An experimental study comparing linguistic phylogenetic reconstruction methods

François Barbançon,[1] Steven N. Evans,[2] Luay Nakhleh[3] Don Ringe,[4], and
Tandy Warnow,[5,*]
[1]Palantir Technologies, 100 Hamilton Street, Suite 300, Palo Alto CA 94301, USA
[2]Dept. of Statistics, Univ. of California at Berkeley, Berkeley CA 94720-3860, USA
[3]Dept. of Computer Sciences, Rice University, Houston TX 77005, USA
[4]Dept. of Linguistics, Univ. of Pennsylvania, Philadelphia, PA 19104, USA
[5]Dept. of Computer Sciences, Univ. Texas at Austin, Austin, TX 78712, USA
[*]To whom correspondence should be sent; tandy@cs.utexas.edu

Appendix

1 Overview

We provide details for how we generated the data for the simulation study, including how we generated the model phylogenetic networks (see Section 6.1), the mathematical model used for simulating linguistic character evolution down the phylogenetic networks (see Section 6.2), how we computed statistically-corrected distances between languages under the model (see Section 6.3), how we computed consensus trees (see Section 6.4), and software versions and commands used (see Section 6.5). We also include some additional discussion in Section 1.6). We direct the interested reader to Felsenstein (2004) for an overview of the related problem of estimating phylogenies for molecular sequences.

1.1 Model network generation

Our simulation generates random binary trees using a Yule process with per individual birth rate 1 conditioned to have the requisite number of terminal taxa at time 1, as implemented in `r8s` (Sanderson, 2006). Thus, the trees we generated by r8s have edge lengths that represent elapsed time, and are normalized so that all paths from root to terminal leaf have length 1. We indicate the elapsed time on edge e by $t(e)$.

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

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

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

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

$$S(\pi) = -\log \frac{(t_1 - t_0)}{(t_2 - t_0)}.$$

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

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

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

1.2 Stochastic model of language evolution

We use the stochastic model of language evolution proposed in Warnow et al. (2006). In that model, there is a fixed collection of linguistic characters, each of which has an infinite collection of possible states. A language is represented by the particular states it exhibits for each of the characters (note, however, that two leaves in the tree *may* be identical with respect to the characters, due to insufficient evolution). Languages evolve down an underlying tree with added reticulate edges that represent contact events between lineages. At a contact event, the state of each character may be instantaneously transferred from the lineage at one end of the edge to the lineage at the other end (that is, one lineage 'borrows' the character state of another), and replaces the character state inherited from its genetic parent.

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

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

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

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

Reticulate evolution occurs through character borrowing via contact edges added to the underlying tree. Each contact edge allows bi-directional borrowing between a pair of languages present in different lineages at the same time. Suppose an edge e_a of the underlying tree is in existence between times t_a and t'_a, and an edge e_b is in existence between times t_b and t'_b. If the time interval $[t_a, t'_a]$ overlaps the time interval $[t_b, t'_b]$, then a contact edge $e_{a,b}$ may be generated between e_a and e_b at any time point t_c in the overlap interval.

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

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

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

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

The rates-across-sites assumption is quite standard in molecular systematics and its underlying models, but is nevertheless also questionable. It states that every two characters evolve proportionally – so that if one character evolves at twice the speed of another character on one branch of the tree, then it evolves at twice the speed of the other character on every branch in the tree. We quantify the deviation from this assumption through the parameter het, which we also define below. (See Evans and Warnow (2005) for a study discussing the rates-across-sites assumption and statistical identifiability of divergence times.)

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

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

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

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

and

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

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

1.3 Corrected distances.

One of the methods we used to estimate the genetic tree is neighbor joining, a distance-based method. When we use distance-based methods, such as neighbor joining, to analyze datasets generated by the stochastic model described above, we need to 'correct' the distances to account for unseen changes. This allows us to estimate, in a statistically rigorous way, a distance matrix that will approach (as the number of characters increases) a matrix that uniquely identifies the true genetic tree that produced the data. Such a matrix is called 'additive', and is defined by the path distances (the sum of the lengths of the edges in a leaf-to-leaf path) in the tree. Given an additive matrix (or a matrix sufficiently close to an additive matrix for the true tree), the true tree can be obtained in polynomial time. Here we show how we calculate this corrected distance from the observed character data.

The corrected distance $D(i,j)$ between two languages i and j is computed by calculating corrected distances for each type of character (i.e., slow lexical (SL), medium lexical (ML), fast lexical (FL) and morphological (Mo)), and then averaging them:

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

where num_X is the number of characters in class X as X ranges over the four classes of characters, $HD_X(i,j)$ is the Hamming Distance between languages i and j computed only on the basis of the characters in the class X, and $D_X(i,j) = -\log(1 - HD_X(i,j)/\mathrm{num}_X)$. Under the model we propose, if we do not allow reticulation, homoplasy or heterotachy (that is, violation of the rates-across-sites assumption), then the $D(i,j)$ will be consistent statistical estimators of genuine tree distances that are concordant with the topology of the underlying genetic tree. That is, when the numbers of replicates num_X are large, the $D(i,j)$ will be close to a collection of leaf-to-leaf distances on a tree with edge lengths whose shape is that of the genetic tree. ('Replicates' are independent random samples drawn from a statistical distribution; in this context the number of characters is the number of replicates.)

1.4 Majority Consensus Trees.

We now define the majority consensus of a set $\mathcal{T}$ of trees, each on the same set of languages. Every edge in a tree defines a bipartition on the set of leaves of the tree, in that deleting the edge (but not its endpoints) splits the leaf set into two parts. Thus, every tree on a set of leaves can be represented by the set of bipartitions induced by its edges. Given a set of trees on the same set of languages, the majority consensus is the tree that contains those edges whose bipartitions appear in more than half of the trees.

1.5 Software Versions and Commands

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

Generating trees with R8s

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

UPGMA using PAUP*

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

Neighbor joining using PAUP*

```
NJ treefile=PAUP/PAUP_nj_out.trees replace;
```

Maximum parsimony or weighted maximum parsimony using PAUP*

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

Gray & Atkinson’s method using MrBayes

```
begin mrbayes;
set autoclose=yes nowarn=yes;
        lset rates=gamma;
        mcmcp ngen=100000 printfreq=10000 samplefreq=500
                nruns=1 nchains=4 savebrlens=yes filename=Bayes_out;
        mcmc;
        set nowarnings=yes;
        sumt filename=Bayes_out burnin=100;
        quit;
end;
```

1.6 Additional Discussion

Various stochastic models of linguistic character evolution have been proposed or implicitly suggested in simulation studies and statistical analyses of language evolution (Gray and Atkinson, 2003; McMahon and McMahon, 2006; Nakhleh et al., 2005; Atkinson et al., 2005). Models of linguistic character evolution differ in several ways, as follows.

(1) A model may assume that all evolution is treelike, so that no borrowing occurs, or it may explicitly model borrowing. Borrowing is modelled with ‘contact edges’ linking edges of the tree which represent direct descent; a borrowed character

evolves partly down descent edges and partly down contact edges, so that a tree which represents its evolution is different from the tree representing true linguistic descent.

(2) A model may assume that evolution is clock-like or not. If evolution is clocklike, characters do not necessarily evolve at the same rate, but the rates at which they evolve will be multiples of a constant.

(3) A model may assume that characters evolve identically or not. In practice a model which assumes identical evolution will not be realistic for linguistic data, because there are many ways in which linguistic characters typically evolve non-identically. For instance, if some characters are borrowed and others are not, their evolution is not identical, since the trees representing their evolution are necessarily different (see 1 above); but word-borrowing is a natural linguistic process. If characters evolve at different rates in different parts of the tree, their evolution is not identical; but it is clear that some languages are more conservative than others.

(4) A model may assume that different characters evolve independently or not. In reality it seems likely that the evolution of some phonological characters and some lexical characters can depend on the evolution of others, but linguists usually try to use independently evolving characters so as to maximize the amount of independent evidence for a single tree.

(5) Finally, a model may allow homoplasy to occur or not. For example, Nakhleh et al. (2005) proposed the non-parametric 'perfect phylogenetic network' model of language evolution in which every character evolves without any homoplasy but borrowing is permitted, while Warnow et al. (2006) proposed a parametric model that allows for borrowing and limited homoplasy.

Parametric stochastic models are those in which the probability distribution of the observed data comes from a given family of possible distributions, with the actual member of the family being determined by a collection of numerical parameters. For example, a parametric model may assume that all characters evolve independently, but without any homoplasy, and require extra parameters to specify the probability that a given character will change its state (and thus evolve into a new state) on a given edge in the tree. Thus the model is fully specified by the underlying tree and the parameter values for the substitution mechanism.

References

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

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

Felsenstein, J. 2004. Inferring Phylogenies. Sinauer.

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

McMahon, A. and R. McMahon. 2006. Why linguists don't do dates: evidence from Indo-European and Australian languages. Pages 153–160 *in* Phylogenetic Methods and the Prehistory of Languages (P. Forster and C. Renfrew, eds.). McDonald Institute for Archaeological Research.

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

Sanderson, M. 2006. r8s version 1.71 software package. http://loco.biosci.arizona.edu/r8s.

Warnow, T., S. Evans, D. Ringe, and L. Nakhleh. 2006. A stochastic model of language evolution that incorporates homoplasy and borrowing. Pages 75–90 *in* Phylogenetic Methods and the Prehistory of Languages (P. Forster and C. Renfrew, eds.). MacDonald Institute for Archaeological Research.