A STOCHASTIC MODEL OF LANGUAGE EVOLUTION THAT INCORPORATES HOMOPLASY AND BORROWING

TANDY WARNOW, STEVEN N. EVANS, DONALD RINGE, AND LUAY NAKHLEH

1. Introduction

The inference of evolutionary history, whether in biology or in linguistics, is aided by a carefully considered model of the evolutionary process and a reconstruction method which is expected to produce a reasonably accurate estimation of the true evolutionary history when the real data match the model assumptions and are of sufficient quantity. In molecular systematics (i.e., the inference of evolutionary histories from molecular data), much of the research effort has focused in two areas: first, the development of increasingly parameter rich models of molecular sequence evolution, and second, the development of increasingly sophisticated software tools and algorithms for reconstructing phylogenies under these models. The plethora of software for reconstructing phylogenies from molecular data is staggering. By comparison, much less has been done in historical linguistics in terms of developing statistical models of character evolution or reconstruction methods, suggesting that there is perhaps much to be gained by doing so.

To date, although some models have been proposed for language evolution, all have failed in some significant ways. In particular, linguistic models either explicitly or implicitly have assumed that no homoplasy (i.e., parallel evolution and/or back–mutation) occurs (see for example (Ringe *et al.* , 2002; Taylor *et al.* , 2000; Warnow, 1997)). Most, but not all, have not modelled borrowing between languages. In this paper, we go beyond earlier models by explicitly incorporating both homoplasy and borrowing into our model. We show that this model is

TW supported by NSF ITR grants 0331453, 0312830, and 012680, by the Program for Evolutionary Dynamics at Harvard University, the Radcliffe Institute for Advanced Study, the Institute for Cellular and Molecular Biology at the University of Texas, and The David and Lucile Packard Foundation.

SNE supported in part by NSF grants DMS-0071468 and DMS-0405778.

DR supported in part by NSF grant BCS-0312911.

not only quite rich, but reflects important properties of real linguistic characters. Our examination of phylogenetic inference under the model therefore has important ramifications for phylogenetic analyses of real data.

The paper is organised as follows. We present a model of language evolution that incorporates homoplasy in Section 2. Computational issues involved with inferring phylogenetic trees under this model, including identifiability and calculating likelihood scores, are discussed in Section 3; proofs of the mathematical results in this section are provided in an appendix. We then discuss how we can incorporate borrowing into our model for homoplasy, and discuss the issues for inferring evolution under both homoplasy and borrowing in Section 4. We compare our model and its ramifications for phylogenetic analysis to biological models in Section 5. We discuss the consequences for phylogenetic analysis in historical linguistics in Section 6 and conclude in Section 7 with a mention of a model similar to the homoplasy-free special case of ours that was proposed and investigated in (Mossel & Steel, 2004) because of its rather simple theoretical properties.

2. A new model of language evolution on trees

Most of the models used in studies of language evolution explicitly or implicitly assume that evolution is treelike, and that linguistic characters evolve without homoplasy. We begin our discussion with a precise statement of what these assumptions mean.

2.1. The standard assumptions of language evolution.

The simplest models about language evolution are expressed in the following two statements:

- Evolution is treelike, i.e. the Stammbaum model applies.
- When a linguistic character changes its state, it changes to a new state not yet in the tree, i.e. there is no “back–mutation” nor parallel evolution.

The first condition is understood in the linguistics community but the second condition is not quite as standard, and so it is worth discussing in greater detail.

The phenomenon of back–mutation and/or parallel evolution is called *homoplasy*. When there is no homoplasy in a character, then all changes of state for that character result in new states. 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.

When characters evolve without homoplasy, it is sometimes very easy to reconstruct the underlying unrooted tree, because each character yields definite information about the branching order within the tree. For example, if a character evolves so as to change only once in the tree, then that character defines a split of the leaves of the tree (i.e., the languages) into two parts, and that split is associated with the unique edge of the tree on which the character changes state. Characters that evolve in this way include practically unrepeatable phonological innovations, and are highly informative about evolutionary history.

The assumption that linguistic characters evolve without homoplasy is made implicitly in simulation studies (see (McMahon & McMahon, 2003) for one such paper), and was also made explicitly by Ringe & Warnow in their early work (Taylor *et al.* , 2000) where they sought a tree on which all the characters were compatible. However, the pairing of these two assumptions, namely that evolution is treelike and that linguistic characters evolve without homoplasy (i.e., that perfect phylogenies exist) is too strong, as our analysed showed definitively that perfect phylogenies do not exist for our Indo-European (IE) dataset.

One possible explanation for the inability to find perfect phylogenies is that the evolution isn't treelike, i.e., some contact between lineages must be inferred in order to explain the evolutionary process. In that case, the "network model" makes sense, as described below. Non-treelike evolution is clearly realistic, since lexemes are transmitted between lineages. However, since sound changes can make the presence of such transmissions apparent, the character states that are assigned to lexemes that are borrowed are not identical (this is a direct consequence of the comparative method). Where reticulate evolution becomes problematic is when the borrowing is not detected, because then the comparative method will assign identical states to lexemes that are not actually cognate. Thus, although lexical characters are particularly vulnerable to borrowing, careful application of the comparative method can detect much - but not necessarily all - of the borrowing, and hence alleviate much of the problem involved in using lexical characters, at least with respect to this issue.

However, another possible explanation of the non-existence of perfect phylogenies is that characters may evolve with back–mutation or parallel evolution. Events of both types have been documented in linguistics; for instance, sound changes can occur repeatedly, resulting in phonological characters that exhibit parallel evolution (Ringe, Warnow, and Taylor 2002: 66-7), and a language can shift a semantic function from its morphology to its syntax, resulting in morphological characters that have back-mutations (see the discussion below).

Models that have incorporated borrowing into linguistic evolution while still assuming homoplasy-free evolution have been used in several simulation studies (McMahon & McMahon, 2003) and in the inference of Indo-European evolution (Nakhleh *et al.* , 2004). Such models thus explicitly assume that all incompatibilities of characters with the genetic tree must be explained by borrowing, since homoplasy is not permitted. Indeed, determining whether incompatibility is due to borrowing or homoplasy is one of the major challenges in historical linguistic reconstruction.

In order to make progress on this difficult question, we have begun by formulating a stochastic model of linguistic evolution that formally models homoplasy in ways that are consistent both with linguistic scholarship and with our own experience with Indo-European characters. We will show that inference of evolutionary trees is both theoretically possible and realistically feasible, even in the presence of homoplasy, provided that homoplasy can be identified and dealt with appropriately.

2.2. **Different types of linguistic characters.** There are three types of characters - lexical, morphological, and phonological. Here we assume that phonological characters are binary, and that all character state assignments are made on the basis of rigorous application of the Comparative Method (Hoenigswald, 1960).

Homoplasy is an aspect of a character's evolution with respect to a particular set of languages, whereby a state for that character arises more than once in the evolutionary history of that set of languages. There are essentially two types of homoplasy: back–mutation (which means the reappearance of an ancestral state) and parallel evolution (whereby two languages have the same state, but no common ancestor of those languages has that state). Examples of both types of homoplasy exist in language evolution.

Homoplasy is possible for any type of linguistic character, although some characters are less likely to evolve with homoplasy than others. Our own study of linguistic characters in Indo-European suggests that true homoplasy (meaning either parallel evolution or back–mutation, not simple incompatibility due to borrowing) is very rare for morphological characters (although see the discussion later), but very likely for phonological characters and somewhat frequent for lexical characters. For example, phonological characters, which are frequently binary, can exhibit parallel evolution if the sound change is at all natural; loss of the consonant h is an example of such a sound change character that can evolve with parallel evolution. However, if phonological characters are based on phonemic mergers they do not exhibit back–mutation, since reversals of phonemic mergers do not occur.

Interestingly, linguistic scholarship makes it possible in many cases to identify the *homoplastic states* (that is, states which can arise more than once) based upon linguistic scholarship alone (i.e., before any phylogenetic analysis is done, and without reference to an estimated phylogeny), at least for morphological and phonological characters within the well-studied Indo-European family. For instance, the loss of /h/ in various Greek dialects and in late Latin is an obvious parallel development (Buck, 1955; Sihler, 1995); so is the merger of long /a:/ and long /o:/ in Germanic and in Slavic (but not in Baltic, which is more closely related to Slavic) (Brugmann, 1897); among lexical characters, the use of "nursery terms" originally meaning 'dad' as the usual words for 'father' in Gothic and Hittite must have occurred independently (Pokorny, 1959), since those languages were never in contact at any time after their ancestors began to diverge. The ability to identify homoplastic states *in advance* of a phylogenetic reconstruction will allow us to infer the evolutionary history both accurately and efficiently, with ambiguity in the phylogenetic estimation only when there is not enough morphological and phonological character data to fully resolve the tree (or when we are unable to identify accurately the homoplastic states).

In addition, the number of homoplastic states is very small, at most one for either morphological or phonological characters, and lexical characters too seem to have a very small number of homoplastic states (most of the time only one, but in principle this number could be unbounded). For some morphological characters, homoplasy can be a back–mutation to the state of "absence" or in parallel via mutations to a "default" state.[1] An example of the former is the superlative in *-ism̥mo- (Brugmann, 1906), which is clearly an innovation of Italo-Celtic that was subsequently lost in many Romance languages. Examples of the latter are scarce in archaic Indo-European languages but easier to find in their more modern descendants; for instance, the spread of the second-person singular ending /-st/ from the present indicative (where it originated) to the past and the subjunctive can be demonstrated to have occurred independently in English and German (Campbell, 1962; Lühr, 1984).

Phonological characters have homoplasy when the sound change occurs sufficiently naturally for it to arise more than once; in this case, the state indicating presence is the homoplastic state. In addition

[1]We are grateful to Bill Poser for reminding us of the latter phenomenon.

to the examples noted above, one can cite the merger of voiced and breathy-voiced stops in various clades of IE (Brugmann, 1897), the merger of "palatal" and "velar" stops in a different but overlapping set of clades (including Hittite but *not* the Luvian subgroup of Anatolian) (Melchert, 1987), and so on.

Thus, homoplastic states (ones that can arise more than once) within morphological and phonological characters can be easily identified, at least when the language family is well understood (even if its phylogeny is still unclear). On the other hand, the case of lexical characters is somewhat more difficult: even for the well-studied Indo-European family, accurate identification of homoplastic states without a given (and robust) phylogeny is not necessarily easy.

2.3. **Modelling character evolution.** We now state our parametric stochastic model of evolution.

In order to simplify the exposition, we will adopt the terminological convention that the term *homoplastic state* means a state that can appear homoplastically (i.e., one that can arise more than once in the tree). Thus the designation of a state as homoplastic is a feature of the model rather than the data: a homoplastic state may or may not appear in a homoplastic event in a particular random realisation of the model (that is, in a particular data set).

We will assume that there is at most one homoplastic state per character (it is trivial to extend the analyses and proofs to the case where there is a fixed finite number of homoplastic states), that each homoplastic state can be identified before a phylogenetic analysis of the data, and that the probability of each substitution depends only upon the type of states that are involved (i.e. whether the states are homoplastic or not).

We now define a very general model of individual site evolution for linguistic characters. We will associate a stochastic substitution matrix to each combination of edge in the tree and each character, as follows. We denote the homoplastic state by h^*, for "homoplastic", and the non-homoplastic states by n. The stochastic substitution matrix for the edge e and character c is defined by the following quintet:

- $p_{e,c}(n, h^*)$: the probability of a substitution of a non-homoplastic state with the homoplastic state.
- $p_{e,c}(n, n')$: the probability of a substitution of a non-homoplastic state with a new non-homoplastic state.
- $p_{e,c}(n, n)$: the probability of not changing, given that we start with a non-homoplastic state.

- $p_{e,c}(h^*, h^*)$: the probability of not changing, given that we start with the homoplastic state.
- $p_{e,c}(h^*, n)$: the probability of a substitution of the homoplastic state with a non-homoplastic state.

Thus $p_{e,c}(n, h^*)+p_{e,c}(n, n')+p_{e,c}(n, n) = 1$ and $p_{e,c}(h^*, h^*)+p_{e,c}(h^*, n) = 1$. Note that this is a very general model, since we do not assume that different characters have the same stochastic substitution matrices on any given edge, nor do we assume that these substitution matrices cannot change as we move across the tree. In this sense the model is highly unconstrained. Note also that we allow states to be "sinks", so that once a language is in that state there is no possibility of changing state (that is, we allow $p_{e,c}(h^*, h^*) = 1$ or $p_{e,c}(n, n) = 1$).

2.3.1. *Modelling how different characters can evolve differently.* How we allow variation between characters in this model involves issues that are familiar in biological phylogenetics. Do we want to assume that the evolution along an edge results from the operation of a dynamic process that differs from character to character only by the rate with which substitutions occur? That is, do we want to impose the analogue of the rates-across-sites assumption from biology (see (Evans & Warnow, 2004) for a discussion of the rates-across-sites assumption and an extensive list of references)? Or do we want to only make the minimum assumption that all sites evolve down the same tree? As we will see, for the conditions we assume - namely, that we can identify the homoplastic states - we do not need to make any assumptions constraining how characters can vary in their evolutionary processes in order to be able to reconstruct the tree. This is a surprising result that distinguishes our model from other models which do not explicitly assume the existence of sufficient homoplasy-free states.

2.3.2. *How the model works for different character types.* There are two different types of characters we will consider: those which represent the presence or absence of a given feature (phonological characters are the main example of this type), and those for which there is an unbounded number of possible homoplasy-free states.
Characters indicating presence/absence: For the first type of character (which reflects our binary phonological characters), the two possible states represent presence or absence, and evolution proceeds from absence to presence. Sound changes can occur more than once, but once a sound change has occurred in the tree, all nodes below the edge on which the sound change occurs will be recognisable as having undergone the sound change (i.e., parallel evolution is possible, but back-mutation

is impossible). Thus, $p_{e,c}(h^*, n) = 0$ and $p_{e,c}(h^*, h^*) = 1$. We will make one quite mild additional assumption, which is that for such characters, $0 < p_{e,c}(n, h^*) < 1$ for all edges e in the tree.
All other characters: For other types of characters (i.e., morphological and lexical), each state represents a different form for the character (semantic slot for the lexical characters), and hence there is an unbounded number of states for these characters. Morphological characters and lexical characters can have both types of homoplasy (back-mutation and parallel evolution), but in both cases we assume that the homoplastic states can be identified. (We acknowledge that in the case of lexical characters this identification may not be as reliable as in the cases of morphological or phonological characters; our mathematical analysis that follows addresses the case where we are able to make this identification.) We again make a mild assumption, which is that $0 < p_{e,c}(n, n') < 1$ for all edges e in the tree.

3. Inference of evolutionary history under our model.

We will now discuss issues involved with inferring evolutionary history under our models, beginning with the theoretical issue of identifiability, and then addressing actual methods for inferring evolutionary histories.

3.1. **Identifiability.**

The first issue is whether the model is *identifiable*. In essence, this is a question that asks whether it is possible to uniquely determine the model, as well as its associated parameters, from the probability of each possible pattern at the leaves. We leave that general question open, but show a positive answer to the fundamental question of whether the evolutionary tree (albeit not the location of the root nor the parameters of the evolutionary process) is identifiable:

Theorem 3.1. *The model tree (modulo the placement of the root and the parameters of evolution) is identifiable, provided that we are able to identify correctly the homoplastic states.*

The proof of this theorem is given in the appendix.

3.2. **Algorithms for inferring evolution under our model.**

Because the model tree is identifiable, this means that it is possible to reconstruct, with complete accuracy, the underlying (unrooted) evolutionary tree for a language family - provided that there are enough data, we use appropriate methods, and the family evolves under the model. Note that this statement does not imply the ability to estimate to any degree of accuracy other features about the evolutionary history – such as the location of the root, the parameters $p_{e,c}(\cdot,\cdot)$, dates at internal nodes if we assume a model in which there is a functional dependence between the $p_{e,c}(\cdot,\cdot)$ and such dates, etc.

Also, we need to qualify our statement about completely accurate reconstruction of the tree. Mathematically, having data on even an infinite number of characters may not be enough to reconstruct the tree perfectly if, as we consider more characters, the corresponding rates of linguistic evolution become slower or faster too precipitously. For example, suppose that we actually have an infinite number of lexical or morphological characters c and for some edge e $\sum_c (1-p_{e,c}(h^*,h^*)) < \infty$ and $\sum_c (1-p_{e,c}(n,n)) < \infty$. Then by a standard result from probability theory (the Borel-Cantelli lemma), with probability one only finitely many characters will exhibit a change of state on the edge e, and there is positive probability that no characters exhibit a change on e. In particular, if this state of affairs holds for every edge, then the data will "freeze" after a certain point and all leaves will exhibit the same state for all but a finite number of characters – implying that we are unable to reconstruct the tree with certainty. Similarly, if $\sum_c p_{e,c}(h^*,h^*) < \infty$ and $\sum_c p_{e,c}(n,n) < \infty$ for an edge e, then (again by the Borel-Cantelli lemma) with probability one all but finitely many characters will exhibit a change on edge e and there is positive probability that every character exhibits a change on e. If this state of affairs holds for every edge, then with probability one only finitely many characters will be informative and for the remaining characters the languages at the leaves of the tree will appear to be completely unrelated.

Note that this problem also occurs for models proposed for molecular evolution, and so this issue is not particular to the linguistic model we propose. However, if such pathologies are not present, then under our model algorithms for reconstructing phylogenetic trees can be designed which will yield reliable estimates of the true tree, as we now show.

3.2.1. *Algorithms for inferring evolution under morphological or lexical characters.* For morphological and lexical characters (i.e., those characters with an unbounded number of *possible* homoplasy-free states), there are two simple algorithms which will reconstruct the true tree. Each uses knowledge of the homoplasy-free states in order to infer explicit constraints on the topology of the underlying tree. The first method infers bipartitions on the leaf-set and is the simplest algorithmically, but also requires (probabilistically) more data in order to resolve the tree completely. The second method infers quartet trees and algorithmically more complex, but can use the data more efficiently.

Algorithm 1 (bipartition-based): The first algorithm seeks bipartitions defined by two distinct homoplasy-free states. If a character exhibits two homoplasy-free states and no other states in the family, then (under the assumptions of the model) the bipartition it defines on the set of languages corresponds to an edge in the tree. We therefore just collect all such bipartitions, and use standard polynomial time methods for constructing the minimal tree consistent with all the bipartitions (see (Gusfield, 1990)). Given enough characters of this type to infer each edge on which there is a change, we can reconstruct the true tree for the language family.
Algorithm 2 (quartet-based): Consider a character in which states 1 and 2 are known not to be homoplastic. Suppose languages A and B both have state 1 and languages C and D both have state 2. In this case, the only possible form for the tree on A, B, C, D is $AB|CD$ (i.e., there must be at least one edge separating the languages A and B from C and D). The algorithm proceeds as follows. First, we examine each character in turn, and for each pair of non-homoplastic states, we construct the trees on four-language sets using this rule. Then, after we have computed the set of all such quartet tree constraints, we seek a tree that is consistent with all the input constraints. Finding the tree that meets all the constraints is a computational problem that is in general NP-hard to solve (i.e., hard to solve efficiently) (it is equivalent to perfect phylogeny which is NP-hard (Bodlaender *et al.* , 1992; Steel, 1992)), but under some conditions is solvable in a time that is polynomial in the number of languages (i.e., "computationally feasible"). In particular, if the correct subtree is given for all quartets of languages, then the problem is solvable in polynomial time.

3.2.2. *Algorithms for phonological characters.* There are two approaches for using binary phonological characters. The first is to use linguistic knowledge to screen the dataset and remove all characters which have evolved with any homoplasy. This is the traditional approach, which tries to only use those phonological characters that represent very unusual sound changes, unlikely to evolve in parallel (Hoenigswald, 1960). The use of complex phonological characters is an example of this type – any single simple phonological character might be likely to evolve in parallel, but the conjunction of independent ones together might represent a highly unusual such character. For instance, the sequence of sound changes (a) Grimm's Law (Streitberg, 1896), (b) Verner's Law (Streitberg, 1896), (c) shift of stress to initial syllables, and (d) merger of unstressed /e/ with /i/ unless /r/ follows immediately – occurring in that order – is a complex phonological character which is probably sufficient to validate the Germanic clade even without further evidence (Ringe, 2005). (On the other hand, each of the sound changes of this complex character can be shown to have close parallels elsewhere: (a) in Armenian, (b) in late Middle English, (c) in Italic, Celtic, Latvian, etc., and (d) in Hittite and various other languages.)

Provided that such a screening can be done, phonological characters can then be quite useful for inferring evolution, since each then represents a binary split that must hold for the true tree.

Analyzing unscreened phonological characters can also be done, using the technique for estimating evolutionary distances we provide in the appendix, and applying methods such as neighbour joining (Saitou & Nei, 1987) which are guaranteed to be correct on *tree-like* distances (also called *additive* distances). Such an approach is guaranteed to be correct, but the conditions under which the approach holds are not necessarily realistic. The conditions for correct reconstruction of the true tree include that there be a fair abundance of phonological characters, and that they all be drawn from the same distribution. That is, unlike the case of either morphological or lexical characters, we cannot use any individual phonological character to yield information about the evolutionary tree, and instead must use the aggregate information among all the phonological characters. Therefore, all the characters must actually be essentially identical in their evolutionary process – a condition we do not impose on morphological or lexical characters. They can have different *rates* of evolution, but the rates of evolution must be drawn from a distribution which we can estimate from the data. These are strong conditions, and will not necessarily hold in practice. (Such issues arise in most statistical models – and in particular, in most statistical models that are used in molecular systematics.)

4. Reticulate evolution in linguistics

In this section we explain how we model borrowing between languages, using phylogenetic *networks*. When there is no contact between languages after they have diverged, the Stammbaum model can be used to describe the evolution of the languages. In this case, a rooted tree is used to model how languages evolve from a common ancestor. However, when there is contact between two languages after they have diverged, a different type of model is needed. One such approach, which is appropriate when an underlying tree (the so-called *genetic* tree) can still be reasonably defined, is a *network* model. In this case, additional edges, representing contact between language communities that co-exist in time and are geographically proximate, are added to the rooted tree. These additional edges indicate the flow of linguistic characters between two groups, and hence are bi-directional (since the transmission of characters can go both ways, in general).

The contact edges are actually pairs of directed edges, one directed edge for each of the two orientations. Every node in the network (other than the root) has a unique parent node, but may also *borrow* a state (i.e., receive a state, and replace its current state with the new state) from a neighbour, to whom it is connected by contact edges. We make two mild assumptions: first, that no state changes occur on a contact edge, and second, that no node in the tree has more than one contact edge incident with it, and so has at most one neighbour.

An important issue in modelling linguistic evolution using networks is whether we will allow a language to inherit a state for a character, as well as borrow a state for that character from one of its neighbours, without replacing its inherited state by the borrowed state. In this paper we only allow replacement rather than allowing the two states to co-exist; thus, we do not allow polymorphism (two or more states for a character in a language) to occur as a result of borrowing. Thus, we assume that the evolutionary process operates first genetically, so that each node receives its state from its genetic parent, but that state will be replaced if the node borrows a state from a neighbour.

This assumption allows us to assert that each character evolves down a tree contained within the network, since we can define the tree by picking, for each node of the network, the node from which it obtains its state (either by inheritance or by borrowing). Thus, every character has a treelike evolution, even if the tree on which it evolves is not the genetic tree. (If the character evolves without any borrowing, then its tree will be the genetic tree, and otherwise its tree will include one or more contact edges, and hence differ from the genetic tree.)

The parameters associated with the evolutionary process involved in borrowing determine the relative probabilities of each character to be borrowed, as well as the degree of contact of each borrowing edge. Thus, we will have the parameter κ_e, where e is a directed contact edge, indicating the probability of transmission via contact in that direction of the most easily transmitted character. We also have the parameter π_c for a character c, which determines its probability of being borrowed. Therefore, the probability of character c being transmitted on edge e is $\pi_c \times \kappa_e$. These parameters allow us to determine for a given character c, the probability of the character evolving down each of the trees contained within the network.

Since every character evolves down a tree contained within the network, and we have described the process by which the tree on which the character evolves is chosen, it suffices to describe the evolutionary process on trees. More generally, it is straightforward to combine any given model of treelike evolution with this reticulate evolution model, since each character evolves in a treelike fashion on some tree contained within the network (in statistical parlance, our model is a *mixture* of treelike models).

However, inferring an accurate reticulate evolutionary scenario presents several difficulties; only some simple situations can be readily handled (in particular, we can infer a network with one contact edge under this model, since such networks are defined by their collections of bipartitions, but this is not generally true). In order to extend the inference to be able to handle more borrowing, it may be necessary to identify which characters evolve on the same evolutionary tree within the network. When this can be done, then if there are enough characters to determine each of the trees contained within the network, the network itself may be identified (under some conditions) from its constituent trees. However, determining the network from its constituent trees is not a trivial matter, though some cases can be handled efficiently (see (Nakhleh, 2004)). On the other hand, the assumption that we can determine which characters evolve on the same tree is potentially unrealistic. Hence, inference under a model that allows both borrowing and homoplasy may be fairly challenging. Eliminating one of these two factors - borrowing or homoplasy - certainly makes inference much easier.

5. Comparison with models of molecular evolution

The model we present here is a fairly simple model which imposes one major assumption (the ability to detect homoplastic states) but is otherwise highly unconstrained. In particular, we allow characters to evolve without any common mechanism, assuming only that they evolve down the same tree. Under this model, we are able to show identifiability, linear time likelihood calculations, and most importantly, inference of the underlying (unrooted) evolutionary tree that is efficient with respect to data and with respect to running time.

By comparison, all models of molecular evolution that are in use make strong assumptions about the common mechanisms governing the different characters; without such strong assumptions, identifiability is lost, and it becomes impossible to reconstruct the true underlying tree from even infinite data.

Since some of the interest in phylogenetic reconstruction for historical linguistics has been for dating internal nodes, a few additional remarks are in order. Our discussion of what is achievable without imposing a common mechanism model shows that we can reconstruct the underlying unrooted tree, but not the parameters of the model. If dating of internal nodes is desired, much more information about the evolutionary processes is needed; in particular, the different characters in the dataset must be assumed to evolve under a common model with either the same quintet of probabilities for all edges, or quintets that are related under a rates-across-sites model, with a known (or estimable) distribution of rates. These are very strong requirements, and may not hold in practice. Making these assertions with any degree of confidence is probably beyond what can be done at this date; inferring dates without having a basis for making these assertions is therefore potentially quite problematic. Indeed, it may be best to avoid making such inferences until the validity of assertions along these lines can be evaluated.

6. Consequences for phylogenetic analysis in linguistics

Inference of evolutionary history under our model rests upon being able to identify homoplastic states, since this allows us to reconstruct the true tree (given enough data) without having to remove any characters from the dataset. In practice, several issues complicate this issue. The first is that even in the case where all the homoplastic states can be identified prior to the phylogenetic analysis, the presence of borrowing can make the inference of evolution difficult, except when the total amount of borrowing is quite low. Therefore, characters that are resistant to borrowing and for which homoplastic states are identifiable, will be much more useful in a phylogenetic analysis. This means that morphological characters are the most valuable, since they are most resistant to borrowing, have a very low incidence of homoplasy, and the homoplastic states (when they exist) are most easily identified, especially in archaic Indo-European languages (Meillet, 1925). The second issue is that the identification of homoplastic states requires a very thoroughly trained and knowledgeable historical linguist, and that even the most skilled linguist may not be able to accurately identify all the homoplastic states. Finally, insufficient data presents the problem of incomplete resolution within a phylogenetic analysis. Therefore, in practice, phylogenetic analyses within historical linguistics are likely to remain somewhat ambiguous, whether due to insufficient data or insufficient identification of homoplastic states.

Note that this discussion assumes that all characters are kept, and none are removed from the dataset. What about the traditional approach in which the data are screened, and all characters suspected of homoplasy are deleted before the phylogenetic analysis (Hoenigswald, 1960; Garde, 1961)? This approach is controversial in part because of its potential to be biased (i.e., it may be that the characters are not really homoplastic so much as inconsistent with a presumed phylogeny), but in any event once a dataset is modified through this process, the resultant *screened* data require somewhat different handling than our methods described in this paper. In particular, our approach for phonological data assumes that the characters evolve identically and independently. Once the character set is modified through this process, while the independence assumption will still hold, the identical distribution of character evolution will not necessarily still hold. Thus, the proposed technique for analyzing phonological characters cannot be applied in this case. Instead, since the characters will be presumed to now evolve without homoplasy (assuming all characters that are homoplastic have been successfully identified and deleted), the traditional approach of treating each binary phonological character as being homoplasy-free can be used. Thus, screening data for homoplasy, and deleting all such characters, can be applied successfully (although great care must be exercised not to delete characters that simply do not agree with one's assumptions), but phylogeny reconstruction on such modified datasets requires different techniques.

This commentary reflects the different issues involved in analyzing each type of data. Lexical characters, being the most easily borrowed, and having the most difficult-to-detect homoplastic states, are the most difficult to use (Porzig, 1954). Phonological characters, on the other hand, are somewhat more interesting to discuss – they are frequently homoplastic, but homoplasy in phonological characters is relatively easily identified; while traditional methods may eliminate all phonological characters suspected of homoplasy, this paper shows how to properly analyse the full set of phonological characters without deleting suspicious characters from the dataset. Morphological characters, being the most resistant to both borrowing and homoplasy, however, are likely to be the most valuable for phylogenetic analysis of languages.

In practice, we must also consider whether the model fits the data, and whether we will have enough data (meaning enough independent characters) in order to obtain a sufficiently accurate estimate of evolution. The assumptions of the model may, of course, not hold for the dataset in question – the most difficult aspect to ensure is that we can identify the homoplastic character states for every character. This requires a great deal of linguistic competence, and also a great deal of knowledge of the particular family, and even then may not be guaranteed to be correct. Hence, from a practical standpoint, this is still a problem area. Finally, if a tree does not fit the dataset, so that contact must be inferred, we can guarantee success but only in a somewhat limited way: if the evolutionary history does not include too much borrowing (specifically, too many contact edges), we should still be able to infer the evolutionary history. Quantifying the limits of how much borrowing can be allowed is part of our ongoing research.

7. Related work

The research most closely related to our work is (Mossel & Steel, 2004), which studies a no-common-mechanism model of evolution in which there is no homoplasy. This model is mathematically equivalent to the special case of ours that obtains when homoplasy is not allowed. The authors present a reconstruction method using quartets similar to the one we described, but improve upon it by using the observation that not all quartets are necessary to determine the tree. They give a precise quantification of when it is possible to reconstruct the tree with high probability very efficiently (in terms of the amount of data required). Their analysis can be carried over to our somewhat more general situation to give the following result.

Theorem 7.1. *Suppose that the tree T is binary and has n leaves (that is, there are n languages). Assume further that there are constants $0 < a < b < 1$ a such that $a \leq p_{e,c}(n, n') \leq b$ for all edges e and characters c. Then for any given $0 < \epsilon < 1$ there are constants γ and δ depending on a, b, ϵ such that if the number of characters is at least $\gamma + \delta \log n$, then the tree can be correctly reconstructed from the data with probability at least $1 - \epsilon$. Moreover, there is a polynomial-time (in n) algorithm for the reconstruction.*

8. Conclusions

In this paper we have presented a new model of linguistic character evolution that allows for homoplasy and borrowing between languages. We have shown that both morphological and lexical characters are sufficient to identify the true tree, even without a rates-across-sites assumption, provided that we can identify homoplastic states and the amount of borrowing is limited. We have also provided a new technique for analyzing phonological characters, which allows us to keep characters that evolve with homoplasy, and which will also identify the true tree provided that all phonological characters evolve identically and independently. Thus, our research extends the current models of linguistic character evolution and provides new techniques for analyzing linguistic data. We have also provided an initial attempt for the inference of reticulate (non-tree) evolution in historical linguistics.

It is worth noting that our research is an extension of linguistic methodology, rather than a radical departure; the techniques we propose are consistent with existing techniques, while allowing for better use of the available data. Furthermore, the research also provides a mathematical explanation for the belief within the historical linguistic research community that the choice of data is extremely important, and that morphological and phonological characters in general are better than lexical characters with respect to phylogeny reconstruction.

References

Bodlaender, H., Fellows, M., & Warnow, T. 1992. Two strikes against perfect phylogeny. *Pages 273–283 of: Proceedings of the 19th International Colloquium on Automata, Languages, and Programming*. Lecture Notes in Computer Science. Springer Verlag.

Brugmann, Karl. 1897. *Grundriss der vergleichenden Grammatik der indogermanischen Sprachen*. 2nd edn. Vol. 1. Strassburg: Trübner.

Brugmann, Karl. 1906. *Grundriss der vergleichenden Grammatik der indogermanischen Sprachen*. 2nd edn. Vol. 2. Strassburg: Trübner.

Buck, Carl D. 1955. *The Greek dialects*. Chicago: U. of Chicago Press.

Buneman, P. 1971. The recovery of trees from measures of dissimilarity. *Mathematics in the Archaelogical and Historical Sciences*, 387–395.

Campbell, Alistair. 1962. *Old English grammar*. Revised edn. Oxford: Clarendon Press.

Evans, Steven N., & Warnow, Tandy 2004. Unidentifiable divergence times in rates-across-sites models. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **1**, 387–395.

Garde, Paul. 1961. Réflexions sur les différences phonétiques entre les langues slaves. *Word* **17**, 34–62.

Gusfield, D. 1990. Efficient algorithms for inferring evolutionary trees. *Networks*, **21**, 19–28.

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

Kim, Junhyong, & Warnow, Tandy. 1999. *Tutorial on Phylogenetic Tree Estimation*. Presented at ISMB (Intelligent Systems for Molecular Biology) 1999, Heidelberg, Germany. Available electronically at `http://kim.bio.upenn.edu/~jkim/media/ISMBtutorial.pdf`.

Lühr, Rosemarie. 1984. Reste der athematischen Konjugation in den germanischen Sprachen. *In:* Untermann, Jürgen, & Brogyanyi, Béla (eds), *Das Germanische und die Rekonstruktion der indogermanischen Grundsprache*. Amsterdam: Benjamins.

McMahon, April, & McMahon, Robert. 2003. Finding families: quantitative methods in language classification. *Transactions of the Philological Society*, **101**, 7–55.

Meillet, Antoine. 1925. *La méthode comparative en linguistique historique*. Oslo: Aschehoug.

Melchert, H. Craig. 1987. PIE velars in Luvian. *In:* Watkins, Calvert (ed), *Studies in memory of Warren Cowgill*. Berlin: de Gruyter.

Mossel, Elchanan, & Steel, Michael. 2004. A phase transition for a random cluster model on phylogenetic trees. *Math. Biosci.*, **187**(2), 189–203.

Nakhleh, L. 2004. *Phylogenetic Networks in Biology and Historical Linguistics*. Ph.D. thesis, The University of Texas at Austin.

Nakhleh, Luay, Ringe, Don, & Warnow, Tandy. 2004. Perfect Phylogenetic Networks: A New Methodology for Reconstructing the Evolutionary History of Natural Languages. *Language*. To appear.

Pokorny, Julius. 1959. *Indogermanisches etymologisches Wörterbuch*. Bern: Francke.

Porzig, Walter. 1954. *Die Gliederung des indogermanischen Sprachgebiets*. Heidelberg: Winter.

Ringe, Don. 2005. *A linguistic history of English. Vol. 1. From Proto-Indo-European to Proto-Germanic*. Accepted for publication by Oxford U. Press.

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

Saitou, N., & Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.

Sihler, Andrew. 1995. *New comparative grammar of Greek and Latin*. Oxford: Oxford U. Press.

Steel, Michael. 1992. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification*, **9**, 91–116.

Streitberg, Wilhelm. 1896. *Urgermanische Grammatik*. Heidelberg: Winter.

Taylor, A., Warnow, T., & Ringe, D. 2000. Character-based reconstruction of a linguistic cladogram. *Pages 393–408 of:* Smith, J.C., & Bentley, D. (eds), *Historical Linguistics 1995, Volume I: General*

issues and non-Germanic languages. Amsterdam: Benjamins.
Tuffley, C., & Steel, M. 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of Mathematical Biology*, **59**(3), 581–607.
Warnow, T. 1997. Mathematical approaches to comparative linguistics. *Proc. Natl. Acad. Sci.*, **94**, 6585–6590.
Waterman, M.S., Smith, T.F., Singh, M., & Beyer, W.A. 1977. Additive evolutionary trees. *Journal of Theoretical Biology*, **64**, 199–213.

9. Appendix

9.1. **The model.** We begin with a summary of the model of evolution. The graphical component of the model is a rooted phylogenetic network N, which is a rooted tree T along with contact edges, where the presence of a contact edge represents the assumption that there is contact between two language communities at a given point in time. Thus, we always work with pairs of contact edges, one edge in each of the two orientations, between two nodes in T, and each of these individual directed edges $e = (v, w)$ is associated with a parameter κ_e which is the probability transmission of character states from v to w. Note that $\kappa_{(v,w)}$ may not be equal to $\kappa_{(w,v)}$. Note also that because contact edges can only take place between nodes that are able to co-exist, not all networks defined as unions of rooted trees with contact edges are feasible – certain additional constraints must also exist. Thus, while it is not necessary otherwise to incorporate time into the model, we may also require that the nodes of the tree be associated with a date, so that these dates decrease as you move from the root towards the leaves, and so that contact edges exist only between nodes that have the same date.

The model also is equipped with a set of characters, and a probability distribution on the set. Thus, each site evolves under a random process, selected at random (under this distribution) from the set. Each character in the set is equipped with a probability distribution for the state type (i.e., homoplastic or non-homoplastic) at the root of the tree, and a collection of of the transition probabilities (the $p_{e,c}(\cdot, \cdot)$ parameters) between the two types of states for each edge in the tree. Letting h^* denote the unique homoplastic state for a given character and n denote a non-homoplastic state, these transition probabilities are given by $p_{e,c}(n, n), p_{e,c}(n, h^*), p_{e,c}(n, n'), p_{e,c}(h^*, h^*)$ and $p_{e,c}(h^*, n)$. Each character c is also equipped with a relative probability π_c of being

borrowed, so that the probability that character c is transmitted across a borrowing edge e is $\pi_c \kappa_e$.
Notes: In our discussion earlier we discussed how the parameters of the different types of characters might be constrained by their type; thus, for example, since binary phonological characters have only two states – one of which is non-homoplastic and the other potentially homoplastic – we would have $p_{e,c}(n, n') = 0$ for any phonological character on any edge e. The description given here is just the more general one. Note also that this model, as described, is a linguistic equivalent of the *no-common-mechanism* model of Tuffley and Steel (Tuffley & Steel, 1997).

In order to fully specify this model, we need to delineate the actual collection of non-homoplastic states and then specify the precise mechanism for picking from the non-homoplastic states when a substitution is to result in such a state. The requirement that any change to a non-homoplastic state always results in a new state forces the set of such states to be uncountable. Since we are treating all non-homoplastic states as being on an equal footing, it suffices to label the set of non-homoplastic states by any nice uncountable set such as the unit interval and to pick a state according to the uniform probability distribution whenever a substitution is to result in a non-homoplastic state.

9.2. **A dynamic form of the model.** In other stochastic models for character evolution in both linguistic and biological phylogenetics, the substitution probabilities for an edge are derived from the net effect of a continuous dynamic substitution process occurring along the edge – typically a Markov chain. We can also introduce such a structure into our framework as follows.

Now each edge e for character c will have a *length* $t_{e,c}$ and we imagine a rate one Poisson process of substitution events running for "time" $t_{e,c}$, so that the expected number of substitution events on the edge is $t_{e,c}$.

What happens at each substitution event is determined by a quintuple of probabilities $q_{e,c}(h^*, h^*)$, $q_{e,c}(h^*, n)$, $q_{e,c}(n, h^*)$, $q_{e,c}(n, n)$, and $q_{e,c}(n, n')$, where $q_{e,c}(h^*, h^*) + q_{e,c}(h^*, n) = 1$ and $q_{e,c}(n, h^*) + q_{e,c}(n, n) + q_{e,c}(n, n') = 1$. The interpretation of, say, $q_{e,c}(n, h^*)$ is that it is the probability that a given substitution event will result in a change from a non-homplastic state n to the unique homoplastic state h^*. If $q_{e,c}(h^*, h^*) \neq 0$ or $q_{e,c}(n, n) \neq 0$, then we can have substitution events that are "spurious" in the sense that they don't actually result in any change in the state of the character.

In order to derive the corresponding quintuple $p_{e,c}(\cdot, \cdot)$, we first observe that if we combine all of the non-homoplastic states into a single state, then the "clumped" evolution is still Markovian and is just a two state Markov chain with rate matrix (that is, infinitesimal generator matrix)

$$\begin{pmatrix} -q_{e,c}(h^*,n) & q_{e,c}(h^*,n) \\ q_{e,c}(n,h^*) & -q_{e,c}(n,h^*) \end{pmatrix}$$

We can explicitly diagonalise this matrix and hence exponentiate it in closed form to get

$$\begin{aligned}
&p_{e,c}(h^*,h^*) \\
&\quad = \frac{q_{e,c}(n,h^*) + \exp(-t_{e,c}(q_{e,c}(h^*,n) + q_{e,c}(n,h^*)))q_{e,c}(h^*,n)}{q_{e,c}(h^*,n) + q_{e,c}(n,h^*)} \\
&p_{e,c}(h^*,n) \\
&\quad = \frac{q_{e,c}(h^*,n) - \exp(-t_{e,c}(q_{e,c}(h^*,n) + q_{e,c}(n,h^*)))q_{e,c}(h^*,n)}{q_{e,c}(h^*,n) + q_{e,c}(n,h^*)} \\
&p_{e,c}(n,h^*) \\
&\quad = \frac{q_{e,c}(n,h^*) - \exp(-t_{e,c}(q_{e,c}(h^*,n) + q_{e,c}(n,h^*)))q_{e,c}(n,h^*)}{q_{e,c}(h^*,n) + q_{e,c}(n,h^*)} \\
&p_{e,c}(n,n) + p_{e,c}(n,n') \\
&\quad = \frac{q_{e,c}(h^*,n) + \exp(-t_{e,c}(q_{e,c}(h^*,n) + q_{e,c}(n,h^*)))q_{e,c}(n,h^*)}{q_{e,c}(h^*,n) + q_{e,c}(n,h^*)}.
\end{aligned}$$

To complete the computation of the quintuple $p_{e,c}(\cdot,\cdot)$, we just need to find $p_{e,c}(n,n)$ and then get $p_{e,c}(n,n')$ by subtraction. Now $p_{e,c}(n,n)$ is the probability that any substitution on the edge is from n to itself. The per-substitution-event rate at which that chain exits the state n is $q_{e,c}(n,n') + q_{e,c}(n,h^*) = 1 - q_{e,c}(n,n)$, and thus

$$p_{e,c}(n,n) = \exp(-t_{e,c}(1 - q_{e,c}(n,n))).$$

This dynamic model could be further constrained to provide linguistic equivalents of standard molecular evolution models. For example, requiring that $t_{e,c}$ has the product form $\alpha_e \times \beta_c$ would result in the analogue of the rates-across-sites assumption.

9.3. **Identifiability.** The following two lemmas provide the fundamental techniques that we will use in proving identifiability of the underlying unrooted tree.

Lemma 9.1. *Let T be an unrooted binary tree, leaf-labelled by the set $\mathcal{L}$ of languages. Let $Q(T)$ denote the set of all the subtrees of T induced by quartets of languages drawn from $\mathcal{L}$. Then T is defined by the set $Q(T)$. That is, if T' is an unrooted binary tree with $Q(T') = Q(T)$, then $T = T'$. Let $C(T)$ denote the set of bipartitions defined by the edges of the tree T. Then if T is an unrooted binary tree with $C(T') = C(T)$,* *then* $T = T'$. *Furthermore, given* $C(T)$ *or* $Q(T)$, *it is possible to recover* T *in polynomial time.*

The proof is well known in the computational biology literature (see, for example, (Kim & Warnow, 1999)), and is omitted.

Lemma 9.2. *Let* T *be an unrooted binary tree with positive edge weights* $w : E \to \mathbb{R}_+$, *where* E *is the set of edges of* T. *Let* $[D_{ij}]$ *be a matrix defined by* $D_{i,j} = \sum_{e \in P_{i,j}} w(e)$, *where* i *and* j *are leaves of* T *and* $P_{i,j}$ *is the collection of edges on the path connecting* i *and* j. *Then* T, *and the associated edge weights, are uniquely determined by* $[D_{i,j}]$, *and can be constructed from the matrix in polynomial time.*

The first part (uniqueness of the tree and edge weights) of this theorem is given in (Buneman, 1971), and the polynomial time algorithm for obtaining T and w is given in (Waterman *et al.* , 1977).

9.3.1. *Morphological and lexical characters.* Recall that in our model we make the assumptions that for morphological and lexical characters we can identify all homoplastic states, and that $0 < p_{e,c}(n, n') < 1$ for all edges e in the tree. Suppose that the probability of a non-homoplastic state at the root is non-zero. Now let c be a morphological or lexical character, and let e be an arbitrary edge in T. Consider the probability that the state of c at the root is non-homoplastic, and that c changes exactly once - on the edge e - to a new non-homoplastic state. By our assumptions the probability of this is strictly positive. Furthermore, given these events, the character c defines a bipartition on the leaf set into two sets, defined by the edge e. Thus, given the states of the leaf set defined by the character c, we can infer the edge e. It is also easy to see that any bipartition of the leaf set defined by two non-homoplastic states has zero probability if that bipartition does not correspond to an edge in the tree. Therefore, given the probability distribution on bipartitions of the leaf-set defined by two non-homoplastic states, we can infer the tree T (but not the root location). Therefore, the underlying unrooted tree T is identifiable under this model. (A similar argument can be used to prove identifiability from the quartet trees defined on the basis of pairs of non-homoplastic states.)

9.3.2. *Binary phonological characters.* We consider binary phonological characters. We will prove identifiability of the full model (the underlying tree and the associated $p_{e,c}(\cdot, \cdot)$ parameters, but not the location of the root). Our proof relies only upon the assumption that the root state is known and is not homoplastic.

In binary phonological characters, the root state is 0 and indicates absence of the sound change, and all transitions are from absence to presence (indicated by 1). Thus, we allow $0 \to 1$ substitutions, but no $1 \to 0$ substitutions.

Let $\rho_{e,c}$ denote the probability of a $0 \to 1$ substitution on the edge e (i.e., $\rho_{e,c} = p_{e,c}(n, h^*)$). Recall that we assume that for all edges e we have $0 < \rho_{e,c} < 1$. Set $\ell_{e,c} = -\log(1 - \rho_{e,c})$ and call $\ell_{e,c}$ the length of the edge e. We define $\ell_c(P_{i,j})$ the length of a path $P_{i,j}$ between leaves i and j to be the sum of the lengths of the edges in the path. To ensure identifiability, we need only to show that we can compute $\ell(P_{i,j})$ for all pairs of leaves i and j in the tree, given the probabilities of the patterns at the leaves.

We know the probability that both i and j are in state 0, and we also know the probabilities that each is individually in state 0. The probability that i is in state 0 is just $\prod_{e \in P_{r,i}}(1 - \rho_{e,c})$, where r is the root, and we can compute the probability that j is in state 0 similarly. The probability that both i and j are in state 0 is $\prod_{e \in P_{r,v}}(1 - \rho_{e,c}) \times \prod_{e \in P_{v,i}}(1 - \rho_{e,c}) \times \prod_{e \in P_{v,j}}(1 - \rho_{e,c})$, where v is the most recent common ancestor of i and j. Therefore,

$$\prod_{e \in P_{i,j}} (1 - \rho_{e,c}) = \frac{\Pr[i = 0 \ \& \ j = 0]^2}{\Pr[i = 0] \ \Pr[j = 0]}.$$

Equivalently, the length of the path $P_{i,j}$ is set by

$$\ell(P_{i,j}) = -2 \log(\Pr[i = 0 \ \& \ j = 0]) + \log(\Pr[i = 0]) + \log(\Pr[j = 0]).$$

Since lengths of paths are identifiable, so (by Lemma 9.2) is the tree. Hence model trees are identifiable from binary phonological characters if we require $0 < \rho_{e,c} < 1$.

9.3.3. *Morphological and lexical characters revisited.* Although identifiability of the tree T was established above for morphological and lexical characters, we would like to point out that a distance argument analogous to that used for binary phonological characters can also be used in this setting if the model doesn't allow homoplasy. This observation has the practical consequence that it enables distance-based reconstruction methods to be used for morphological and lexical characters when there is no homoplasy permitted.

We therefore assume for some character c that the root state is non-homoplastic and that $p_{e,c}(n, h^*) = 0$ for all edges e. Then for two leaves

i and j we have, in the notation above, that

$$-\log \Pr[i \text{ and } j \text{ in the same state}] = \sum_{e \in P_{i,j}} \{-\log(p_{e,c}(n,n))\}$$

and we can apply Lemma 9.2 to establish identifiability of the tree provided $0 < p_{e,c}(n,n) < 1$ for all e and c (which is equivalent to $0 < p_{e,c}(n,n') < 1$ for all e and c because of our assumption that $p_{e,c}(n,h^*) = 0$).

If we are in the setting of Section 9.2, then the edge weight $-\log(p_{e,c}(n,n))$ is $t_{e,c}(1 - q_{e,c}(n,n))$. In particular, if we impose the rates-across-sites assumption that $t_{e,c} = \alpha_e \times \beta_c$, then the vectors of edge weights for different characters are just scalar multiples of each other.

9.4. **Likelihood calculations.** Because of the independence of the characters, it suffices to show how to compute likelihoods for single characters. Hence, let c be a single character, and (as usual) let $c(x)$ denote the state of x under the character c.

We begin by preprocessing the tree in order to assign states (when possible) to internal nodes. This is possible for every internal node that is on a path between two leaves with the same non-homoplastic state. The result of this labelling yields rooted subtrees, denoted by t, so that within each subtree t every two leaves that have the same state have state h^*. The question is how to calculate the likelihood for this special case.

9.4.1. *Notation.* We begin with some notation. We distinguish between the case where a node is labelled with h^*, and where a node has a non-homoplastic state, denoted by n.

A *marking* of a rooted subtree t is a set of edges of t along with the kind of event (mutation to h^*, or homoplasy-free mutation) that occurs on each edge. A marking allows us to determine the equivalence relation on the nodes of the tree defined by the character. Thus, some such markings will have probability 0 since they will be incompatible with the pattern at the leaves, and others will have non-zero probability. Given a marking of a subtree t, the probability of the data given the marking will be either 0 (if the marking is incompatible) or 1 (if the marking is compatible). Hence we only need to compute the probabilities of the markings which are compatible with the pattern at the leaves.

We let the set of all markings of edges of the subtree t be denoted by $M(t)$.

We let $M^{h^*}(t)$ denote all markings of the subtree t which have the root labelled by h^*, and $M^n(t)$ denote all markings of the subtree in which the root is labelled by a non-homoplastic state (i.e., something other than h^*).

We let $DM^n(t)$ denote the markings of t which have the root labelled by a non-homoplastic state, but labelled distinctly from all leaves below the root. We let $SM^n(t)$ denote the markings of t which have the root labelled by a non-homoplastic state, and identically labelled as some leaf below the root.

When we write $\Pr[M^n(t)], \Pr[SM^n(t)], \Pr[DM^n(t)]$, or $\Pr[M^{h^*}(t)]$, we mean the probability of the character states at the leaves of the tree t, over all markings of the tree t with the properties defined by the referenced set.

Before we show how we compute the various probabilities, we need to define two more quantities. The subtrees we will work with are always rooted subtrees, but have two different forms. Let v be a node in the tree T. We denote the subtree of T rooted at v by T_v. However, if v is not a leaf and if a is one of $v's$ children, then we denote $T(v,a)$ the tree rooted at v with one child a, along with T_a. Thus, there will be two types of subtrees t – those whose root has one child, and those whose root has two children.

We are now ready to show how we compute all the probabilities we need.

9.4.2. *The base case: t is a leaf.* If t is a leaf then $c(t)$ is already defined, and in this case we can compute the various probabilities we need to compute. Thus, $\Pr[M^{h^*}(t)] = 1$ if $c(t) = h^*$, and $\Pr[M^{h^*}(t)] = 0$ if $c(t) \neq h^*$. (We set $\Pr[M^n(t)]$ in the opposite way.) Similarly, $\Pr[SM^n(t)] = 1$ if $c(t) \neq h^*$, and $\Pr[SM^n(t)] = 0$ if $c(t) = h^*$. Finally, $\Pr[DM^n(t)] = 0$ for all leaves t.

9.4.3. *The inductive case: t is not a leaf.* We can then establish the following identities.

1. $\Pr[M^n(t)] = \Pr[SM^n(t)] + \Pr[DM^n(t)]$ (definition)
2. $\Pr[M(t)] = \Pr[M^{h^*}(t)] + \Pr[M^n(t)]$ (definition)
3. $\Pr[M^{h^*}(T(v,a))] = p_{e,c}(h^*,h^*)\Pr[M^{h^*}(T_a)] + p_{e,c}(h^*,n)\Pr[M^n(T_a)]$
4. $\Pr[M^{h^*}(T_v)] = \Pr[M^{h^*}(T(v,a))]\Pr[M^{h^*}(T(v,a'))]$, where a and a' are the two children of v
5. $\Pr[SM^n(T(v,a))] = p_{e,c}(n,n)\Pr[SM^n(T_a)]$
6. $\Pr[SM^n(T_v)] = \Pr[SM^n(T(v,a))]\Pr[DM^n(T(v,a'))] + \Pr[DM^n(T(v,a))]\Pr[SM^n(T(v,a'))]$, where a and a' are the two

children of v. Note that this suffices because of our preprocessing step – which results in the case that in every subtree, two leaves which share the same state of a given character must both have the homoplastic state h^*.

(7) $\Pr[DM^n(T(v,a))] = p(n,h^*)\Pr[M^{h^*}(T_a)] + p_{e,c}(n,n')\Pr[M^n(T_a)] + p_{e,c}(n,n)\Pr[DM^n(T_a)]$

(8) $\Pr[DM^n(T_v)] = \Pr[DM^n(T(v,a))]\Pr[DM^n(T(v,a'))]$, where a and a' are the two children of v

Hence, the probability of the states at the leaves can be computed in linear time, given the probabilities of substitution on the edges, for all characters.

10. Acknowledgments

The authors would like to thank the McDonald Institute for inviting them to submit this paper.

Tandy Warnow, Department of Computer Sciences, University of Texas at Austin, Austin, TX 78712, U.S.A.
E-mail address: tandy@cs.utexas.edu

Steven N. Evans, Department of Statistics #3860, University of California at Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, U.S.A
E-mail address: evans@stat.Berkeley.EDU

Don Ringe, Department of Linguistics, 619 Williams Hall, University of Pennsylvania, Philadelphia, PA 19104-6305, U.S.A.
E-mail address: dringe@unagi.cis.upenn.edu

Luay Nakhleh, Department of Computer Science, Rice University, 6100 Main St., MS 132, Houston TX, 77005-1892, U.S.A.
E-mail address: nakhleh@cs.rice.edu