape is that of the genetic tree. (Note that using NJ with uncorrected distances is not a statistically consistent estimator of phylogenies, except for cases where the lexical-clock assumption holds.)

Maximum parsimony and weighted maximum parsimony *Maximum Parsimony*, or MP, is an optimization problem which seeks a tree on which a minimum number of character state changes occurs. When the characters are weighted, then the objective is to find a tree in which the total weighted number of character state changes is minimized. Both MP and WMP are NP-hard problems, which has the consequence that exact solutions cannot be guaranteed using polynomial time algorithms. Hence, we use heuristics in the PAUP* [37] software package to find good (though not provably optimal) solutions. Since there can be many equally good solutions, the majority consensus tree of the set of optimal solutions is returned. In our experiments, we used a weighting scheme where the weight of every morphological character is 50, and the weight of every lexical character is 1; this weighting scheme was selected on the basis of the perceived relative resistance of the *screened* datasets we analyzed in [30], and so reflects the expectation that screened morphological characters will be much more resistant to homoplasy and borrowing than screened lexical characters.

Weighted maximum compatibility When all the characters evolve without homoplasy down a tree, then the tree is called a "perfect phylogeny", and each of the characters is said to be "compatible" on the tree. Weighted Maximum Compatibility, or WMC, is the optimization problem which seeks a tree with the maximum weighted compatibility score, which is computed by adding up all the weights of each character which is compatible on the tree. WMC is an NP-hard problem, which we "solve" heuristically through the use of the WMP (weighted maximum parsimony) analysis – by taking all the trees which are optimal for WMP, scoring each one under the WMC criterion, and then returning those trees which are optimal under WMC. Once again, we return the majority consensus of the best trees we find. Since WMC (like MP and WMP) is NP-hard, these solutions are not guaranteed to be globally optimal solutions.

Gray & Atkinson's method (G&A) The method (originally presented in [14]) designed by Russell Gray and Quentin Atkinson operates as follows. First, each multistate character is replaced by a binary encoded version of the character, and these binary characters are then interpreted as restriction sites and analyzed under a rates-across-sites model in the MrBayes software [20]. MrBayes uses a Markov chain Monte Carlo exploration of tree and parameter space to simulate the Bayesian posterior distribution of the tree and parameters under its model. The run of the Markov chain is divided into a *burn-in* and a *stationary* phase of equal length. Each phase contains 75,000 iterations. During the second, *stationary* phase, 100 simulated values are recorded at regular intervals. We report the majority consensus tree of those 100 values.

Comments These six methods are most of the ones that have been used in phylogenetic reconstructions on linguistic datasets: UPGMA is the standard method used in lexico-statistics, maximum parsimony has been used in several dataset analyses (see for example the analysis of the Bantu language family in [17]), and Gray & Atkinson used their method to analyze an Indo-European dataset [14] and to analyze the Bantu language family [18]. In our own analyses [25, 30, 32] of IE datasets, we have used methods designed to find trees that optimize weighted maximum compatibility; most recently, we have modified this approach by looking for "perfect phylogenetic networks" which use the obtained trees as candidates for the underlying genetic tree. Thus, WMC is included in order to represent a technique that is closely allied to our approaches. Neighbor joining is included in order to

provide a method from the biological systematics toolkit (although it has also been used in phylogenetic analyses for language families).

Some comments should be made about the use of weighting in maximum parsimony or maximum compatibility. The weights in these methods are supposed to reflect the relative resistance to borrowing and homoplasy, with higher weights given to characters that are believed to be more resistant to borrowing and homoplasy. In our studies, we have used WMC *only after* the data have been screened to remove clearly homoplastic characters. In our simulation study, we have the weights for all lexical characters set to 1 and weights for all morphological characters set to 50, to reflect the expectation that morphological characters, after screening, will have a very low incidence of homoplasy and borrowing, as compared to lexical characters. Thus, WMC and WMP should *not* be used in this way on unscreened data. However, we include data showing how WMC and WMP perform on unscreened data in order to show how the use of extremely poor estimates of character weights impacts phylogenetic accuracy.

Software We used PAUP* [37] for all the phylogeny reconstruction methods we studied, except for Gray & Atkinson. For our implementation of Gray & Atkinson, we used MrBayes [20]. We used the r8s program [35] to generate our model trees. See the appendix for the commands we used.

2.3 Model network generation

Our simulation generates random binary trees using a Yule process with per individual birth rate 1 conditioned to have the requisite number of terminal taxa at time 1, as implemented by Sanderson's `r8s` software [35]. Thus, the trees we generated by r8s have edge lengths that represent elapsed time, and are normalized so that all paths from root to terminal leaf have length 1. We indicate the elapsed time on edge e by $t(e)$.

In our model of evolution, the implementation of borrowing requires the existence of contact edges between lineages. Those contact edges must be added to the generated binary tree and the resulting structure is no longer a tree but a network. Two languages must be in existence at the same absolute time to borrow from each other. Thus contact edges can only be generated between points that are equidistant from the root.

Suppose we have a pair of tree edges in different lineages that overlap for some interval of time $[t_1, t_2]$. Let $t_0 \leq t_1$ be the time of the most recent common ancestor of the points in the two edges. We begin by laying down *candidate* contact edges according to an inhomogeneous Poisson process – some of these candidate contact edges will be removed to form the final reticulate network via a procedure that we describe below. The infinitesimal probability that a candidate contact edge occurs during the time interval $[t, t+dt]$ between the two edges is initially $\mu(t-t_0)^{-1}\,dt$ for $t_1 \leq t \leq t_2$, where μ is some parameter controlling the initial laying down of candidate contact edges. This prescription has the two features that the probability a pair of edges will be connected by a contact edge is increasing with the length of the overlap of the edges in time and decreasing from the time at which the lineages containing the edges diverged.

The inclusion of contact edges between two edges that issue from the same branch point doesn't introduce reticulation and so we discard such candidate edges. We would then like to condition the contact edge generation process to create exactly n contact edges (for some specified integer n) between edges that don't issue from the same branch point. This conditioning eliminates the parameter μ and gives a network with a prescribed number of possibilities for borrowing. We may approximate the effect of such a conditioning by the following procedure that allows at most one contact edge between any two tree edges.

- For each pair π of tree edges that overlap for some non-empty time interval $[t_1, t_2]$ and have their most recent common ancestor at time $t_0 < t_1$ (so that the edges don't issue from the same branch point), assign a score $S(\pi)$ given by
$$S(\pi) = -\log \frac{(t_1 - t_0)}{(t_2 - t_0)}.$$

- Draw without replacement n pairs of edges, such that each pair π is drawn with probability equal to its normalized score $S(\pi)/\sum_{\pi'} S(\pi')$.
- Once n such edge pairs have been drawn, the corresponding contact edges are drawn by generating a time of contact t_c for the edge via

$$t_c = t_0 + (t_1 - t_0)\exp\left(U\log(\frac{(t_2 - t_0)}{(t_1 - t_0)})\right),$$

where U is a random variable uniformly distributed on $[0, 1]$ and these random variables are independent for different edge pairs. In particular, each pair of edges in the tree is connected by at most one contact edge.

2.4 Stochastic model of language evolution

We use the stochastic model of language evolution proposed in [39]. In that model, there is a fixed collection of linguistic characters, each of which has an infinite collection of possible states. A language is represented by the particular states it exhibits for each of the characters (note, however, that two leaves in the tree *may* be identical with respect to the characters, due to insufficient evolution). Languages evolve down an underlying tree with added reticulate edges that represent contact events between lineages. At a contact event, the state of each character may be instantaneously transferred from the lineage at one end of the edge to the lineage at the other end (that is, one lineage "borrows" the character state of another), and replaces the character state inherited from its genetic parent.

The set of possible states for a given character consists of a distinguished state h^*, which we call the homoplastic state, that may arise at several points in time in the same or different lineages, and an inexhaustible set of states denoted n, n', n'', ..., which we call the non-homoplastic states, each of which may arise no more than once across all times and all lineages as the result of a transition from another (homoplastic or non-homoplastic) state.

Given an edge in a model tree with edge lengths $t(e)$ indicating elapsed time on the edge e, the transition events along the edge follow a homogeneous Poisson process with a rate to be described later.

In this paper we simplify the model of single character evolution by taking the transition probabilities to be identical for all edges and all characters and to depend on a single parameter $0 \leq$ **homoplasy_factor(c)** ≤ 1 which depends upon the character c, as follows:

- $\Pr(h^*, h^*) = \Pr(n, n) = 0$
- $\Pr(n, h^*) =$ **homoplasy_factor(c)**
- $\Pr(n, n') = 1 -$ **homoplasy_factor(c)**
- $\Pr(h^*, n) = 1$

The probability that the state of a character c is transferred along a contact edge e depends upon two parameters, one which depends upon the edge, and one which depends on the character. The parameter that depends upon the edge is **edge_borrowing(e)**, which is the probability that the most easily borrowed character transmits a state in one of the two directions for the edge. This parameter can depend upon the edge, to reflect the possibility that some contact events are more extensive than others; however, in our simulation study we set **edge_borrowing(e)** to the same value for all edges. The other parameter is **character_borrowing(c)**, which reflects the probability that the character will transmit its state across a contact edge. This parameter depends upon the character since some character types are more easily borrowed than others (in particular, some lexical characters and morphological characters are not readily borrowed, but other lexical characters and some phonological characters are easily borrowed). In our simulations, we set **character_borrowing(c)** for each of the different character classes, but set it to 0 for the morphological characters since we do not permit them to be borrowed. For a given edge and character, the probability of borrowing in one direction along the edge is the same as the probability of borrowing in the other direction. Thus, the probability of character c transmitting its state in one direction on the edge e is given by $\frac{1}{2}\mathbf{edge\_borrowing(e)} \times \mathbf{character\_borrowing(c)}$.

2.5 Character evolution

The phylogenetic network consists of an underlying genetic tree with additional contact edges, whose edge lengths $t(e)$ represent the elapsed time on edge e (so that contact edges have $t(e) = 0$). We now describe additional parameters so that we can describe how each character evolves down this network, independently of the other characters.

We begin by defining the expected number of changes of a given character on a given edge. This expected number of changes will depend upon the edge e (and specifically on $t(e)$), but also on some additional parameters which we need to define. However, before we define these parameters we need to describe the concepts of *ultrametricity* and *rates-across-sites*.

The condition of ultrametricity is that the path length from the root to each leaf is identical; when all taxa are current-day and path lengths represent time, ultrametricity is immediate. However, when path lengths represent the expected number of changes of a random site, then ultrametricity depends upon the *lexical clock* hypothesis, which is generally discounted. We quantify the deviation from the lexical clock through the use of a parameter σ_0 which we define below.

The rates-across-sites assumption is quite standard in molecular systematics and its underlying models, but is nevertheless also questionable. It states that every two characters evolve proportionally – so that if one character evolves at twice the speed of another character on one branch of the tree, then it evolves at twice the speed of the other character on every branch in the tree. We quantify the deviation from this assumption through the parameter σ_1, which we also define below. (See [10] for a study discussing the rates-across-sites assumption and statistical identifiability of divergence times.)

We now define the expected number of transitions on edge e for character c to be:

$$t(e) \times V_e \times \textbf{ height\_factor(c)} \times W_{c,e},$$

where **height_factor(c)** is a parameter that only depends on the class of the character c, and V_e and $W_{c,e}$ are random variables with

$$V_e = \exp(X_e - \sigma_0^2/2), \quad X_e \sim N(0, \sigma_0^2)$$

and

$$W_{c,e} = \exp(X_{c,e} - \sigma_1^2/2), \quad Y_{c,e} \sim N(0, \sigma_1^2).$$

The normal random variables X_e and $Y_{c,e}$ are independent over all choices of edge e and character c. Note that V_e and $W_{c,e}$ both have mean 1. The parameter σ_0 controls the degree to which the model deviates from a lexical clock (that is, fails to be ultrametric). The parameter σ_1 controls the degree to which the *rates-across-sites* assumption fails.

Model conditions Certain parameters of the model are specific to the phylogenetic network but vary with the experiments; these include the model phylogeny topology (in particular the number of contact edges) and the elapsed time on each edge. We fix the parameter **edge_borrowing(e)** which indicates the probability of a character state being transmitted on the edge. In addition to these network-specific parameters, there are parameters that can change according to the character; these include **homoplasy_factor(c)**, **character_borrowing(c)**, **height_factor(c)**, *deviation from lexical clock* (represented by the parameter σ_0), and *heterotachy* (represented by the parameter σ_1).

We add the following constraints to the parameter system to suppress additional degrees of freedom unnecessary for the purpose of our experiments:

- We set the parameters σ_0 and σ_1 identically for all characters within any one simulation, but vary these parameters between different experiments.
- The other parameters have one set value for each of the four character classes we consider.
- The value of **height_factor(c)** only depends on the class of the character and increases as we go from slow to medium to fast lexical characters. Its value for morphological characters is the same as that for slow lexical characters.

- The values of **homoplasy_factor(c)** and **character_borrowing(c)** are identical across the three classes of slow, medium and fast lexical characters. They only differ between lexical and morphological characters.

- Because morphological characters will not undergo borrowing, **character_borrowing(c)** is a parameter for lexical characters only. We are therefore able to add the constraint that for all contact edges and for all character classes **edge_borrowing(e)** = **character_borrowing(c)**, which reduces the parameterization related to borrowing to a single parameter.

For each experiment, we set the above stochastic parameters partly by targeting measurable model conditions such as observed homoplasy and borrowing, as well as other considerations such as the number of contact edges, number and type of characters analyzed, etc. For each experiment, we generate 32 random networks by taking a tree generated by r8s and adding the contact edges. For each of these networks, we make three random draws of the random variables (V_e and $W_{c,e}$). For each of these draws, we generate three random sequences of characters at the root and simulate their evolution. In total, for each each experiment, we generate a data point averaged over 288 measurements: 32 topological networks $\times 3$ draws of the random variables $\times 3$ randomly evolved datasets.

The state at the root of each character is drawn as h^* (with probability **homoplasy_factor(c)**) or n (with probability 1-**homoplasy_factor(c)**). After each run of the simulation process, we obtain a set of sequences, one for each leaf in the phylogenetic tree or network, where each sequence represents the states of the language represented by that leaf for each of the characters in the simulation process. This resulting character state matrix is used by each reconstruction method to produce an estimated tree, which can then be compared with the genetic tree within the model phylogenetic network.

Preliminary experiments showed that most of the variability in the estimated trees was due to variability in the network, and this is why we have many more replicates of the network itself, rather than evolving many datasets down any given network.

2.6 Error rates for phylogeny reconstruction methods

We compute two types of error rates: "false negatives" and "false positives", which we now define. Recall that each phylogeny reconstruction method produces a tree, which is compared to the "genetic tree" contained within the model network. (That is, although a phylogenetic network can contain many trees, there is an underlying binary tree to which the contact edges are added in order to produce the network; it is this tree that we will make our comparisons to.)

Every edge in a tree defines a bipartition of the leaves of the tree, and hence can be identified with that bipartition. Two trees on the same leaf set can thus be compared on the basis of their bipartitions. A bipartition in the genetic tree that is missing from the estimated tree is said to be a "false negative", while a bipartition that appears in an estimated tree that does not appear in the genetic tree is a "false positive". The number of false negatives is bounded by $n - 3$, where there are n leaves, and so the "false negative rate" (FN rate) is defined to be the number of false negatives, divided by $n - 3$. Similarly, the false positive rate (FP rate) is the number of false positives, divided by $n - 3$. Genetic trees are always binary, but estimated trees may not be - consensus trees, in particular, will often not be fully resolved. However, when estimated trees are binary, then their false negative rates and false positive rates are identical. In general, though, we can only assert that the false positive rate is always no more than the false negative rate. We focus our attention on False Negative rates, but provide information about false positive rates as well. (The average of these two rates is often referred to as the Robinson-Foulds rate [33].)

Each data point represents an average of 288 measurements. We report the average false negative and false positive rates between the majority consensus tree for the reconstruction methods and the genetic tree generated by r8s (the genetic tree).

3 Experimental results

3.1 Preliminary discussion

We now describe our experimental results. We begin by noting some conditions that hold throughout all the experiments. Parameter settings (specifically **character_borrowing(c)** and **homoplasy_factor(c)**) are set so that on the low homoplasy or screened datasets, 1% of the lexical characters and none of the morphological characters evolve homoplastically, and 6% of the lexical characters and none of the morphological characters evolve with borrowing, while on the moderate homoplasy or unscreened datasets, 13% of the lexical and 24% of the morphological characters are homoplastic, and 7% of the lexical and none of the morphological evolve with borrowing. All these settings were made to reflect the empirical data analyses in [32] for low homoplasy datasets ("screened datasets" in [32]) and moderate homoplasy datasets ("unscreened datasets" in [32]).

The borrowing parameter in our experiments replicates not all the lexical borrowing to be found in a real-language dataset, but only borrowings that *are not detected as borrowings*. Thus, we take no account of borrowings from languages not in the dataset, nor of borrowings between languages in the dataset that can be detected using the usual criteria (such as failure to reflect the regular sound changes diagnostic of the borrowing language).

One of the consequences of the settings we chose is that before "screening", the morphological characters are *much more* likely to be homoplastic than the lexical characters, and after screening they are much less likely. The weighting we use for the weighted parsimony and weighted compatibility methods are identical for both screened and unscreened datasets, where morphological characters are weighted 50 and lexical characters are weighted 1.

Summary of experiments We begin by describing the 28 different basic experiments we ran, each consisting of a model condition (parameters for the evolutionary process) and the number and type of characters simulated under each condition. For each of these basic experiments, we produced 288 datasets. Thus, all in all we created 9216 datasets, each of which was analyzed by the six phylogeny reconstruction methods we studied.

The 28 different experiments we ran can be grouped into four sets.

- Basic experiment: We fixed $\sigma_0 = 0.3$ and $\sigma_1 = 1.2$, reflecting intermediate values for these parameters. We then allowed the number of contact edges to vary from 0 to 3, and the homoplasy level to vary from low (to reproduce the conditions of screened data) to moderate (to reproduce the conditions of unscreened data). For each experiment, we generated 300 lexical and 60 morphological characters, with the 300 lexical grouped evenly between slow, medium, and fast evolving characters. This produced eight different model conditions.

- Experiment 2: We set the number of contact edges to three, and $\sigma_0 = 0.3$. We set σ_1 to be either 0.6 or 1.8, and homoplasy levels to be either low or moderate. For each experiment, we generated 300 lexical and 60 morphological characters, with the 300 lexical grouped evenly between slow, medium, and fast evolving characters. This produced four different model conditions.

- Experiment 3: We set the number of contact edges to three, and $\sigma_1 = 1.2$. We let σ_0 be either 0.15 or 0.45, and we let the homoplasy level be either low or moderate. For each experiment, we generated 300 lexical and 60 morphological characters, with the 300 lexical grouped evenly between slow, medium, and fast evolving characters. This produced four possible model conditions.

- Experiment 4: We fixed $\sigma_0 = 0.3$ and $\sigma_1 = 1.2$, and we let the number of contact edges be 0 or 3, the homoplasy level be either low or moderate, and we varied the number and type of characters in three ways: 360 lexical and 1 morphological, 300 lexical and 1 morphological, and 300 lexical and 20 morphological. Regardless of their number, the lexical characters remain grouped evenly between slow, medium, and fast evolving characters. This produced 12 possible model conditions.

Our discussion, provided below, explores the impact of various model conditions (homoplasy levels, deviations from a lexical clock, deviations from the rates-across-sites assumption, and choice of dataset) on the performance of the six phylogeny reconstruction methods. In each of these experiments, we report the false negative rate.

False positive rates are not shown due to space limitations, but can be summarized as follows. UPGMA and NJ produce binary trees, and hence for these two methods their false positive and false negative rates are identical. The remaining methods (G&A, MP, WMP, and WMC) all use the majority consensus method to produce their output, and for these the false positive rates are lower than their false negative rates. In general, we see that the false positive rates are quite low for these four methods – often below 1%, but almost always below 5%. Furthermore, much of the time the false positive rates of these four methods are very close, and don't really help distinguish between them (the cases where there is a difference are generally restricted to the low homoplasy settings where G&A tends to do less well with respect to both false negative and false positive rates than the other methods).

3.2 Impact of homoplasy

We begin by considering the impact of the level of homoplasy on a phylogenetic analysis. Recall that we set the parameter values for our "low homoplasy" and "moderate homoplasy" datasets to reflect what we observed for our screened and unscreened datasets, respectively, in [30], and this has the consequence that morphological characters are *more* homoplastic than lexical characters for unscreened data, but less homoplastic than lexical characters for screened data. However, the weighting we use for the weighted parsimony and weighted compatibility methods (where morphological characters receive higher weight than lexical characters) is identical for both conditions, and is therefore not appropriate for unscreened data.

In Figure 1 we show the results when the model phylogeny is a tree, and in Figure 2 we show the results when the model phylogeny is a network with three contact edges. We see that screening improves weighted parsimony and weighted compatibility the most, which is not surprising since the weighting scheme is inappropriate for the unscreened data. Thus, the improvement in accuracy of the weighted MP and weighted MC methods obtained as a result of screening is to be expected.

We also see an improvement in MP's performance from unscreened to screened, and this too is to be expected since maximum parsimony will tend to improve as the homoplasy level decreases (in particular, maximum parsimony should be accurate when the characters evolve without any homoplasy).

However, there is no change in performance for the other methods between screened and unscreened data, indicating that these methods are not designed to extract better phylogenetic signal under conditions of low homoplasy.

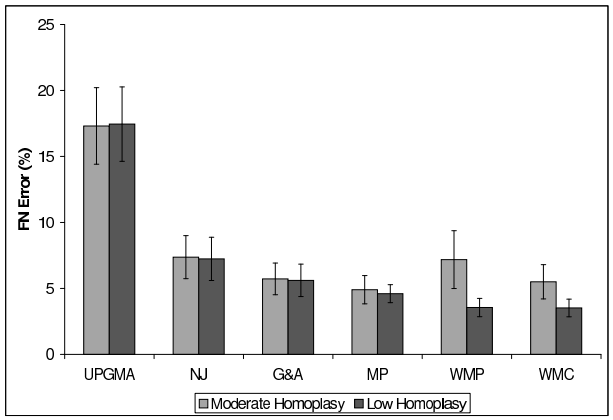


Figure 1: Impact of homoplasy on accuracy of phylogeny reconstruction methods for 300 lexical characters and 60 morphological characters evolved down a phylogenetic tree under a moderate deviation from a lexical clock ($\sigma_0 = 0.3$) and moderate deviation from the rates-across-sites assumption ($\sigma_1 = 1.2$).

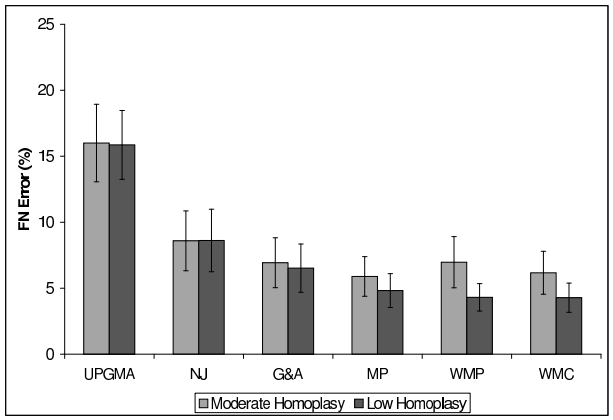


Figure 2: Impact of homoplasy on accuracy of phylogeny reconstruction methods for 300 lexical characters and 60 morphological characters evolved down a phylogenetic network with three contact edges under a moderate deviation from the lexical clock ($\sigma_0 = 0.3$) and moderate deviation from the rates-across-sites assumption ($\sigma_1 = 1.2$).

3.3 Impact of deviation from a lexical clock

We now examine the impact of varying the deviation from a lexical clock, from almost clock-like behavior (with $\sigma_0 = 0.15$) to a moderate deviation (with $\sigma_0 = 0.45$). We show the results on the screened datasets obtained from a phylogenetic network with three contact edges, and with moderate deviation from the rates-across-sites assumption ($\sigma_1 = 1.2$); results for other conditions (including unscreened datasets) were similar in terms of the impact of this parameter on performance. Error rates increase for all methods as the deviation from the lexical clock increases, but this is most pronounced for UPGMA and quite slight for the other methods.

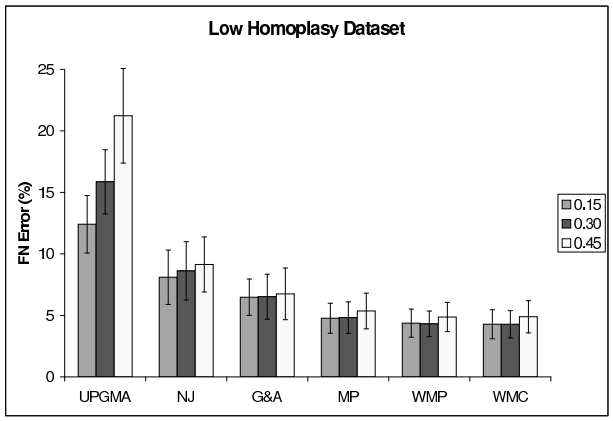


Figure 3: Impact of the deviation from a lexical clock on phylogenetic analyses of a 30-taxon phylogenetic network with three contact edges, from 300 lexical characters and 60 morphological characters evolved under low levels of homoplasy and with a moderate deviation from the rates-across-sites assumption ($\sigma_1 = 1.2$). We vary the deviation from a lexical clock from low ($\sigma_0 = 0.15$) to moderate ($\sigma_0 = 0.45$).

3.4 Impact of heterotachy

In Figure 4 we show the effect on phylogenetic analyses of deviating from the rates-across-sites assumption to various degrees, by exploring the difference in accuracy obtained as σ_1 varies from 0.6 (which is close to the rates-across-sites) to $\sigma_1 = 1.8$ (which is further away), on data simulated on a phylogenetic network with three contact edges and low homoplasy; the same trends are observed for other model conditions. The rates-across-sites assumption is critical to statistical models that attempt to estimate parameters under the assumption that all the sites evolve as *multiples* of each other (i.e., some faster and some slower, but with a constant ratio held between all sites). This is a standard assumption in phylogenetic analyses since it enables distance-based methods to be statistically consistent under suitable conditions, and it also enables dating of internal nodes.

Interestingly, we see that as σ_1 increases - i.e., as we relax the rates-across-sites assumption - methods *improve* in accuracy. The degree of improvement is small for UPGMA, and largest for the character-based methods. One explanation for this is that as the rates-across-sites assumption is relaxed, the range of rates-of-change exhibited by the set of characters on any given edge will also increase (with high probability); this, in particular, increases the probability that edges that are quite "short" (i.e., edges e for which $t(e)$ is small) will exhibit some changes by some characters, making these edges more likely to be inferred by a phylogeny reconstruction method.

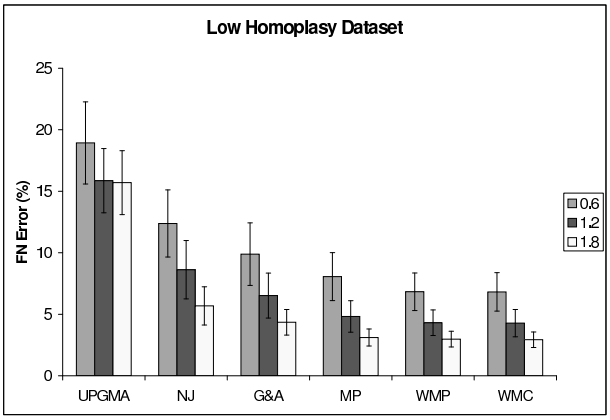


Figure 4: Impact of heterotachy (deviation from the rates-across-sites assumption) on the accuracy of phylogenetic reconstruction methods on data (300 lexical characters and 60 morphological characters) evolved down a phylogenetic network with three contact edges with low homoplasy, and with moderate deviation from a lexical clock ($\sigma_0 = 0.3$). The bars refer to the different values for σ_1.

3.5 Varying the proportion of lexical and morphological characters

Our next analysis considered the impact of using combined datasets (both morphological and lexical together) versus lexical-only datasets, for low homoplasy levels (set to reflect the estimated homoplasy levels in [30] for the "screened" datasets). Recall that in our simulations, we set the parameters for screened morphological characters so that there is no borrowing (this is true even of unscreened morphological characters) and so that they exhibit much less homoplasy than lexical characters. The inclusion of morphological characters into a dataset thus reduces the rate of homoplasy and borrowing. We look at four different possibilities: (a) 360 lexical and one morphological, (b) 300 lexical and one morphological, (c) 300 lexical and 20 morphological, and (d) 300 lexical and 60 morphological. The comparison between (a) and (b) mostly addresses the impact of adding more data of the same type (lexical); the comparison between (b), (c) and (d) reflects the consequence of adding morphological characters to a dataset which is primarily lexical. Finally, the comparison between (a) and (d) allows us to see the consequence of choosing between lexical and morphological characters, when the total amount of data is kept fixed. In Figure 5,

we see the result of this experiment, on screened datasets obtained by simulating down a phylogenetic network with three contact edges.

In comparing (a) and (d), we see the consequence of choosing 360 lexical versus 300 lexical and 60 morphological characters. Here we see that the distance-based methods and Gray & Atkinson's method are not improved by exchanging lexical characters for morphological characters (in fact, neighbor joining even gets worse), while maximum parsimony (weighted and unweighted) and weighted maximum compatibility improve - with WMC improving the most. Comparisons between (a) and (b) are as expected: decreasing the number of characters while not changing their nature decreases the accuracy of all methods. Comparisons between (b), (c), and (d) show all methods improving and thus indicate that all methods are able to improve in accuracy by adding characters (confirming the earlier inference), but show some differences between methods. In particular, the distance-based methods show much less improvement between (b) and (d) (i.e., as the number of morphological characters is increased from 1 to 60), possibly because the estimated distances are not becoming substantially more accurate with the increasing data.

The results for other model settings for screened data are similar; however, as expected, on the unscreened datasets weighted MP and weighted MC methods get worse as additional morphological characters are added rather than better. The obvious explanation is the weighting scheme used by these methods, as it is inappropriate for unscreened data.

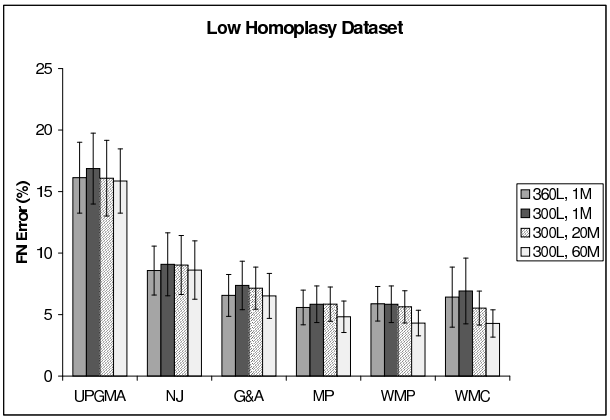


Figure 5: Impact of data selection on the accuracy of phylogenetic reconstructions on data evolved down a phylogenetic network with three contact edges, under low homoplasy ("screened data"), moderate deviation from a lexical clock ($\sigma_0 = .3$), and moderate deviation from the rates across sites assumption ($\sigma_1 = 1.2$).

3.6 Impact of the number of contact edges

In Figure 6, we show the results of our experiment in which we vary the number of contact edges from 0 (for tree-like evolution) to 3, for low homoplasy datasets ("screened data"), with moderate deviation from the lexical clock ($\sigma_0 = 0.3$) and moderate deviation from the rates-across-sites assumption ($\sigma_1 = 1.2$). Mostly what we see here is as expected: most methods return better estimates of the genetic tree when there is no borrowing (or less borrowing) between lineages. Two aspects of this study are surprising, however: first, reticulation, in the form of contact edges, does generally lead to increased error, but not as much as might be expected (though some methods are more adversely impacted than others). The other surprising observation is that UPGMA gets *better* with added contact edges. Understanding why this is so will require further investigation. Similar trends exist for other model conditions for screened data, but results for unscreened data differ only in degree: UPGMA gets better and the other methods get worse. However, the significance of these trends isn't clear, given the magnitude of the standard deviations.

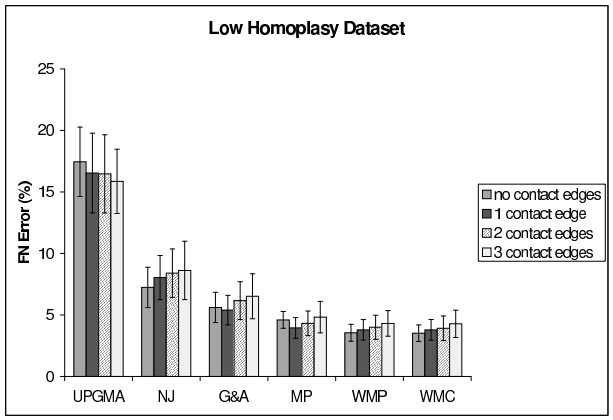


Figure 6: Impact of the number of contact edges on phylogenetic reconstructions of a phylogenetic network with three contact edges, from 360 characters (300 lexical and 60 morphological) evolved under low homoplasy, moderate deviation from a lexical clock ($\sigma_0 = 0.3$), and moderate deviation from the rates-across-sites assumption ($\sigma_1 = 1.2$).

3.7 Relative performance of different methods

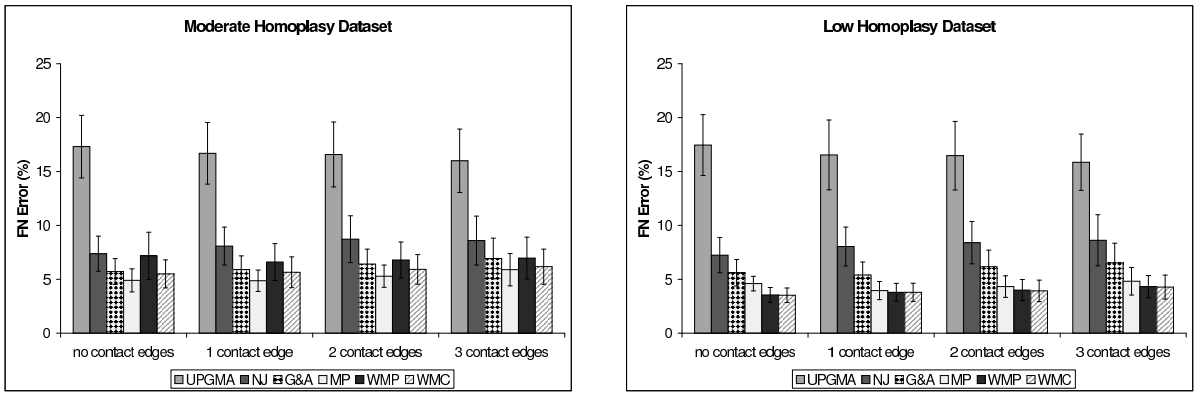
We turn now to the question of relative performance of different methods. Interestingly, if we exclude weighted maximum parsimony and weighted maximum compatibility, the relative performance of the remaining methods is consistent across all model conditions, with UPGMA the worst, NJ the next, Gray & Atkinson next, and finally MP. The difference between the methods depends upon the model condition, but the gaps between UPGMA and NJ and between NJ and G&A are generally large, while the differences between G&A and MP are relatively small. Here we show experiments to demonstrate the conditions in which the gaps in performance between these methods are smallest, and where they are largest.

In general, the gap in performance G&A and MP is only small when working with unscreened data (i.e., moderate levels of homoplasy instead of low), since G&A doesn't improve with reductions in homoplasy but MP does. The cases where the gap between G&A and MP is extremely small are for the unscreened data, with few morphological characters, and with a low deviation from the rates-across-sites assumption – see, for example, Figure 8a and Figure 10a. More generally, however, the gap between G&A and MP is smallest for those model conditions in which all methods have a harder time, whereas as the model conditions improve (for example, by increasing the number of characters, or deviating from the rates-across-sites assumption), the gap increases. Thus, in particular, on the screened datasets, MP is clearly better than G&A.

3.8 Summary

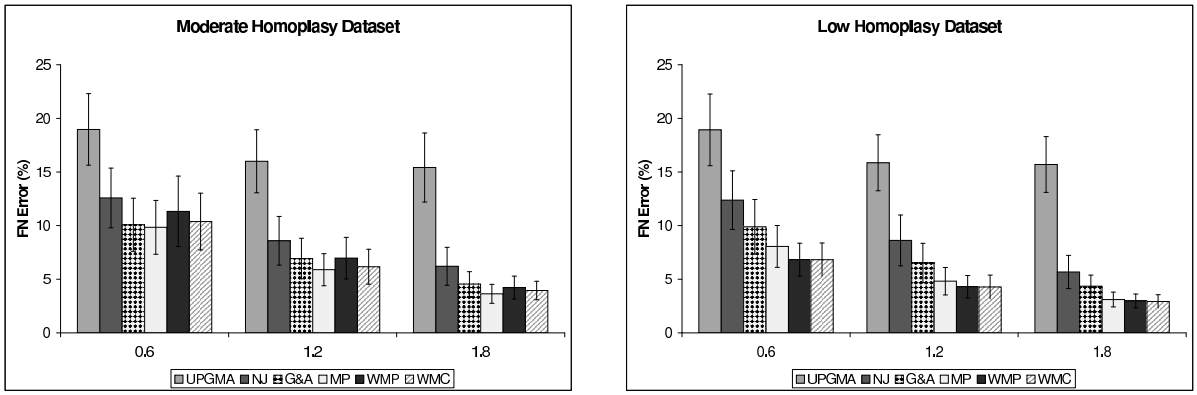
Our study showed the following:

- There was a consistent pattern of relative accuracy of phylogenies reconstructed using these methods, with the two distance-based methods (UPGMA and neighbor joining) less accurate than the character-based methods (maximum parsimony, weighted maximum parsimony, weighted maximum compatibility, and Gray & Atkinson). UPGMA was always the worst by far.
- The relative performance within the character-based methods was often quite close, but Gray & Atkinson's method was always the least accurate. The only conditions under which Gray & Atkinson's (G&A) method was close to Maximum Parsimony (MP) were for unscreened data.
- Deviating from the lexical clock made all methods somewhat worse, but had the biggest impact on UPGMA.



(a) (b)

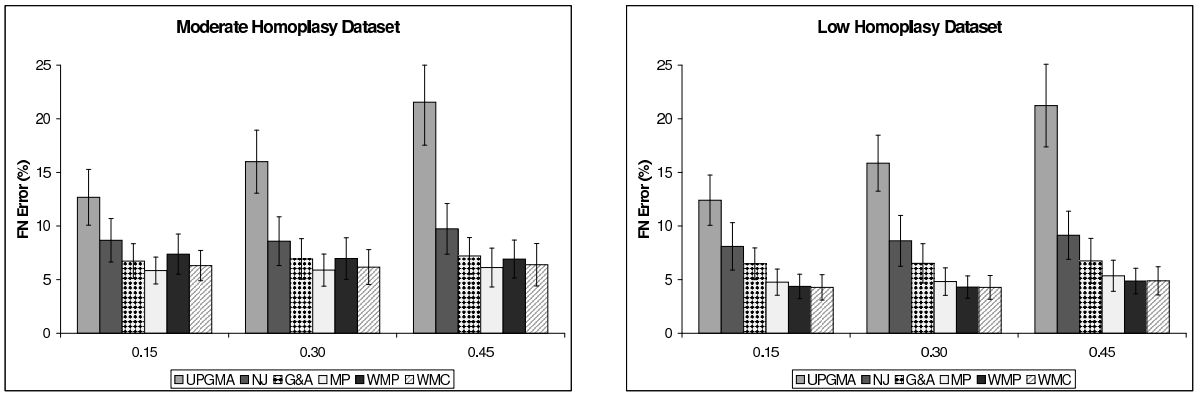
Figure 7: Impact of the number of contact edges on phylogenetic reconstruction methods for 300 lexical characters and 60 morphological characters, under two levels of homoplasy (moderate in (a), and low in (b)). All datasets evolve under a moderate deviation from a lexical clock ($\sigma_0 = 0.3$) and moderate deviation from the rates-across-sites assumption ($\sigma_1 = 1.2$).



(a) (b)

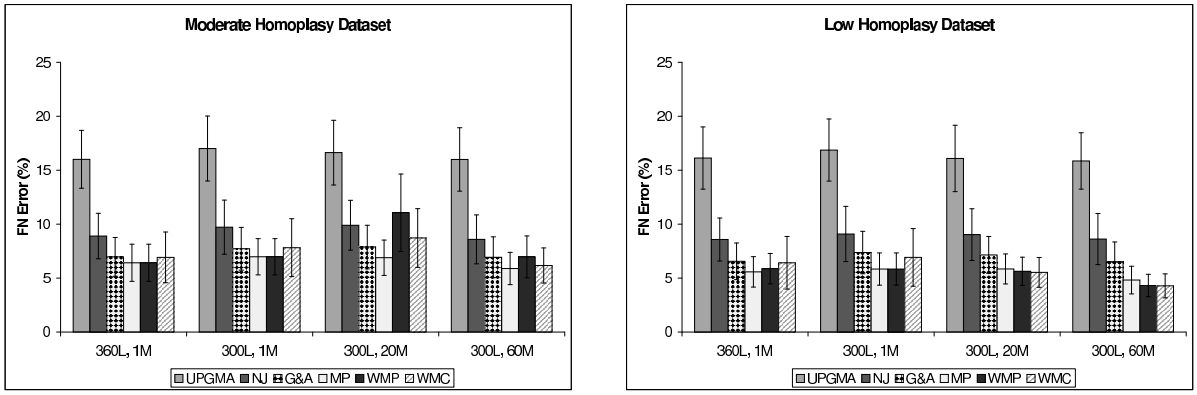
Figure 8: Impact of the deviation from the rates across sites assumption on phylogenetic reconstruction methods, for 300 lexical characters and 60 morphological characters, under two levels of homoplasy (moderate in (a) and low in (b)). All characters evolve down a phylogenetic network with three contact edges under a moderate deviation from a lexical clock ($\sigma_0 = 0.3$). We vary σ_1, the parameter for deviating from the rates-across-sites assumption, from low (0.6) to moderate (1.8).

- Deviating from the rates-across-sites assumption improved the character-based methods but had little impact on the distance-based methods.
- The incidence of borrowing between languages generally made reconstructions less accurate, but not dramatically so; surprisingly, it made UPGMA somewhat more accurate.
- The inclusion of screened morphological characters with low levels of homoplasy improves the accuracy of all phylogeny reconstruction methods, but especially MP, WMP, and WMC.
- Using WMP and WMC on data with high levels of homoplasy produced poor results, but using WMP and WMC on data with lower levels of homoplasy (and with weights reflecting the relative resistance to homoplasy) improved accuracy.



(a) (b)

Figure 9: Impact of deviating from the lexical clock on phylogenetic reconstruction methods for 300 lexical characters and 60 morphological characters, under two homoplasy levels (moderate in (a) and low in (b)). All characters evolve down a phylogenetic network with three contact edges under a moderate deviation from the rates-across-sites assumption ($\sigma_1 = 1.2$). We vary the deviation from the lexical clock from low ($\sigma_0 = 0.15$) to moderate ($\sigma_0 = 0.45$).

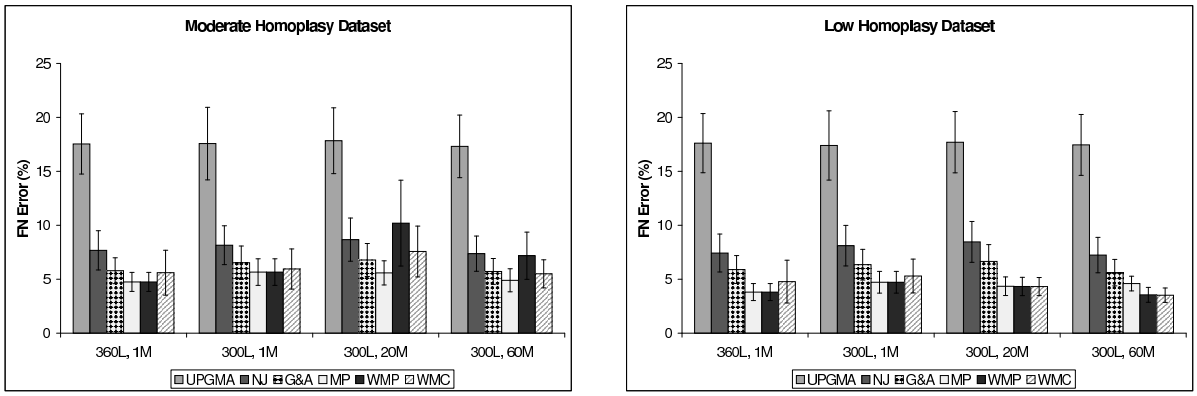


(a) (b)

Figure 10: Impact of data selection on phylogenetic reconstruction of a phylogenetic network with three contact edges under two homoplasy levels (moderate in (a) and low in (b)). All characters evolve under a moderate deviation from a lexical clock ($\sigma_0 = 0.3$) and moderate deviation from the rates-across-sites assumption ($\sigma_1 = 1.2$).

4 Discussion

What does our study imply about the choice of phylogeny reconstruction method, or about the choice of dataset for a phylogenetic analysis? At a minimum, the study indicates that phylogenies estimated using distance-based methods (e.g, the UPGMA used in lexico-statistics, and neighbor joining) are much less accurate than phylogenies estimated using character-based methods. However, stronger statements can also be made. It is clear that data selection has the potential to make a very big impact on the accuracy of the phylogenies that are constructed. In particular, careful screening of datasets so as to reduce homoplasy and/or borrowing, and using characters which are more resistant to homoplasy and borrowing (i.e., screened morphological and phonological characters), *can* yield significantly improved results, although not all methods are able to take advantage of these modifications.



(a) (b)

Figure 11: Impact of data selection on the relative performance of reconstruction methods of a tree, under two levels of homoplasy (moderate in (a) and low in (b)). All characters evolve under moderate deviation from a lexical clock ($\sigma_0 = 0.3$) and moderate deviation from the rates-across-sites assumption ($\sigma_1 = 1.2$).

Furthermore, *when* screened datasets that include morphological characters as well as lexical characters are analyzed, then the best analyses are clearly obtained by using weighted maximum parsimony or weighted maximum compatibility, and in these cases the difference in performance between these methods and other methods can be quite substantial.

On the other hand, with the exception of UPGMA, under most conditions we studied, all the remaining methods (even neighbor joining) were able to reconstruct all but (about) 10% of the edges of the true tree. In other words, probably all the methods (except for lexicostatistics, which uses UPGMA) will agree on a substantial portion of the tree, and probably succeed in reconstructing the major subgroups. The differences between methods really come down to finer details of the phylogenetic analysis. In IE terms, these questions might be: where does Germanic lie in the Indo-European family tree, is Italo-Celtic a subgroup, are Greek and Armenian sisters? These "fine details", in other words, are where much of the intense debate lies within the historical linguistics community.

On the other hand, our study did not address the performance of phylogenetic *network* reconstruction methods, although the use of these methods for phylogeny reconstruction of language families is of increasing interest; recent studies [9, 12, 13, 18, 22, 25] have used diverse techniques to produce these estimates, including SplitsTree [2, 21], Neighbor Net [8], Median-Joining [3, 4, 5], and our Perfect Phylogenetic Network [25] method. However, the relative performance of these methods has not been studied, due in part to a lack of accepted criteria by which to evaluate the performance of phylogenetic network reconstruction methods (see, however, [23, 29]) and lack of simulation tools. These studies will also require a range of models to cover the wide range of "reticulation" in language evolution, from the end where the underlying "genealogical" tree is clearly defined (even if contact occurs), to the other end where there is no underlying genealogical tree, but rather a dialect continuum.

We now briefly touch upon some of the outstanding theoretical questions. Currently methods for phylogenetic analysis are fundamentally limited to using characters which exhibit at most one state on each language, and hence cannot be used for "polymorphic" characters which exhibit two or more states on some languages. Polymorphism is, unfortunately, quite common - especially among lexical characters. Thus, clearly one of the outstanding problems in linguistic phylogenetics is to develop methods which can utilize polymorphic characters, and to do this we need to begin with appropriate models of how polymorphism arises. Some simple examples of polymorphism are arise from semantic shift, whereby two characters with different meaning gradually become indistinguishable within one language with respect to meaning, so that the language then has two words for the same basic meaning. English examples of this include *big* and *large*, or *rock* and *stone*. In our initial work [7] on modelling polymorphism, we considered the case where polymorphism arises only from semantic shift, but no homoplasy is permitted. However, polymorphism can also arise from borrowing, through the incorporation of a loan word into

a language, as well as from other processes; in addition, we now have good evidence that while morphological characters may generally evolve with little (or no) homoplasy, the same is not true for lexical characters. Hence, our first model for polymorphism is incomplete, and must be extended.

Another issue that must be addressed comes about because a speech community is not comprised of a single individual speaking the language, but a community of speakers, and thus *population effects* must be considered. In effect, the basic problem of estimating phylogenies in languages that still confronts historical linguistics is that models of linguistic character evolution are too simple in that they do not take population effects into consideration. This is obvious in polymorphism, but it holds as well for the modelling of all characters.

It is worth noting that the same issue arises in biology. There is a divide between the "between-species" stochastic models of biological character evolution typically used in phylogenetic analysis, which usually assume monomorphism and also do not take population heterogeneity into consideration, and the "within-species" models of population genetics, in which there is only partial geographical or reproductive separation between sub-populations, leading to polymorphism within sub-populations and the possibility that different samples of individuals from each of the sub-populations may exhibit varying evolutionary trees.

Mathematical models of evolution that would take these population effects explicitly into consideration would have to include modifications of the underlying graphs (so that vertices and edges in the phylogenies would represent populations of speakers, rather than a single individual speaker), as well as of the stochastic processes that operate on the characters. As important as this is to historical linguistics, little has yet been done.

For many researchers, the question of estimating dates at internal nodes is of central importance. However, from a mathematical point of view, estimating dates at internal nodes is extremely difficult without significant constraints on the deviation from a lexical clock (the linguistic equivalent of a molecular clock). Thus, our viewpoint on this matter is that it's best to limit phylogenetic reconstruction to estimating the underlying branching process, rather than also estimating the dates. See [10, 22, 36] for more on this topic.

5 Acknowledgments

This work began while Tandy Warnow was a Fellow at the Radcliffe Institute for Advanced Research, and was supported by the Institute. Tandy Warnow, Luay Nakhleh, and François Barbançon were supported in part by NSF grant BCS-0312830. Tandy Warnow was also supported by the Radcliffe Institute for Advanced Research, by the Program for Evolutionary Dynamics at Harvard University, and the Institute for Cellular and Molecular Biology at the University of Texas. Steve Evans was supported in part by NSF grant DMS-0405778. Don Ringe was supported in part by NSF grant BCS-0312911.

References

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

[2] H-J. Bandelt and A. W. M. Dress. Split decomposition: a new and useful approach to phylogenetic analaysis of distance data. *Mol. Phylogenet. Evol.*, 1:242–252, 1992.

[3] H. J. Bandelt, P. Forster, and A. Rohl. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.*, 16:37–48, 1999.

[4] H. J. Bandelt, P. Forster, B.C. Sykes, and M.B. Richards. Mitochondrial portraits of human populations using median networks. *Genetics*, 141:743–753, 1995.

[5] H. J. Bandelt, V. Macaulay, and M.B. Richards. Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. *Mol. Phyl. Evol.*, 16:8–28, 2000.

[6] O. Bininda-Emonds, S. Brady, J. Kim, and M. Sanderson. Scaling of accuracy in extremely large phylogenetic trees. In *Proc. 6th Pacific Symposium on Biocomputing (PSB01)*, pages 547–557. World Scientific, 2001.

[7] M. Bonet, C.A. Phillips, T. Warnow, and S. Yooseph. Constructing evolutionary trees in the presence of polymorphic characters. *SIAM J. Computing*, 29(1):103–131, 1999. (A preliminary version appeared in the ACM Symposium on the Theory of Computing, 1996.).

[8] D. Bryant and V. Moulton. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, 21:255–265, 2003.

[9] D. Bryant and M. Steel. Fast algorithms for constructing optimal trees from quartets. *J. Algs.*, 38(1):xxx–xxx, 2001.

[10] S.N. Evans and T. Warnow. Unidentifiable divergence times in rates-across-sites models. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1:130–134, 2005.

[11] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts, 2004.

[12] P. Forster, T. Polzin, and A. Rohl. Evolution of English basic vocabulary within the network of Germanic languages. In P. Forster and C. Renfrew, editors, *Phylogenetic Methods and the Prehistory of Languages*, pages 131–138. McDonald Institute for Archaeological Research, 2006.

[13] P. Forster and A. Toth. Towards a phylogenetic chronology of ancient Gaulish, Celtic, and Indo-European. *Proceedings of the National Academy of Sciences*, 100(15):9079–9084, 2003.

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

[15] D. M. Hillis, J. P. Huelsenbeck, and D. L. Swofford. Hobgoblin of phylogenetics. *Nature*, 369:363–364, 1994.

[16] H.M. Hoenigswald. *Language Change and Linguistic Reconstruction*. University of Chicago Press, Chicago, 1960.

[17] C.J. Holden. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. *Proceedings of the Royal Society of London, Series B*, 269:793–9, 2002.

[18] C.J. Holden and R. Gray. Rapid radiation, borrowing, and dialect continua in the Bantu languages. In P. Forster and C. Renfrew, editors, *Phylogenetic Methods and the Prehistory of Languages*, pages 19–32. McDonald Institute for Archaeological Research, 2006.

[19] J. P. Huelsenbeck and D. M. Hillis. Success of phylogenetic methods in the four-taxon case. *Syst. Biol.*, 42:247–264, 1993.

[20] J.P. Huelsenbeck and R. Ronquist. MrBayes: Bayesian inference of phylogeny. *Bioinformatics*, 17:754–755, 2001.

[21] D. H. Huson. SplitsTree: A program for analyzing and visualizing evolutionary data. *Bioinformatics*, 14(1):68–73, 1998.

[22] A. McMahon and R. McMahon. Why linguists don't do dates: evidence from Indo-European and Australian languages. In P. Forster and C. Renfrew, editors, *Phylogenetic Methods and the Prehistory of Languages*, pages 153–160. McDonald Institute for Archaeological Research, 2006.

[23] B.M.E. Moret, L. Nakhleh, T. Warnow, C.R. Linder, A. Tholse, A. Padolina, J. Sun, and R. Timme. Phylogenetic networks: modeling, reconstructibility, and accuracy. *IEEE/ACM Transactions on Computational Biology and Biocomputing*, 1(1), 2004.

[24] L. Nakhleh, B. M. E. Moret, U. Roshan, K. St. John, J. Sun, and T. Warnow. The accuracy of fast phylogenetic methods for large datasets. In *Proc. 7th Pacific Symposium on Biocomputing (PSB02)*, pages 211–222. World Scientific, 2002.

[25] L. Nakhleh, D. Ringe, and T. Warnow. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language (Journal of the Linguistic Society of America)*, 81(2):382–420, 2005.

[26] L. Nakhleh, U. Roshan, K. St. John, J. Sun, and T. Warnow. The performance of phylogenetic methods on trees of bounded diameter. In *Proceedings of the 1st Workshop on Algorithms in Bioinformatics (WABI)*, 2001. Aarhus, Denmark.

[27] L. Nakhleh, U. Roshan, K. St. John, J. Sun, and T. Warnow. The accuracy of phylogenetic methods for large datasets. In *Proceedings of the Pacific Symposium on Biocomputing (2002)*, pages 211–222, 2002. Kauai, Hawaii.

[28] L. Nakhleh, U. Roshan, K. St. John, J. Sun, and T. Warnow. Designing fast converging phylogenetic methods. *Bioinformatics*, 17:190S–198S, 2001.

[29] L. Nakhleh, J. Sun, T. Warnow, C.R. Linder, B.M.E. Moret, and A. Tholse. Towards the development of computational tools for evaluating phylogenetic network reconstruction methods. In *Proc. 8th Pacific Symp. on Biocomputing (PSB 2003)*, 2003.

[30] L. Nakhleh, T. Warnow, D. Ringe, and S.N. Evans. A comparison of phylogenetic reconstruction methods on an IE dataset. *The Transactions of the Philological Society*, 3(2):171–192, 2005.

[31] G.K. Nicholls and R.D. Gray. Dated ancestral trees from binary trait data. *Unpublished*, 2006.

[32] D. Ringe, T. Warnow, and A. Taylor. Indo-European and computational cladistics. *Transactions of the Philological Society*, 100(1):59–129, 2002.

[33] D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.

[34] N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425, 1987.

[35] M. Sanderson. r8s software package. http://loco.ucdavis.edu/r8s/r8s.html.

[36] D. Ringe S.N. Evans and T. Warnow. Inference of divergence times as a statistical inverse problem. In P. Forster and C. Renfrew, editors, *Phylogenetic Methods and the Prehistory of Languages*, pages 119–130. MacDonald Institute for Archaeological Research, 2006.

[37] D. Swofford. PAUP*: Phylogenetic analysis using parsimony (and other methods), version 4.0. 1996.

[38] D.L. Swofford, G.J. Olsen, P.J. Waddell, and D.M. Hillis. Phylogenetic inference. In D.M. Hillis, C. Moritz, and B.K. Mable, editors, *Molecular Systematics*. Sinauer Associates, Sunderland, Massachusetts, 1996.

[39] T. Warnow, S.N. Evans, D. Ringe, and L. Nakhleh. A stochastic model of language evolution that incorporates homoplasy and borrowing. In P. Forster and C. Renfrew, editors, *Phylogenetic Methods and the Prehistory of Languages*, pages 75–90. MacDonald Institute for Archaeological Research, 2006.

Appendix

Software commands We provide the details about the commands we used with each software package.

Generating trees with R8s

```
#nexus
begin r8s;
    simulate diversemodel=yule_c T=1.0 ntaxa=30 nreps=1 seed=1965807332 speciation=1
        charevol=yes ratemodel=normal startrate=1.0 changerate=0.0 infinite=yes
        minrate=1.0 maxrate=1.0;
    describe plot=phylo_description;
end;
```

UPGMA using PAUP*

```
begin paup;
    UPGMA treefile=PAUP/PAUP_up_out.trees replace;
quit;
```

Neighbor joining using PAUP*

```
begin paup;
    NJ treefile=PAUP/PAUP_nj_out.trees replace;
quit;
```

Maximum parsimony or weighted maximum parsimony using PAUP*

```
begin paup;
    set criterion=parsimony maxtrees=100 increase=no;
    weights 1:1-300, 50:301-360;
    hsearch start=stepwise addseq=random nreps=25 swap=tbr;
    filter best=yes;
    set maxtrees=100 increase=no;
    hsearch start=current swap=tbr hold=1 nbest=100;
    filter best=yes;
    pscores all/ ci ri rc hi scorefile=PAUP_wmp_out.scores replace=yes;
    savetrees file=PAUP_wmp_out.trees replace=yes format=nexus;
quit;
end;
```

Gray & Atkinson's method using MrBayes

```
begin mrbayes;
    set autoclose=yes nowarn=yes;
    lset rates=gamma;
    mcmcp ngen=150000 printfreq=10000 samplefreq=750
        nruns=1 nchains=4 savebrlens=yes filename=Bayes_out;
    mcmc;
    set nowarnings=yes;
    sumt filename=Bayes_out burnin=100;
quit;
end;
```

Parameter settings We used the following settings for our simulations.

The parameter height_factor was set to 1.0, 2.0 and 3.0 for slow, medium and fast lexical characters, and 1.0 for morphological characters. In addition, we set the remaining parameters for each of the two homoplasy levels, as follows:

- Moderate homoplasy dataset: Lexical - 13.0% incompatible due to homoplasy is achieved with homoplasy_factor = 0.05788. Lexical - 7.0% incompatible due to borrowing is achieved with edge_borrowing = character_borrowing = 0.3035. Morphological - 24% incompatible due to homoplasy is achieved with homoplasy_factor = 0.1215.

- Low homoplasy dataset: Lexical - 1.0% incompatible due to homoplasy is achieved with homoplasy_factor = 0.01321. Lexical - 6.0% incompatible due to borrowing is achieved with edge_borrowing = character_borrowing = 0.281425. Morphological - no borrowing, no homoplasy, so homoplasy_factor - 0.0 = edge_borrowing= character_borrowing.