MOLECULAR PHYLOGENETICS AND EVOLUTION
Vol. 8, No. 3, December, pp. 398–414, 1997
ARTICLE NO. FY970452

General Time-Reversible Distances with Unequal Rates across Sites: Mixing Γ and Inverse Gaussian Distributions with Invariant Sites

Peter J. Waddell*,[1] and M. A. Steel†

*School of Biological Sciences, Massey University, Palmerston North, New Zealand; and †Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch, New Zealand

Received February 25, 1997; revised August 1, 1997

A series of new results useful to the study of DNA sequences using Markov models of substitution are presented with proofs. General time-reversible distances can be extended to accommodate any fixed distribution of rates across sites by replacing the logarithmic function of a matrix with the inverse of a moment generating function. Estimators are presented assuming a gamma distribution, the inverse Gaussian distribution, or a mixture of either of these with invariant sites. Also considered are the different ways invariant sites may be removed and how these differences may affect estimated distances. Through collaboration, we implemented these distances into PAUP* in 1994. The variance of these new distances is approximated via the delta method. It is also shown how to predict the divergence expected for a pair of sequences given a rate matrix and a distribution of rates across sites, allowing iterated ML estimates of distances under any reversible model. A simple test of whether a rate matrix is time reversible is also presented. These new methods are used to estimate the divergence time of humans and chimps from mtDNA sequence data. These analyses support suggestions that the human lineage has an enhanced transition rate relative to other hominoids. These studies also show that transversion distances differ substantially from the overall distances which are dominated by transitions. Transversions alone apparently suggest a very recent divergence time for humans versus chimps and/or a very old (>16 myr) divergence time for humans versus organgutans. This work illustrates graphically ways to interpret the reliability of distance-based transformations, using the corrected transition to transversion ratio returned for pairs of sequences which are successively more diverged. © 1997 Academic Press

INTRODUCTION

Failure to allow for unequal substitution rates at different sites in two aligned sequences can lead to serious underestimates of the true distance between them (Golding, 1983). Furthermore, this underestimation becomes progressively worse the larger the true distance, which in turn compromises the additivity necessary for transformed distance phylogenetic methods to be guaranteed consistent (Felsenstein, 1982, 1984, 1993). If this error becomes serious enough, parallelisms and convergences due to multiple substitutions at a site (which occur predominantly between long edges of a tree) can outweight parsimony informative characters (Felsenstein, 1978; Hendy and Penny, 1989). This effect is often termed "long edges attract," as such edges (or internodes, as defined below) may be spuriously joined together by tree reconstruction methods (including distance methods), even when all other aspects of the model are correct (e.g., Hasegawa and Fujiwara, 1993; Lewis and Gaut, 1995; Waddell, 1995; Chang, 1996; Lockhart *et al.*, 1996). Failure to account for unequal base compositions in the sequence also leads to a progressive underestimate of the true distance (e.g., Tamura, 1992), with similar effects expected upon tree selection.

The general time-reversible distance is the most general transformation that can be applied to a pair of DNA sequences which aims to return the expected average number of substitutions per site. This distance was first described by Lanave *et al.* (1984), and in a different, but numerically and algebraically equivalent, form by Tavaré (1986), Barry and Hartigan (1987), and Rodríguez (1990). [Gillespie (1986) and Zharkikh (1994) note that the Lanave *et al.* distance is time reversible and not a more general 12-parameter model; Waddell (1995) notes the numerical identity, while Swofford and Lewis (1997) provide a proof.] This paper aims to explain to biologists the assumptions of these distances and to clarify some earlier misconceptions. Importantly, nearly all of the currently used distance estimates (including those of Tamura, 1992; Tamura and Nei, 1994) are special cases (restrictions) of the general time-reversible distance (see Zharkikh, 1994; Swofford *et al.*, 1996).

A general time-reversible distance assumes a general time-reversible model of evolution, which is a model where the probability (or likelihood) of the data is independent of the placement of the root on the tree

[1] To whom correspondence should be addressed. E-mail: waddell@ism.ac.jp.

1055-7903/97 $25.00

(Felsenstein, 1981; Adachi and Ahasegawa, 1994; Yang, 1994). With the exception of some special matrices (the Kimura 3 ST being the most general; e.g., Hendy *et al.*, 1994), this further impilies the relative rates of all substitutions remain constant across the tree (that is, a homogeneous model) and that the root base composition is in equilibrium. This in turn implies that all states in the model (e.g., the four nucleotides A, C, G, and T) remain at the same frequencies; that is, they (and the model) are said to be stationary. Given this model, it is possible to make estimates of the rates of all types of substitution using just pairs of sequences (Lanave *et al.,* 1984; Tavaé, 1986).

A variety of specific distances have been modified to take unequal substitution rates across sites into account. These include the Jukes–Cantor (1-parameter) and Kimura (2-parameter) distances (Golding, 1983; Olsen, 1987; Jin and Nei, 1990) and a specific 6-parameter distance (Tamura and Nei, 1993; each allowing a *t* distribution of site rates). In addition, a variety of methods to calculate likelihoods of the data under similar conditions have been used (Hasegawa *et al.*, 1985; Churchill *et al.*, 1992; Reeves, 1992; Sidow *et al.*, 1992; Steel *et al.*, 1993; Yang, 1993; Waddell, 1995; Felsenstein and Churchill, 1996; Waddell and Penny, 1996; Waddell *et al.*, 1997a).

An important case of unequal rates across sites is the existence of some sites which are incapable of changing due to biological constraints. These invariant (invariable) sites lead to distortions of estimated distances (Shoemaker and Fitch, 1989), and in some cases to inconsistency of tree selection (e.g., Hasegawa and Fujiwara, 1993; Waddell, 1995; Chang, 1996; Lockhart *et al.*, 1996; Waddell *et al.*, 1997a). Here we consider how time-reversible distances may be modified to take these sites into account, especially when the base composition of these sites does *not* reflect that of the variable sites (Waddell, 1995).

A primary motivation for this work was to have distances to both infer trees and more accurately estimate the edge lengths on trees. Such weighted trees (i.e., trees with estimated edge lengths) are critical for inferring the divergence times of many taxa (e.g., Hillis *et al.*, 1996; Waddell and Penny, 1996). We use an example from Waddell and Penny (based on 5 kb of hominoid mtDNA sequences from Horai *et al.*, 1992) to illustrate the new methods and to infer the divergence time of human versus chimp lineages.

Recently an extension of the general time-reversible distance has been proposed to accommodate any distribution of rates across sites (Waddell, 1995; Swofford *et al.*, 1996; Waddell and Steel, 1996). Here we describe a biological and mathematical framework for the use of this distance. Our work with Dr. David Swofford has incorporated this general distance into the phylogenetics package PAUP 4.0 (Swofford, 1997), where it can be used with a wide variety of tree selection and evaluation criteria (Waddell *et al.*, 1997b).

Some terms and abbreviations used in this paper:

c the sequence length

CSR constant site removal (from the data)

δ a transformed distance estimate

F a matrix of proportions of aligned paired nucleotides

$\mathbf{F}^{\#}$ matrix **F** symmetrized

i.r. assuming all site are evolving at an identical intrinsic rate

p_{inv} the proportion of invariant (invariable) sites

ti/tv transition to transversion ratio

A rate matrix, **R**, contains the underlying relative rates of substitution, unaffected by multiple hits. Rates of change form base *i* to *j* are always nonnegative and, biologically, can be regarded as always positive. Rates of change from *i* to *i* (i.e., no change) are negative numbers (shown on the diagonal as *). As such, their magnitude is minus the sum of the rates at which *i* is changing to other bases; consequently all rows in **R** will sum to zero.

Last, please note the rationale for the use of the term "edge," our preferred term to define an "internode" or "link" in a tree. Some phylogeneticists call an edge a "branch," but branch has multiple meanings, so can be an ambiguous or misleading term (e.g., Penny *et al.*, 1992; Waddell, 1995). For example, when the founders of phylogenetics refer to a branch, they often mean all the descendants of an ancestor. Darwin (1859) follows this convention. He also has an especially dynamic usage, where some branches (groups of species) keep growing and shading (out-competing) other branches (those with fewer tips or extant taxa). We follow Darwin's usage, which is clearly *incompatible* with those who would equate branch with internode.

MATERIALS AND METHODS

Time Reversible Distances: Their Form and Assumptions

If all sites evolve at an identical rate (i.r.) the general time-reversible distance can be written (Rodríguez *et al.,* 1990) in the form

$$\delta_{ij} = -\text{trace}(\mathbf{\Pi}\ ln\ [\mathbf{\Pi}^{-1}\mathbf{F}]), \qquad (1)$$

where δ_{ij} is the distance between sequences *i* and *j* measured as the expected number of substitutions per site (including multiple changes at a site), **Π** is a diagonal matrix of the nucleotide base composition of the sequences, **F** is the divergence matrix of sequences *i* and *j*, and *ln* is the matrix logarithm function. The divergence matrix is just the expected proportion of times one state is aligned next to another state in the two sequences (Fig. 1). The logarithm of a matrix **X** is defined as *ln* (**X**) = $\Sigma_{n=1}^{\infty}$ $(\mathbf{I} - \mathbf{X})^n/n$, where **I** is the identity matrix, provided this limit exists. Under a time-reversible process of evolution, all **F** matrices are symmetric in expectation (e.g., Tavaré, 1986; Barry and Hartigan, 1987). In dealing with finite samples, **Π** and **F** are replaced with their sample estimates (these are denoted by $\hat{\mathbf{\Pi}}$ and $\hat{\mathbf{F}}$). Furthermore, $\hat{\mathbf{F}}$ is then replaced by $\mathbf{F}^{\#}$, a symmetrized form of $\hat{\mathbf{F}}$. This is done to reduce sampling errors, and $\mathbf{F}^{\#}$ is shown to be a ML estimator of **F** under the model (see Appendix 1a). This result is convenient, since a useful way to evaluate the matrix logarithm function of the matrix $\mathbf{\Pi}^{-1}\mathbf{F}^{\#}$ (if defined) is via diagonalization (see Fig. 1). The symmetry of both $\mathbf{F}^{\#}$ and **Π** implies that their product is always diagonalizable and will have real eigenvalues (e.g., Keilson, 1979). Should any of these eigenvalues be negative (e.g., with real data), this implies an infinite distance when the distance assessment is considered in a ML framework (i.e., infinity would be the ML estimate of the distance).

These distances are consistent (i.e., become exactly correct as sequence lengths go to infinity) and so aditive in expectation on a tree, provided all sites have the same rate of substitution (i.r.) and the process of evolution is time-reversible across all paths in the tree. The most general form of the rate matrix, **R**, then has nine parameters and can be written as $\mathbf{\Pi}^{-1}\mathbf{S}$, where **S** is a symmetric matrix of relative rates, and **Π** is the diagonal matrix of the stationary base compositions of the nucleotide states (Tavaré, 1986). An equivalent parameterization of **R** is **SΠ**, for a different but still symmetric rate matrix **S** (Tavaré, 1986; Zharkikh, 1994; for a proof see Appendix 6). Thus **Π** has three free parameters (since nucleotide proportions must sum to 1), while **S** has up to six (since rows of a rate matrix sum to zero), making a total of nine free parameters in the model.

A special case where Eq. (1) is also exact, but the model may not be strictly time-reversible, is when the base composition is equal-frequency [0.25, 0.25, 0.25, 0.25] (Rodríguez *et al.*, 1990). To meet this requirement the **R** matrix must have both its rows and its columns summing to zero; this gives nine free parameters but not necessarily a time-reversible model (see Waddell and Steel, 1996, for a counterexample). The restrictive assumption of a molecular clock made by Lanave *et al.* (1984) and Rodríguez *et al.* (1990) in deriving Eq. (1) is not necessary (e.g., Tavaré, 1986; Barry and Hartigan, 1987).

An example of calculating the general time-reversible distance is given in Fig. 1. It considers the divergence matrix between the human and chimp sequences of Horai *et al.* (1992) (as edited by the removal of all sites with insertions or deletions).

Checking the Data

Evidence of nonstationarity in the Horai *et al.* data was sought using an exhaustive set of pairwise tests of base compositions using the X^2 statistic, as implemented in PAUP* 4.0 (Swofford, 1997). The overall result was nonsignificant (i.e., no evidence of nonstationarity), and removal of constant sites did not alter the result (a precaution against the presence of invariant sites, i.e., sites unable to vary).

It is also useful to test the expectation that $c\mathbf{F}$ is symmetric (e.g., Tavaré, 1986) and using either a X^2 or G^2 test statistic is reasonable (Read and Cressie, 1988). The X^2 test statistic is

$$\sum_{i\neq j} \frac{(c\hat{\mathbf{F}}_{ij} - c\mathbf{F}^{\#}_{ij})^2}{c\mathbf{F}^{\#}_{ij}},$$

while the G^2 test uses

$$\sum_{i\neq j} c\hat{\mathbf{F}}_{ij} \ln\left(\frac{\hat{\mathbf{F}}_{ij}}{\mathbf{F}^{\#}_{ij}}\right).$$

Both test statistics asymptotically have a χ^2 distribution with degrees of freedom (*d.f.*) equal to the number of entries $i \neq j$ (12) minus the number of estimates made in $\mathbf{F}^{\#}$ (6), equalling 6 *d.f.* For the comparison of human and chimp sequences, the X^2 and G^2 values are 8.86 ($P = 0.18$) and 9.67 ($P = 0.14$), respectively (or 6.80 ($P = 0.15$) and 6.84 ($P = 0.14$) when grouping cells with expected values of less than 5). Of all such tests only the comparisons of African apes to orangutan were significant ($P < 0.05$), suggesting a change in the substitution process in the orangutan lineage (consistent with Adachi and Hasegawa's 1996 findings). While there is a multiple test problem here, robust Bonferroni type corrections suggest that the orangutan indeed violates the model expectations at the 95% level.

A potential problem with this type of test is its lack of power when a molecular clock is likely, since this too implies that **F** is symmetric (see proof in Appendix 1b). The symmetry of **F** under a molecular clock is understandable as "the expectation of the same frequency of evolutionary events in each of two lineages of precisely the same duration, each evolving by exactly the same stochastic process." However, the test is not affected by invariant sites. Thus, overall, most of the data appear to conform reasonably well to the expectations of a reversible model, with the exception of the orangutan lineage.

RESULTS

The Time-Reversible Distance with Any Distribution of Rates across Sites

Distances estimated under stationary time-reversible models (with up to nine parameters in their transition matrices) can be extended to allow for unequal rates across sites using the same general approach used in Steel *et al.* (1993) and Waddell *et al.*

(1997a) (for the Hadamard conjugation). As explained below, the extension allows correction for a variety of site rate distributions, including the commonly used Γ and lognormal. Our new distance formula (Waddell, 1995; Swofford *et al.*, 1996; Waddell and Steel, 1996; and all test versions of PAUP* since 1994) estimating the expected number of substitutions per site is

$$\delta_{ij} = -\text{trace}(\mathbf{\Pi} M^{-1}[\mathbf{\Pi}^{-1}\mathbf{F}]), \qquad (2)$$

where M^{-1} is the inverse of the moment generating function of the distribution of rates across sites (defined below and see a proof in Appendix 2). The application of M^{-1} to $\mathbf{\Pi}^{-1}\mathbf{F}$ (here taken as matrix $\mathbf{Z}$) is defined as

$$M^{-1}(\mathbf{Z}) = \mathbf{\Omega} M^{-1}[\mathbf{\Psi}]\mathbf{\Omega}^{-1}, \qquad (3)$$

where $\mathbf{\Omega}$ is a matrix containing, as columns, the right eigenvectors of $\mathbf{Z}$ (i.e., $\mathbf{Z\Omega} = \mathbf{\Omega D}$), $\mathbf{\Omega}^{-1}$ is its inverse, and function M^{-1} is applied componentwise to the diagonal entries of the diagonal matrix $\mathbf{\Psi}$ of the associated eigenvalues of $\mathbf{Z}$. As with the time-reversible i.r. model, we symmetrize $\hat{\mathbf{F}}$ to give $\mathbf{F}^{\#}$ when dealing with sampled data.

The function $M[x]$ is defined as the expectation, $M[x] = E[e^{\lambda_j x}]$, the moment generating function of the statistical distribution the λ_j site rates (Table 2 of Waddell *et al.*, 1997a, gives relevant examples; see also Steel *et al.*, 1993). Note that

$$M[x] \approx \frac{1}{c}\sum_{i=1}^{c} e^{\lambda_i x},$$

the average value of the $e^{\lambda_i x}$ over the sites (where c is the sequence length). Here, the argument of M will always be ≤0 (rather than positive as in most statistical applications). Consequently, function M will always be defined in our applications and will lie in the range from 0 to 1. M^{-1} denotes the left functional inverse (the standard inverse) of M (i.e., $M^{-1}[M[x]] = x$), which always exists since $M[x]$ is a monotone increasing function (again see Waddell *et al.*, 1997a, for full details). In real applications we do not know the function M exactly for any given sequence, so its form is inferred with an ML method that compares more than two sequences at a time, as discussed later.

It is proven that any distance based on only the observed dissimilarity (e.g., that of Tajima-Nei; see Swofford *et al.*, 1996) and assuming identical site rates will (asymptotically as $c \to \infty$) always underestimate the true distance if there is any site-to-site rate variation (see Waddell and Steel, 1996; Appendix 3).

Due to sampling error when $\mathbf{F}$ is estimated from a finite number of sites, the eigenvalues of $\mathbf{P} = (\text{-}\mathbf{\Pi}^{-1}\mathbf{F}^{\#})$ (which are expected to lie in the range [0, 1]), may be negative, making M^{-1} undefined (basically the distance appears too large or infinite given the expectations of the model). This is a commonly encountered problem with all model-based distance transformations, which may also be caused by nonstationarity of base composition (Waddell, 1995).

In cases where an eigenvalue of $\mathbf{F}^{\#}$ is negative, a useful rule of thumb in estimating phylogenetic relationships is set all undefined distances to twice the value of the largest defined distance from the distance matrix of species being compared. This is justified since the largest distance (path) on a tree can never be more than twice the size of the second largest value (Waddell, 1995). Given more information about the tree, it may also be possible to refine the expected range for inapplicable distance estimates. The use of reduced bias estimators based on Taylor series-like approximations (e.g., Tajima, 1993) does not appear feasible as the eigenvalues do not have a simple sampling distribution.

Our general approach also provides a quick way of calculating the transition matrix, $\mathbf{P}$, along any edge or path moving down a tree (from root to tips) when rates at sites vary and $\mathbf{R}$ is given (e.g., when modeling sequence evolution). Let τ be equal to the total expected number of substitutions on an edge or along a path, while $\mathbf{R}$ is scaled so that the positive entries of $\mathbf{\Pi R}$ sum to 1; then

$$\mathbf{P} = M[\mathbf{R}\tau]. \qquad (4)$$

Under reversibility the relation $\mathbf{F} = \mathbf{P}^{t}\mathbf{\Pi P}$ (e.g., see Barry and Hartigan, 1987) then allows us to quickly calculate the divergence matrix ($\mathbf{F}$) given any distribution of rates across sites (later this is used to calculate iterated pairwise ML distances). As with Eqs. (2) and (3), it is assumed sites evolve independently. A proof of the last equation is given in Appendix 3. It is useful to note that if $\mathbf{R}$ defines a time-reversible process it can always be diagonalized and has real eigenvalues (Keilson, 1979, section 3.2). For convenience we will label the eigenvalues of $\mathbf{R}\tau$ as entries ξ_{ii} of the diagonal matrix Ξ (e.g., Figs. 1 and 2).

Specifically, the moment-generating functions, M, are for standardized distributions (e.g., Stuart and Ord, 1987, p. 192), where the mean of the underlying distribution has been set to 1 (i.e., $E_\lambda[\lambda] = 1$), so that inferred distances are recovered as the expected number of substitutions per site and not some other multiple of this number (e.g., Golding, 1983; Jin and Nei, 1989; Steel *et al.*, 1993). Two distributions are particularly useful because they both have closed forms for both M and M^{-1}. The first of these is for the much used gamma (Γ) distribution (e.g., Golding, 1983; Jin and Nei, 1990; Steel *et al.*, 1993), where $M[x] = ((k - x)/k)^{-k}$, while $M^{-1}[x] = k(1 - x^{-1/k})$, where k is the shape parameter. When $k \to \upsilon$, the Γ distribution tends to the delta distribution (i.e., identical rates), and M tends to

$$\begin{array}{c|cccc} & A & C & G & T \\ \hline A & 1415 & 8 & 55 & 2 \\ C & 4 & 1371 & 1 & 144 \\ G & 73 & 0 & 578 & 0 \\ T & 3 & 117 & 1 & 1126 \end{array} \quad \begin{bmatrix} 0.2889 & 0.0012 & 0.0131 & 0.0005 \\ 0.0012 & 0.2799 & 0.0001 & 0.0266 \\ 0.0131 & 0.0001 & 0.1180 & 0.0001 \\ 0.00005 & 0.0266 & 0.0001 & 0.2299 \end{bmatrix} \quad \begin{bmatrix} 0.3037 & 0 & 0 & 0 \\ 0 & 0.3079 & 0 & 0 \\ 0 & 0 & 0.1313 & 0 \\ 0 & 0 & 0 & 0.2571 \end{bmatrix}$$

$$\mathrm{c}\hat{\mathbf{F}} \qquad \mathbf{F}^{\#} = (\hat{\mathbf{F}} + \hat{\mathbf{F}}^{t})/2 \qquad \hat{\Pi}\ (\hat{\Pi}_{ii} = \text{row sum of } \mathbf{F}^{\#})$$

$$\begin{bmatrix} 0.9513 & 0.0040 & 0.0430 & 0.0017 \\ 0.0040 & 0.9092 & 0.0003 & 0.0865 \\ 0.0995 & 0.0008 & 0.8989 & 0.0008 \\ 0.0030 & 0.1036 & 0.0004 & 0.8940 \end{bmatrix} \quad \begin{bmatrix} 0.4088 & 0.5473 & -0.5000 & 0.0138 \\ 0.0020 & -0.4236 & -0.5000 & -0.6449 \\ -0.9134 & 0.5770 & -0.5000 & -0.0159 \\ -0.0164 & -0.4337 & -0.5000 & 0.7640 \end{bmatrix} \quad \begin{bmatrix} 0.8546 & 0 & 0 & 0 \\ 0 & 0.9922 & 0 & 0 \\ 0 & 0 & 1.0000 & 0 \\ 0 & 0 & 0 & 0.8066 \end{bmatrix}$$

$$\hat{\mathbf{P}} = \hat{\Pi}\mathbf{F}^{\#} \qquad \mathbf{\Omega} = \text{right eigenvectors of } \hat{\mathbf{P}} \qquad \mathbf{\Psi}\ (\psi_{ii} = \text{eigenvalues of } \hat{\mathbf{P}})$$

$$\begin{bmatrix} 0.7728 & 0.0038 & -0.7502 & -0.0264 \\ 0.6975 & -0.5473 & 0.3179 & -0.4681 \\ -0.6074 & -0.6158 & -0.2626 & 0.5143 \\ 0.0151 & 0.7137 & -0.0075 & 0.7661 \end{bmatrix} \quad \begin{bmatrix} -0.1571 \\ -0.0079 \\ 0.0000 \\ -0.2150 \end{bmatrix} \quad \begin{bmatrix} -0.0524 & 0.0042 & 0.0466 & 0.0016 \\ 0.0042 & -0.1008 & 0.0002 & 0.0963 \\ 0.1078 & 0.0006 & -0.1091 & 0.0007 \\ 0.00019 & 0.1154 & 0.0004 & -0.1176 \end{bmatrix}$$

$$\Omega^{-1} \qquad \ln(\psi_{ii}) \qquad \hat{\mathbf{R}}\tau = \Omega\ln(\Psi)\Omega^{-1}$$

$$\begin{bmatrix} -0.0159 & 0.0013 & 0.0142 & 0.0005 \\ 0.0013 & -0.0310 & 0.0001 & 0.0297 \\ 0.0142 & 0.0001 & -0.0143 & 0.0001 \\ 0.0005 & 0.0293 & 0.0001 & -0.0302 \end{bmatrix} \quad \begin{bmatrix} -78.0 & 6.3 & 69.3 & 2.4 \\ 6.3 & -152.0 & 0.4 & 145.3 \\ 69.3 & 0.4 & -70.2 & 0.5 \\ 2.4 & 145.3 & 0.5 & -148.2 \end{bmatrix}$$

$$\hat{\Pi}\hat{\mathbf{R}}\tau \text{ (subs. of each type per site)} \qquad c\hat{\Pi}\hat{\mathbf{R}}\tau \text{ (estimated totals)}$$

δ_{hc} is the expected number of substitutions per site between human and chimp sequences, which gives rise to the estimate $\hat{\delta}_{hc}$ = -trace($\hat{\Pi}\hat{\mathbf{R}}\tau$) = - (-0.0159 + -0.0310 + -0.0143 + -0.0302) = 0.09154 (0.09152 with full precision).

FIG. 1. The steps in calculating the time-reversible distance [Eq. (1)]. The observed divergence matrix $c\hat{\mathbf{F}}$ (where c is the sequence length) is for the comparison of human and chimp mtDNA sequences (Horai *et al.*, 1992). Starting with the observed matrix of aligned paired-nucleotide frequencies ($c\hat{\mathbf{F}}$) we estimate $\mathbf{R}\tau$ and other quantities. Entries in $\hat{\mathbf{R}}\tau$ are inferred relative rates, whereas entries in $\hat{\Pi}\hat{\mathbf{R}}\tau$ are estimated numbers of each type of substitution divided by the sequence length. The observed (Hamming) distance from $\hat{\mathbf{F}}$ is

$$\sum_{i \neq j} F_{ij} = (8 + 55 + \cdots + 1)/4898 = 0.0833,$$

whereas the distance for the data corrected under the i.r. time reversible model is (6.3 + 69.3 + . . . + 0.5)/4898 = 0.0915. The matrix $c\hat{\Pi}\hat{\mathbf{R}}\tau$ (which is analogous to the $\hat{\mathbf{F}}$ matrix with corrections for multiple hits) shows the estimated number of transversions almost unchanged. In contrast, the number of multiple hits is estimated as [(69.3 + 69.3)/(55 + 73) − 1] × 100% = 8.3% among the A ↔ G transitions or [(145.3 + 145.3)/(117 + 144) − 1] × 100% = 11.3% among the more numerous C ↔ T transitions. This in turn has increased the overall transition to transversion ratio from 20.47 for the observed data to 22.50 for the i.r. time-reversible model corrected data, an increase of 9.9%. Note: A worked example in Rodríguez *et al.* (1990) does not symmetrize **F**, and has serious round off errors, making it unsuitable for checking computations.

the *ln* function. When k decreases, the distribution assumes a skewed normal, then exponential shape (at $k = 1$); for $k < 1$ the distribution becomes ever more L-shaped (e.g., see Golding, 1983; Jin and Nei, 1990; Swofford *et al.*, 1996).

The second distribution (Waddell, 1995; Waddell *et al.*, 1997a) is the inverse Gaussian distribution, which is shaped more like the lognormal distribution (introduced with genetic distances by Olsen, 1987). For the inverse Gaussian, $M[x] = \exp(d\{1 - [1 - (2x/d)]^{0.5}\})$,

M^{-1} eigenvalues — Inferred rate matrix per nucleotide — Overall inferred substitutions

$M^{-1}[\psi_{ii}]$ — $\hat{\mathbf{R}}\tau$ — $c\hat{\Pi}\hat{\mathbf{R}}\tau$

(a) M^{-1} for inverse Gaussian with $d = 0.213$, $\delta = 0.13274$.

$$\begin{bmatrix} -0.2151 \\ -0.0080 \\ 0.0000 \\ -0.3235 \end{bmatrix} \begin{bmatrix} -0.0707 & 0.0053 & 0.0643 & 0.0012 \\ 0.0052 & -0.1507 & -0.0002 & 0.1457 \\ 0.1487 & -0.0004 & 0.1489 & -0.0006 \\ 0.0014 & 0.1745 & 0.0003 & -0.1762 \end{bmatrix} \begin{bmatrix} -105.2 & 7.8 & 95.6 & 1.8 \\ 7.8 & -227.3 & -0.3 & 219.7 \\ 95.6 & -0.3 & -95.7 & 0.4 \\ 1.8 & 219.7 & 0.4 & -221.9 \end{bmatrix}$$

(b) M^{-1} for Γ with $k = 0.351$, $\delta = 0.12205$.

$$\begin{bmatrix} -0.1982 \\ -0.0080 \\ 0.0000 \\ -0.2966 \end{bmatrix} \begin{bmatrix} -0.0654 & 0.0050 & 0.0591 & 0.0013 \\ 0.0049 & -0.1384 & -0.0001 & 0.1335 \\ 0.1368 & -0.0002 & -0.1373 & 0.0007 \\ 0.0015 & 0.1598 & 0.0004 & -0.1617 \end{bmatrix} \begin{bmatrix} -97.3 & 7.4 & 87.9 & 1.9 \\ 7.4 & -208.6 & -0.1 & 201.3 \\ 87.9 & -0.1 & -88.3 & 0.4 \\ 1.9 & 201.3 & 0.4 & -203.6 \end{bmatrix}$$

(c) M^{-1} for with $p_{inv} = 0.592$, Π_{inv} estimated from $\mathbf{F}_{hc}$, $\delta = 0.26713$ (0.10899).

$$\begin{bmatrix} -0.4406 \\ -0.0194 \\ 0.0000 \\ -0.6427 \end{bmatrix} \begin{bmatrix} -0.1461 & 0.0115 & 0.1312 & 0.0034 \\ 0.0113 & -0.3003 & 0.0001 & 0.2888 \\ 0.3034 & 0.0003 & -0.3056 & 0.0018 \\ 0.0040 & 0.3458 & 0.0009 & -0.3508 \end{bmatrix} \begin{bmatrix} -88.6 & 7.0 & 79.6 & 2.1 \\ 7.0 & -184.8 & 0.1 & 177.7 \\ 79.6 & 0.1 & -80.2 & 0.5 \\ 2.1 & 177.7 & 0.5 & -180.3 \end{bmatrix}$$

(d) M^{-1} for with $p_{inv} = 0.592$, Π_{inv} estimated from constant sites, $\delta = 0.26635$ (0.10867).

$$\begin{bmatrix} -0.5772 \\ -0.0202 \\ 0.0000 \\ -0.5575 \end{bmatrix} \begin{bmatrix} -0.1687 & 0.0125 & 0.1523 & 0.0040 \\ 0.0098 & -0.2543 & 0.0001 & 0.2445 \\ 0.3993 & 0.0002 & -0.4018 & 0.0023 \\ 0.0042 & 0.3251 & 0.0009 & -0.3301 \end{bmatrix} \begin{bmatrix} -93.2 & 6.9 & 84.1 & 2.2 \\ 6.9 & -179.4 & 0.0 & 172.4 \\ 84.1 & 0.0 & -84.6 & 0.5 \\ 2.2 & 172.4 & 0.5 & -175.1 \end{bmatrix}$$

FIG. 2. The effect of different forms of the distribution of rates across sites on the transformed distances. (a) An inverse Gaussian (with shape parameter $d = 0.213$), (b) Γ (with shape $k = 0.351$), (c) CSR(F), and (d) CSR(cons) are invariant sites/variable sites distributions. Averaging over all 6 hominoid mtDNA sequences in Horai *et al.* (1992) gives $\pi = [0.30, 0.31, 0.13, 0.26]$, whereas just the unvaried sites have base composition $\pi^c = [0.32, 0.28, 0.15, 0.25]$, which is significantly different (by a X^2 statistic test). For both invariant sites models, the estimated rate matrix is for just the variable sites. Likewise, δ, measured over just the variable sites is shown first and then δ averaged over all sites [i.e., multiplied by $(1 - p_{inv})$] is given in brackets.

while $M^{-1}[x] = 0.5d(1 - \{1 - (ln[x]/d)\}^2)$. Here d is the shape parameter, and the coefficient of variation (i.e., ratio of s.d. to mean) for site rates is $d^{0.5}$. Here, as $d \to \infty$, M tends to the natural logarithm function. As d decreases below 1, the rates across sites follow a highly skewed lognormal-like distribution (see Fig. 1, Waddell *et al.*, 1997a). Apart from the notable shape difference, the inverse Gaussian distribution tends to have a flatter tail than the Γ, inferring more of the most rapidly evolving sites (e.g., the sites evolving more than 40 times the mean rate). Distance formulas assuming flat-tailed distributions will thus often infer more multiple hits, larger distances, and accordingly often higher ti/tv ratios (see Waddell *et al.*, 1997a).

Constant Site Removal and Invariant Sites Distributions

Not accounting for invariant (here equals invariable) sites leads to distortions of estimated distances (Shoemaker and Fitch, 1989), reduced statistical efficiency (Hasegawa and Fujiwara, 1993), and sometimes inconsistency of tree selection (Waddell, 1995; Lockhart *et al.*, 1996). Failure to identify and account for the distinct base composition of invariant sites relative to the base composition of variable sites can, in itself, also lead to inconsistency (Waddell, 1995). Invariant sites are often ambiguously identified, so they cannot be directly edited out of the sequences. An effective alternative modification of the **F** matrices is possible, given estimates of the overall proportion of invariant sites (p_{inv}) and the proportions of the invariant base (in the diagonal matrix $\mathbf{\Pi}_{\text{inv}}$). So we arrive at $\mathbf{F}_{\text{var}} = (\mathbf{F} - p_{\text{inv}}\,\mathbf{\Pi}_{\text{inv}})/(1 - p_{\text{inv}})$, where $\mathbf{F}_{\text{var}}$ is the **F** matrix of just the variable sites (Waddell, 1995).

There are a number of ways to specify $\mathbf{\Pi}_{\text{inv}}$, the base composition of the invariant sites: (1) The invariant sites will have base composition equal in the four bases, i.e., $\boldsymbol{\pi}_{\text{inv}} = [0.25, 0.25, 0.25, 0.25]$. (2) The invariant sites reflect the base composition of the sequence as a whole. This can be estimated in two ways, as either $\boldsymbol{\pi}_{\text{inv}} = \boldsymbol{\pi}$ (for a particular **F** matrix) or, more logically, as the average base composition across all sequences. (3) Often $\mathbf{\Pi}_{\text{inv}}$ is better reflected in the base composition of the sites which are unvaried or constant (e.g., as tested on the Horai data in the legend to Fig. 2 and usually most sensibly estimated from the sites constant across all sequences). (4) Direct optimization of entries in $\boldsymbol{\pi}_{\text{inv}}$ by ML or some other criteria (i.e., separately optimize the base compositions of the varied and unvaried sites). This last option is most computationally intensive, but desirable (Waddell, 1995). The first three ways of making these modifications have been implemented in PAUP 4.0 (Swofford, 1997).

An interesting feature is that the more the base compositions of the varied sites and the unvaried sites differ, often the more pronounced the amount of correction made when π_{inv} is inaccurately estimated (i.e., overestimation of distances if the model were to hold exactly). To avoid this we tend to prefer estimating π_{inv} as the base composition of the sites which are constant across all sequences (via method 3 or 4).

The term "constant site removal" (CSR) modification for these modification seems appropriate, since they can also be invoked to give more robust distance estimates when the site to site rate variation follows a continuous distribution (see examples in Waddell, 1995). *Three CSR methods are used later in data analysis:* CSR(F) (a form of method 2), where the base composition of the invariant sites are estimated separately as $\boldsymbol{\pi}$ for each pairwise **F** matrix; CSR(all sites) (again a variant of method 2), where $\boldsymbol{\pi}_{\text{inv}}$ is estimated as the unweighted averaged across all sequences; and CSR-(cons.) (method 3), where $\boldsymbol{\pi}_{\text{inv}}$ is estimated from the sites constant across all taxa.

The various CSR modifications are easily made before forming $\mathbf{F}_{\text{var}}$ from the observed paired nucleotide counts. However, with $\boldsymbol{\pi}_{\text{inv}}$ estimated at $\boldsymbol{\pi}$ for each pair of sequences, the modification to the transformation of the eigenvalues can be made by replacing *ln* [x] with *ln* [$(x - p_{\text{inv}})/(1 - p_{\text{inv}})$] or, more generally, $M^{-1}[(x - p_{\text{inv}})/(1 - p_{\text{inv}})]$ (Waddell *et al.*, 1997a). Thus, either way, it is straightforward to allow for mixed variable/invariable site distributions (e.g., Waddell, 1995; Gu *et al.*, 1995; Waddell and Penny, 1996; Waddell *et al.*, 1997a).

Note that like other forms of our distance, the CSR distance is a pairwise distance ML estimator under the nine-parameter reversible model, given p_{inv}, $\mathbf{\Pi}_{\text{inv}}$, and M^{-1}. Since any symmetrized $\mathbf{F}$ matrix with all positive eigenvalues will yield a valid $\mathbf{R}$ matrix, it is not possible to simultaneously estimate p_{inv}, $\mathbf{\Pi}_{\text{inv}}$, or M^{-1}. All these "site rate" parameters may be estimated using an appropriate ML model applied to three or more sequences. Waddell (1995) uses eight different methods to estimate some of these parameters. ML or a parsimony-based approximation is often best (e.g., easiest to implement and robust) if all parameters are to be estimated simultaneously, while generalized least-squares fitting of the distances to a tree has attractions also (Waddell, 1995; Waddell *et al.*, 1997b).

ML Estimators and Iterated ML Distances

Simulations and analytic calculations with Eq. (2) show it to yield an ML estimate of the true distance given just $\hat{\mathbf{F}}$ and a distribution of rates across sites. By ML estimate, we mean the distance which will minimize the G^2 statistic between $\mathbf{F}$ and $\hat{\mathbf{F}}$ when all entries in $\mathbf{R}$ (i.e., both components of $\mathbf{R} = \mathbf{\Pi}^{-1}\mathbf{S}$) are simultaneously optimized (a formal proof is given in Waddell, 1997). ML estimators often have the desirable property that as c becomes large, they have the minimum possible sampling variance of all estimators for that model. Consistent with this, under models known to have ML distance estimators (e.g., those of Jukes and Cantor, 1969; kimura, 1980, 1981, as shown in Saitou, 1990), Eq. (5) (see below) returns identical variance estimates to the delta method variances of these estimators (Waddell, 1997). In our simulations and bootstrap analyses, Eq. (2) often has less variance than distance estimators which are now known not to be ML estimators (Waddell, 1995). These include the three-parameter distance of Tamura (1992) and the six-parameter distance of Tamura and Nei (1993) (see Zharkikh, 1994). In this sense Eq. (2) is often a better distance to use for estimating evolutionary trees than these estimators, especially when distances become larger and/or base composition unequal.

It is also possible to use Eq. (4) to predict what the expected divergence matrix is, given any time-reversible model and any distribution of rates across sites. That is, the expected divergence matrix is just $\mathbf{F}_{\text{exp}} = \mathbf{\Pi}\mathbf{P}_{\text{exp}} = \mathbf{\Pi}M[\mathbf{R}\tau]$ (here $\mathbf{R}$ is linearly scaled so that all off-diagonal entries sum to 1, making variable τ the distance in expected number of substitutions). This allows the likelihood of the observed pairwise data, $c\hat{\mathbf{F}}$, to be calculated and then optimized. Felsenstein (1993) uses iterated ML distances with the more specific i.r. Kimura 2P (1980) and Felsenstein (1984) distances (in the program DNADIST). Note that with the new equations, all parameters relating to site rates must be supplied by the user from elsewhere, as explained earlier. If you wish to add invariant sites with a base composition distinct from the variable sites, then $\mathbf{F}_{\text{exp}}$ becomes $\mathbf{F}_{\text{exp}}(1 - p_{\text{inv}}) + p_{\text{inv}}\mathbf{\Pi}_{\text{inv}}$, and M with its associated shape parameters, p_{inv}, and $\mathbf{\Pi}_{\text{inv}}$ must all be supplied from elsewhere.

This approach will always give lower sampling errors of the distance (about the sample mean) than those made by Eq. (2). It is often reasonable to assume a fully homogeneous and reversible model. In this circumstance, if the unknown parameters in $\mathbf{R}$ can be estimated by a statistically efficient method (such as ML applied to a set of sequences) and then fixed for all iterated pairwise distance estimates, then not only the sampling errors but the absolute error about the true value will decrease relative to Eq. (2). However, if there are factors such as a truly variable ti/tv ratio, then even if these violate the reversible model assumption, use of Eq. (2) may be more robust (i.e., give smaller errors about the true distances). Equation (2) will also allow detection of violations (as in Fig. 3 below), while homogenization of $\mathbf{R}$ will prevent their identification.

Reducing the number of parameters in $\mathbf{R}$ will also reduce sampling variance. Further, if the differences in the parameters homogenized are small enough (e.g., when entries are statistically indistinguishable), and the reversibility assumptions hold, then this will reduce total error not just about the sample mean but about the true distance values also. Should it be decided to use iterated ML distances based on, say, the Tamura (1992) model, then again we have a choice: (I) to homogenize $\mathbf{R}$ for all distance estimates; (II) to estimate the Tamura (1992) form of $\mathbf{R}$ independently for each pair of species. The former can be done with ML (Felsenstein, 1982) based on more than three species at a time, while the latter can be done using the previous equations for $\mathbf{F}_{\text{exp}}$ for each pair of sequences separately.

Thus, the user needs to decide, first, whether to fix $\mathbf{R}$ for all comparisons (and reduce sampling error) or to estimate it via iterated ML for each comparison (to improve robustness), and second, whether to use all possible free parameters (for robustness) or to go for a reasonable submodel to reduce sampling error. A subfamily of iterated ML distances based on Eq. (4) is implemented in PAUP 4.0 (Swofford, 1997), applicable when $\mathbf{R}$ is fixed, while a non-ML solution (Lewis and Swofford, 1997) is used when $\mathbf{R}$ has less than its full generality and is not fixed for all comparisons. (However, it is preferable, regarding sampling errors, to use iterated pairwise ML instead of other solutions, as already mentioned here and in Waddell, 1997.) For related issues of choosing a suitable distance, see Waddell (1997) and Waddell *et al.* (1997b).

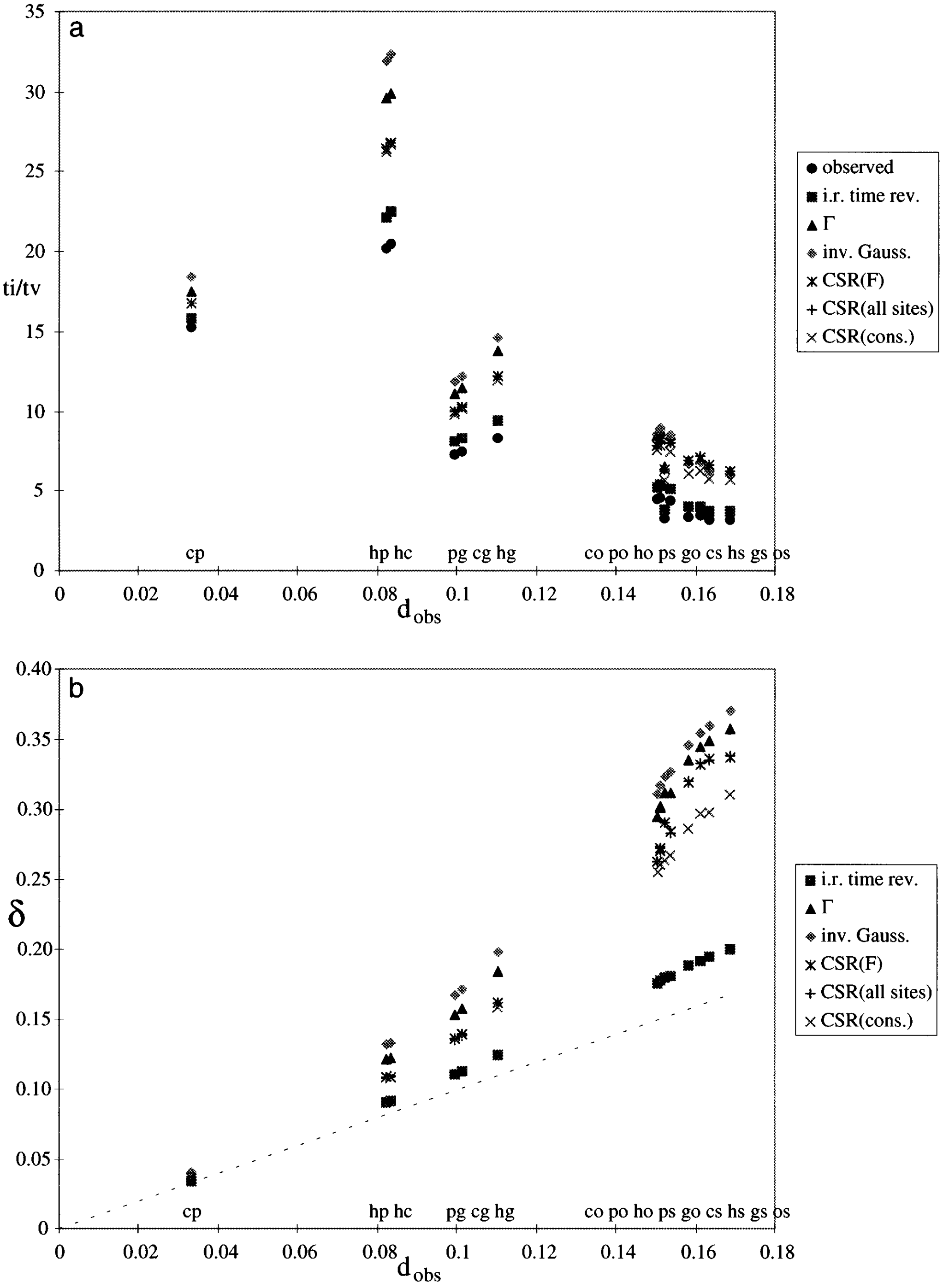

FIG. 3. (a) Transition versus transversions ratios and (b) total distance with rates across sites time-reversible distances versus the observed (Hamming) distance (d_{obs}). The pair of sequences being compared is indicated in rank order by the symbols c (chimpanzee), p (pygmy chimp), h (human), g (gorilla), o (orangutan), and s (siamang). Using simpler distances, such as that of Kimura (2ST), yields even lower ti/tv ratios than the i.r. general time-reversible used in (a). The dotted line in (b) shows the expected distance if there was no correction for multiple hits being made. The addition of unequal rates across sites has more than tripled the number of "corrections" being made, relative to the standard i.r. time-reversible distances.

A Computational Example

Application of Eq. (2) to the mtDNA sequences of Horai *et al.* (1992) allows correction for an unequal distribution of rates across sites, a very high transition to transversion ratio, and a skewed base composition (see legend to Fig. 2) of mammalian mtDNA. To illustrate some consequences of using Eq. (2) and different distributions of rates across sites, the same human–chimp comparison as described in the legend to Fig. 1 is used. The shape parameter for each distribution is calculated with maximum likelihood (ML) tree estimation on sequences allowing for a distribution of rates across sites, using the ML methods and models of Waddell and Penny (1996) (for a proof of such site likelihood calculations see Steel *et al.*, 1993; Yang, 1993; or Waddell *et al.*, 1997a). Using the generalized Kimura 3ST model is both computationally convenient and unlikely to result in overestimates of the spread of site rates (Waddell, 1995).

With the edited Horai *et al.* (1992) data, the ML sequence based method estimated the shape parameter of the inverse Gaussian distribution as $d = 0.213$. The optimal fit of data to model, measured by the likelihood ratio, G^2 (Ritland and Clegg, 1987; Stuart and Ord, 1991, p. 1160) was 334.8 (P. J. Waddell, unpublished). For the Γ distribution, k was estimated to be 0.351 and G^2 was 303.2, while an invariant sites/i.r. distribution yielded $p_{\mathrm{inv}} = 0.592$ and the best fit of G^2 at 279.4 (Waddell and Penny, 1996). Allowing a mixture of invariant sites with either a Γ or an inverse Gaussian distribution did not further improve the fit under the Kimura 3P model (Waddell and Penny, 1996; Waddell, 1995). [However, it does when using the same data but more general models; e.g., allowing unequal base composition (P. J. Waddell, unpublished data).]

As Fig. 2 and Table 1 show, taking a distribution of rates across sites into account increases substantially the estimated distance between these sequences. Importantly, the inferred distance is dependent upon the assumed distribution, here being minimal with the invariant sites models and largest with the inverse Gaussian distribution. As the distance between sequences increases, the importance of the site rate distribution becomes more pronounced (see later, Fig. 3).

A partial solution to the uncertainty as to which distribution to use is to ignore those models which have a substantially lower likelihood than the optimal model. Here, this would suggest ignoring the Γ, inverse Gaussian, and i.r. estimates. This solution is only partial since our optimal model may still not be the true model, which might suggest a distinct distribution of site rates and could therefore infer quite different distances again. Additionally, the model being used to measure likelihoods may be critically deficient in some way and consequently may not be accurately measuring the rank of models attributable to the distribution of rates over sites. Third, we implicitly assume the relative rates of sites are fixed, but they are unlikely to be (e.g., a covarion model like that of Fitch and Markowitz, 1970, seems more likely; Waddell and Penny, 1997a), so all fixed site rate estimates could be severely misleading at larger distances. Different forms of site rate distribution also give distinct ratios of transitions to transversions. In this case the largest ratio for the human–chimp comparison was with the inverse Gaussian distribution, consistent with this distribution having the flattest tail (described later).

Two distinct invariant sites models are evaluated in Fig. 2. While both give similar estimates of δ, the matrix $c\mathbf{\Pi R}_{\tau}$ shows this is partly coincidence. This second model is suggesting there is a smaller proportion of variable sites with A or G than there is with C or T. This in turn leads to more corrections of AG transitions and fewer for CT transitions, but coincidentally these two effects nearly cancel (they do not always, as we see later in Fig. 3). We do not show a mixed invariant sites/continuous distribution (such as CSR with Γ), since for these data such a mixture did not improve the likelihood of the models considered (Waddell and Penny, 1996). However, the properties of such corrections (e.g., inferred distance or ti/tv ratio) are studied and tend to be near an average of their component parts (Waddell, 1995, and unpublished).

Overall then, the type of distribution of rates across sites is an important parameter for gaining both more precise estimates of absolute distances (and hence exact edge lengths on trees) and ti/tv ratios. While here the invariant sites model fits the data better than either a Γ distribution or an inverse Gaussian distribution, this is not always the case (Waddell, 1995; Waddell *et al.*, 1997a).

TABLE 1

Distances and ti/tv Estimates with Different Distributions of Rates across Sites

Rate distribution	δ_{hc}	Increase over i.r.	ti/tv	Increase observed	Ratio tr(AG)/tr(CT)	Percentage of multiple hits in tv's
i.r.	0.0915	—	22.50	9.9%	0.477	0.4%
Inverse Gaussian	0.1327	45.0%	32.34	58.0%	0.435	2.6%
Γ	0.1221	33.4%	29.90	46.1%	0.437	1.8%
CSR(F)	0.1090*	19.1%	26.77	30.8%	0.448	1.2%
CSR(cons.)	0.1087*	18.7%	26.68	30.3%	0.488	1.2%

A Delta Method Approximation of the Variance

The previous sections have dealt with estimating the expected distance between a pair of sequences, but the variances of these estimates are also important. Barry and Hartigan (1987) derive a delta method approximation to the variance of the i.r. time-reversible distance. Extending this approach to allow unequal rates across sites [i.e., Eq. (2)] gives the following result (proven in Appendix 4),

$$Var\ [\hat{\delta}] \approx \frac{1}{c}\left[\sum_{k=1}^{4} \pi_k \left(R_{kk} - \sum_{i=1}^{4} \pi_i R_{ii}\right)^2 + \sum_{k=1}^{4} \pi_k \left\{\sum_{l=1}^{4} P_{kl}\left(G_{kl} - \sum_{j} P_{kj} G_{kj}\right)^2\right\}\right] + O(c^{-2}), \quad (5)$$

where (G_{kl}) are elements of the matrix

$$G = -\sum_{r=1}^{\infty} a_r \sum_{s=0}^{r-1} \mathbf{B}^s (\mathbf{B}^t)^{r-1-s}, \quad \mathbf{B} = \mathbf{I} - \mathbf{P},$$

where $\mathbf{B}^t$ is the transpose of $\mathbf{B}$, while $\mathbf{P} = \mathbf{\Pi}^{-1}\mathbf{F}^{\#}$, $\mathbf{R} = M^{-1}[\mathbf{P}]$ (replaced by their estimates when working from a finite sample) and c is the sequence length. Simulations (see Waddell *et al.*, 1997b) confirm the accuracy and consistency of Eq. (5).

The term in matrix **G** which changes given different assumed distributions of rates across sites is a_r. This term is given by the coefficients of the Taylor series expansion of $M^{-1}[1 - x] = -\Sigma_i\, a_i x^{\,i}$. For example, in the case of the i.r. distribution, a_r is the coefficient of x^r in $-ln\,(1 - x)$, and so $a_r = 1/r$ (i.e., $a_1 = 1$, $a_2 = 1/2$, $a_3 = 1/3, \ldots$). For the Γ distribution with shape parameter k the series becomes, $a_r = [(k + 1)\ (2k + 1) \ldots ((r - 1)k + 1)]/r!k^{r-1})$, for example, with shape parameter $k = 0.351$, $a_1 = 1/(1 \times 0.351^0) = 1$; $a_2 = (0.351 + 1)/(2!(0.351^1)) = 1.92$; $a_3 = [(0.351 + 1)(0.702 + 1)]/(3!(0.351^2)) = 3.11$. As k goes to infinity, this series converges to that for the i.r. distribution, as expected.

For the inverse Gaussian distribution,

$$a_r = \frac{1}{r} + \frac{1}{2d} \sum_{m=1}^{r-1} \frac{1}{m(r - m)},$$

as derived in Appendix 5. As an example, with shape parameter $d = 0.213$, this gives

$$a_1 = 1 + 1/0.426 \times \sum_{m=1}^{l-1} \frac{1}{m(1 - m)} = 1 + 0$$

(since the summation does not take effect) = 1;

$$a_2 = 1/2 + 1/0.426 \times \sum_{m=1}^{2-1} \frac{1}{m(2 - m)} = 1/2 + 1/0.426 \times (1) = 2.85;$$

$a_3 = 1/3 + 1/0.426 \times (1/2 + 1/2) = 2.68$; $a_4 = 1/4 + (1/3 + 1/4 + 1/3)/0.426 = 0.250 + 1.49 = 2.40$, etc.

If we look at the terms for this distribution, there are two parts. The first part, i.e., $1/r$, is the same as the standard i.r. log transform, while the second part can be thought of as extra uncertainty due to unequal rates. As $d \to \infty$, this second term goes to zero, and the variance converges to that of the i.r. model, as expected.

With the CSR distances, the easiest way to make the computation of this variance is to redefine **P**, **R**, and **Π** as $\mathbf{P}_{\mathrm{CSR}}$, $\mathbf{R}_{\mathrm{CSR}}$, and $\mathbf{\Pi}_{\mathrm{CSR}}$, i.e., their values after the removal of constant sites. For the mixed invariant sites/Γ distribution (Gu *et al.*, 1995; Waddell, 1995; Waddell and Penny, 1996; Waddell *et al.*, 1997a) or a mixed invariant sites/inverse Gaussian distribution (Waddell, 1995; Waddell *et al.*, 1997a) do as for the CSR distances, except apply the appropriate power series for the term a_r in Eq. (5).

Application of Eq. (5) shows that the major cost in calculating $Var\ [\hat{\delta}]$ is the evaluation of matrix **G**. For example, in the HC comparison, **G** calculated with r up to 4 is accurate to the third decimal place due to a fourfold rate of decrease of the products involving matrix **B**. However, large distance comparisons, e.g., to siamang, require more (here 11) terms for the same accuracy. Premature truncation biases the result of Eq. (5) toward underestimation of the variance.

The increase of the standard error with unequal rates across sites can be substantial. With the i.r. time reversible distance the estimated HC distance has a standard error of 0.0048, whereas the CSR(F) distribution of site rates ($p_{\mathrm{inv}} = 0.592$) gives a SE of 0.0066, a slight decrease in accuracy. Assuming the Γ ($k = 0.351$) increases the inferred standard error to 0.00837, the inverse Gaussian gives 0.00915. In contrast, the simpler i.r. Kimura (1980) 2ST (SE = 0.0047) or Jukes–Cantor (1969) (SE = 0.0044) distances give only a slight reduction in stochastic error. Thus, the distribution of rates across sites is causing a far greater increase in sampling error than going from an i.r. one-parameter to an i.r. nine-parameter model. This is should be offset by a proportionately greater reduction in bias (about the true distance) when more general distance estimates, that allow unequal rates across sites are used.

The decrease in the sampling accuracy of distances estimated taking unequal site rates into account often translates to decreased bootstrap support for many nodes in a tree. Thus, many real data sets which have been estimated under i.r. assumptions and look highly informative need to be reevaluated under more realistic assumptions. This can lead to both a collapse of support for widely accepted results and the emergence of good support for new hypotheses (e.g., Waddell, 1995; Lockhart *et al.*, 1996). Note that fixing all other input, Eq. (5) returns a monotonically increasing variance the more unequal site rates are. Thus it follows that site rate equality, as well as the form of site rate distribution, is an important criterion in assessing sequence suitability for phylogenetic analysis.

Last, as is generally the case with delta method approximations of the variance of a distance, Eq. (5) assumes that the form of the distribution, the shape parameter(s), and the base composition of invariant sites are known. The variability of distances, due to estimation of these parameters, appears best taken into account by a bootstrapping procedure which includes reestimation of the shape parameter(s) for each bootstrap sample.

Applying Time-Reversible Transformations to mtDNA from Apes and Humans

The first application is to estimate the ti/tv ratio between pairs of taxa (Fig. 3) from the six-taxon 5-kb hominoid mtDNA of Horai *et al.* (1992). When unequal rates across sites are allowed for, the inferred ratio increases substantially over either that observed or that estimated by i.r. time-reversible distances. When an inverse Gaussian distribution is assumed, the increase is from 20 to about 100% for the largest distances (comparisons to siamang).

The data in Fig. 3 show an anomalously high ti/tv ratio for the HC and HP comparisons and also less strikingly for the HG comparison (taxa abbreviations as given in Fig. 3). This suggests an increased ti/tv ratio in the human lineage, which is all the more striking as it goes against the trend of decreasing ti/tv ratio with increasing distance. This observation is consistent with the claims of Adachi and Hasegawa (1996). They arrived at their conclusion using an ML model allowing unequal ti/tv ratios in different edges of the tree. They also show a way of comparing observed distances with predicted observed distances based on the ML model. This is a useful technique also. Given the present lack of programs to implement their methods, our distance methods offer a useful alternative. Either approach can be extended to break the distances up not just into ti versus tv, but into all six independent categories in $c\hat{\mathbf{H}}\hat{\mathbf{R}}\tau$ (as shown in Fig. 1).

With regard to the overall trend of decreasing ti/tv ratio, we doubt this is really as strong as it appears; rather it is most likely an artifact of these estimators in relation to the true model of evolution. That this trend is so pronounced, yet so consistent among all the assumed distributions of rates across sites, suggests that in general it may be very difficult to obtain accurate estimates of the ti/tv transition rate across all sites when taxa are even moderately diverged. Such a plot may give an indication of how accurately a distance transformation is correcting for multiple hits among the transitions. Note how the inverse Gaussian ti/tv ratios generally tend to be larger than those from the Γ. We expect this is due to a "flat-tails effect" (Waddell *et al.*, 1997a).

The flat-tails effect states that some distributions (e.g., the "*F* distribution" relative to the lognormal) predict many more sites with the highest rates, even though the bulk of the sites in the two distributions have similar rates. This results in the more flat-tailed distribution predicting more multiple and "multiple–multiple" hits. Accordingly, the *F* distribution gives a larger correction factor for things like ti/tv ratio relative to the less flat-tailed distribution, even though overall fit may be similar.

The methods of constant site removal show another trend described in Waddell *et al.* (1997). Here, as the removal of constant sites brings the "infinite distance" (i.e., one or more of the eigenvalues tend to zero) closer (i.e., it approaches an asymptote), total distance and relative rates and ratios can experience a great increase, with relatively small changes in d_{obs}. The effect here is not dramatic, but it is expected to become more pronounced for "deeper" comparisons, e.g., human to monkey sequences.

There are way to alleviate, but not eliminate, the downward bias in estimating ratios (due no doubt to an underestimation of the total number of transitions). One is further editing of the data; e.g., separating out the three codon positions and the structural RNA coding regions. However, a plot like Fig. 3a using just third positions, while much less pronounced, still suggests a downward trend of the ratoi for larger distances (P. J. Waddell, unpublished). Another possibility, mentioned earlier, is that the shape parameters of the various distributions are underestimated. However, replotting the figure with k as low as 0.2 or d as low as 0.15 does not alter this trend. While careful data editing can reduce the bias shown, our purpose here is mainly to illustrate just how substantial this effect can be among taxa diverged for less than 20 million years. The implication is that many estimates of ti/tv ratios at older times may be of dubious accuracy.

Figure 3b shows that all of these transformations assuming unequal rates across sites are making increasingly large corrections for multiple hits compared to the i.r. methods, as the observed distance becomes larger (as expected from the proof in Appendix 3). There is also a noticeable difference between the different distributions. While the removal of constant sites makes a lesser difference initially, CSR(all sites) and CSR(F) transformations show signs of becoming suddenly larger as d_{obs} increases.

A simple way to infer the divergence time of a group (e.g., human–chimpanzee) is the ratio of a distance going through the node of interest (e.g., human–pygmy chimp) to a distance going through a node of more reliably known age (e.g., chimpanzee–orangutan). For the six taxa in this study, all such ratios are shown for all possible pairs of distances (Fig. 4a) estimated with the various distances. This figure shows the substantial, and closely agreeing, drop in divergence time estimates achieved by all the methods modeling a

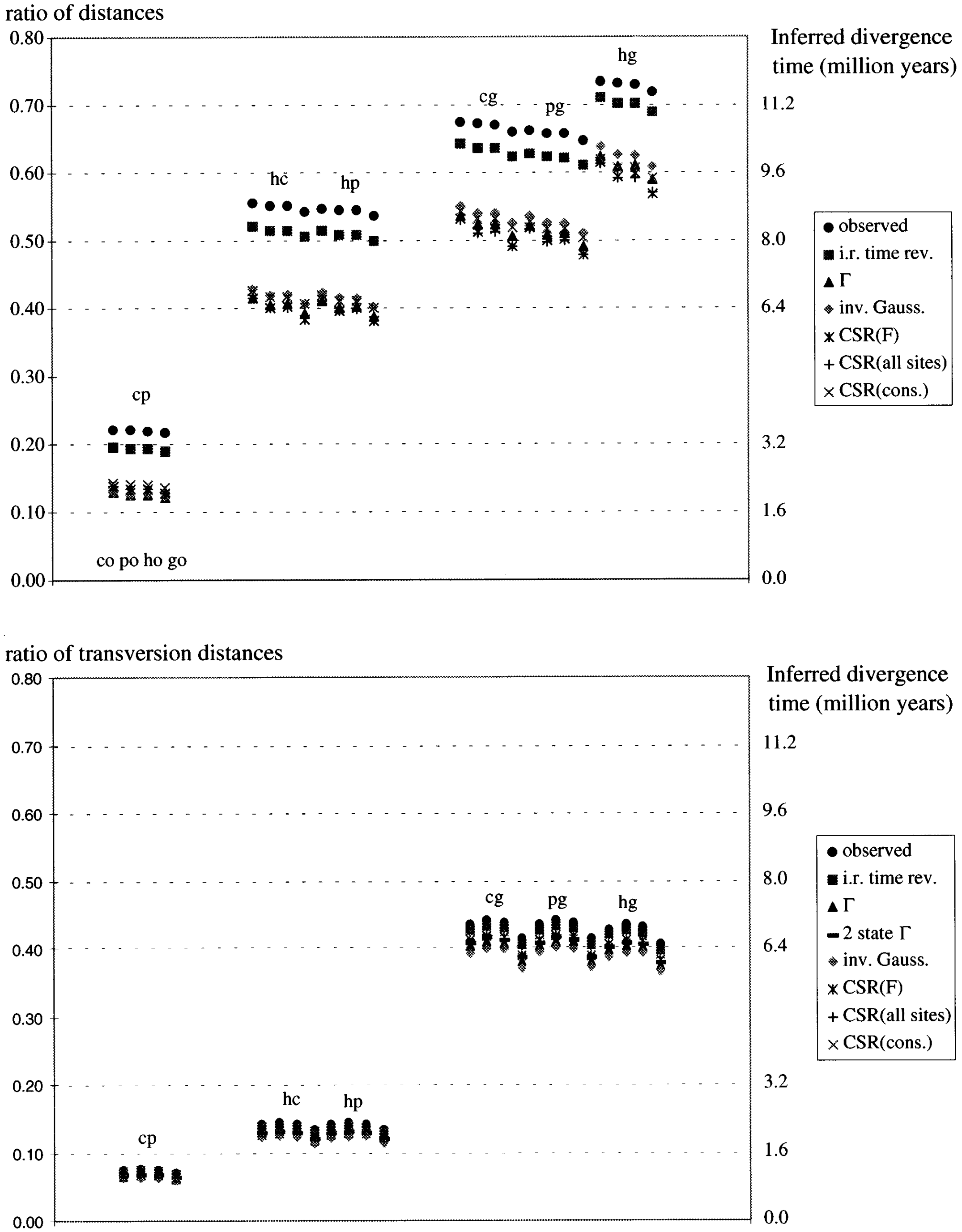

FIG. 4. (a) Estimates of divergence times using time-reversible distances counting all types of substitution and (b) just transversional changes. In (a) all possible pairs of distances (for the more recent and the older divergence) are shown (*x* axis is arbitrary). The distance used corresponding to the more recent divergence (e.g., cp) is shown above the block of four values, and the denominator for this comparison (e.g., the distance co, po, ho, or go) is shown below the first set, with the same order in all instances. (b) Follows the same pattern of pairs of distances, but uses only the transversional changes. The distance transformations are the same as (a), except for the mapping down to purines/pyrimidines (AG versus CT). The 2-state Cavender Γ distance is also shown. The divergence time is calibrated by assuming the orangutan–African apes split occurred 16 million years ago for the reasons outlined in Waddell and Penny (1996).

distribution of rates across sites. The dates fluctuate little with respect to the distance to orangutan used (due no doubt to a high correlation of paths through the tree), slightly more with respect to the use of chimpanzee or pygmy chimp sequence, but substantially when there is a choice of distance between human–gorilla and chimp–gorilla. Clearly there can be expected to be even greater fluctuations with choice of species when estimating older divergence times.

The ages on the right *y* axis are made assuming a 16-million-year-old divergence of orangutans from African apes. This is a date preferred due to biogeographic and fossil evidence as interpreted in Waddell and Penny (1996). The ML divergence dates in Waddell and Penny (1996), made under a Kimura 3P model with unequal rates across sites, fall between the i.r. and unequal rates times shown in Fig. 4a (excluding instances with the hg distance). Otherwise, the divergence times reported here tend to be proportionately older than those based on other recent ML analyses (e.g., Adachi and Hasegawa, 1996) for two main reasons: Their data were edited into a class of sites (fourfold degenerate) which had very similar substitution properties; ML has some robustness (e.g., Felsenstein and Kuhner, 1994) to pick up extra multiple hits due to its inherent ability to infer probable states at internal nodes.

Given the appearance of at least close to a molecular clock for these data (Horai *et al.*, 1992; Adachi and Hasegawa, 1994; Waddell and Penny, 1996; although see Adachi and Hasegawa, 1996), it is surprising that the divergence times estimates from just the transversion substitutions are so different (Fig. 4b). Since the model used to correct the transversion distance makes little difference to these times, they are not easily as systematically biased. Taken at face value they indicate either incredibly recent divergence times for these taxa (especially humans from chimps) or a very ancient divergence of orangutan. Either interpretation runs into conflict with the fossil evidence. While there may be an accelerated rate of transversion in the orangutan lineage (e.g., Horai *et al.*, 1995; Adachi and Hasegawa, 1996), this does not explain why the human–chimp versus human–gorilla dates are so distinct.

Importantly, the patterns reported here are still apparent to some degree in the analysis of the complete mtDNA sequences of Horai *et al.* (1995), suggesting that they are not just artifacts due to data editing or sequence length. The nature of this apparent conflict is considered further elsewhere (Waddell, in preparation). Indeed, testing whether the more conservative changes agree with the picture painted by all sites is a general and useful way of testing the coherency of a phylogenetic model. This includes tests (e.g., by likelihood ratio) of the relative lengths of edges in the trees produced by each category of site or change (Waddell, 1995). In summary, the main point of this section is to illustrate the ways these new distances may be used, especially so that they will highlight features which appear anomalous (irrespective of the ultimate cause).

A Quick "Test" of the Reversibility of a Specified Rate Matrix

Often rate matrices are derived *de novo* without being written explicitly as $\mathbf{\Pi}^{-1}\mathbf{S}$ or the equivalent form $\mathbf{S\Pi}$ (e.g., see Zharkikh, 1994, for a table of reversible and nonreversible rate matrices), so it is handy to have a quick way of checking the reversibility property. In this case, a simple "test" (or diagnosis) is that

$$\mathbf{R} = \begin{bmatrix} * & A & B & C \\ D & * & E & F \\ G & H & * & I \\ J & K & L & * \end{bmatrix}$$

(All nondiagonal elements > 0 and row sums $= 0$), defines a time reversible model if, and only if, the following three conditions hold:

$$\text{(E1) } AGE = BDH, \tag{6}$$

$$\text{(E2) } AJF = CDK, \tag{7}$$

$$\text{(E3) } EKI = FHL. \tag{8}$$

A proof of this test is given in Appendix 6.

DISCUSSION

Choosing a distance with which to estimate divergence times is not always straightforward. The only pairwise distance estimator which is linearly related to time under a stationary i.r. 12-parameter model is the LogDet (Steel, 1994; Lockhart *et al.*, 1994; Swofford *et al.*, 1996) or paralinear distance (Lake, 1994). This makes it suitable for divergence time estimates using phylogenetic trees, given these conditions, plus a molecular clock and stationary base frequencies (Waddell, 1995; Swofford *et al.*, 1996). Importantly, though, if all sequences have a base composition close to equal frequency, then Eq. (1) may well return a distance just as additive (in practice) as any under a nonreversible stationary i.r. model, with the added advantage that Eq. (2) can be used to accommodate a variety of distributions of rates across sites.

Of course, as base composition becomes unequal, but stationary, with site rates i.r., the LogDet will become more useful for estimating relative divergence times. If, in addition, site rates are unequal, the CSR-LogDet (Waddell, 1995; see also Swofford *et al.*, 1996; Swofford, 1997) may become the best measure of relative divergence times using just pairwise distances (see Waddell, 1995, for examples). A major problem with nonstationarity is that it becomes necessary to make a largely subjective judgment of how severely the molecular clock is violated and how much the LogDet distance deviates from giving an unweighted estimate of the number of substitutions per site.

An additional use of the distance in Eq. (2) is for obtaining a first approximation to the length of edges on a tree when performing ML with unequal rates across sites (e.g., Yang, 1993; Waddell and Penny, 1996; Swofford, 1997). This could be done by building a tree with a method such as weighted least-squares (e.g., Fitch and Margoliash, 1967) based on these pairwise distances (e.g., Adachi, 1995; Swofford, 1997). An alternative would be to use generalized parsimony (e.g., Sankoff, 1975) to estimate the **P** matrix for each edge in the tree. Application of the transformation $-\text{tr}\,(\mathbf{\Pi}(M^{-1}[\mathbf{P}])$ to each edge could give a revised estimate of the length of that edge, better taking into account multiple hits, prior to the first iteration with likelihood. This is a more general application of a method being implemented in PAUP* (Swofford and Rogers, in preparation; Swofford, 1997), where an important preparatory step is to assign ambiguous parsimony changes to their most likely positions; i.e., on the longest edges.

All the methods suggested here can easily be extended to fewer or to more states, e.g., purines versus pyrmidines (2 states), amino acids (20 states), the first 2 sites of protein codons (16 states or 162 entries in **F**), or all 61 nonstop codons. The basic assumptions remain the same: the process is time-reversible, or the base composition is equal-frequency, so all matrices have a double stochastic form. The use of 20 or 61 states removes much of the local correlation between site substitutions caused by the genetic code, and for this reason (plus the reduced likelihood of convergences), such distances may be preferable with more diverged sequences.

While it is mathematically nearly equivalent to treat sites that evolve very slowly as invariant, there is an important biological difference. That is, very slowly evolving sites will often contain important phylogenetic information for deep divergences, while invariant sites cannot (by definition). If the former is true, then there are times when it is much better to edit the data to remove the more rapidly evolving sites than to simply apply rates across sites distance corrections (Waddell, 1995). This "editing" approach reduces both the variances and biases of estimated distances, as the more rapidly evolving sites contribute disproportionately large amounts to both forms of error (for examples of each factor, see Waddell, 1995). This argument holds in a similar manner even if the site rate distribution is continuous (e.g., Γ or i.G.).

There is a tendency to accept that current models are adequate, often for the simple reason we like to believe something is solid and certain. Herein, as in Adachi and Hasegawa (1996), there is evidence this may not be the case even with divergences less than 20 mya. While ML and careful editing of sequences offer hope for more accurate estimates when the model is violated, we suggest that plots such as those in Figs. 3 and 4 be employed to help detect such diases (perhaps with bootstrapping to show the fluctuations expected due to sequence length). Since currently available ML programs do not generally allow detection of altering substitution rate matrices (nonhomogeneity), pairwise distance comparisons should be useful for alerting biologists to such possibilities in their data.

APPENDIX 1

(a) Proof that symmetrizing $\hat{F}$ gives the ML estimate of ***F*** *as* $\boldsymbol{F}^{\#}$*.* Since $c\mathbf{F}$ is expected to by symmetric under the model, $cF_{ij} = cF_{ji}$. As $c\mathbf{F}$ is expected to have a multinomial distribution (giving binomial marginal distributions for each entry cF_{ij}), the ML estimator of $cF_{ij} + cF_{ji}$ is $c(\mathbf{F}_{ij} + \mathbf{F}_{ji})$, so it follows that the ML estimator of F_{ij} or F_{ji} is 1/2 $(\mathbf{F}_{ij} + \mathbf{F}_{ji})$, i.e., entry $\mathbf{F}^{\#}_{ij}$. This applies jointly for all entries in $\mathbf{F}^{\#}$ since all symmetrized pairs are nonoverlapping.

(b) Proof that for any Markov process operating on a rooted tree obeying a molecular clock, ***F*** *is symmetric (without assuming that the root base composition is necessarily in equilibrium).* Under a molecular clock,

$$\mathbf{F} = \mathbf{P}^{t}\mathbf{\Pi}\mathbf{P}, \quad \text{and so } \mathbf{F}^{t} = \mathbf{P}^{t}\mathbf{\Pi}(\mathbf{P}^{t})^{t} = \mathbf{F}, \text{ as claimed.}$$

APPENDIX 2: PROOF OF EQ. (2)

We have $\mathbf{F} = E\lambda[\mathbf{\Pi} \exp[\mathbf{R}\tau\lambda]] = \mathbf{\Pi}E\lambda[\exp[\mathbf{R}\tau\lambda]] = \mathbf{\Pi}M[\mathbf{R}\tau]$, by the relationship $\mathbf{P} = M[\mathbf{R}\tau]$ of Eq. (4). Thus, $\mathbf{R}\tau = M-1[\mathbf{\Pi}-1\mathbf{F}]$, and since $\delta ij = -tr\,(\mathbf{\Pi}\mathbf{R})\tau E\lambda[\lambda]$, and we are assuming that $E\lambda[\lambda] = 1$, we have

$$\delta_{ij} = -tr\,(\mathbf{\Pi}M^{-1}[\mathbf{\Pi}^{-1}\mathbf{F}]),$$

as claimed.

APPENDIX 3: PROOF OF EQ. (4)

First we recall how the domain of the moment-generating function M is extended so that the function is defined on matrices. Namely, if $M[x] = 1 + \sum_{k=1}^{\infty} \lambda_k x^k$, then for a matrix $\mathbf{X}$,

$$M[\mathbf{X}] := \mathbf{I} + \sum_{k=1}^{\infty} \lambda_k \mathbf{X}^k.$$

We may assume, without loss of generality, that the rate matrix $\mathbf{R}$ is diagonalizable, so that we can write

$$\mathbf{R}\tau = \mathbf{A}\mathbf{D}\mathbf{A}^{-1}.$$

Then $\mathbf{P} = E_\lambda[\exp(\mathbf{R}\tau\lambda)] = E_\lambda[\mathbf{A}\ \exp(\lambda\mathbf{D})\mathbf{A}^{-1}] = \mathbf{A}E_\lambda[\exp(\lambda\mathbf{D})]\mathbf{A}^{-1}$.

Now, $\exp(\lambda\mathbf{D})$ is the diagonal matrix with iith entry $\exp(\lambda D_{ii})$; thus $E_\lambda[\exp(\lambda\mathbf{D})] = M(\mathbf{D})$, by the above definition of $M(D)$. Thus, again invoking this definition,

$$\mathbf{P} = \mathbf{A}M[\mathbf{D}]\mathbf{A}^{-1} = M[\mathbf{A}\mathbf{D}\mathbf{A}^{-1}] = M[\mathbf{R}\tau],$$

APPENDIX 4: DELTA METHOD APPROXIMATION

A delta-method variance is an approximation that becomes exact as sequences become infinitely long. As a simple example, the delta-method variance of the "transversion" distance $\delta = -\frac{1}{2}\ln[1–2d_{\text{obs}}]$ is obtained by:

(1) calculating the variance of the term in brackets. It is a scaled binomial so variance = $(1–2d_{\text{obs}})(2d_{\text{obs}})/c$;

(2) deriving a linear approximation for the effect of the power function (ln) by measuring the gradient of the log function at $1–2d_{\text{obs}}$ (equals $1/(1–2d_{\text{obs}})$ or the first derivative of the logarithmic function);

(3) scaling the variance by the first derivative squared;

(4) accommodating the initial multiplication of $-1/2$ by squaring, yielding $\text{Var}[\delta] = d_{\text{obs}}/\{2c \times (1–2d_{\text{obs}})\}$.

Barry and Hartigan (1987) do essentially the same thing, but must take into account the variance and covariance of the nine free parameters in $\mathbf{F}$ (equals six free parameters in $\mathbf{S}$ and three in $\mathbf{\Pi}$). This makes the variance more complicated in derivation, but conceptually the same. Workers such as Tamura and Nei (1993) take a shortcut by ignoring the contribution of the terms in $\mathbf{\Pi}$ (this is further explained in Waddell, 1997).

To adjust for unequal site rates, the gradient is measured for the new power function (M^{-1}). For example, with the inverse Gaussian (i.G.) distribution, the simple transversion distance becomes

$$\delta = -1/2(b\{1 - (1 - ln\,[1–2d_{\text{obs}}]/b)^2\}/2),$$

where b is the i.G. shape parameter.

For this M^{-1} function, the gradient at d_{obs} equals $1/(1–2d_{\text{obs}}) \times (1 - \ln[1–2d_{\text{obs}}]/b)$ (Waddell, 1995; Waddell *et al.*, 1997a), giving $\text{Var}[\delta] = 1/2d_{\text{obs}}(1 - \ln[1–2d_{\text{obs}}]/b)^2/(1–2d_{\text{obs}})(1 - \ln[1–2d_{\text{obs}}]/b)$. By replacing the power series of the logarithmic function in the variance formula of Barry and Hartigan (1987) with the power series of M^{-1}, we are affecting the same type of modification for unequal site rates. More formally, we have the following proof of the delta method approximation, Eq. (5), for the variance of Eq. (2). The proof is a direct extension of Barry and Hartigan's (1986) proof of the special case where $M[x] = e^x$ to a general M. In particular, by Eqs. (2) and (4), we have

$$\delta_{ij} = -tr\,(\mathbf{\Pi}M^{-1}[\mathbf{P}]) = -\sum_{r=1}^{\infty} a_r tr\,(\mathbf{\Pi}\mathbf{B}^r),$$

and the remainder of Barry and Hartigan's proof applies upon substitution of their term $1/r$ for a_r.

APPENDIX 5: DERIVATION OF A DISTRIBUTION COEFFICIENT

The term a_r is given by the equation $M^{-1}[1 - x) = -\Sigma_i\, a_i\ x^i$. For the inverse Gaussian distribution, $M^{-1}[1 - x]$ can be written as

$$y = \frac{d}{2}\left[1 - \left\{1 - \frac{ln\,[1 - x]}{d}\right\}^2\right].$$

The function

$$ln\,[1 - x] = -\sum_{i=1}^{\infty}\frac{x^i}{i} \text{ so } y = \frac{d}{2}\left[1 - \left\{1 + \frac{1}{d}\sum\frac{x^i}{i}\right\}^2\right]$$

$$= -\sum\frac{x^i}{i} - \frac{1}{2d}\left(\sum\frac{x^i}{i}\right)^2.$$

Thus

$$a_i = \frac{1}{i} + 2d\sum_{j=1}^{i-1}\frac{1}{j(i - j)}.$$

APPENDIX 6

Proof of the "test" of time reversibility via equalities $E1$–$E3$ (marked as Eqs. (6) to (8)).

It is easily checked that a rate matrix

$$\mathbf{R} = \begin{bmatrix} * & A & B & C \\ D & * & E & F \\ G & H & * & I \\ J & K & L & * \end{bmatrix}$$

forms a reversible model precisely if we can write $\mathbf{R}$ in form (2), where

$$\mathbf{R} = \begin{bmatrix} * & x_1a & x_1b & x_1c \\ x_2a & * & x_2d & x_2e \\ x_3b & x_3d & * & x_3f \\ x_4c & x_4e & x_4f & * \end{bmatrix}.$$

Here, all entries are positive, except the "*" entries, which are chosen so that each row sums to 0. But then $AGE = (x_1a)(x_3b)(x_2d) = x_1x_2x_3abd$ and $BDH = (x_1b)(x_2a)(x_3d) = x_1x_2x_3abd$; that is, $AGE = BDH$, while similarly $AJF = CDK$ and $EKI = FHL$ ($E1$–$E3$).

Conversely, suppose the three equations hold. We show that $\mathbf{R}$ can be written in the form (2) and therefore it forms a reversible model.

Set

$$x_1 = \frac{C}{J}, \quad x_2 = \frac{EI}{HL}, \quad x_3 = \frac{I}{L}, \quad x_4 = 1, \text{ and}$$

$$a = \frac{AJ}{C}, \quad b = \frac{BJ}{C}, \quad c = J, d = \frac{HL}{I}, \quad e = \frac{FHL}{EI}, \quad f = L.$$

Then clearly $A = x_1 a$, $B = x_1 b$, $C = x_1 c$, $H = x_3 d$, $I = x_3 f$, $J = x_4 c$, $L = x_4 f$, $E = x_2 d$, $F = x_2 e$.

It remains to check that

$$D = x_2 a, \quad G = x_3 b, \quad K = x_4 e.$$

We have

$$x_2 a = \frac{EI}{HL}\frac{AJ}{C} \overset{(E3)}{=} \frac{FAJ}{KC} \overset{(E2)}{=} \frac{CDK}{KC} = D,$$

$$x_3 b = \frac{I}{L}\frac{BJ}{C} \overset{(E1)}{=} \frac{AGEIJ}{DHLC} \overset{(E2)}{=} \frac{GEIK}{HLF} \overset{(E3)}{=} \frac{GFHL}{HLF} = G,$$

$$x_4 e = \frac{FHL}{EI} \overset{(E3)}{=} \frac{EKI}{EI} = K,$$

as required, completing the proof.

Note. If $\mathbf{R}$ is reversible, then we can write $\mathbf{R} = \mathbf{Q\Pi}$ (for some symmetric $\mathbf{Q}$) in place of $\mathbf{R} = \mathbf{\Pi}^{-1}\mathbf{S}$.

Proof. Set $\mathbf{Q} = \mathbf{R\Pi}^{-1}$. We need to show $\mathbf{Q} = \mathbf{Q}^t$. Since $\mathbf{\Pi R} = \mathbf{R}^t\mathbf{\Pi}$, if we pre- and postmultiply this equation through by $\mathbf{\Pi}^{-1}$ we get, $\mathbf{Q} = \mathbf{R\Pi}^{-1} = \mathbf{\Pi}^{-1}\mathbf{R}^t = \mathbf{Q}^t$. Thus, $\mathbf{Q} = \mathbf{Q}^t$ as claimed. Further, this shows $\mathbf{R}$ is reversible if and only if $\mathbf{R} = \mathbf{Q\Pi}$ (as Zharkikh, 1994, also notes, but with some ambiguity due to his following text and simulations).

ACKNOWLEDGMENTS

Thanks to the Smithsonian Fellowships (1994) for paying travel and stipend for P.J.W. to collaborate with David Swofford in implementing these methods into PAUP*; to the Massey University Graduate Program (1991–1994 to P.J.W.) for support deriving the main results; to David Penny for his comments, encouragement, and support in completing the manuscript (not to mention checking our Darwin reference); to the N. Z. Marsden Fund; to Masami Hasegawa for kindly supplying the data sets used; the editor, Joe Felsenstein, and an anonymous reviewer for their valuable comments.

REFERENCES

Adachi, J., and Hasegawa, M. (1994). Time scale for the mitochondrial DNA tree of human evolution. *In* "The Origin and Past of Modern Humans as Viewed from DNA" (S. Brenner, and K. Hanihara, Eds.), pp. 46–68, World Scientific, Singapore.

Adachi, J., and Hasegawa, M. (1996). Tempo and mode of synonymous substitutions in Mitochondrial DNA of primates. *Mol. Biol. Evol.* **13:** 200–208.

Barry, D., and Hartigan, J. A. (1987). Asynchronous distance between homologous DNA sequences. *Biometrics* **43:** 261–276.

Chang, J. T. (1996). Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. *Math. Biosci.* **134:** 189–215.

Churchill, G. A., von Haeseler, A., and Navidi, W. C. (1992). Sample size for phylogenetic inference. *Mol. Biol. Evol.* **9:** 753–769.

Darwin, C. (1859). "On the Origin of Species by Means of Natural Selection," J. Murray, London.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27:** 401–410.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17:** 368–376.

Felsenstein, J. (1982). Numerical methods for inferring evolutionary trees. *Quart. Rev. Biol.* **57:** 379–404.

Felsenstein, J. (1984). Distance methods for inferring phylogenies: A justification. *Evolution* **38:** 16–24.

Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package) and Manual, Version 3.5c. Department of Genetics, University of Washington, Seattle.

Felsenstein, J., and Churchill, G. A. (1996). A hidden Markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* **13:** 93–104.

Fitch, W. M., and Margoliash, M. (1967). Construction of phylogenetic trees. *Science* **155:** 279–284.

Fitch, W. M., and Markowitz, E. (1970). An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem. Genetics* **4:** 579–593.

Gaut, B. S., and Lewis, P. O. (1995). Success of maximum likelihood in the four-taxon case. *Mol. Biol. Evol.* **12:** 152–162.

Gillespie, J. H. (1986). Rates of molecular evolution. *Annu. Rev. Ecol. Syst.* **17:** 637–665.

Golding, G. B. (1983). Estimates of DNA and protein sequence divergence: An examination of some assumptions. *Mol. Biol. Evol.* **1:** 125–142.

Gu, X., Fu, Y.-X., and Li, W.-H. (1995). Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol. Biol. Evol.* **12:** 546–557.

Hasegawa, M., and Fujiwara, M. (1993). Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. *Mol. Phyl. Evol.* **2:** 1–5.

Hendy, M. D., and Penny, D. (1989). A framework for the quantitativ study of evolutionary trees. *Syst. Zool.* **38:** 297–309.

Hendy, M. D., Penny, D., and Steel, M. A. (1994). Discrete Fourier analysis for evolutionary trees. *Proc. Natl. Acad. Sci. USA* **91:** 3339–3343.

Hillis, D. M., Mable, B. K., and Moritz, C. (1996). Applications of molecular systematics. *In* "Molecular Systematics" (D. M. Hillis, C. Moritz, and B. K. Mable, Eds.), 2nd ed., pp. 515–543, Sinauer, Sunderland, MA.

Horai, S., Satta, Y., Hayasaka, K., Kondo, R., Inoue, T., Ishida, T., Hayashi, S., Takahata, N. (1992). Man's place in the Hominoidea revealed by mitochondrial DNA genealogy. *J. Mol. Evol.* **35:** 32–43.

Horai, S., Hayasaka, K., Kondo, R., Tsugane, K., and Takahata, N. (1995). Recent origin of modern human revealed by complete sequences of hominoid mitochondrial DNAs. *Proc. Natl. Acad. Sci. USA* **92:** 532–536.

Jin, L., and Nei, M. (1990). Limitations of the evolutionary parsimony method of phylogenetic analysis. *Mol. Biol. Evol.* **7:** 82–102.

Jukes, T. H., and Cantor, C. R. (1969). Evolution of protein molecules. *In* "Mammalian Protein Metabolism" (H. Munro, Ed.), pp. 21–132, Academic Press, New York.

Keilson, J. (1979). "Markov Chain Models—Rarity and Exponentiality," *Appl. Math. Sci.*, Vol. 28, Springer-Verlag, New York.

Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16:** 111–120.

Kimura, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA* **78:** 454–458.

Lake, J. A. (1994). Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. *Proc. Natl. Acad. Sci. USA* **91:** 1455–1459.

Lanave, C., Preparata, G., Saccone, C., Serio, G. (1984). A new method for calculating evolutionary substitution rates. *J. Mol. Evol.* **20:** 86–93.

Lewis, P. O., and Swofford, D. L. (1996). A general method for estimating evolutionary distances under any specific time reversible model. [Submitted]

Lockhart, P. J., Larkum, A. W., Steel, M. A., Waddell, P. J., Penny, D. (1996). Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis. *Proc. Natl. Acad. Sci. USA* **93:** 1930–1934.

Lockhart, P. J., Steel, M. A., Hendy, M. D., Penny, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.* **11:** 605–612.

Olsen, G. J. (1987). The earliest phylogenetic branchings: Comparing rRNA-based evolutionary trees inferred with various techniques. *Cold Spring Harbor Symp. Quant. Biol.* **52:** 825–837.

Penny, D., Hendy, M. D., and Steel, M. A. (1992). Progress with methods for constructing evolutionary trees. *Trend. Ecol. Evol.* **7:** 73–79.

Read, T. R. C., and Cressie, N. A. C. (1988). "Goodness-of-Fit Statistics for Discrete Multivariate Data," Springer-Verlag, New York.

Reeves, J. H. (1992). Heterogeneity in the substitution process of amino acid sites of proteins coded for by Mitochondrial DNA. *J. Mol. Evol.* **35:** 17–31.

Ritland, K., and Clegg, M. T. (1987). Evolutionary analysis of plant DNA sequences. *Am. Nat.* **130**(Suppl.): S74–S100.

Rodríguez, F., Oliver, J. L., Marin, A., and Medina, J. R. (1990). The general stochastic model of nucleotide substitution. *J. Theor. Biol.* **142:** 485–501.

Saitou, N. (1990). Maximum likelihood methods. *In* "Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences" (R. F. Doolittle, Ed.), Methods in Enzymology, Vol. 183, pp. 584–598, Academic Press, San Diego.

Sankoff, D. (1975). Minimal mutation trees of sequences. *SIAM J. Appl. Math.* **28:** 35–42.

Shoemaker, J. S., and Fitch, W. M. (1989). Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. *Mol. Biol. Evol.* **6:** 270–289.

Sidow, A., Nguyen, T., and Speed, T. P. (1992). Estimating the fraction of invariable codons with a capture-recapture method. *J. Mol. Evol.* **35:** 253–260.

Steel, M. A. (1994). Recovering a tree from a leaf colourations it generates under a Markov model. *Appl. Math. Lett.* **7:** 19–23.

Steel, M. A., Székely, L., Erdös, P. L., and Waddell, P. J. (1993). A complete family of phylogenetic invariants for any number of taxa under Kimura's 3ST model. *N. Z. J. Bot.* **31:** 289–296.

Stuart, A., and Ord, J. K. (1987). "Kendall's Advanced Theory of Statistics," 5th ed., Vol. 1, Edward Arnold, London.

Stuart, A., and Ord, J. K. (1991). "Kendall's Advanced Theory of Statistics, Vol. 2, Distribution Theory: Classical Inference and Relationship," 5th ed., Edward Arnold, London.

Swofford, D. L. (1997). PAUP* Version 4.0, Sinauer, Sunderland, MA (In press). (Note: cited calculations made with test versions from 1994 to 1995.)

Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1996). Phylogenetic inference. *In* "Molecular Systematics" (D. M. Hillis, C. Moritz, and B. K. Mable, Eds.), pp. 407–514, Sinauer, Sunderland, MA.

Tajima, F. (1993). Unbiased estimation of evolutionary distance between nucleotide sequences. *Mol. Biol. Evol.* **10:** 677–688.

Tamura, K. (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G + C-content biases. *Mol. Biol. Evol.* **9:** 678–687.

Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10:** 512–526.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci.* **17:** 57–86.

Waddell, P. J. (1995). Statistical methods of phylogenetic analysis: Including Hadamard conjugations, LogDet transforms, and maximum likelihood. Ph.D. thesis, Massey University.

Waddell, P. J. (1997). Improved distance and variance estimation based on the VR-GTR transformation. [Submitted]

Waddell, P. J., and Penny, D. (1996). Evolutionary trees of apes and humans from DNA sequence. *In* "Handbook of Human Symbolic Evolution" (A. J. Lock and C. R. Peters, Eds.), pp. 53–73, Oxford Univ. Press, Oxford.

Waddell, P. J., Penny, D., and Moore, T. (1997a). Extending Hadamard conjugations to model sequence evolution with variable rates across sites. *Mol. Phylogenet. Evol.* **8**:33–50.

Waddell, P. J., and Steel, M. A. (1996). General time reversible distances with unequal rates across sites. Research Report 143, Dept. of Mathematics, University of Canterbury, New Zealand. [For an electronic copy, refer to the directory "waddell" at "onyx.-si.edu" (the "PAUP" site)]

Waddell, P. J., Lewis, P. O., and Swofford, D. L. (1997b). Distance based methods of inferring evolutionary trees. *In* Phylogenetic Analysis Under Parsimony, Version 4.0 (PAUP* 4.0), Computer Program (by D. L. Swofford)," chap. 4, Sinaur, Sunderland, MA (in press).

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10:** 1396–1401.

Yang, Z. (1994). Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* **39:** 105–111.

Zharkikh, A. (1994). Estimation of evolutionary distances between nucleotide sequences. *J. Mol. Evol.* **39:** 315–329.