STOCHASTIC MODELS OF LANGUAGE EVOLUTION AND AN APPLICATION TO THE INDO-EUROPEAN FAMILY OF LANGUAGES

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

ABSTRACT. We propose several models of how languages evolve, and discuss statistical estimation of evolution under these models. We also discuss issues of identifiability and statistical consistency under these models.

1. INTRODUCTION

In recent months several methods for estimating evolutionary histories of languages have been described and used on Indo-European (IE) datasets in order to estimate dates at which languages diversified. Implicit in these methods are stochastic models of how languages evolve (Forster & Toth, 2003; Gray & Atkinson, 2003). We agree that a carefully considered stochastic model can be of tremendous use to historical linguistics: if sufficiently realistic, inference under the model can reveal much about the history of the language family, and examinations of how reconstruction methods perform under these models (via simulation, in particular) can help us quantify the reliability of a reconstruction method. Since our own interest in this is primarily motivated by the IE family, we will formulate this model so as to reflect what we believe is likely to be true about IE's evolution. Much, however, should be appropriate for other families, and we will discuss extensions to other families at the end of the paper.

2. MODELS

In this section we explain what is meant by a stochastic model of language evolution, and we present some specific models that are worth examining in the context of IE evolution.

We begin by explaining what linguistic "characters" are, since the evolutionary model describes how each character evolves.

Date: April 16, 2004.

TW supported by NSF grant BCS-0312830.

SNE supported in part by NSF grant DMS-0071468.

DR supported in part by NSF grant BCS-0312911.

2.1. **Linguistic 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. Thus, our characters include lexical characters, where the different states are the 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. Other characters include phonological characters (the appearance of a sound change within the language or its ancestry) and morphological characters (e.g., inflectional markers). Thus, a character defines an equivalence relation on the language family, where two languages are equivalent if they exhibit the same state for the character. Given a partition of a set into disjoint subsets, we can define an equivalence relation by making two languages equivalent if and only if they are in the same subset; thus, a partition of a set into disjoint subsets defines an equivalence relation (and the converse holds as well).

Our first simplifying assumption is that all the characters are "monomorphic", which means that every language exhibits only one state of each character. The contrasting phenomenon is a character which has two or more states for some languages; examples of such characters include the semantic slot "rock" for which English contains at least two equivalents: "rock" and "stone". Because we do not understand in enough detail how polymorphism arises, we will exclude polymorphic characters from our model.

- *Simplifying assumption #1*: there is no polymorphism (i.e, the appearance of two or more states for a given character in a given language).

For each character, we can assign numbers to the states of the character so that the character is defined to be a function that assigns every language in a set $\mathcal{L}$ of languages a real number; the number assigned to the language is called the "state" of the character for that language. Thus, the states of all our characters are real numbers, and when we write $c(L)$ for a language L and a character c, we mean the state of the character c exhibited by the language L. However, the particular real number used to label a state is irrelevant, and all that matters is whether two states are equal or different.

2.2. **Tree models.** Languages can evolve in a purely treelike fashion (the *Stammbaum model*), or with enough contact between languages that undetected (or undetectable) borrowing occurs between lineages, so that it becomes difficult (or inappropriate) to define a "genetic tree" for the family. Many conditions can make evolution non-treelike; creoles (hybrid languages) are one, dialect continua are another, but more generally contact itself between divergent lineages can also lead to trees being inappropriate (or just difficult to infer). All of these conditions can be loosely grouped under the category of "reticulate evolution".

We will initially describe the model for the case where there is no reticulate evolution, since most of the concepts are more familiar in that context; later we will show how the model extends to the case where we permit reticulate evolution.

In the case where there is no reticulate evolution, the evolutionary history of the languages is described by a rooted tree T, in which the leaves represent the languages in the family, and the internal nodes represent ancestral languages at particular points in time; this is the "genetic tree" for the family. Every node v in T has a time $t(v)$ associated to it, with times at nodes increasing as one moves away from the root of the tree. All of the internal nodes in the tree will have at least two edges issuing from them (that is, they will have *out-degree* at least two) so that nodes can also be thought of as representing diversification events. Therefore, an edge within the tree represents the development of the language over a period of time between diversification events.

2.3. **The evolution of characters down trees.** Now we will look at how the characteristics of languages evolve down the genetic tree for the language family. These characteristics, as described by the character states at each node in the tree, evolve down the tree, changing state on the edges of the tree as they evolve. Thus, we will refer to the character state assigned to each node, using the notation $c(v)$ for character c and node v. We can model the evolution of each character probabilistically, by assigning for each character and each edge a probability of the character changing its state on that edge. If we assume that the characters evolve independently of each other, then we can define the joint stochastic model for the characters by simply specifying the model for each character. In linguistics this seems to be a reasonable assumption, provided that the characters analyzed are chosen with care; we will let this be our second simplifying assumption:

- *Simplifying assumption #2*: The characters evolve independently.

Homoplasy. A substitution of character c on edge $e = (u, v)$ starting at u and ending at v is one in which $c(u) \neq c(v)$ (see above for the meaning of $c(v)$). When a substitution occurs on an edge, it can result in a new state (one that does not appear in the tree yet), or one that has already occurred. The first type of substitution is said to be *non-homoplaseous*, and the second type is said to be *homoplaseous*. The difference is easy to explain. Recall first that every node v in T has a time $t(v)$ associated to it. A homoplaseous substitution of character c on edge $e = (u, v)$ starting at u and ending at v is one where $c(v) = c(w)$ for some node w for which $t(w) \leq t(v)$. In other words, a homoplaseous substitution results in the reappearance of a character state in the tree – so that the substitution produces a state that is

either present currently on the tree, or occurred at an earlier time (although possibly on another lineage). By contrast, a non-homoplaseous substitution always results in a new state.

Most of the homoplasy that occurs in linguistic evolution is relatively obvious and can be detected using traditional methods; however, we do not know how much of the "homoplasy" is actually due to undetected borrowing between languages, or potentially even polymorphism, rather than true homoplasy (i.e., backmutation or parallel evolution). One way of handling characters that exhibit homoplasy is to sequester them, or process them (for example, by coding clearly borrowed lexemes as unique states). However, we can never be certain that all homoplasy has been discovered, and in any case in may be advisable for the purposes of statistical inference not to remove homoplaseous characters from the dataset. Therefore stochastic models should allow for characters to evolve with homoplasy. However, modelling homoplasy is somewhat tricky, as it involves factors that we do not understand about the structure of, for example, lexical space that affect the probability of the backmutation or parallel evolution inherent in homoplaseous substitutions. Just as we excluded polymorphism for the moment because we do not understand it sufficiently to model it, we will exclude homoplasy from the models we posit (though we will include a section at the end discussing the issues involved in modelling homoplasy).

Thus our third simplifying assumption is:

- *Simplifying assumption #3*: All characters evolve without homoplasy; thus, all substitutions of states result in new states.

Under this assumption, it becomes possible to infer the evolutionary history of a set of languages with some level of accuracy. Consider, for example, the case of inferring the evolutionary history of four languages L_1, L_2, L_3, and L_4, under the assumption that each character evolves without homoplasy. In this case, if there is any character that produces a two-two split (i.e., a character that has two states on these languages, and groups the four languages into two sets with two languages each), then the true tree (but not the location of the root) is immediately known! For example, if a character c has $c(L_1) = c(L_2) = 1$ and $c(L_3) = c(L_4) = 2$, then the tree *must* contain an edge separating L_1 and L_2 from L_3 and L_4. Since a tree on four leaves only contains one internal edge, the tree is uniquely determined by that single character. Similarly, if the language family has more than four languages, then as soon as there are enough two-state characters to define each of the edges, the tree is uniquely determined. This observation is well understood in historical linguistics, and is based upon the assumption of homoplasy-free evolution.

2.4. **Stochastic processes operating on characters.** Now that we have described the kinds of events that happen to characters as they evolve down the tree, we will explicitly model this as a stochastic process. We will assume that for every edge e and character c, the character will change its state on the edge with a probability that will depend on both the character and the edge. We will denote that probability by $p_{e,c}$ to reflect the dependency on the edge/character pair. (That this probability depends upon the character should be obvious; for example, some lexical characters change more easily than others. The dependency on the edge is also natural, for example since edges can have different time durations.) Our fourth simplifying assumption (coupled with our second one) implies says that the probabilities $p_{e,c}$ suffice to specify the complete stochastic structure of the model:

- *Simplifying assumption #4*: Whether or not a substitution occurs on an edge for a given character is statistically independent of what happens on other edges for that character.

Our first simplifying assumption (that there is no polymorphism) means that changes of state always replace the previous state by a new state, rather than adding a new state to the current set of states. Our second and fourth simplifying assumptions mean that whether or not a given character changes on an edge is independent of the changes for other character/edge pairs. The third simplifying assumption means that every change of state results in a new state – i.e., there is no substitution of states that results in a previous state (one that is already in the tree) appearing.

Given a rooted tree T down which k characters evolve (under our simplifying assumptions), and a specific state defined at the root of the tree T for each of the k characters, we can think of these substitution probabilities $p_{e,c}$ (one for each edge/character combination) as defining a random process that generates random data at the leaves of the tree.

As a very simple example, suppose we have only one character, and the model tree is as given in Figure 1 below. Suppose the state at the root is 1, and we use the model parameters to generate states at each node of the tree, and thus also at the leaves. The tree has only four leaves, and so it is not too hard to calculate the probability of each of the possible "patterns" we can observe at the leaves. The easiest pattern to analyze is where all the leaves have the same state as the root, i.e., all leaves have state 1. In this case, the *only* way this pattern can be obtained is if there is absolutely no change on any edge (since all changes result in new states, and there is no backmutation). Hence, the probability of this particular pattern – all leaves having the same state – is just the product of the probabilities of no change on each edge, or $0.5 \times 0.8 \times 0.8 \times 0.8 \times 0.9 \times 0.7 = 0.16128$.

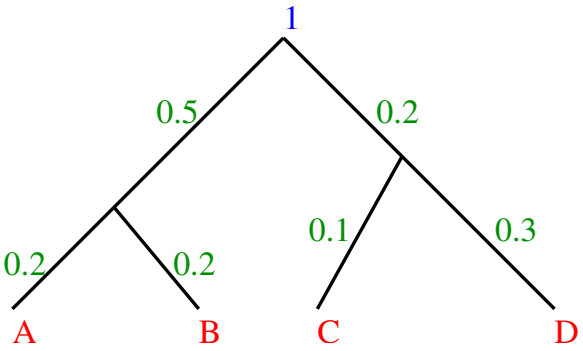

FIGURE 1. A tree with probabilities of change on the edges.

Extending this calculation to other patterns involves more complicated arithmetic, as we have to calculate the sum of the probabilities of each sequence of events (change or no change on each edge for each character) that would produce that pattern. However, in essence, the idea is straightforward (we discuss the details in Appendix 8). As a consequence, the probabilistic model defines the probability of every given dataset. More interesting perhaps is that it also suggests how reconstruction methods might try to use the properties of the model and the observed character states for each language, in order to figure out the evolutionary history of the dataset.

For example, while the location of the root of the evolutionary tree may be hard or impossible to infer, the unrooted version of the evolutionary tree can be reconstructed quite accurately, under some circumstances. Referring again to the same tree in Figure 1, note that the pattern that groups leaves A, C together with the same state, and groups B, D together (but with a state different from the state assigned to A and C) has zero probability, because of the assumption that there is no homoplasy. Similarly, the pattern grouping A, D together and B, C together also has zero probability. On the other hand the pattern that groups A, B together and C, D together has non-zero probability. This observation is true for *any* model tree that contains an edge separating the leaves A and B from leaves C and D. Thus, given a dataset of four languages described by characters, if there is any character in which A and B have the same state, and C and D have the same state - but one that differs from the state shared by A and B, then the *only* possible candidate phylogenies will contain the split $\{A, B\}|\{C, D\}$. That means that the *underlying unrooted tree* can be identified from one such character – it is a simple, but somewhat non-obvious, fact that the topology of an unrooted tree can be reconstructed from a knowledge of the topologies of the subtrees spanned by each set of four leaves. The consequence in terms of phylogeny reconstruction is significant: even when the characters evolve with different

substitution probabilities, the underlying unrooted tree can be reconstructed if there are enough random data. Less obvious, but also significant, is the following: if every character has the same evolutionary process (that is, $p_{e,c}$ does not depend on c, so we can denote the common value by p_e), then as the number of characters evolving down this tree increases, not only will the underlying unrooted tree be reconstructed correctly, but we can even estimate the shared values p_e of substitution parameters on the edges as well, with arbitrarily high accuracy. In general, however, the location of the root will still not be identifiable under this model.

The point of this discussion is to show that under certain assumptions about the stochastic process operating on the tree, much about the underlying model tree can be estimated quite accurately. The point of this paper is to examine the conditions under which these assumptions allow for estimating the different aspects of the model tree – its topology, the substitution probabilities, the times on the internal nodes, etc.

What we have just described is an evolutionary model in which for every edge e and every character c we have a substitution probability $p_{e,c}$, which denotes the probability of the character changing its state to a new state on the edge. If we make no further assumptions (relating, for example, the probabilities of substitution so that different characters evolve in similar ways), then this is the **no common mechanism model**. Since there are $2n - 2$ edges in a rooted binary tree with n leaves, each *model tree* in the no common mechanism model can be specified by at most $(2n - 2)k$ parameters, where there are k different characters and n is the number of leaves, since we assume that all interior nodes have out-degree at least two.

On the other hand, we may wish to constrain the model further by assuming that all the characters evolve under exactly the same process. In this case, the probability of a substitution for a character on an edge depends only on the edge and not on the character. Thus, $p_{e,c} = p_{e,c'}$ for all pairs of characters c, c', and so it makes sense to define p_e to be the probability that any given character changes its state on the edge e. This model is the **simplest homoplasy-free model**. The number of parameters in this model is thus just the the number of edges in the tree (i.e., at most $2n - 2$).

It should be clear from this discussion that under neither of these models is it possible to infer anything about times at nodes, because there is no linkage between the probability of a substitution occurring on an edge and the amount of time that has elapsed on that edge. Thus, we now turn to describing more elaborate models that make such linkage possible by first incorporating a description of the evolutionary dynamics and then relating the probability of observing a net change between the endpoints to this evolutionary process.

We begin with a discussion of the properties of the number of times a character changes on an edge, since these properties will determine in great part the extent to which we can estimate dates at internal nodes. The first assumption is that the number of times each character changes on an edge is Poisson distributed. This assumption defines the probability of each number of substitutions in a particular way that seems reasonably appropriate for evolutionary events such as these, as well as for mutational events in molecular systematics, where it is standard. In the next section we provide the mathematics for Poisson random variables.

Poisson random variables. The Poisson assumption can be heuristically described as follows. Fix a character c and an edge $e = (u, v)$. It is reasonable to posit that in a tiny time sub-interval $[t, t + \Delta t]$ with $t(u) \leq t \leq t + \Delta t \leq t(v)$ the character c has probability approximately $1 - q_{e,c}(t)\Delta t$ of not changing in the sub-interval for some positive number $q_{e,c}(t)$, probability approximately $q_{e,c}(t)\Delta t$ of undergoing a single change of state in the sub-interval, and probability of making two or more changes in the sub-interval that is negligible when compared to Δt. It is also reasonable to suppose that the random number of changes in any such subinterval is independent of the ensemble of changes that occur outside the subinterval. Under these conditions, it is well-known that the number of changes for character c in the interval $[t(u), t(v)]$ is Poisson distributed with parameter

$$\lambda_{e,c} := \int_{t(u)}^{t(v)} q_{e,c}(t)\, dt.$$

That is, if we denote the number of changes by $X_{e,c}$, then

$$\Pr\{X_{e,c} = k\} = e^{-\lambda_{e,c}} \frac{\lambda_{e,c}^k}{k!}, \quad k = 0, 1, 2, \ldots$$

The quantity $q_{e,c}(t)$ is the *instantaneous rate of evolution* for character c on edge e at time t, and the parameter $\lambda_{e,c}$ is equal to $\mathrm{E}[X_{e,c}]$, the expected number of changes on edge e for character c.

Using these results, it is possible to show that $p_{e,c} = 1 - e^{-\lambda_{e,c}}$; this relates the two types of edge parameters (substitution probabilities, and edge lengths).

Relating lengths of edges to elapsed time. We now discuss how to set up a linkage between the edge lengths $\lambda_{e,c}$ and the time duration of edges, and hence the time depth of interior nodes. For the edge $e = (u, v)$, write $T(e) := t(v) - t(u)$ for its time duration and set

$$r_{e,c} := \lambda_{e,c}/T(e) = \frac{1}{t(v) - t(u)} \int_{t(u)}^{t(v)} q_{e,c}(t)\, dt.$$

Thus $r_{e,c}$ is the *average rate of evolution* for character c on edge e. Equivalently, it follows that $\lambda_{e,c} = r_{e,c}T(e)$. If we have information regarding the ways rates can vary across characters and edges, we may be able to estimate the $\lambda_{e,c}$ from observed character data; however, direct estimations of the times at nodes requires being able to factor the edge length $\lambda_{e,c}$ as the product of $T(e)$ and the average rate of evolution of the character c on the edge e. In other words, there is a *degeneracy* issue here: we could obtain the same probability distribution for the data if we pick any constant $a > 0$ and set $T'(e) = aT(e)$ and $r'_{e,c} = r_{e,c}/a$, since then $\lambda_{e,c} = r_{e,c}T(e) = r'_{e,c}T'(e)$. That is, we can't tell which factorization is right.

Relative versus absolute times. A question that is of interest to many researchers is whether *absolute times* at nodes can be estimated with acceptable accuracy. By "absolute times" we mean the $t(v)$ values assigned to each node v in the model tree; in a phylogenetic analysis, these values are estimates of the true historical dates of diversification events for the language family. It is clearly impossible to do this without explicit information about rates of evolution (which are controversial) and/or at least one "calibration point". There are thus significant challenges to estimating absolute times at internal nodes.

There are alternatives to absolute dates that may be feasible. Lacking a calibration point, we may be able to estimate what are called *relative times* at nodes. By this we mean dates that are correct, but only up to a constant multiple; in other words, all estimated dates are off by a factor of d, but the precise value of d is not known. It is easy to see that with relative times at nodes and one calibration point, we can obtain absolute times at nodes.

Finally, we may be interested in just the *order* in which the different diversifications happened – so that we can determine for each pair of nodes in the tree, which one occurred earlier. Relative times at nodes suffice to provide this ordering of internal nodes.

When the lexical clock holds, and we know how rates can vary across characters, then we can establish relative times, and hence also absolute times if we have also a calibration point. However, in the most general setting, none of these can be estimated reliably. If we have a lexical clock but no knowledge about how rates vary across characters, we can obtain an ordering on internal nodes, but not relative times.

The impossibility of inferring even relative dates is most obvious under the *no common mechanism* assumption: in essence we only have a single observation for each edge/character pair, and we cannot expect to have any statistical power to estimate the associated parameters. The more constraints we imply the more feasible it is that we can estimate relative dates

with some degree of statistical power. On the other hand, the assumption of these constraints may make the model unrealistic.

Therefore, it makes sense to examine those models that incorporate some constraints on the variation among these $r_{e,c}$ parameters, so as to understand the conditions under which we can estimate either relative or absolute times at internal nodes. We begin with a discussion of the lexical clock assumption.

Lexical clock assumption. The lexical clock hypothesis is comparable to the molecular clock in biology, and has the same abstract statement: for all characters c and edges e, $\lambda_{e,c}$ is proportional to $T(e)$, with a constant of proportionality that is the same for all edges. Using our notation, this is equivalent to the assertion that $r_{e,c} = r_{e',c}$ for all pairs of edges e, e' and all characters c, so that it makes sense to define the *rate r_c of the character c.* Note that the lexical clock hypothesis does not imply that all characters have the same rate for all edges, only that the rate is constant for each character.

The lexical clock hypothesis is sufficient to establish the rooted tree, using simple lexicostatistical techniques (as long as we are given enough data). Furthermore, the same mathematical argument that establishes the correctness of the tree reconstruction also allows us to establish the ordering on the dates at internal nodes. However, it is not possible to estimate the relative dates at nodes, unless we have additional information about how rates vary. Thus, even with the lexical clock assumption we have only limited capability. Furthermore, the lexical clock is likely to be violated by real linguistic evolution.

We direct the interested reader to (Evans *et al.* , 2004) for more about dating on internal nodes, and (Bergsland & Vogt, 1962) for a discussion of the lexical clock.

2.5. **Four stochastic models of treelike evolution for languages.** Earlier we described two basic models: the *simplest homoplasy-free model* and the *no common mechanism model*. We can describe a model tree in each of these basic models as a rooted tree T, times $t(v)$ for every internal node v (and hence also the duration of edge e, given by $T(e)$), and rates of change $r_{e,c}$ for every character c and edge e; these parameters then allow us also to define $\lambda_{e,c}$, the expected number of changes of a character c on edge e, by setting $\lambda_{e,c} = r_{e,c}T(e)$. The difference between the two models is that in the no common mechanism model the $r_{e,c}$ (and hence $\lambda_{e,c}$) can be arbitrary, but under the simplest model we require that $r_{e,c} = r_{e,c'}$ (so that also $\lambda_{e,c} = \lambda_{e,c'}$) for all pairs of characters c, c' (that is, all characters evolve under the same underlying process). With the additional consideration of

whether a lexical clock is also presumed, we obtain four basic models as follows:

- **The simplest model without a lexical clock.** Thus, $\lambda_{e,c} = \lambda_{e,c'}$ for all edges, but no additional constraints are implied. This model can be described by specifying λ_e for every edge since the value does not depend upon the character, and so requires as many parameters as there are edges in the tree, or $O(n)$ parameters.
- **The simplest model with a lexical clock.** Thus, $r_{e,c} = r_{e',c'}$ for all pairs of edges e, e' and all pairs of characters c, c'. This model is the most constrained of the four models. Under this model we can define the rate of evolution r, since it does not depend upon the edge or character. This model can be defined therefore by the times on the internal nodes and the rate of evolution, for a total of $O(n)$ parameters. Although it requires only one additional parameter than the previous model, because of the lexical clock we should be able to estimate relative times under this model.
- **The no common mechanism model without a lexical clock.** In this model we have no constraints at all on the edge parameters $\lambda_{e,c}$. We need $O(nk)$ parameters to define this model.
- **The no common mechanism model with a lexical clock.** In this case we will presume that for every character c there is a rate r_c so that $\lambda_{e,c} = r_c T(e)$. This model requires $O(n + k)$ parameters: the times at every node and the rates for each character. It differs from the simplest model with a lexical clock by not assuming that $r_c = r_{c'}$ (i.e., the characters can evolve at arbitrarily different rates in this model).

In the next section we will examine phylogeny estimation under each of these models.

3. PHYLOGENY ESTIMATION

Phylogeny estimation (as it is called in statistics) or phylogeny reconstruction (as it is called in computer science) addresses the issues of estimating or constructing a tree, along (possibly) with the associated parameters of evolution, from data that evolved down the tree. In this section we will examine the possibility of estimating the model tree (under each of the four basic models described in the previous section) from data generated on the tree under the associated random process operating on the tree.

There are several issues we will wish to address. The first is what we want to estimate – the underlying tree topology, or the parameters of evolution as well? As we will show, it is quite difficult to estimate dates at internal nodes, except under the most constrained model. However, estimating the

tree topology (modulo the location of the root) may be quite feasible, even under the least constrained model. Thus, parameter estimation – especially of times at internal nodes – is much harder than tree estimation.

The second issue is what we mean by the quality of an estimation procedure. In statistical inference, it is traditional to discuss what is *possible* under a model, given "enough data". In this context, we would be interested in understanding whether it is possible to estimate the tree and its associated parameters (perhaps modulo the location of the root) with error going to 0 as the amount of data goes to infinity. Here, the data are characters, so the question is asking about essentially an infinite number of characters. We can weaken the question, and instead of asking whether everything about the model can be reconstructed given infinite data, we ask what can be estimated exactly, given infinite data. In other words, if times cannot be inferred exactly, can the edge lengths ($\lambda_{e,c}$ parameters) be inferred? And if these edge lengths cannot be inferred from infinite data, can the underlying tree topology be inferred exactly? Finally, since data are never infinite, we will have to also address the quality of a reconstruction (measured in a precise quantitative way) on a finite number of characters.

3.1. **Statistical consistency and identifiability.** We begin with some comments about statistical estimation. We begin by introducing the concept of "patterns", and then the probability of a pattern for a given model tree. Recall that a model tree is a rooted binary tree along with associated parameters of evolution (times at internal nodes and rates of evolution of characters on edges). Suppose that the model tree has n leaves, labelled $s_1, s_2, \ldots, s_n$. A *pattern* (for our model) is just an equivalence relation on the leaves. Thus, we would consider a character that assigns all leaves the state 1 to define the same *pattern* as a different character that assigns all leaves the state 2. Thus, in a tree with four leaves, A, B, C, and D, there are only a finite number of patterns. For ease we will use a 4-tuple of integers to represent each pattern. For example, if each leaf is in its own class, we can represent this by $1, 2, 3, 4$. Or if all four leaves are in the same class, we can represent it by $1, 1, 1, 1$. The set of possible patterns is given as follows:

Patterns.

(1) $1, 1, 2, 2$
(2) $1, 2, 1, 2$
(3) $1, 2, 2, 1$
(4) $1, 1, 1, 1$
(5) $1, 1, 1, 2$
(6) $1, 1, 2, 1$

(7) $1, 2, 1, 1$
(8) $2, 1, 1, 1$
(9) $1, 1, 2, 3$
(10) $1, 2, 1, 3$
(11) $1, 2, 3, 1$
(12) $2, 1, 1, 3$
(13) $2, 1, 3, 1$
(14) $2, 3, 1, 1$
(15) $1, 2, 3, 4$

It is important to remember that in our model the actual states do not matter; all that matters is the equivalence relation imposed on the set of leaves (i.e., which leaves are assigned the same state, and not what state they are assigned). That is why this is the full set of patterns on a four-leaf tree.

It should be clear that because we have not allowed any homoplasy, if the model tree has an edge separating leaves A and B from C and D, then two of the patterns listed above have zero probability: namely, patterns 2 and 3 will never appear – no matter how the substitution probabilities are defined. On the other hand, as long as all $p_{e,c} \neq 0$, the first pattern has non-zero probability. This means that the underlying tree is *identifiable*. We summarize this discussion with the definition of identifiable, as follows:

Definition 1. *A model is said to be* identifiable *if the model can be distinguished from every other model by the probability it defines on every pattern.*

Since a model tree is more than just an unrooted tree (it contains a root, and associated edge parameters), we can ask about the identifiability of the full model. In fact, it can be shown that for each of our models, the underlying unrooted tree and the edge lengths $\lambda_{e,c}$ are identifiable, but that the location of the root and the times at the internal nodes are not identifiable in general. On the other hand, when the lexical clock assumption holds, then the location of the root can also be obtained correctly – given enough data.

Saying that a model is identifiable does not imply that any particular method will perform well, even given infinite data, however. So we now turn to questions about the performance of reconstruction methods. We begin with the definition of statistical consistency:

Definition 2. *A phylogeny estimation method is said to be* **statistically consistent** *under a model of evolution if the probability of recovering the tree (and its associated parameters) converges to* 1 *as the number of characters increases.*

This is a strong requirement, since it requires the ability to accurately estimate all the parameters of the evolutionary process. We may instead focus on whether a reconstruction method is statistically consistent with respect to just the underlying tree. However, if we are interested in estimating dates at internal nodes, we will need to look at performance issues involved with parameter estimation as well.

Statistical consistency is concerned with performance in the limit, and not actually with performance on finite data. Therefore, concrete performance studies, largely based upon simulation, are also appropriate (and common in the molecular phylogenetics literature).

3.2. **Perfect Phylogenies.** We now consider phylogeny (i.e., tree) reconstruction under the models introduced in Section 2.3. In all these models, we assumed that there is no homoplasy. Consequently, the true tree can, for all of the characters, be labelled with states on the internal nodes so that it is a *Perfect Phylogeny* in the sense of the following definition.

Definition 3. *A tree T on a set $\mathcal{L}$ of languages is a perfect phylogeny for a set C of characters if it is possible to label all internal nodes with character states so that all characters evolve without backmutation or parallel evolution in T. In this case the tree T is said to be compatible with the states observed at the leaves for all of the characters. (Kannan & Warnow, 1997).*

A labelling of internal nodes establishes a perfect phylogeny if and only if for every character and every pair of leaves with the same state of that character all nodes in the path through the tree between the two leaves share that same state.

As we remarked above, given a set of languages described by characters under a homoplasy-free model (any such homoplasy-free model, including ones we have not described), the only possible candidates for the true tree must be perfect phylogenies. Therefore, under the assumption that the data evolve under a homoplasy-free model, there must be at least one perfect phylogeny for the data. Furthermore, if there is a unique perfect phylogeny, then it must be the true tree.

We can formalize this statement as follows. Given a model tree (that is, a rooted tree T with associated numerical parameters), we denote the probability of the data $\mathcal{L}$ at the leaves by $\Pr\{\mathcal{L} \mid T, \text{parameters}\}$. For a given realisation of the data, this probability is a function of the tree T and the various numerical parameters (edge lengths, rates, time-durations etc., depending on the particular model class we are considering). This function is called the *likelihood* of the data.

Theorem 1. *The following dichotomy holds for the models considered here.*

- *If T is not a perfect phylogeny, then for all ways of assigning values to the parameters we have*

$$\Pr\{\mathcal{L} \mid T, \text{ parameters}\} = 0.$$

- *If T is a perfect phylogeny, then it is possible to assign parameter values so that*

$$\Pr\{\mathcal{L} \mid T, \text{ parameters}\} \neq 0.$$

3.3. **Maximum Likelihood.** *Maximum likelihood* is a general framework for constructing statistical estimators. In our setting, the maximum likelihood estimator of the tree T and the associated numerical parameters is the choice of these quantities that maximises the likelihood $\Pr\{\mathcal{L} \mid T, \text{ parameters}\}$. We will call the model tree that maximizes the likelihood the *maximum likelihood tree*.

It follows from Theorem 1 that the maximum likelihood tree is a perfect phylogeny.

If we are only interested in estimating T, then the maximum likelihood approach can be thought of as criterion for choosing between competing perfect phylogenies. If we are also interested in estimating both the tree and the numerical parameters, then we can fit them in a two step procedure that first restricts attention to the trees that are perfect phylogenies and then carries out the required maximisation over the resulting smaller set of possible trees and associated numerical parameters.

3.4. **Algorithms for solving perfect phylogeny.** Here we address the problem of estimating the true evolutionary tree under the homoplasy-free assumption. As discussed earlier in Section 2.4, given enough characters, it may be possible to infer the tree on the basis of its four-leaf subtrees. We describe here a specific polynomial time algorithm to do this which is guaranteed to be correct *if* enough characters exist to define the quartet on every tree, but it may fail to solve the problem otherwise. (Alternatively, algorithms which provably solve perfect phylogeny – always successfully constructing perfect phylogenies when they exist – are computationally expensive; each runs in time exponential in some parameter for the input – such as the number of characters or the maximum number of states per character, since the problem is NP-complete (Steel, 1992; Bodlaender *et al.* , 1992). Of the various such algorithms, the algorithm by Kannan and Warnow can enumerate all perfect phylogenies on a given dataset (Kannan & Warnow, 1997).)

Heuristic for Perfect Phylogeny Reconstruction For each quartet of languages L_1, L_2, L_3, L_4, see if there is a character in the dataset that splits the quartet into two sets; for example, a character c such that $c(L_1) = c(L_2) \neq$

$c(L_3) = c(L_4)$. If such a character exists, note that the true tree on these four languages has to split L_1, L_2 on one side, and L_3, L_4 on the other. (There cannot exist two characters splitting the four languages differently, because then no perfect phylogeny exists.) Record the constraints on all quartets. These constraints have to be then combined into a tree on which all the constraints hold. Constructing a tree consistent with all these constraints is itself an NP-hard problem, but can often be done in practice in a greedy fashion:

- Step 1: find a pair of languages A, B which are always grouped together, and make them siblings. If no such pair exists, return *fail*.
- Step 2: remove A from the set of languages, and run the algorithm recursively on the remaining languages.
- Step 3: introduce A back into the tree on the other languages by making it sibling to B within the tree.

This algorithm will work as long as the first step can be executed (such a pair of languages may not exist in some cases). Thus, although a perfect phylogeny may exist, the algorithm may fail and not produce it; any tree produced by the algorithm will, however, be a perfect phylogeny.

Considering the problem from a theoretical perspective, given enough data there will be (with high probability) a unique perfect phylogeny, and so these perfect phylogeny reconstruction methods are statistically consistent techniques for estimating the underlying unrooted tree, for all homoplasy free models, provided that rates of evolution are bounded away from 0 and infinity. On the other hand, estimating the remaining parameters (edge lengths in particular) requires additional techniques.

3.5. **Reconstruction under a lexical clock model.** We now turn to the problem of estimating evolution under a lexical clock. In this case, rather than using the computationally intensive perfect phylogeny techniques, we can instead use fast lexicostatstical methods to estimate the true tree. If there are enough data (i.e., enough characters), then the reconstructed rooted tree will be correct. At this point, we may wish to estimate relative times at internal nodes. However, to do this we will need additional information about how rates can vary across characters. Under the simplest model homoplasy-free evolution there is no variation of rates across characters, and so under the combined assumptions (all characters evolve under the same clock-like model), we can both estimate the rooted tree and the relative times at internal nodes. Under more general models, however, these estimations may not be statistically well-founded. For example, we indicate in Appendix 9 why it can be impossible to obtain relative times for the no common mechanism model with a lexical clock.

4. PHYLOGENETIC MODELS THAT INCORPORATE BORROWING

After evolving from a common ancestor, languages may remain in close contact and borrow from each other. To the extent that all such borrowing can be clearly identified, a genetic tree can still be an appropriate model of the evolutionary development of the family; however, if the contact cannot be clearly identified, then trees are inadequate for modeling the evolutionary history of such languages. In this section we describe how we can extend the tree models of linguistic evolution to the case where languages continue to remain in contact.

4.1. **Networks - the graphical model of language evolution.** When it is reasonable to define an underlying genetic tree, so that evolution by contact can be discriminated from genetic inheritance, the appropriate graphical model is a rooted *network* N, consisting of two components: the underlying "genetic tree" T, and a set E_c of additional "contact edges". Since borrowing between two languages may occur in both directions, contact edges are bidirectional. See Figure 2 for an example of such a network.

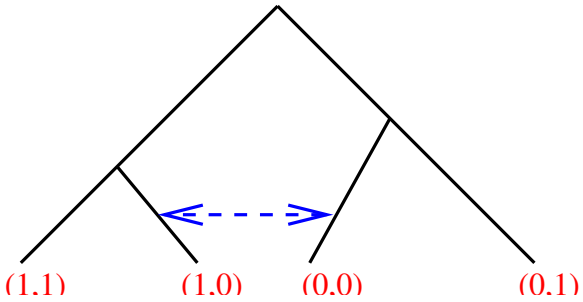

FIGURE 2. A phylogenetic network with a single contact edge. Both characters labeling the leaves are compatible on this network, since each is compatible on at least one of the three trees contained inside the network (the three trees shown in Figure 3).

Since contact occurs between two languages that co-exist in time, the existence of a contact edge between two nodes u and v in a network N implies that $t(u) = t(v)$ (this is why contact edges are always drawn horizontally).

4.2. **Character evolution down networks.** There are two modes of evolution on networks:

(1) "genetic evolution": evolution down the edges of the genetic tree, and
(2) "horizontal transfer": transmission of character states by contact.

When the state i of character c is transmitted horizontally from language L_1 to language L_2, then L_2 replaces its current state of character c by state i. Therefore, the state i of a character c at a node u that has two parents is transmitted from exactly one parent. Hence, a character is compatible on a network N if it is compatible on at least on the trees contained inside the network. Figure 3 shows the three trees contained inside the network of Figure 2. The first character is compatible on the tree in Figure 3(a), while the second character is compatible on both trees in Figures 3(b) and 3(c). Hence, both characters are compatible on the network in Figure 2.

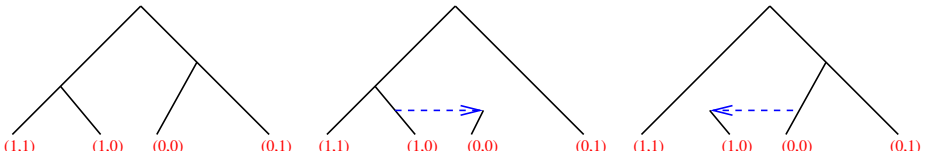

FIGURE 3. The three trees contained inside the phylogenetic network in Figure 2.

If we extend the homoplasy-free assumption to the network case, we will assert that every character evolves without backmutation or parallel evolution, but may evolve down either tree edges, or contact edges. Note therefore that every character evolves down a tree contained within the network. We can therefore extend the perfect phylogeny concept (which is defined only for trees) to the network case, as follows:

Definition 4. *We say N is a perfect phylogenetic network (PPN) for a set L of languages described by a set C of characters if every character in C is compatible on at least one of the trees contained inside N.*

The network in Figure 2 is a perfect phylogenetic network.

4.3. **Parameters of network models.** In addition to the parameters describing how characters evolve down the genetic tree, a full description of a network model requires some additional parameters. In particular, for each character and orientation of a contact edge, we need to define the probability that the character will be transmitted on that contact edge in that orientation. We will make the assumption that such probabilities can be written as the product of two values:

- p_e: the probability of transmission via contact on edge e of the most easily borrowed characters, and
- $ptrans_c$: the probability of transmission via contact of the character c.

(In other words, the probability that a character c is transmitted on a contact edge e is $ptrans_c \times p_e$.)

Allowing for contact edges means we will need additional $2|E_c| + |C|$ parameters, where E_c denotes the set of contact edges, and C denotes the set of characters.

Since every character must evolve down a tree contained within the phylogenetic network for the family, as long as we assume that evolution is homoplasy-free, by definition the true phylogenetic network will by necessity be a perfect phylogenetic network. This also means that the maximum likelihood network for a dataset will be a perfect phylogenetic network.

4.4. **Estimating the true phylogenetic network.** Inferring perfect phylogenetic networks is a more complicated issue than the comparable problem of inferring perfect phylogenies, for a number of reasons. One reason is that we do not yet now the conditions under which the homoplasy-free network models are identifiable; our initial research shows that for a very small number of contact edges the model is identifiable, but we do not know about the general case. Another reason is computational: while finding perfect phylogenies (when they exist) is a computationally hard problem, it is not clear how to go about constructing a good perfect phylogenetic network. (It is easy to construct a perfect phylogenetic network with a lot of contact edges, but, for example, constructing one with a minimum number of contact edges is computationally hard.) Further research will need to investigate how to address both those issues.

4.5. **An Indo-European analysis.** In (Nakhleh *et al.* , 2004), we analyzed a dataset of 24 Indo-European languages, described by a set of 292 characters. The methodology we used to analyze the dataset consisted of two steps:

(1) Find the genetic tree
(2) Add a minimum number of contact edges to make all characters compatible.

We examined several different candidate genetic trees (including two suggested by Craig Melchert) during this analysis. Our analysis then compared each of the minimal completions of each genetic tree to a perfect phylogenetic network, with respect to several mathematical criteria: number of contact edges added, number of characters that must evolve on the contact edges, and number of borrowing events; we also examined each perfect phylogenetic network with respect to its feasibility with regard to established historical records. The best of these perfect phylogenetic networks, with respect to each of the criteria, is shown in Figure 4. Of the full set of 292 characters, 278 characters (or more than 95%) are compatible with the

genetic tree underlying the network. Two of the contact edges (both involving Germanic) are well-supported in this analysis, but the remaining contact edge is questionable (it has less support, and is less feasible from a historical perspective). Further research will attempt to clarify the IE history.

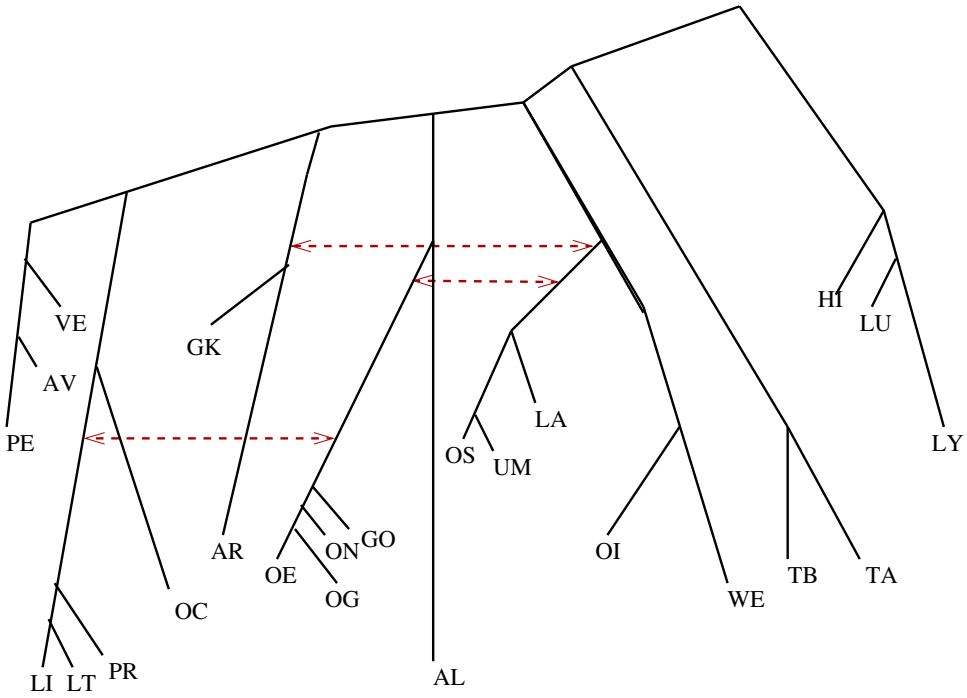

FIGURE 4. A perfect phylogenetic network for the IE dataset described in (Nakhleh *et al.* , 2004). The solid lines represent the genetic tree, and the dashed lines represent the three contact edges.

5. EXTENSIONS

Missing from this article is a discussion of how these models might be extended to address various situations that arise in phylogenetic analyses in historical linguistics:

- Characters that evolve with homoplasy (specifically evolving in parallel, or with back mutation),
- Polymorphic characters (exhibiting two or more states on some language), and
- Families that evolve in so much contact that our network models are inappropriate.

Modelling homoplasy is challenging; we need to understand the mechanisms of real homoplasy, rather than simply identifying characters that are incompatible with the genetic tree, since incompatibility can arise from borrowing, undetected polymorphism, or other factors that we have not identified. One clear cause of polymorphism is simply insufficient understanding of the family, so that true cognates cannot be established accurately. Clearly, modelling homoplasy will require a serious linguistic study, before its features can be then modelled mathematically. In an earlier paper we proposed a model of how polymorphism arises (Bonet *et al.* , 1999), based upon semantic shift; however, borrowing of lexemes provides an alternative explanation that needs to be considered. The problem of inferring evolution when there is a great deal of contact between lineages is formidable, and clearly will require a different approach than what we have taken in (Nakhleh *et al.* , 2004). A critical issue there is to determine the conditions under which the underlying genetic tree can be clearly (or fairly well) identified, even in the presence of significant borrowing. Here, techniques from molecular systematics may be worth examining (in particular, attempts to identify a phylogenetic tree for bacteria, despite all the horizontal gene transfer, may have been successful - see (Lerat *et al.* , 2003)).

6. CONCLUSIONS

This paper has several purposes. First, we wish to make explicit what is meant by a probabilistic model of evolution, so that the assumptions in each model can be examined, and the consequences for evolutionary history reconstruction examined in a scientific way. We hope that our discussion of the difficulties in estimating times at nodes should make it clear that even under these models, which are designed to reflect linguistic evolution, such estimations may not be realistically sought – at least not until we have a much better understanding of the stochastic processes underlying linguistic characters. We therefore hope to caution researchers seeking to estimate dates, and also help readers of the scientific literature to critique such attempts. Finally, we hope that we have also demonstrated the potential for a careful statistical inference, based upon a reasonable model, to elucidate evolutionary histories better than purely traditional means have been able.

7. ACKNOWLEDGMENTS

TW would like to acknowledge the Radcliffe Institute for Advanced Study and the Program in Evolutionary Dynamics at Harvard University, which provided generous support for this research. We also thank the MacDonald Institute for inviting us to the conference, and to submit this paper.

8. APPENDIX: LIKELIHOOD COMPUTATIONS

We show how it is possible to compute likelihoods for the general model introduced in Subsection 2.4.

Because of our simplifying assumption that characters evolve independently, it suffices to calculate the likelihood for a single character. Recall that the data for that character consist of a partition of the set S of languages. (Recall further that a partition of S is a collection of disjoint, nonempty subsets of S called *blocks* whose union is S.) We declare that two languages are in the same block if they exhibit the same state of the character.

Now the collection of partitions of S is a *partially ordered set*, where we declare that a $\pi \leq \sigma$ if every block of π is contained in some block of σ (that is, if the blocks of σ are unions of blocks of π) and we say that π is a *refinement* of σ.

Write Σ for our observed partition (that is, our data). We want to compute the probability of the event $\{\Sigma = \sigma\}$ for each partition σ of S.

Now $\pi \leq \Sigma$ if and only if for each block of π there is no change on any of the edges of the smallest subtree containing that block. For each block A of π, let $T(A)$ be the smallest subtree containing A and denote by $T(\pi)$ the union of the subtrees $T(A)$. Denoting by $\Pi(\pi)$ the product of the $p_{e,c}$ over edges in $T(\pi)$, we have

$$\sum_{\pi \leq \sigma} \Pr\{\sigma = \Sigma\} = P\{\pi \leq \Sigma\} = \Pi(\pi).$$

We can now use Moebius inversion (see, for example, (Stanley, 1997)) on the partially ordered set of partitions of S to obtain the probabilities $\Pr\{\sigma = \Sigma\}$. If $\pi \leq \sigma$, σ has k blocks, and the ith block of σ is the union of n_k blocks of π, then the the value $\mu(\pi, \sigma)$ of the Moebius function is $\prod_{i=1}^{k} (-1)^{n_i - 1}(n_i - 1)!$. By the dual form of the Moebius inversion formula (see Proposition 3.7.2 of (Stanley, 1997)), we have

$$\Pr\{\pi = \Sigma\} = \sum_{\pi \leq \sigma} \mu(\pi, \sigma) \Pr\{\sigma \leq \Sigma\} = \sum_{\pi \leq \sigma} \mu(\pi, \sigma) \Pi(\sigma).$$

9. APPENDIX: UNIDENTIFIABILITY OF THE NO COMMON MECHANISM MODEL WITH A LEXICAL CLOCK

Recall that under the no common mechanism model with a lexical clock we have that the probability of no substitution on an edge is $e^{-\lambda_{e,c}}$, where $\lambda_{e,c} = r_c T(e)$. It might seem reasonable that if we have an infinite amount of data, then we can determine the relative edge durations $T(e')/T(e'')$ for any pair of edges e', e''. This is in fact not the case.

The crux of the matter is contained in the following set-up. Suppose that $(X_1, Y_1), (X_2, Y_2), \ldots$ is an infinite sequence of pairs of random variables such that:

- each X_i and each Y_i take the value 0 or 1,
- for some parameters τ and $\rho_1, \rho_2, \ldots$,
$$\Pr\{X_i = 0\} = e^{-\rho_i}$$
and
$$\Pr\{Y_i = 0\} = e^{-\rho_i \tau},$$
- the random variables $X_1, Y_1, X_2, Y_2, \ldots$ are independent.

We think of X_i (respectively, Y_i) as the random variable that takes the value 0 if there is no subsitution on edge e' (respectively, edge e'') for the i^{th} character and 1 otherwise. Because of our freedom to multiply rates and divide edge durations by the same constant, we take edge e' to have length 1 and edge e'' to have lenth τ. The "nuisance parameter" ρ_i is the rate of substitution for character i.

The question is, "To what extent can we recover τ with arbitrary amounts of data without any knowledge of the behaviour of the ρ_i?" As the following shows, the answer is "Not at all."

To see this, consider the situation where the ρ_i are actually realisations of an (unobserved) independent, identically distributed sequence R_i. Then, incorporating this randomness, we have that the likelihood of the sequence $(X_1, Y_1), (X_2, Y_2), \ldots$ is a product of terms for each i and the i^{th} term just involves the four terms

$$\begin{aligned}
\Pr\{X_i = 0, Y_i = 0\} &= \mathrm{E}[e^{-R}\, e^{-\tau R}], \\
\Pr\{X_i = 0, Y_i = 1\} &= \mathrm{E}[e^{-R}\, (1 - e^{-\tau R})], \\
\Pr\{X_i = 1, Y_i = 0\} &= \mathrm{E}[(1 - e^{-R})\, e^{-\tau R}],
\end{aligned}$$

and

$$\Pr\{X_i = 1, Y_i = 1\} = \mathrm{E}[(1 - e^{-R})\, (1 - e^{-\tau R})],$$

where R is a random variable with the same distribution as the R_i. These terms are in turn linear combinations of the three quantities $\mathrm{E}[e^{-R}]$, $\mathrm{E}[e^{-\tau R}]$, and $\mathrm{E}[e^{-(1+\tau)R}]$. It is clear that estimation of τ will be impossible if we can find another rate $\upsilon \neq \tau$ and another random variable S such that

$$\begin{aligned}
\mathrm{E}[e^{-R}] &= \mathrm{E}[e^{-S}], \\
\mathrm{E}[e^{-\tau R}] &= \mathrm{E}[e^{-\upsilon S}],
\end{aligned}$$

and

$$\mathrm{E}[e^{-(1+\tau)R}] = \mathrm{E}[e^{-(1+\upsilon)S}].$$

We will take R and S to each be the sum of two independent gamma distributed random variables. By choosing the parameters of the gamma

distributions appropriately, we can arrange for R and S to have Laplace transforms

$$\mathrm{E}[e^{-\zeta R}] = (\alpha\zeta + 1)^{-a}(\beta\zeta + 1)^{-a}$$

and

$$\mathrm{E}[e^{-\zeta S}] = (\gamma\zeta + 1)^{-1}(\delta\zeta + 1)^{-1}$$

for $a, \alpha, \beta, \gamma, \delta > 0$. The question is thus whether we can find values of $\tau \neq \upsilon$ and $a, \alpha, \beta, \gamma, \delta$ such that

$$(\alpha + 1)^a(\beta + 1)^a = (\gamma + 1)(\delta + 1),$$

$$(\alpha\tau + 1)^a(\beta\tau + 1)^a = (\gamma\upsilon + 1)(\delta\upsilon + 1),$$

and

$$(\alpha(1 + \tau) + 1)^a(\beta(1 + \tau) + 1)^a = (\gamma(1 + \upsilon) + 1)(\delta(1 + \upsilon) + 1),$$

.

Fix $\tau_0, \alpha_0, \beta_0$ with $\alpha_0 \neq \beta_0$. We can clearly solve the above equations for υ, γ, δ when $\tau = \tau_0$, $\alpha = \alpha_0$, $\beta = \beta_0$, and $a = 1$ by setting $\upsilon = \tau_0$, $\gamma = \alpha_0$, and $\delta = \beta_0$. It is not hard to check that the Jacobian matrix of the transformation

$$(\upsilon, \gamma, \delta) \mapsto ((\gamma+1)(\delta+1), (\gamma\upsilon+1)(\delta\upsilon+1), (\gamma(1+\upsilon)+1)(\delta(1+\upsilon)+1))$$

is non-singular, and so for any choice of (τ, α, β, a) near $(\tau_0, \alpha_0, \beta_0, 1)$ the above equations have a solution. The only thing we have to rule out is that all such solutions have $\upsilon = \tau$. If we again fix τ at τ_0, then another Jacobian calculation shows that the transformation

$$(\alpha, \beta, a) \mapsto ((\alpha+1)^a(\beta+1)^a, (\alpha\tau_0+1)^a(\beta\tau_0+1)^a, (\alpha(1+\tau_0)+1)^a(\beta(1+\tau_0)+1)^a)$$

maps any open neighourhood of $(\alpha_0, \beta_0, 1)$ into a set with non-empty interior. Because the image of the transformation

$$(\gamma, \delta) \mapsto ((\gamma+1)(\delta+1), (\gamma\tau_0+1)(\delta\tau_0+1), (\gamma(1+\tau_0)+1)(\delta(1+\tau_0)+1))$$

is a two-dimensional surface, it is clear that not all solutions of the above equations will have $\tau = \upsilon$.

REFERENCES

Bergsland, Knut, & Vogt, Hans. 1962. On the validity of glottochronology. *Current Anthropology*, **3**, 115–153.

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.

Bonet, M., Phillips, C.A., Warnow, T., & Yooseph, S. 1999. Constructing evolutionary trees in the presence of polymorphic characters. *SIAM J. Computing*, **29**(1), 103–131.

Evans, Steven, Ringe, Don, & Warnow, Tandy. 2004. *Inference of divergence times as a statistical inverse problem*. Submitted.

Forster, Peter, & Toth, Alfred. 2003. Toward a phylogenetic chronology of ancient Gaulish, Celtic, and Indo-European. *Proc. Natl. Acad. Sci. USA*, **100**, 9079–9084.

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

Kannan, Sampath, & Warnow, Tandy. 1997. A fast algorithm for the computation and enumeration of perfect phylogenies when the number of character states is fixed. *SIAM J. Computing*, **26**(6), 1749–1763.

Lerat, Emmanuelle, Daubin, Vincent, & Moran, Nancy A. 2003. From Gene Trees to Organismal Phylogeny in Prokaryotes:The Case of the γ-Proteobacteria. *PLoS Biology*, **1**(1), e19. DOI: 10.1371/journal.pbio.0000019.

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

Stanley, Richard P. 1997. *Enumerative combinatorics. Vol. 1*. Cambridge Studies in Advanced Mathematics, vol. 49. Cambridge: Cambridge University Press. With a foreword by Gian-Carlo Rota, Corrected reprint of the 1986 original.

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

E-mail address: tandy@cs.utexas.edu

TANDY WARNOW, DEPARTMENT OF COMPUTER SCIENCES, UNIVERSITY OF TEXAS AT AUSTIN, AUSTIN, TX 78712, U.S.A.

E-mail address: evans@stat.Berkeley.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: dringe@unagi.cis.upenn.edu

DON RINGE, DEPARTMENT OF LINGUISTICS, 619 WILLIAMS HALL, UNIVERSITY OF PENNSYLVANIA, PHILADELPHIA, PA 19104-6305, U.S.A.

E-mail address: nakhleh@cs.utexas.edu

LUAY NAKHLEH, DEPARTMENT OF COMPUTER SCIENCES, UNIVERSITY OF TEXAS AT AUSTIN, AUSTIN, TX 78712, U.S.A.