Statistical Applications in Genetics and Molecular Biology

Volume 10, Issue 1

2011

Article 16

A Variance-Components Model for Distance- Matrix Phylogenetic Reconstruction

Walter R. Gilks, *University of Leeds and Rothamsted
Research*

Tom M.W. Nye, *University of Newcastle upon Tyne*

Pietro Lio, *University of Cambridge*

Recommended Citation:

Gilks, Walter R.; Nye, Tom M.W.; and Lio, Pietro (2011) "A Variance-Components Model for Distance-Matrix Phylogenetic Reconstruction," *Statistical Applications in Genetics and Molecular Biology*: Vol. 10: Iss. 1, Article 16.

DOI: 10.2202/1544-6115.1574

A Variance-Components Model for Distance-Matrix Phylogenetic Reconstruction

Walter R. Gilks, Tom M.W. Nye, and Pietro Lio

Abstract

Phylogenetic trees describe evolutionary relationships between related organisms (*taxa*). One approach to estimating phylogenetic trees supposes that a matrix of estimated evolutionary distances between taxa is available. Agglomerative methods have been proposed in which closely related taxon-pairs are successively combined to form ancestral taxa. Several of these computationally efficient agglomerative algorithms involve steps to reduce the variance in estimated distances. We propose an agglomerative phylogenetic method which focuses on statistical modeling of variance components in distance estimates. We consider how these variance components evolve during the agglomerative process. Our method simultaneously produces two topologically identical rooted trees, one tree having branch lengths proportional to elapsed time, and the other having branch lengths proportional to underlying evolutionary divergence. The method models two major sources of variation which have been separately discussed in the literature: noise, reflecting inaccuracies in measuring divergences, and distortion, reflecting randomness in the amounts of divergence in different parts of the tree. The methodology is based on successive hierarchical generalized least-squares regressions. It involves only means, variances and covariances of distance estimates, thereby avoiding full distributional assumptions. Exploitation of the algebraic structure of the estimation leads to an algorithm with computational complexity comparable to the leading published agglomerative methods. A parametric bootstrap procedure allows full uncertainty in the phylogenetic reconstruction to be assessed. Software implementing the methodology may be freely downloaded from StatTree.

KEYWORDS: agglomerative method, distance matrix, generalized least-squares regression, phylogenetic tree, variance-components model

Erratum

On page 3 and 28 the URLs should direct as follows:

Our methodology is implemented in STATTREE, which may be freely downloaded from <http://www.mas.ncl.ac.uk/~ntmwn/stattree>. Examples of the application of STATTREE can be found at <http://www.cl.cam.ac.uk/~pl219/GilksNyeLioSAGMB.html>.

1 Introduction

Phylogenetic trees describe evolutionary relationships between related organisms, or *taxa*. Currently, the most sophisticated statistical methods for reconstructing phylogenetic trees depend on probabilistic models of DNA or protein sequence evolution, estimated using maximum likelihood or Markov chain Monte Carlo; see for example Felsenstein (2004). These approaches require many evolutionary assumptions but concentrate on small-scale events in sequences such as point mutations or deletions. Such methods do not accommodate large-scale genomic rearrangements or supplementary non-sequence data such as phenotype or habitat. Moreover, their need to explore the vastness of tree space may render them computationally infeasible for problems involving very large numbers of taxa.

For large or complex phylogenetic data, methods based on distance matrices, such as NJ (Saitou and Nei, 1987), BIONJ (Gascuel, 1997), WEIGHBOR (Bruno, Socci, and Halpern, 2000), MVR (Gascuel, 2000) and FASTME (Desper and Gascuel, 2004), offer greater flexibility and computational efficiency. Such methods involve a distance metric which scores the dissimilarity between each pair of taxa. The resulting inter-taxon distance matrix is then processed by an agglomerative algorithm which progressively combines pairs of taxa, forming ancestral taxa and estimating branch lengths in the tree, concomitantly reducing the size of the working distance matrix. In general, for more than 3 taxa, the imputed phylogeny cannot reproduce observed distances exactly. Thus, a distance matrix contains uncertainty or *noise*, and might be consistent with many alternative phylogenies. Ideally, this uncertainty should be taken into account during the agglomeration in estimating branch lengths, in making topological decisions and in monitoring the amounts of information available for those decisions.

Each of the published agglomerative methods addresses the issue of uncertainty in its own style. NJ employs ordinary least-squares (OLS) and WEIGHBOR uses weighted least squares (WLS) estimates of branch lengths, which might be appropriate when distance estimates are uncorrelated (see, for example, Felsenstein (1987)). However, phylogenetic distances are often derived from DNA or protein sequence data, based on simple models of sequence evolution. Such models imply correlated distance estimates with variances which increase exponentially with branch length (Nei, Stephens, and Saitou, 1985, Bulmer, 1991). For such distances, Gascuel (1997) developed BIONJ, introducing the idea of estimating branch lengths by minimizing a sum of branch-length variances at each agglomeration. Rather different optimization criteria involving variances of branch-lengths are employed by WEIGHBOR, MVR and FASTME.

To estimate phylogenetic trees, correlations between distances should be taken into account (Chakraborty, 1977). For this, Hasegawa, Kishino, and Yano

(1985) and Bulmer (1991) have proposed the use of generalized least-squares (GLS). However, these non-agglomerative methods require estimation of the full tree in a single computationally intensive step, and hence are unsuitable for exploring large phylogenies. On the other hand, for computational efficiency, agglomerative methods such as NJ, WEIGHBOR and FASTME employ OLS or WLS, and thus do not take formal account of distance correlations. The full implementation of MVR correctly takes account of induced distance correlations, but greater computational efficiency is obtained if they are ignored (Gascuel, 2000).

Some phylogenetic methods are *ultrametric*. An ultrametric tree assumes a molecular clock beating out a constant pace of evolution in all branches. If branch lengths are drawn proportional to the amount of evolution occurring along them, then an ultrametric tree may be drawn with a physical time axis such that each ancestral taxon is aligned at its extant time, and each leaf taxon at the present time, as in Figure 1(a). Non-ultrametric trees, on the other hand, allow evolution to deviate from time-proportionality, so that leaf taxa become unaligned, as in Figure 1(b). Such varying amounts of evolution may be due to positive or negative selection pressures, to interspecific differences in generation times, to random variation in the rate of evolution (Thorne, Kishino, and Painter, 1998, Kishino, Thorne, and Bruno, 2001), or simply to the randomness of evolutionary events under neutral selection. Currently, the practitioner must decide whether to conduct an ultrametric or a non-ultrametric analysis. However, underlying any non-ultrametric *divergence tree* must lie a topologically identical but ultrametric *time tree* in which branch lengths are drawn proportional to elapsed physical time. This assumes that each speciation event occurs at a specific (but generally unknown) physical time. Thus an ultrametric time-tree and a non-ultrametric divergence tree can be simultaneously true. They estimate different aspects of evolution, sharing the same topology but differing in branch lengths.

Here we introduce a new distance-based agglomerative phylogenetic method, based on a variance-components model involving only means, variances and covariances of distances. Thus our method takes account of distance correlations but avoids making full probabilistic evolutionary assumptions. Our method simultaneously produces both a non-ultrametric divergence tree and its underlying ultrametric time tree. Consequently, without needing to specify an outgroup, our divergence tree is rooted. The true divergence tree may, to a greater or lesser extent, conform to a molecular clock assumption; our method allows the degree of this non-ultrametricity to be estimated from the data.

Our approach uses a sequence of staged GLS regressions to build the estimated tree, each stage formally incorporating variances and covariances in distances observed or estimated at the previous stage. The GLS computations involve inversion of square matrices with dimension of order $\mathcal{O}(m^2)$, where m is

Figure 1: Illustrating (a) an ultrametric tree with time axis recording years before present (bp) and leaves aligned at the present time; and (b) a non-ultrametric tree of the same topology.

the number of taxa. If tackled head-on, each GLS regression would have computational complexity of order $\mathcal{O}(m^6)$. However, thorough exploitation of the algebraic structure of these matrices leads to a GLS algorithm with computational complexity of order $\mathcal{O}(m)$, and the full phylogenetic reconstruction has computational complexity on a par with the most efficient distance-based algorithms. A parametric bootstrap procedure allows full phylogenetic uncertainty to be evaluated. Our methodology is implemented in STATTREE, which may be freely downloaded from <http://www.mas.ncl.ac.uk/~ntmwn/stattree>. Examples of the application of STATTREE can be found at <http://www.cl.cam.ac.uk/~pl219/GilksNyeLioSAGMB.html>.

2 Method

2.1 Time, divergence and distance

We distinguish between physical time t , divergence d , and distance $\hat{d}$.

Time: Let t_j denote the time before present (bp) that taxon j was extant. Let k denote the most recent common ancestor of taxa j and ℓ . Then the total physical time over which taxa j and ℓ have independently evolved is $t_{j\ell} = (t_k - t_j) + (t_k - t_\ell) = 2t_k - t_j - t_\ell$.

Divergence: This is the concept of *branch length*, which we define here to be the gross amount of change in sequence or other variables, summed along a path in the evolutionary tree. Each change makes a positive contribution to divergence, the amount being determined in some way appropriate to the problem

in hand. Thus a change immediately followed by its reversal both contribute positively. Since such reversed changes are generally undetectable, divergences are intrinsically unobservable. Let $d_{jk} = d_{kj}$ denote the divergence between any two taxa j and k . We expect d_{jk} to be equal on average to t_{jk} (subsuming an arbitrary scaling of time), with some random variation around this mean.

Distance: Although divergences are unobservable, we may be able to estimate them from available sequence or other data. We use the term *distance* and the notation $\hat{d}_{jk}$ to denote a measurement or estimate of d_{jk} . We expect $\hat{d}_{jk}$ to be equal on average to d_{jk} , with some random variation around this mean reflecting measurement imprecision.

The distinction between distance and divergence is seen more sharply when considering tree-additivity. Divergences are tree-additive, that is, for any three taxa j, k, ℓ where k lies on the path in the phylogenetic tree from j to ℓ , we have

$$d_{j\ell} = d_{jk} + d_{k\ell}. \quad (1)$$

However, tree-additivity is generally only an approximate property of distances, that is $\hat{d}_{j\ell} \approx \hat{d}_{jk} + \hat{d}_{k\ell}$.

For present-day taxa, we assume sequence or other data are available from which distances may be calculated directly. However, such data are in general unavailable for ancient taxa, so distances between them must be computed indirectly. Indeed, agglomerative methods recursively compute distances between ancient taxa on the basis of previously computed distances between more recent taxa (see below). That is, divergences are estimated recursively on the basis of previous divergence estimates. Thus, as we consider taxa j and k increasingly far back in time, distances $\hat{d}_{jk}$ will tend to become increasingly imprecise estimates of their underlying divergences d_{jk} . On the other hand, the number of pairs of present-day taxa descending from j and k and therefore contributing to the calculation of $\hat{d}_{jk}$ will concomitantly increase, which should tend to reduce imprecision (although there has been some debate on this issue; see Zwickl and Hillis (2002)). Therefore, as the agglomeration proceeds, it is unclear how much imprecision accumulates in the working distance matrix. It would be desirable to know whether, or under what conditions, imprecision accumulates or dissipates in agglomerative methods.

As noted above, current agglomerative methods employ a variety of strategies to take account of imprecision in distances. Here we consider the question of imprecision through a more coherent statistical-modeling framework. We show how imprecision can be estimated and monitored as the tree reconstruction proceeds; how it can be factored into the tree reconstruction; and how all this can be done without introducing unacceptable computational overheads.

2.2 Agglomeration

In common with most other distance-matrix methods, our approach is agglomerative. We begin with a set of currently extant taxa (or other genetic objects) of interest, and a matrix of distances between them. Recursively, we choose two taxa to be replaced by a new taxon representing their last common ancestor (LCA). This taxon pair is chosen to minimize the estimated total tree length, according to the neighbor-joining principle of Saitou and Nei (1987) (see Section 2.11). Distances from the new taxon to each of the remaining taxa are then computed. This agglomerative process continues until a single taxon remains.

We start the agglomerative process at stage 0 with $m^{(0)}$ taxa all extant at time 0, and a distance matrix containing $n^{(0)} = m^{(0)}(m^{(0)} - 1)/2$ elements in its upper triangle. There will be $m^{(0)}$ agglomerative stages. At a generic stage $i \geq 0$ of the agglomerative process, we consider a set of $m^{(i)}$ taxa, labeled $1^{(i)}, 2^{(i)}, \dots, m^{(i)}$. These taxa may have been extant at different times; we use $t_j^{(i)}$ to denote the extant time of taxon $j^{(i)}$. At this stage we will have a working distance matrix between these $m^{(i)}$ taxa, stored for convenience in a vector $\hat{d}^{(i)}$ of length $n^{(i)} = m^{(i)}(m^{(i)} - 1)/2$, whose elements are indexed by taxon-pairs in upper-triangular row order, which we refer to as *standard pair order*, as follows:

$$(1^{(i)}, 2^{(i)}), (1^{(i)}, 3^{(i)}), \dots, (1^{(i)}, m^{(i)}), (2^{(i)}, 3^{(i)}), \dots, (2^{(i)}, m^{(i)}), \dots, (m^{(i)} - 1, m^{(i)}). \quad (2)$$

Our principal tasks at stage i are to select, on the basis of $\hat{d}^{(i)}$, the two taxa to be agglomerated; to determine the extant time of their LCA; and to calculate a vector $\hat{d}^{(i+1)}$ containing distances between the resulting $m^{(i+1)} = m^{(i)} - 1$ taxa. Other quantities to be estimated are described below. The LCA represents a new internal taxon in the tree under construction, as illustrated in Figure 2.

2.3 Divergence relationships

To simplify the presentation, without loss of generality, we assume until further notice that the two taxa to be agglomerated at stage i are $1^{(i)}$ and $2^{(i)}$, as in Figure 2. Then, at stage $i + 1$, we label the new LCA as taxon $1^{(i+1)}$, and relabel each remaining taxon as follows:

stage	taxon label					
i	$3^{(i)}$	$4^{(i)}$	$\dots$	$k^{(i)}$	$\dots$	$m^{(i)}$
$i + 1$	$2^{(i+1)}$	$3^{(i+1)}$	$\dots$	$(k - 1)^{(i+1)}$	$\dots$	$m^{(i+1)}$

Figure 2: Illustrating a generic stage i of the agglomerative process, showing (a) the ultrametric time tree with its time axis, and (b) the corresponding non-ultrametric divergence tree whose edge lengths are proportional to the amount of evolution taking place along them. The $m^{(i)} = 4$ current taxa are denoted by black circles, and the LCA of taxa $1^{(i)}$ and $2^{(i)}$, which are to be agglomerated next to form taxon $1^{(i+1)}$, is denoted by a white circle. Taxon $1^{(i+1)}$ will replace taxa $1^{(i)}$ and $2^{(i)}$ at the next stage $i + 1$. Broken lines indicate earlier stages of the agglomeration.

From our assumption (1) of tree-additivity of divergences, we may then write

$$d_{jk}^{(i)} = \begin{cases} a_j^{(i)} + a_k^{(i)}; & j = 1; k = 2 \\ a_j^{(i)} + d_{1,k-1}^{(i+1)}; & j = 1, 2; k = 3, \dots, m^{(i)} \\ d_{j-1,k-1}^{(i+1)}; & j = 3, \dots, m^{(i)} - 1; k = j + 1, \dots, m^{(i)} \end{cases}, \quad (3)$$

where $d_{jk}^{(i)}$ denotes the divergence between taxa $j^{(i)}$ and $k^{(i)}$, and where $a_1^{(i)}$ and $a_2^{(i)}$ denote the divergences along the edges from taxa $1^{(i)}$ and $2^{(i)}$ to their LCA, taxon $1^{(i+1)}$. The first condition in (3) states that the divergence between the two taxa being agglomerated is equal to the sum of the divergences from each of them to taxon $1^{(i+1)}$, their LCA. The second condition states that the divergence between either of these two taxa and any other taxon is the sum of the divergences from each of them to taxon $1^{(i+1)}$ (noting the change in taxon labels at stage $i + 1$, described above). The final condition states that the divergence between any other pair of taxa remains unaltered (again noting the change in taxon labels).

Equation (3) may be expressed more concisely in vector notation:

$$d^{(i)} = A^{(i)} S^{(i)} a^{(i)} + B^{(i)} d^{(i+1)}, \quad (4)$$

where $d^{(i)}$ is the $n^{(i)}$ -vector of divergences $\{d_{jk}^{(i)}\}$ stored in standard pair order (2); $a^{(i)}$ is the 2-vector $(a_1^{(i)}, a_2^{(i)})^T$, where T denotes transposition; and $A^{(i)}$, $S^{(i)}$ and $B^{(i)}$ are the following $(n^{(i)} \times m^{(i)})$, $(m^{(i)} \times 2)$ and $(n^{(i)} \times n^{(i+1)})$ binary matrices. Matrix $S^{(i)}$ contains a 1 in its $(1, 1)$ and $(2, 2)$ elements, and 0 elsewhere. The rows of $A^{(i)}$ and $B^{(i)}$ correspond to taxon pairs of stage i in standard pair order. Each column $k = 1, \dots, m^{(i)}$ of $A^{(i)}$ indicates which taxon-pairs of stage i contain taxon $k^{(i)}$. The columns of $B^{(i)}$ correspond to taxon pairs of stage $i + 1$ in standard pair order, the column for pair $(j^{(i+1)}, k^{(i+1)})$ indicating which taxon pairs of stage i are connected on the tree via the path connecting taxa $j^{(i+1)}$ and $k^{(i+1)}$. For example, from Figure 2 we have

$$A^{(i)} = \begin{array}{c} \begin{matrix} & 1^{(i)} & 2^{(i)} & 3^{(i)} & 4^{(i)} \end{matrix} \\ \begin{matrix} (1^{(i)}, 2^{(i)}) \\ (1^{(i)}, 3^{(i)}) \\ (1^{(i)}, 4^{(i)}) \\ (2^{(i)}, 3^{(i)}) \\ (2^{(i)}, 4^{(i)}) \\ (3^{(i)}, 4^{(i)}) \end{matrix} \end{array} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}, \quad B^{(i)} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad S^{(i)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

where the row and column labels of A are shown, the rows labels of B are as for A , and the column labels of B correspond to taxon pairs $(1^{(i+1)}, 2^{(i+1)})$, $(1^{(i+1)}, 3^{(i+1)})$ and $(2^{(i+1)}, 3^{(i+1)})$.

2.4 Statistical assumptions

We now state our statistical assumptions concerning divergences $d^{(i)}$ and distances $\hat{d}^{(i)}$. Our approach is moment-based: we argue entirely in terms of conditional expectations, variances and covariances, and thus avoid the need to make explicit distributional assumptions, unlike WEIGHBOR, for example, which assumes Gaussian distances.

We adopt the convention that all expectations, variances and covariances implicitly condition on the true but unknown tree topology and taxon times, including the assertion (for now) that the LCA of taxa $1^{(i)}$ and $2^{(i)}$ is the most recent of any i th-stage taxon pair. However, where expectations and variances condition on divergences, we will make this explicit.

2.4.1 Non-ultrametricity

Our first set of assumptions concerns the relationship between the divergences and times which we consider at each stage $i \geq 0$ of the algorithm. Let $u^{(i)}$ be the 2-vector

containing the time-intervals from taxa $1^{(i)}$ and $2^{(i)}$ to their LCA, taxon $1^{(i+1)}$:

$$\begin{aligned} u^{(i)} &= \left[t_1^{(i+1)} - t_1^{(i)}, t_1^{(i+1)} - t_2^{(i)} \right]^T \\ &= \mathbf{1}_2 t_1^{(i+1)} - S^{(i)T} t^{(i)}, \end{aligned} \quad (5)$$

where $\mathbf{1}_k$ denotes a k -vector of 1s and $t^{(i)} = [t_1^{(i)}, t_2^{(i)}, \dots, t_m^{(i)}]^T$. Conditioning on $d^{(i+1)}$ and (implicitly) on $t_1^{(i+1)}$, we assume:

$$\mathbb{E}[a^{(i)} \mid d^{(i+1)}] = u^{(i)}, \quad (6)$$

$$\text{Var}[a^{(i)} \mid d^{(i+1)}] = \nu u_D^{(i)}, \quad (7)$$

where $\nu \geq 0$ is a scalar constant and $u_D^{(i)}$ denotes the diagonalization of vector $u^{(i)}$. (It may seem strange to condition here on $d^{(i+1)}$, which we do not know, but formally $d^{(i+1)}$ will take the role of a parameter vector in a regression model, as described below.) Thus, we assert that the variance of the divergence along a path is proportional to that path's time-length, and that divergences along non-overlapping paths are uncorrelated.

For example, suppose we have a DNA sequence of length n bases, where each base has one of the four states $\{A, C, G, T\}$. If we assume the Jukes–Cantor model of sequence evolution (Jukes and Cantor, 1969), in which events (state-changes) occur independently and with equal probability according to a Poisson process at rate λ , then the total number of events (including reversals of previous state-changes) in the sequence in a time-interval u will be Poisson with mean λnu . Defining divergence a to be the total number of events *per base* in time-interval u , scaling time so that $\lambda = 1$, and setting $\nu = n^{-1}$, leads to divergence a having mean u and variance νu , in agreement with (6,7). This simple example helps to provide some intuition about ν . However, our methodology is not tied to such simple evolutionary processes. With suitable scaling of time and choice of ν , any homogenous independent-increments process will satisfy (6,7), including more general sequence evolution models (Kimura, 1980, Hasegawa et al., 1985, Lanave, Preparata, Saccone, and Serio, 1984), with or without rate-variation across sites (Yang, 1993). The assumptions might be equally applicable to much more complex processes including large-scale genomic rearrangements; see, for example, Wang and Warnow (2005). Our moment-based approach has the advantage that details of the stochastic process need not be specified beyond (6,7).

Equation (7) allows divergences to depart from a strict molecular clock assumption. Thus the divergence tree is not constrained to be ultrametric, unlike the time tree. We call this kind of variation *distortion*. The amount of molecular clock distortion is controlled by parameter ν . From (6,7), setting $\nu = 0$ would assert that

$a^{(i)} = u^{(i)}$ for all i , so branch lengths in the divergence tree would be identical to those in the time tree and the divergence tree would then be ultrametric, obeying a strict molecular clock. Contrastingly, setting $\nu > 0$ confers some elasticity in the branches of the divergence tree, as illustrated in Figure 1(b), larger ν conferring greater elasticity. In the simple Poisson process example above, we have $\nu = n^{-1}$. However, our methodology does not require ν to be set externally, as we provide methodology for estimating it (see Section 2.12).

Taking expectations and variances over $a^{(i)}$ in equation (4) using (6,7), it follows immediately that divergences have the following conditional moments, conditioning on $d^{(i+1)}$, the part of the divergence tree that remains after agglomerating taxa $1^{(i)}$ and $2^{(i)}$:

$$\mathbb{E}[d^{(i)} \mid d^{(i+1)}] = A^{(i)} S^{(i)} u^{(i)} + B^{(i)} d^{(i+1)}, \quad (8)$$

$$\text{Var}[d^{(i)} \mid d^{(i+1)}] = \nu A^{(i)} S^{(i)} u_D^{(i)} S^{(i)T} A^{(i)T}. \quad (9)$$

A recursive expression for the marginal variance of $d^{(i)}$ then follows from (8,9):

$$\text{Var}[d^{(i)}] = \nu A^{(i)} S^{(i)} u_D^{(i)} S^{(i)T} A^{(i)T} + B^{(i)} \text{Var}[d^{(i+1)}] B^{(i)T}. \quad (10)$$

In particular, (10) implies that

$$\text{Var}[d_{12}^{(i)}] = \nu(u_1^{(i)} + u_2^{(i)}) \quad (11)$$

$$\text{Cov}[d_{12}^{(i)}, d_{13}^{(i)}] = \nu u_1^{(i)} \quad (12)$$

$$\text{Cov}[d_{12}^{(i)}, d_{34}^{(i)}] = 0, \quad (13)$$

where $u_k^{(i)}$ denotes the k th element of $u^{(i)}$. This illustrates several consequences of our assumptions: that the variation in divergence along a tree path is proportional to the time-length of that path (11); that the covariation of divergences along two overlapping paths is proportional to their shared path length (12); and that divergences along non-overlapping paths are uncorrelated (13). Such relationships were first proposed by Chakraborty (1977), and underpin BIONJ.

Although our assumptions (6,7) accommodate a great variety of models, they do not formally allow for evolving rates of evolution (Thorne et al., 1998, Kishino et al., 2001), as this would induce correlations between non-overlapping paths in the tree.

2.4.2 Non-tree-additivity

Our second set of assumptions concerns the relationship between distances and divergences. We begin by considering the observed distances $\hat{d}^{(0)}$ between $m^{(0)}$ currently extant taxa. There are $n^{(0)}$ elements in $\hat{d}^{(0)}$, but the corresponding (rooted)

divergence tree contains only $(2m^{(0)} - 2)$ branches, so for $m^{(0)} > 4$ it will generally be impossible to find a divergence tree whose branches exactly replicate the distances in $\hat{d}^{(0)}$. We refer to this lack of fit to tree-additivity as *noise*, and model it as follows:

$$\mathbb{E}[\hat{d}^{(0)} \mid d^{(0)}] = d^{(0)}, \quad (14)$$

$$\text{Var}[\hat{d}^{(0)} \mid d^{(0)}] = \sigma^2 I_{n^{(0)}}, \quad (15)$$

where I_k denotes the $(k \times k)$ identity matrix. The amount of noise is controlled by variance parameter σ^2 . Setting $\sigma^2 = 0$ implies strict tree-additivity in distances. Thus we assert that observed distances are unbiased, uncorrelated, equivariant estimates of the tree-additive divergences underlying them. We emphasize that this is not an abandonment of previous work expounding the genesis of correlations between distances (Chakraborty, 1977, Hasegawa et al., 1985, Bulmer, 1991), since these relate to the marginal distribution of distances whereas (14,15) relate to their conditional distribution, conditioning on underlying divergences $d^{(0)}$. Equations (14,15) refer only to the relationship between distances and the divergences they measure, not to the relationship between divergences and time intervals, which we have already discussed in Section 2.4.1. In particular, (15) does not imply an underlying star-tree topology. As Chakraborty (1977) and others have shown, correlations between distances arise precisely through their overlapping paths in the divergence tree. Below we show how this plays out within the framework of equations (6,7,14,15). The independent, equivariant, distance errors assumed by NJ are represented here by (15), but the branch-length dependent variances and covariances assumed by BIONJ correspond here to equation (9). There is no conflict in simultaneously assuming both sources of variation, as we do.

At each subsequent agglomerative stage $i > 0$, our algorithm produces a vector of inter-taxon distances $\hat{d}^{(i)}$, estimating underlying divergences $d^{(i)}$. As we show by induction in Section 2.8, assumptions (6,7,14,15), together with the estimation procedures described below, result in the following conditional moments at each stage $i \geq 0$, conditioning on unknown divergences $d^{(i)}$:

$$\mathbb{E}[\hat{d}^{(i)} \mid d^{(i)}] = d^{(i)}, \quad (16)$$

$$\text{Var}[\hat{d}^{(i)} \mid d^{(i)}] = \sigma^2 V^{(i)}, \quad (17)$$

where $V^{(i)}$ is a symmetric, positive-definite $(n^{(i)} \times n^{(i)})$ matrix of a particular algebraic structure. Thus, at each stage, our model implies that distances are unbiased measures of divergence, with *conditional* variances and covariances which do not depend on the lengths of the divergences they measure. At stage $i = 0$, equation (15) asserts that $V^{(0)}$ is the identity matrix, so that distances at stage 0 have a common

variance and are uncorrelated. However, these simple variance-covariance properties necessarily disappear at later stages, because divergences estimated at stage $i > 0$ *jointly* depend on imprecisely estimated divergences from previous stages, thereby altering variances of divergence estimates and inducing correlations between them, ultimately affecting subsequent divergence estimates. NJ, BIONJ and FASTME take no account of such cumulative effects of estimation; WEIGHBOR takes account of estimation effects on variances, but ignores effects on covariances; whilst the full, computationally intensive, version of MVR (Gascuel, 2000) correctly accounts for such effects in its minimum-variance criterion. Our methodology likewise takes full account of these estimation effects to preserve statistical efficiency, but without sacrificing computational efficiency.

2.4.3 Combining distortion and noise

It follows from equations (6,7,16,17), marginalizing over $d^{(i)}$ while retaining the conditioning on $d^{(i+1)}$, that distances have the following conditional variance property:

$$\text{Var}[\hat{d}^{(i)} \mid d^{(i+1)}] = \sigma^2 W^{(i)}, \quad (18)$$

where

$$W^{(i)} = V^{(i)} + \theta A^{(i)} S^{(i)} u_D^{(i)} S^{(i)T} A^{(i)T}, \quad (19)$$

where $\theta = v/\sigma^2$. Notice the differences between equations (17) and (18). They differ on the left because in (17) we condition on $d^{(i)}$, while in (18) we condition only on $d^{(i+1)}$, a smaller condition as it involves less of the tree. They differ on the right because the marginalization has introduced an extra variance component. However, we emphasize that these are not alternative models, they are different aspects of the same model.

Equation (19) succinctly demonstrates that our model contains both noise (controlled by σ^2) and distortion (controlled by v). The first term in (19) represents the noise in $\hat{d}^{(i)}$ as a measurement of $d^{(i)}$; and the second term represents the effects of $a_1^{(i)}$ and $a_2^{(i)}$ being independent evolutionary distortions of the same underlying time-interval $u^{(i)}$.

2.5 Regression model

From (4,5,6,16), we may write

$$\hat{d}^{(i)} = A^{(i)} S^{(i)} a^{(i)} + B^{(i)} d^{(i+1)} + \varepsilon^{(i)} \quad (20)$$

$$a^{(i)} = \mathbf{1}_2 t_1^{(i+1)} - S^{(i)T} t^{(i)} + \epsilon'^{(i)}, \quad (21)$$

where the error terms $\epsilon^{(i)}$ and $\epsilon'^{(i)}$ accommodate departures of $\hat{d}^{(i)}$ and $a^{(i)}$ from their means. Equations (20,21) correspond to a random-effects regression model where $\hat{d}^{(i)}$ is regressed on $A^{(i)}$ and $B^{(i)}$ with regression parameters $a^{(i)}$ and $d^{(i+1)}$ and the random effects $a^{(i)}$ are regressed on $\mathbf{1}_2$ with parameter $t_1^{(i+1)}$ and a fixed offset $-S^{(i)T} t^{(i)}$. The error terms $\epsilon^{(i)}$ and $\epsilon'^{(i)}$ have zero mean and are uncorrelated with variance-covariance matrices

$$\text{Var}[\epsilon^{(i)} | d^{(i)}] = \sigma^2 V^{(i)}, \quad (22)$$

$$\text{Var}[\epsilon'^{(i)} | d^{(i+1)}] = v u_D^{(i)}, \quad (23)$$

from (7,17). For now, we assume that variance parameters σ^2 and v are known; we will relax this assumption in Section 2.12.

2.6 Generalized Least Squares

The regression parameters in (20,21) may be estimated by GLS. (For a general introduction to GLS, see Mardia, Kent, and Bibby (1979); for its application in non-agglomerative phylogenetics, see Hasegawa et al. (1985), Bulmer (1991), Susko (2003)). The GLS objective function at stage i , based on (20–23), is the Mahalanobis distance:

$$M_C^{(i)} = \sigma^{-2} \epsilon^{(i)T} V^{(i)-1} \epsilon^{(i)} + v^{-1} \epsilon'^{(i)T} u_D^{(i)-1} \epsilon'^{(i)}, \quad (24)$$

where, from (20,21),

$$\epsilon^{(i)} = \hat{d}^{(i)} - A^{(i)} S^{(i)} a^{(i)} - B^{(i)} d^{(i+1)}, \quad (25)$$

$$\epsilon'^{(i)} = a^{(i)} - \mathbf{1}_2 t_1^{(i+1)} + S^{(i)T} t^{(i)}. \quad (26)$$

We refer to (24) as the *conditional* objective function at stage i . Parameter estimates, $\hat{t}_1^{(i+1)}$, $\hat{a}^{(i)}$, $\hat{d}^{(i+1)}$, are obtained by minimizing $M_C^{(i)}$ with respect to $t_1^{(i+1)}$, $a^{(i)}$, $d^{(i+1)}$, for given σ^2 , v and $t^{(i)}$. Note that $t^{(i)}$ is known from the previous stage; that is, we set $t^{(i)} = \hat{t}^{(i)}$ if $i > 0$ or $t^{(i)} = 0$, otherwise.

It is instructive to consider an alternative approach to estimation, derived by substituting equation (21) into (20) to give:

$$\hat{d}^{(i)} = A^{(i)} S^{(i)} \mathbf{1}_2 t_1^{(i+1)} + B^{(i)} d^{(i+1)} - A^{(i)} S^{(i)} S^{(i)T} t^{(i)} + \epsilon^{\star(i)}, \quad (27)$$

where

$$\epsilon^{\star(i)} = \epsilon^{(i)} + A^{(i)} S^{(i)} \epsilon'^{(i)}. \quad (28)$$

Equation (27) represents a marginal regression model, in which $\hat{d}^{(i)}$ is regressed on $A^{(i)}S^{(i)}\mathbf{1}_2$ and $B^{(i)}$, with regression parameters $t_1^{(i+1)}$ and $d^{(i+1)}$, and a fixed offset $-A^{(i)}S^{(i)}S^{(i)T}t^{(i)}$. From (22,23,25,26), the error term $\epsilon^{*(i)}$ has zero mean and variance

$$\text{Var}[\epsilon^{*(i)} \mid d^{(i+1)}] = \sigma^2 W^{(i)}, \quad (29)$$

where $W^{(i)}$ is given by (19). From (27–29), we can construct a marginal GLS objective function:

$$M_M^{(i)} = \sigma^{-2} \epsilon^{*(i)T} W^{(i)-1} \epsilon^{*(i)}, \quad (30)$$

and estimate parameters $t_1^{(i+1)}$ and $d^{(i+1)}$ by minimizing $M_M^{(i)}$ w.r.t. $t_1^{(i+1)}$ and $d^{(i+1)}$. It can be shown, using the Sherman–Morrison–Woodbury matrix-inversion identity (see, for example, Golub and Van Loan (1996)), that $M_C^{(i)} \equiv M_M^{(i)}$, and hence the conditional and marginal GLS approaches lead to identical estimates of $t_1^{(i+1)}$ and $d^{(i+1)}$ for given σ^2 , v and $t^{(i)}$.

2.7 Parameter estimates

Minimization of $M_C^{(i)}$ w.r.t. $t_1^{(i+1)}$, $a^{(i)}$ and $d^{(i+1)}$ leads to the following GLS estimates:

$$\hat{t}_1^{(i+1)} = (Q_{tt} - Q_{td}Q_{dd}^{-1}Q_{dt})^{-1} (P_t - Q_{td}Q_{dd}^{-1}P_d)y^{(i)}, \quad (31)$$

$$\hat{d}^{(i+1)} = Q_{dd}^{-1}P_d \left(y^{(i)} - A^{(i)}S^{(i)}\mathbf{1}_2\hat{t}_1^{(i+1)} \right), \quad (32)$$

$$\hat{u}^{(i)} = \mathbf{1}_2\hat{t}_1^{(i+1)} - S^{(i)T}t^{(i)}, \quad (33)$$

$$\hat{a}^{(i)} = \left(\theta Q_{aa} + u_D^{(i)-1} \right)^{-1} \left(\theta P_a \{ \hat{d}^{(i)} - B^{(i)}\hat{d}^{(i+1)} \} + u_D^{(i)-1}\hat{u}^{(i)} \right), \quad (34)$$

where

$$y^{(i)} = \hat{d}^{(i)} + A^{(i)}S^{(i)}S^{(i)T}t^{(i)} \quad (35)$$

and

$$\begin{aligned} Q_{tt} &= \mathbf{1}_2^T S^{(i)T} A^{(i)T} W^{(i)-1} A^{(i)} S^{(i)} \mathbf{1}_2, & Q_{td} &= \mathbf{1}_2^T S^{(i)T} A^{(i)T} W^{(i)-1} B^{(i)}, \\ Q_{dt} &= B^{(i)T} W^{(i)-1} A^{(i)} S^{(i)} \mathbf{1}_2, & Q_{dd} &= B^{(i)T} W^{(i)-1} B^{(i)}, \\ P_t &= \mathbf{1}_2^T S^{(i)T} A^{(i)T} W^{(i)-1}, & P_d &= B^{(i)T} W^{(i)-1}, \\ Q_{aa} &= S^{(i)T} A^{(i)T} V^{(i)-1} A^{(i)} S^{(i)}, & P_a &= S^{(i)T} A^{(i)T} V^{(i)-1}, \end{aligned} \quad (36)$$

suppressing index $^{(i)}$ in the notation $Q_{tt}, \dots, P_a$.

Now $\hat{d}^{(i+1)}$ in (32) estimates divergence vector $d^{(i+1)}$ imprecisely, so in the terminology of Section 2.1 it is a vector of *distances*. Indeed, $\hat{d}^{(i+1)}$ is the distance vector which will be input to the next stage $i + 1$ of our algorithm. Equation (32) admits the possibility that elements of this distance vector may be negative. Although numerical methods for non-negative GLS are available at greater computational cost, they frustrate the algebraic approach which is particularly valuable at subsequent stages of our analysis (see Sections 2.10 and 2.12). We deal with negative branch lengths in a minimally perturbing post-processing stage, described in Section 2.11.

The following variances of the parameter estimates may be derived from (24) and (30), according to standard GLS theory:

$$\text{Var}[\hat{t}_1^{(i+1)} \mid d^{(i+1)}] = \sigma^2 (Q_{tt} - Q_{td} Q_{dd}^{-1} Q_{dt})^{-1}, \quad (37)$$

$$\text{Var}[\hat{d}^{(i+1)} \mid d^{(i+1)}] = \sigma^2 (Q_{dd} - Q_{dt} Q_{tt}^{-1} Q_{td})^{-1}, \quad (38)$$

$$\text{Var}[\hat{a}^{(i)} \mid \hat{t}_1^{(i+1)}, \hat{d}^{(i+1)}, d^{(i+1)}] = \sigma^2 \theta (Q_{aa} \theta + u_D^{(i)})^{-1}. \quad (39)$$

It can be shown that the matrices to be inverted in equations (31–39) are all positive-definite, provided $V^{(i)}$ is positive definite and no elements of $u_D^{(i)}$ are negative. Noting (40), it follows by induction that $V^{(i)}$ is positive-definite for all i .

2.8 Progression

For the next stage $i + 1$ of the agglomeration, we need to know the matrix $V^{(i+1)}$ assumed in (17). This is provided by equation (38), giving:

$$V^{(i+1)} = (Q_{dd} - Q_{dt} Q_{tt}^{-1} Q_{td})^{-1}. \quad (40)$$

Thus $V^{(i+1)}$ depends on $V^{(i)}$ through equations (19,36,40). These equations suggest that, from one stage to the next, the variance-covariance structure of distances becomes increasingly intricate, as a result of regression parameter estimates at each stage being recursively built upon correlated and imprecise regression parameter estimates from the previous stage. This happens even though in (15) we assumed conditionally independent distances initially. Despite this loss of independence, we show below that $V^{(i+1)}$ remains within a relatively simple class of structured covariance matrices. This is a central result of this paper.

We assumed above that equations (16,17) hold for given i , and consequently we have from the unbiasedness of GLS that $\mathbb{E}[\hat{d}^{(i+1)} \mid d^{(i+1)}] = d^{(i+1)}$, and from (38,40) that $\text{Var}[\hat{d}^{(i+1)} \mid d^{(i+1)}] = \sigma^2 V^{(i+1)}$. Therefore, (16,17) hold with i replaced by $i + 1$. From (14,15), we see that (16,17) hold for $i = 0$. Hence, by induction, (16,17) hold for all $i \geq 0$.

2.9 Dependence of variance on mean

Estimates $\hat{t}_1^{(i+1)}$ and $\hat{d}^{(i+1)}$ in (31,32) assume $W^{(i)}$ is known. However, equations (6,7) specify that the variance of $a^{(i)}$ depends on its mean, and hence $W^{(i)}$ depends on unknown regression parameter $t_1^{(i+1)}$ through (5). To deal with this, we propose an iterative solution, assuming variance parameters σ^2 and v are known. Estimation of σ^2 and v is considered in Section 2.12. At each stage i , beginning with an initial guess $\hat{t}_1^{(i+1)}$ for $t_1^{(i+1)}$, we iterate the calculations for $u^{(i)}, W^{(i)}, y^{(i)}, \hat{t}_1^{(i+1)}$ and $\hat{d}^{(i+1)}$ in (5,19,31–36) until convergence.

2.10 Computation

We have seen that we may proceed from one stage i to the next $i+1$ by computing the distance vector $\hat{d}^{(i+1)}$ from (32) and its variance structure $V^{(i+1)}$ from (40). However, both (32) and (40) appear computationally formidable, requiring inversions of matrices of dimension $n^{(i+1)} \times n^{(i+1)}$. Such calculations, if tackled directly, would have time-complexity of order $\mathcal{O}(m^{(i+1)6})$. Fortunately, we may exploit the algebraic structure of these matrices to perform each inversion in only $\mathcal{O}(m^{(i+1)})$.

As we describe below, these more efficient calculations involve, at stage i , two nonnegative vectors $w^{(i)}$ and $s^{(i)}$, each of length $m^{(i)}$, defined recursively as follows:

$$w^{(i+1)} = \left[w_1^{(i)} + w_2^{(i)}, w_3^{(i)}, w_4^{(i)}, \dots, w_{m^{(i)}}^{(i)} \right]^T, \quad (41)$$

$$s^{(i+1)} = \left[s_1^{(i+1)}, s_3^{(i)}, s_4^{(i)}, \dots, s_{m^{(i)}}^{(i)} \right]^T, \quad (42)$$

where $s_1^{(i+1)}$ is a scalar depending only on $s_1^{(i)}, w_1^{(i)}, u_1^{(i)}, s_2^{(i)}, w_2^{(i)}, u_2^{(i)}$ and $m^{(0)}$ (see equation (B5) in Appendix B). At the initial stage, we set $w^{(0)} = \mathbf{1}_{m^{(0)}}$ and $s^{(0)} = \mathbf{0}_{m^{(0)}}$, the zero vector. Recursion (41) implies that taxon weight $w_j^{(i)}$ is the number of leaves of the tree (*i.e.* extant taxa) descended from taxon $j^{(i)}$. In particular, it follows that

$$\sum_{j=1}^{m^{(i)}} w_j^{(i)} = \sum_{j=1}^{m^{(0)}} w_j^{(0)} = m^{(0)}.$$

It is a remarkable fact, and a central result of this paper, that the complicated recursion for $V^{(i)}$ through (19,36,40) reduces at each stage i to the following simple form (see Appendix B):

$$V^{(i)} = p_D^{(i)-1} + A^{(i)} s_D^{(i)} A^{(i)T}, \quad (43)$$

where $s_D^{(i)}$ diagonalizes $s^{(i)}$ and $p_D^{(i)}$ diagonalizes an $n^{(i)}$ -vector $p^{(i)}$ whose elements are products of pairs of taxon weights in standard order (2), thus:

$$p^{(i)} = \begin{bmatrix} w_1^{(i)} w_2^{(i)}, w_1^{(i)} w_3^{(i)}, \dots, w_1^{(i)} w_{m^{(i)}}^{(i)}, \\ w_2^{(i)} w_3^{(i)}, \dots, w_2^{(i)} w_{m^{(i)}}^{(i)}, \dots, w_{m^{(i)}-1}^{(i)} w_{m^{(i)}}^{(i)} \end{bmatrix}^T. \quad (44)$$

Interpreting (43), the variance in $\hat{d}^{(i)}$ can be thought of as arising from two uncorrelated sources of variation. The first source is a collection of $n^{(i)}$ uncorrelated random variables, one for each taxon pair, with variances inversely proportional to taxon weights, reflecting the pooling of distance information with each agglomeration. Thus we refer to $p^{(i)}$ as a vector of *pooling* coefficients. The second source is a much smaller set of $m^{(i)}$ uncorrelated random variables, one for each taxon. The variance of the j th random variable in this second set is proportional to $s_j^{(i)}$ and arises through agglomerations leading to taxon $j^{(i)}$. This is not purely a consequence of distortion; positive $s_j^{(i)}$ arise even when $v = 0$. This source impacts on $V^{(i)}$ only where taxon-pairs share a taxon. Thus we refer to $s^{(i)}$ as a vector of *sharing* coefficients. A particular consequence of (43) is that inter-taxon distances are uncorrelated where they have no taxon in common. Thus, for example, $\hat{d}_{1,2}^{(i)}$ and $\hat{d}_{3,4}^{(i)}$ are uncorrelated whereas $\hat{d}_{1,2}^{(i)}$ and $\hat{d}_{2,3}^{(i)}$ are not.

Equation (43) leads to substantial computational simplifications, in particular in computing $\hat{d}^{(i+1)}$ in (32). We can show that distances at stage i which do not involve taxa $1^{(i)}$ or $2^{(i)}$ (*i.e.* most of $\hat{d}^{(i)}$) are passed unaltered to the next stage. That is, $\hat{d}_{j-1,k-1}^{(i+1)} = \hat{d}_{j,k}^{(i)}$ for $3 \leq j < k \leq m^{(i)}$. This intuitive result is critical in terms of reducing the time-complexity of our algorithm.

2.11 Building the tree

We have seen above how we may proceed from each stage i to the next, $i + 1$, by agglomerating taxa $1^{(i)}$ and $2^{(i)}$. Of course, by appropriate taxon relabeling, we can agglomerate any two taxa $j^{(i)}$ and $k^{(i)}$ using this methodology. We now turn our attention to consider which pair of taxa to agglomerate.

We begin at stage $i = 0$ with observed distances $\hat{d}^{(0)}$, weights $w^{(0)} = \mathbf{1}_{m^{(0)}}$ and sharing-coefficients $s^{(0)} = \mathbf{0}_{m^{(0)}}$. At each generic stage $i \geq 0$, we agglomerate that pair of taxa which would minimize the estimated tree-length, according to the neighbor-joining principle of Saitou and Nei (1987). That is, we choose $(j^{(i)}, k^{(i)})$ to minimise:

$$L_{jk}^{(i)} = (m^{(i)} - 2)d_{jk}^{(i)} - \sum_{h=1}^{m^{(i)}} (d_{jh}^{(i)} + d_{kh}^{(i)}), \quad (45)$$

which is the simplified NJ criterion of Studier and Keppler (1988). We then add the LCA of taxa $j^{(i)}$ and $k^{(i)}$ to our tree reconstruction, with edges of length $\hat{u}^{(i)}$ in the time tree and $\hat{a}^{(i)}$ in the divergence tree computed from the agglomeration of $j^{(i)}$ and $k^{(i)}$, using the formulae in the foregoing sections (with appropriate taxon relabeling). We then proceed to the next stage, armed only with $\hat{d}^{(i+1)}$, $w^{(i+1)}$ and $s^{(i+1)}$.

Thus, from leaves to root, the complete time tree is built entirely from edges of lengths $\hat{u}^{(i)}$, and the complete divergence tree is built entirely from edges of lengths $\hat{a}^{(i)}$. After completing both trees, any negative branch length $u < 0$ in the time tree is removed by setting it to zero and simultaneously adding $-u$ to the sister branch and $+u$ to the parent branch (if any). These local adjustments, which are applied recursively from tips to root, have the property of preserving the ultrametric property of the time tree. The same form of adjustment is applied to the divergence tree.

2.12 Estimating variance parameters

In general, variance parameters σ^2 and $\theta = v/\sigma^2$ will be unknown. Here we propose an iterative scheme for their estimation, starting with initial guesses $\hat{\sigma}_0^2, \hat{\theta}_0$. Within each iteration $h = 1, 2, \dots$, using current estimates $\hat{\sigma}_{h-1}^2, \hat{\theta}_{h-1}$, we proceed through all agglomerative stages $i = 0, 1, \dots, m^{(0)} - 2$ to recompute the entire time and divergence trees and their associated quantities. These results are then substituted into estimating equations, derived in Appendix A, which may be solved iteratively for σ^2 and θ , giving updates $\hat{\sigma}_h^2, \hat{\theta}_h$. This process of recomputing the tree and updating σ^2 and θ is continued until convergence.

2.13 Putting it together

Putting together the methods of the foregoing sections, the complete algorithm is as follows.

1. Initialize $\hat{\sigma}_0^2, \hat{\theta}_0$, and set $h = 0$.
2. Set $\sigma^2 = \hat{\sigma}_h^2$ and $v = \hat{\sigma}_h^2 \hat{\theta}_h$.
3. For each stage $i = 0, 1, \dots, m^{(0)} - 2$,
 - (a) for each possible join $(j^{(i)}, k^{(i)})$, calculate $\hat{\beta}^{(i,j,k)}$ and $\hat{W}^{(i,j,k)}$ via the iterative method described in Section 2.9;
 - (b) choose the join $(j^{(i)}, k^{(i)})$ with smallest $L_{jk}^{(i)}$ (45), as described in Section 2.11, and calculate $\hat{t}_1^{(i+1)}, \hat{d}^{(i+1)}, w^{(i+1)}$ and $s^{(i+1)}$ accordingly;

4. Calculate $\hat{\sigma}_h^2$, $\hat{\theta}_h$, using the estimating equations (A7,A8) derived in Appendix A.
5. If $\hat{\sigma}_h^2 \not\approx \hat{\sigma}_{h-1}^2$ or $\hat{\theta}_h \not\approx \hat{\theta}_{h-1}$, increment h and return to Step 2.
6. Any negative branch lengths in the time or divergence trees are dealt with by the adjustment procedure of Section 2.11.

2.14 Consistency

Both NJ and BIONJ are *consistent* in the sense that the estimated topology will be correct if distances converge to divergences (which would happen, for example, if the length of the multiple sequence alignment from which distances are calculated tends to infinity). The same property holds for our methodology, the proof of which we now sketch.

Suppose distances are arbitrarily close to divergences: $\hat{d}^{(0)} \approx d^{(0)}$. If we set $\hat{\theta}_0$ arbitrarily large at Step 1 of the algorithm in Section 2.13, then the second term in the objective function (24) will be negligible. It can then be shown by induction that at each stage i , we minimise (24) by setting the unknown parameters $a^{(i)}$, $d^{(i+1)}$ equal to their true values under the true topology. Hence the true topology will be recovered at Step 3 of the algorithm. Then at Step 4, we will again obtain arbitrarily large $\hat{\theta}_0$, so no changes will occur in the parameter or topology estimates in the next iteration of the algorithm, and Step 5 will confirm convergence.

2.15 Sampling full uncertainty

Although equations (38,39) allow us to evaluate uncertainty in divergence estimates, they do not allow us to assess uncertainty in the estimated tree topology. To better understand the full uncertainty in the phylogenetic reconstruction, we propose the following simple parametric bootstrap procedure.

First, a vector of divergences $d^{(0)}$ between the $m^{(0)}$ extant taxa is read off from the complete estimated divergence tree, as constructed in Section 2.11. Second, a bootstrapped vector of distances $\hat{d}^{(0,\text{boot})}$ is generated by sampling independently each element $\hat{d}_{jk}^{(0,\text{boot})}$ from a distribution with mean $d_{jk}^{(0)}$ and variance $\hat{\sigma}^2$, for consistency with (16,17). A natural choice for this would be a Gamma distribution, which is preferable to a Gaussian, as it produces only positive distances. However, alternative distributions could be explored to evaluate their impact. Third, using $\hat{d}^{(0,\text{boot})}$ instead of $\hat{d}^{(0)}$, a bootstrapped version of the time-tree, divergence-tree, $\hat{\sigma}^2$ and $\hat{v}$, is estimated using the methods described above. Many bootstrapped versions may be generated by repeating steps two and three. The frequency of different tree

topologies, histograms of $\hat{\sigma}^2$ and $\hat{\nu}$, and any other statistical summaries of interest, may then be computed from the collection of bootstrapped versions.

3 Simulations

To demonstrate key features of our approach, we present results obtained from STATTREE on simulated data. Distance matrices were randomly generated according to two different schemes. In the first, distance matrices were obtained by distorting trees with fixed topology, and then adding noise. This will be referred to as ‘direct simulation’ below. In second scheme, referred to as ‘sequence-based simulation’, sequence alignments were randomly generated from some fixed trees and distance matrices were then estimated from the alignments. For both schemes, trees were constructed from the distance matrices and compared with the original underlying tree.

3.1 Direct simulation

1. A base ultrametric time tree was selected together with a choice of parameters σ^2 and ν . Two different topologies for the time tree, each with $m^{(0)} = 64$ leaf taxa, were used. These topologies are illustrated in Figure 3, and represent extremes of the kind of branching behaviour that can arise in rooted bifurcating trees. We will refer to these as the *balanced* and *unbalanced* topologies.
2. The selected time tree was then distorted, as follows, to obtain a divergence tree with inter-leaf divergences $\{d_{jk}^{(0)}\}$. The length of each edge on the divergence tree was assigned a value independently sampled from a Gamma distribution with mean equal to the edge length u on the time tree and variance νu , ensuring consistency with equations (6) and (7).
3. A distance matrix was then obtained from the divergence tree and perturbed according to the process described in Section 2.15: each inter-leaf distance $\hat{d}_{jk}^{(0)}$ was assigned a value independently sampled from a Gamma distribution with mean $d_{jk}^{(0)}$ and variance σ^2 . This process ensures consistency with equations (16) and (17).

Various values of σ^2 and ν were used to simulate distance matrices, and these values are specified as follows. The standard deviation σ is best expressed as a proportion of the mean inter-taxon distance in the time tree. For example, if the mean inter-taxon distance in the time tree was δ we might take $\sigma = 0.2 \times \delta$ and refer to this as a value of 20% for σ . The two basic time trees shown in Figure 3 have

Figure 3: Two topologies for the underlying time trees used in the simulations. The tree on the left is referred to as having a *balanced* topology, while the tree on the right is referred to as *unbalanced*. These represent two extremes of the pattern of branching that can be exhibited in rooted bifurcating trees. All internal branches in the trees are taken to have equal length, and the trees were scaled so that the maximum distance between two taxa was 1.0. For both topologies, trees with 16 and 64 taxa were used, as described in the text.

quite different mean inter-taxon distances, and so by specifying σ as a proportion of the mean we ensure that distance matrices are obtained from the different topologies in a comparable way. Similarly, the parameter v needs to be specified in a way that is consistent between different trees. If time trees with a fixed topology are scaled to have different overall lengths and then distorted to obtain divergence trees, it is necessary to vary v with the scale of the tree in order to produce a collection of divergence trees with a similar visual appearance and level of distortion. This is achieved by fixing a proportion p , and taking

$$v = p^2 \times \delta \quad (46)$$

where δ is the mean inter-taxon distance in the time tree. This ensures that when a branch of length u on a time tree is distorted, the ratio of the standard deviation to the branch length is given by

$$\frac{\sqrt{vu}}{u} = p \times \sqrt{\delta/u}. \quad (47)$$

This is a dimensionless quantity that is proportional to p . Fixing v in this way for different tree topologies, tree scales, and numbers of taxa ensures that the results are comparable. The proportion p is usually expressed as a percentage in what follows, and for brevity we will write ' $v = 20\%$ ' to mean $v = 0.2^2 \times \delta$, for example.

To assess the performance of our algorithm, distance matrices were simulated using the scheme in Steps 1–3 above, and phylogenies were estimated according to the steps specified in Section 2.13, as implemented in our software STAT-TREE. Specifically, values of 5%, 10% and 20% for σ and ν were used in the simulations. For each of the 9 corresponding combinations of parameters, 100 distance matrices were generated using the balanced topology on 64 taxa, and another 100 distance matrices were generated using the unbalanced topology on 64 taxa. Tree construction with estimation of σ^2 and ν was then carried out for each of the simulated distance matrices. Means and standard deviations of the estimated parameters across each set of 100 simulations were calculated to assess the accuracy of estimation. In addition, each estimated tree was compared to the true underlying tree via two scores that measure the topological accuracy of the estimated tree. These scores are described in the next paragraph. For each set of parameter values the average scores were calculated for the set of 100 simulations. Trees were also constructed with the BIONJ algorithm and compared to the true tree in the same way. The results are given in Tables 1 and 2.

Two topological scores, the Robinson-Foulds and quartet distance, were used to measure accuracy of estimated trees. The Robinson-Foulds distance (Robinson and Foulds, 1981) is the proportion of bi-partitions that two trees share (each branch in each tree, upon being cut, induces one such bi-partition of the leaves). The Robinson-Foulds distance is very sensitive, in that for certain topologies a single incorrectly positioned taxon can result in a score of zero. For this reason a second score, the quartet distance (Brodal, Faberberg, and C.N.S., 2004), was also used. The quartet distance between two trees is defined as the proportion of subsets of 4 taxa for which the trees induce the same topology.

An immediate observation from Tables 1 and 2 is that σ is estimated with a relatively high degree of accuracy while the estimates for ν show considerable bias and variability. The distortion parameter ν is generally over-estimated, and the bias is greater for the unbalanced topology. This bias seems not to be directly related to the topological accuracy of tree estimates: descending any of the columns in the tables generally shows an approximately constant level of bias while the topological accuracy decreases. As σ and ν increase, the topological accuracy of the estimated trees decreases. This effect is present for both topologies, though the unbalanced topology is generally harder to estimate.

The topological scores $\bar{T}_{RF}$, $\bar{T}_Q$ for trees built with our method are compared, in Tables 1 and 2, with the scores for trees built using BIONJ, based on the same distance matrices. For the unbalanced topology, the Robinson-Foulds score $\bar{T}_{RF}$ can be very low and the quartet score $\bar{T}_Q$ is more useful for making comparisons. Our method produces more accurate trees uniformly, and the contrast between the two methods increases with σ and ν . Even though BIONJ makes no

	Method	Statistic	$\nu = 5\%$	$\nu = 10\%$	$\nu = 20\%$
$\sigma = 5\%$	STATTREE	$\bar{\sigma}$ (s.d.)	5.0 (0.1)	5.0 (0.1)	5.0 (0.1)
		$\bar{\nu}$ (s.d.)	11.1 (2.7)	14.2 (2.2)	22.7 (2.4)
		$\bar{T}_{\text{RF}}, \bar{T}_{\text{Q}}$	99.6, 100	98.9, 100	94.8, 99.6
	BIONJ	$\bar{T}_{\text{RF}}, \bar{T}_{\text{Q}}$	99.6, 100	99.1 , 100	94.7, 99.5
$\sigma = 10\%$	STATTREE	$\bar{\sigma}$ (s.d.)	10.0 (0.2)	10.0 (0.2)	10.0 (0.2)
		$\bar{\nu}$ (s.d.)	12.3 (3.4)	15.5 (3.2)	23.6 (3.0)
		$\bar{T}_{\text{RF}}, \bar{T}_{\text{Q}}$	92.1, 99.8	90.7, 99.7	86.0, 99.1
	BIONJ	$\bar{T}_{\text{RF}}, \bar{T}_{\text{Q}}$	92.1, 99.8	90.8 , 99.6	84.1, 98.2
$\sigma = 20\%$	STATTREE	$\bar{\sigma}$ (s.d.)	19.9 (0.4)	19.9 (0.4)	20.1 (0.4)
		$\bar{\nu}$ (s.d.)	12.7 (2.8)	15.1 (3.3)	24.2 (4.3)
		$\bar{T}_{\text{RF}}, \bar{T}_{\text{Q}}$	76.0, 98.4	72.0, 97.9	65.2, 95.5
	BIONJ	$\bar{T}_{\text{RF}}, \bar{T}_{\text{Q}}$	73.4, 97.4	68.7, 95.5	61.4, 88.2

Table 1: Results from STATTREE of direct simulations for an underlying tree with the balanced topology on 64 taxa, for several true values of parameters σ^2 and ν . The means $\bar{\sigma}$ and $\bar{\nu}$ of the parameter estimates (in units of %) over 100 simulations are given, together with their standard deviations in parentheses. Also given for each case is the average Robinson–Foulds percentage score, $\bar{T}_{\text{RF}}$, and the quartet percentage score $\bar{T}_{\text{Q}}$. For comparison, the average scores are also given for trees estimated from the same distance matrices by BIONJ. In each case, the method with the better score is indicated in bold.

assumptions about the rate of evolution in each branch of the tree, the topological accuracy of its estimates decreases as the level of distortion in the underlying tree increases. WEIGHBOR was not used for these simulations since it requires a parameter that corresponds to the length of the sequence alignment from which the distance matrix was estimated. No such parameter was available for the direct simulations, but WEIGHBOR was incorporated into the sequence-based simulations described below.

	Method	Statistic	$\nu = 5\%$	$\nu = 10\%$	$\nu = 20\%$
$\sigma = 5\%$	STATTREE	$\bar{\sigma}$ (s.d.)	5.1 (0.1)	5.1 (0.1)	5.0 (0.3)
		$\bar{\nu}$ (s.d.)	26.1 (14.3)	29.1 (11.2)	32.1 (9.8)
		$\bar{T}_{\text{RF}}, \bar{T}_{\text{Q}}$	38.1, 89.7	33.2, 87.1	23.2, 79.0
	BIONJ	$\bar{T}_{\text{RF}}, \bar{T}_{\text{Q}}$	36.3, 89.2	31.8, 86.8	22.0, 78.1
$\sigma = 10\%$	STATTREE	$\bar{\sigma}$ (s.d.)	10.3 (0.3)	10.2 (0.2)	10.2 (0.4)
		$\bar{\nu}$ (s.d.)	22.9 (10.9)	24.2 (9.5)	28.1 (9.1)
		$\bar{T}_{\text{RF}}, \bar{T}_{\text{Q}}$	12.7, 77.6	11.9, 75.1	11.2, 68.0
	BIONJ	$\bar{T}_{\text{RF}}, \bar{T}_{\text{Q}}$	12.1, 77.3	11.5, 74.9	10.5, 66.7
$\sigma = 20\%$	STATTREE	$\bar{\sigma}$ (s.d.)	20.4 (0.5)	20.4 (0.5)	20.4 (0.8)
		$\bar{\nu}$ (s.d.)	22.5 (5.5)	26.2 (5.7)	30.5 (7.4)
		$\bar{T}_{\text{RF}}, \bar{T}_{\text{Q}}$	2.0, 57.5	2.0, 55.1	2.9, 48.9
	BIONJ	$\bar{T}_{\text{RF}}, \bar{T}_{\text{Q}}$	1.7, 57.2	2.2 , 53.4	3.3 , 45.2

Table 2: Results from STATTREE of direct simulations for an underlying tree with the unbalanced topology on 64 taxa, for several true values of parameters σ^2 and ν . The means $\bar{\sigma}$ and $\bar{\nu}$ of the parameter estimates (in units of %) over 100 simulations are given, together with their standard deviations in parentheses. Also given for each case is the average Robinson–Foulds percentage score, $\bar{T}_{\text{RF}}$, and the quartet percentage score $\bar{T}_{\text{Q}}$. For comparison, the average scores are also given for trees estimated from the same distance matrices by BIONJ. In each case, the method with the better score is indicated in bold.

3.2 Sequence-based simulations

Our algorithm was also evaluated using distance matrices estimated from sequence alignments, using a scheme very similar to that used to evaluate BIONJ by Gascuel (1997). Distance matrices were randomly generated in the following way.

1. A base tree of divergences was selected. Two different trees were used: one with a balanced topology and the other unbalanced, as illustrated in Figure 4 below. These were obtained by randomly distorting ultrametric balanced and unbalanced trees on 16 taxa with $\nu = 10\%$.

2. DNA sequence alignments were generated from the base tree using the program SEQGEN (Rambaut and Grassly, 1997) according to the scheme specified by Gascuel (1997). Kimura's two-parameter model of substitution (Kimura, 1980) was used with a transition/transversion ratio of 2.0. Standard parameter values were used in SEQGEN to specify rate variation between sites. Sequence alignments with 300 and 600 sites were simulated. In addition, three different scales for the underlying base tree were used: low, medium, high substitution rates with (respectively) a maximum of 0.1, 0.5 and 1.0 substitutions per site.
3. Distance matrices were estimated from each alignment using the program DNADIST (part of Phylip), using the same parameter values as SEQGEN, although the transition/transversion ratio was estimated each time.

One hundred distance matrices were generated in this way for each of the two base trees, three scale factors, and two sequence lengths, totaling 12 different simulation conditions. Trees were constructed for each distance matrix using our algorithm (as specified in Section 2.13), BIONJ, and WEIGHBOR. The Robinson–Foulds score and quartet score were calculated to evaluate the topological accuracy of the estimated trees. The accuracy of estimated branch lengths was measured using the following statistic:

$$B = \sum_{j,k} \frac{(\hat{d}_{jk} - d_{jk})^2}{d_{jk}} \quad (48)$$

where $\hat{d}_{jk}$ is the estimated divergence between taxa j, k and d_{jk} is the true divergence. Table 3 shows the results.

Table 3 shows that all three algorithms are broadly comparable and generally achieve a high degree of topological accuracy. The best topological scores and branch-length scores occur for the longer sequences and the balanced trees, where BIONJ performs slightly better than STATTREE. The more difficult scenarios, where the sequences are short and the rate of substitution is low or the topology unbalanced, tend to favor STATTREE. Taking these results together with those of Section 3.1, it appears that STATTREE offers some advantages in harder inference problems where the signal-to-noise ratio in the distance matrix is low.

4 Application to HIV evolution

Rapidly evolving viral species, such as influenza and HIV, have become important laboratory tools for studying evolution in real time, providing a test-bed for evolutionary theories and inference methodologies. The high genetic diversity of HIV

Figure 4: Balanced and unbalanced trees used for sequence-based simulations.

results from about 10^5 mutations per day per infected individual. Longitudinal clinical studies can provide information on within-patient HIV evolution under varying treatment regimes, with potential to reveal genetic bottlenecks and resurgence of rare wild-type variants post-therapy. Here we reanalyze data from a recent study by Salazar-Gonzalez, Salazar, Keele, Learn, Giorgi, Li, et al. (2009), in which a total of 121 full-length HIV-1 genome sequences were obtained from samples taken longitudinally in 12 infected individuals. The authors reported a rooted, multi-furcating, maximum-likelihood phylogeny of these sequences, based on a bespoke mathematical model of HIV sequence evolution (Keele, Giorgi, Salazar-Gonzalez, Decker, Pham, Salazar, et al., 2008).

It is of interest to determine not only the tree describing the evolution of these sampled HIV genomes, but also the physical times of their diversification. For these tasks our methodology, producing both a divergence tree and its underlying time tree, is directly relevant. Our reanalysis of these data was based on a distance matrix for all 121 genomes, constructed using program DNADIST of Phylip assuming Kimura's two-parameter model of base substitution (Kimura, 1980). The results from STATTREE are shown in Figure 5. Similar results (not shown) were obtained with alternative substitution models. Within-subject branch lengths were all, with one exception, very short, and so for concision are not shown. The one exception is the subject identified in the figure as *ZM.247*, for whom two viral clades, *ZM.247a* and *ZM.247b*, are shown.

As in the original study (Salazar-Gonzalez et al., 2009), the divergence tree in Figure 5b demonstrates clear between-subject differences in clonal lineages, the generally short internal branches suggesting only small differences between their common ancestors. The major exception concerns the three Zambian subjects,

Scale	Method	Length 300		Length 600	
		Balanced $\bar{T}_{RF}, \bar{T}_Q, \bar{B}$	Unbalanced $\bar{T}_{RF}, \bar{T}_Q, \bar{B}$	Balanced $\bar{T}_{RF}, \bar{T}_Q, \bar{B}$	Unbalanced $\bar{T}_{RF}, \bar{T}_Q, \bar{B}$
Low	STATTREE	94.8, 94.1 , 76.3	62.8 , 77.3 , 61.8	98.8, 98.5, 77.2	78.1, 88.0, 61.5
	BIONJ	95.3 , 93.7, 77.8	62.8 , 76.9, 64.3	99.2 , 98.6 , 77.7	78.8 , 88.5 , 64.2
	WEIGHBOR	95.2, 93.7, 77.8	59.3, 74.0, 64.3	98.8, 97.9, 77.7	76.9, 86.8, 64.2
Medium	STATTREE	99.2, 99.2, 21.0	73.2, 84.2, 14.9	100, 100, 21.5	85.3, 92.0, 14.5
	BIONJ	99.5 , 99.3 , 22.3	73.3 , 84.7 , 18.1	100, 100, 22.6	85.7 , 92.1 , 18.5
	WEIGHBOR	99.2, 98.9, 22.4	70.7, 82.4, 18.1	100, 100, 22.6	85.2, 91.8, 18.5
High	STATTREE	98.6, 97.8, 2.8	72.3 , 83.7 , 22.9	100, 99.9, 1.8	82.8, 90.8, 15.5
	BIONJ	98.8 , 97.9 , 2.1	71.8, 83.1, 2.2	100, 99.9, 1.2	83.6 , 91.4 , 1.1
	WEIGHBOR	98.6, 97.5, 2.1	70.2, 81.7, 2.3	100, 99.9, 1.2	82.5, 90.7, 1.1

Table 3: Results of sequence-based simulations generated under balanced and unbalanced tree topologies. For each tree, 100 distance matrices were generated for each combination of two different sequence lengths (300 and 600 bp) and three different scales (corresponding to maxima of 0.1, 0.5 and 1.0 substitutions per site). The table reports the average of the Robinson–Foulds percentage score $\bar{T}_{RF}$, quartet percentage score $\bar{T}_Q$, and branch length score $\bar{B}$, for each of STATTREE, BIONJ and WEIGHBOR. In each case, the method with the better score is indicated in bold.

whose viral strains are quite distinct from those of the USA subjects. The clear separation between the two viral subclades of subject *ZM.247* is interpreted by Salazar-Gonzalez et al. (2009) as evidence of simultaneous infection from an individual carrying both strains. The major difference between our analysis and that of Salazar-Gonzalez et al. (2009) is that in the latter the tree is rooted between the Zambian patients and the rest.

The corresponding time tree in Figure 5a estimates the times of diversification of these viral lineages. This time tree suggests relatively long periods of shared HIV ancestry between some of the USA patients: an insight not available from the divergence tree. Moreover, the relatively short estimated time since divergence of the two subclades of subject *ZM.247* might call into question the authors' explanation for their existence. These observations underline the potential utility of estimating time-trees.

Figure 5: (a) Time and (b) divergence trees of 121 HIV-1 genomes sampled from 12 subjects studied by Salazar-Gonzalez et al. (2009), estimated by STATTREE. Within-subject viral clades are shown as tips. Tip-label prefixes denote subjects' geographical location (AL = Alabama; NC = North Carolina; NY = New York; ZM = Zambia).

5 Discussion

We have developed a variance-components model for phylogenetic reconstruction from distance matrices, incorporating several novel features. Two components of variance are distinguished and incorporated: *noise* (as modeled by NJ) reflecting the lack-of-fit of distances to an underlying divergence tree, and *distortion* (as modeled by BIONJ) reflecting the departure of this divergence tree from its underlying ultrametric time tree. Unlike previous methods, our approach exploits the coherence of both trees, simultaneously estimating them through a computationally efficient procedure which recursively descends through the phylogeny. Variances and covariances of distances and divergences are propagated from one agglomerative stage to the next, and taken into account in building the tree.

Focussing only on means, variances and covariances of distances, our methodology avoids the need for full parametric assumptions about evolutionary and other processes involved. The advantage is that inferences do not depend on the validity of substitution models, thus admitting the possibility of straightforwardly incorporating non-standard sources of phylogenetic information other than sequence data.

As shown in Section 3, even small phylogenies contain enough data to estimate the noise component σ accurately, but information on the distortion compo-

ment v is weak and estimates tend to be positively biased. Our simulation results show that STATTREE performs comparably with WEIGHBOR and BIONJ, STATTREE slightly outperforming BIONJ when distance matrices have a low signal-to-noise ratio such as when sequences and branch lengths are short and the topology unbalanced.

There is scope for further development of the methodology, in particular for a more robust estimation of v . Generalizations of the methodology might allow for non-binary trees and for different amounts of noise in different parts of the tree. A more substantial advance would be required to accommodate evolution in the rate of evolution (Thorne et al., 1998, Kishino et al., 2001), which induces correlations in divergence between non-overlapping branches.

Software and applications

We have implemented the methodology in STATTREE, which may be freely downloaded from <http://www.mas.ncl.ac.uk/~ntmwn/stattree>. Illustrative applications of the methodology to bacterial, viral and mitochondrial datasets can be found at <http://www.cl.cam.ac.uk/~pl219/GilksNyeLioSAGMB.html>.

Appendix A Estimating equations for σ^2 and v

Our estimating equations for σ^2 and v are derived from a global objective function M_G comprising the marginal objective function $M_M^{(i)}$ in (30) summed over all stages i :

$$M_G(\sigma^2, \theta) = \sigma^{-2} \sum_{i=0}^{m^{(0)}-2} \hat{\epsilon}^{*(i)T} W^{(i)-1} \hat{\epsilon}^{*(i)}, \quad (\text{A1})$$

where the dependence on θ is through $W^{(i)}$ given in (19). The estimated marginal residuals $\hat{\epsilon}^{*(i)}$ in (A1) are given by

$$\hat{\epsilon}^{*(i)} = \hat{d}^{(i)} - A^{(i)} S^{(i)} \left(\mathbf{1}_2 \hat{t}_1^{(i+1)} - S^{(i)T} \hat{t}^{(i)} \right) - B^{(i)} \hat{d}^{(i+1)}, \quad (\text{A2})$$

from (27), for the optimal taxon-pair agglomeration $(j^{(i)}, k^{(i)})$, as described in Section 2.11.

We construct our estimating equations by setting partial derivatives of (A1) equal to their expectations (see, for example, Crowder (2001)). These partial derivatives are:

$$\frac{\partial M_G}{\partial \sigma^2} = -\frac{1}{\sigma^4} \sum_i \text{tr} \left[\hat{\boldsymbol{\varepsilon}}^{*(i)} \hat{\boldsymbol{\varepsilon}}^{*(i)T} W^{(i)-1} \right], \quad (\text{A3})$$

$$\frac{\partial M_G}{\partial \theta} = -\frac{1}{\sigma^2} \sum_i \text{tr} \left[\hat{\boldsymbol{\varepsilon}}^{*(i)} \hat{\boldsymbol{\varepsilon}}^{*(i)T} W^{(i)-1} A^{(i)} S^{(i)} u_D^{(i)} S^{(i)T} A^{(i)T} W^{(i)-1} \right], \quad (\text{A4})$$

using (19), where tr denotes trace and the residuals $\{\hat{\boldsymbol{\varepsilon}}^{*(i)}\}$ are considered fixed. Now, it can be shown that $\mathbb{E} \left[\hat{\boldsymbol{\varepsilon}}^{*(i)} \hat{\boldsymbol{\varepsilon}}^{*(i)T} \right] = \sigma^2 P^{(i)} W^{(i)}$ where $P^{(i)}$ is the matrix projecting into the residual space of the marginal regression model (27):

$$P^{(i)} = I_{n^{(i)}} - X^{(i)} \left(X^{(i)T} W^{(i)-1} X^{(i)} \right)^{-1} X^{(i)T} W^{(i)-1},$$

where $X^{(i)} = [A^{(i)} S^{(i)} \mathbf{1}_2, B^{(i)}]$. Therefore, taking expectations in (A3,A4),

$$\begin{aligned} \mathbb{E} \left[\frac{\partial M_G}{\partial \sigma^2} \right] &= -\frac{1}{\sigma^2} \sum_i \text{tr} [P^{(i)}] \\ &= -\frac{1}{\sigma^2} \sum_i n^{(i)} - (1 + n^{(i+1)}) = -\frac{1}{\sigma^2} (n^{(0)} - m^{(0)} + 1), \end{aligned} \quad (\text{A5})$$

$$\mathbb{E} \left[\frac{\partial M_G}{\partial \theta} \right] = -\sum_i \text{tr} \left[P^{(i)} A^{(i)} S^{(i)} u_D^{(i)} S^{(i)T} A^{(i)T} W^{(i)-1} \right]. \quad (\text{A6})$$

Equating observed (A3,A4) and expected (A5,A6) partial derivatives, we obtain the following estimating equations:

$$\sum_i \hat{\boldsymbol{\varepsilon}}^{*(i)T} W^{(i)-1} \hat{\boldsymbol{\varepsilon}}^{*(i)} = \sigma^2 (n^{(0)} - m^{(0)} + 1), \quad (\text{A7})$$

$$\begin{aligned} \sum_i \hat{\boldsymbol{\varepsilon}}^{*(i)T} W^{(i)-1} A^{(i)} S^{(i)} u_D^{(i)} S^{(i)T} A^{(i)T} W^{(i)-1} \hat{\boldsymbol{\varepsilon}}^{*(i)} \\ = \sigma^2 \sum_i \text{tr} \left[P^{(i)} A^{(i)} S^{(i)} u_D^{(i)} S^{(i)T} A^{(i)T} W^{(i)-1} \right], \end{aligned} \quad (\text{A8})$$

which may be solved iteratively for σ^2 and θ .

Appendix B Efficient computation

Using equations (31–40,A7,A8) directly would be computationally prohibitive. However, the algebraic structure of the vectors and matrices involved admit substantial computational savings, as we now describe.

The key result, in Lemma B.1 below, is an efficient formula for computing Q_{dd}^{-1} in (31,32), thus avoiding the double inversion of large matrices involved in $(B^{(i)T}W^{(i)-1}B^{(i)})^{-1}$. Lemma B.3 then shows that the $V^{(i)}$ matrices stay within the class defined by (43). Lemmas B.1–B.3 open the way for efficient computation of equations (31–40,A7,A8).

We first introduce the following scalars, notationally suppressing their dependence on i :

$$\begin{aligned}\pi_j &= w_j^{(i)}(m^{(0)} - 2w_j^{(i)}), \quad j = 1, \dots, m^{(i)} \\ \tilde{\pi} &= (w_1^{(i)} + w_2^{(i)}) \left(m^{(0)} - 2(w_1^{(i)} + w_2^{(i)}) \right) \\ \tau_j &= \begin{cases} s_j^{(i)} + \theta u_j^{(i)}, & j = 1, 2 \\ s_j^{(i)}, & j = 3, \dots, m^{(i)} \end{cases} \\ r_j &= (1 + \tau_j \pi_j)^{-1}, \quad j = 1, \dots, m^{(i)} \\ h &= \frac{w_1^{(i)2} r_1 \tau_1 + w_2^{(i)2} r_2 \tau_2}{(w_1^{(i)} + w_2^{(i)})^2} \\ g &= \frac{w_1^{(i)} r_1 + w_2^{(i)} r_2}{w_1^{(i)} + w_2^{(i)}} \\ f &= \frac{g}{1 - h \tilde{\pi}} \\ b &= \left(1 + \sum_{j=1}^{m^{(i)}} w_j^{(i)2} \tau_j r_j \right)^{-1} \\ \alpha_0 &= - \frac{w_1^{(i)} r_1 \tau_1 + w_2^{(i)} r_2 \tau_2}{(w_1^{(i)} + w_2^{(i)})(1 - h \tilde{\pi})} \\ \xi_0 &= r_1 + r_2 - \alpha_0 g \tilde{\pi}.\end{aligned}$$

Then we have the following results:

Lemma B.1 *If $W^{(i)}$ satisfies (19) and $V^{(i)}$ satisfies (43), then*

$$W^{(i)} = p_D^{(i)-1} + A^{(i)} \tau_D A^{(i)T}, \quad (\text{B1})$$

and

$$\left(B^{(i)\text{T}}W^{(i)-1}B^{(i)}\right)^{-1} = p_D^{(i+1)-1} + A^{(i+1)}\zeta_D^{(i+1)}A^{(i+1)\text{T}}, \quad (\text{B2})$$

where $p_D^{(i)}$, τ_D and ζ_D are the diagonalizations of (44), $(\tau_1, \tau_2, \tau_3, \dots, \tau_{m(i)})$ and $(\frac{h}{1-h\tilde{\pi}}, \tau_3, \dots, \tau_{m(i)})$, respectively.

Proof Equation (B1) follows immediately from (19,43). Using the Sherman–Morrison–Woodbury matrix-inversion identity (see, for example, Golub and Van Loan (1996)), we obtain

$$W^{(i)-1} = p_D^{(i)} - p_D^{(i)}A^{(i)}(e_D - be_Dw^{(i)}w^{(i)\text{T}}e_D)A^{(i)\text{T}}p_D^{(i)},$$

where e_D is the diagonalization of $(\tau_1r_1, \tau_2r_2, \dots, \tau_{m(i)}r_{m(i)})$. Elementary manipulations eventually lead to:

$$B^{(i)\text{T}}W^{(i)-1}B^{(i)} = p_D^{(i+1)} - p_D^{(i+1)}A^{(i+1)}(\tilde{e}_D - b\tilde{e}_Dw^{(i+1)}w^{(i+1)\text{T}}\tilde{e}_D)A^{(i+1)\text{T}}p_D^{(i+1)}, \quad (\text{B3})$$

where $\tilde{e}_D$ is the diagonalization of $(h, \tau_3r_3, \dots, \tau_{m(i)}r_{m(i)})$ and $w^{(i+1)}$ is given by (41). Equation (B2) may then be obtained by again applying the Sherman–Morrison–Woodbury identity, but this time to (B3) and in the reverse direction. $\square$

Lemma B.2 *With the assumptions of Lemma B.1, expression (40) simplifies to*

$$V^{(i+1)} = p_D^{(i+1)-1} + A^{(i+1)}s_D^{(i+1)}A^{(i+1)\text{T}}, \quad (\text{B4})$$

where $p_D^{(i+1)}$ and $s_D^{(i+1)}$ are the diagonalizations of $p^{(i+1)}$ and $s^{(i+1)}$ given by (44) and (42), where in (42) we define:

$$s_1^{(i+1)} = \frac{h}{1-h\tilde{\pi}} + \frac{f^2}{2w_1^{(i)}w_2^{(i)}\xi_0}. \quad (\text{B5})$$

Proof The result may be shown through elementary but extremely lengthy and laborious manipulations, using the results of Lemma B.1. $\square$

Lemma B.3 *Expression (43) for $V^{(i)}$ holds for all $i = 0, 1, \dots, m^{(0)}$.*

Proof By induction from Lemma B.2, noting from (15) that (43) holds for $i = 0$. $\square$

References

- Brodal, G., R. Faberberg, and P. C.N.S. (2004). Computing the quartet distance between evolutionary trees in time $O(n \log^2 n)$. *Algorithmica* 38(2), 377–395.
- Bruno, W., N. Socci, and A. Halpern (2000). Weighted Neighbour Joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Molecular Biology and Evolution* 17, 189–197.
- Bulmer, M. (1991). Use of the method of generalized least squares in reconstructing phylogenies from sequence data. *Molecular Biology and Evolution* 8, 868–883.
- Chakraborty, R. (1977). Estimation of time of divergence from phylogenetic studies. *Canadian Journal of Genetics and Cytology* 19, 217–223.
- Crowder, M. (2001). On repeated measures analysis with misspecified covariance structure. *Journal of the Royal Statistical Society, Series B* 63, 55–62.
- Desper, R. and O. Gascuel (2004). Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. *Molecular Biology and Evolution* 21(3), 587–598.
- Felsenstein, J. (1987). Estimation of hominoid phylogeny from a DNA hybridization data set. *Journal of Molecular Evolution* 26, 123–131.
- Felsenstein, J. (2004). *Inferring Phylogenies*. Massachusetts: Sinauer Associates, Inc.
- Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution* 14, 685–695.
- Gascuel, O. (2000). Data model and classification by trees: the minimum variance reduction (MVR) method. *Journal of Classification* 17, 67–99.
- Golub, G. and C. Van Loan (1996). *Matrix Computations* (3rd ed.). Baltimore: The Johns Hopkins University Press.
- Hasegawa, M., H. Kishino, and T. Yano (1985). Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 22, 160–174.
- Jukes, T. and C. Cantor (1969). Evolution of protein molecules. In M.N.Munro (Ed.), *Mammalian Protein Metabolism*, Volume III, pp. 21–132. New York: Academic Press.
- Keele, B., E. Giorgi, J. Salazar-Gonzalez, J. Decker, K. Pham, M. Salazar, et al. (2008). Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc. Natl. Acad. Sci. USA*. 105, 7552–7557.
- Kimura, M. (1980). A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16, 111–120.

- Kishino, H., J. Thorne, and W. Bruno (2001). Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Molecular Biology and Evolution* 18, 352–361.
- Lanave, C., G. Preparata, C. Saccone, and G. Serio (1984). A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution* 20, 86–93.
- Mardia, K., J. Kent, and J. Bibby (1979). *Multivariate Analysis*. New York: Academic Press.
- Nei, M., J. Stephens, and N. Saitou (1985). Methods for computing the standard errors of branching points in an evolutionary tree and their applications to molecular data from human and apes. *Molecular Biology and Evolution* 2, 66–85.
- Rambaut, A. and N. Grassly (1997). Seq-gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Algorithmica* 13(3), 235–238.
- Robinson, D. and L. Foulds (1981). Comparison of phylogenetic trees. *Mathematical Bioscience* 53, 131–147.
- Saitou, N. and M. Nei (1987). The neighbour-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4, 406–425.
- Salazar-Gonzalez, J., M. Salazar, B. Keele, G. Learn, E. Giorgi, H. Li, et al. (2009). Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. *J. Experimental Medicine* 206(6), 1273–1289.
- Studier, J. and K. Keppler (1988). A note on the neighbor-joining method of Saitou and Nei. *Molecular Biology and Evolution* 5, 729–731.
- Susko, E. (2003). Confidence regions and hypothesis tests using generalized least squares. *Molecular Biology and Evolution* 20, 862–868.
- Thorne, J., H. Kishino, and I. Painter (1998). Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution* 15, 1647–1657.
- Wang, L.-S. and T. Warnow (2005). Distance-based genome rearrangement phylogeny. In O. Gascuel (Ed.), *Mathematics of Evolution & Phylogeny*, Chapter 13, pp. 353–383. Oxford University Press.
- Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution* 10, 1396–1401.
- Zwickl, D. and D. Hillis (2002). Increased taxon sampling greatly reduces phylogenetic error. *Systematic Biology* 51(4), 588–598.