Available online at www.sciencedirect.com
SCIENCE DIRECT®
Molecular Phylogenetics and Evolution 32 (2004) 495–503
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

A comparison of methods for estimating the transition:transversion ratio from DNA sequences

A.K. Kristina Strandberg and Laura A. Salter*

Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131, USA

Received 7 July 2003; revised 29 November 2003
Available online 5 March 2004

Abstract

Estimation of the ratio of the rates of transitions to transversions (TI:TV ratio) for a collection of aligned nucleotide sequences is important because it provides insight into the process of molecular evolution and because such estimates may be used to further model the evolutionary process for the sequences under consideration. In this paper, we compare several methods for estimating the TI:TV ratio, including the pairwise method [TREE 11 (1996) 158], a modification of the pairwise method due to Ina [J. Mol. Evol. 46 (1998) 521], a method based on parsimony (TREE 11 (1996) 158), a method due to Purvis and Bromham [J. Mol. Evol. 44 (1997) 112] that uses phylogenetically independent pairs of sequences, the maximum likelihood method, and a Bayesian method [Bioinformatics 17 (2001) 754]. We examine the performance of each estimator under several conditions using both simulated and real data.

Keywords: Transition/transversion ratio; Transition bias; Rate variation; Maximum likelihood; Bayesian estimation; Independent contrasts

1. Introduction

It is a well-known fact that during DNA sequence evolution the rate of transitional changes differs from the rate of transversional changes, with transitions generally occurring more frequently than transversions. This difference is often referred to as transition bias, and estimation of the extent of transition bias may be of interest, since it may vary for different organisms and for different genes within a collection of organisms. Because knowledge of this quantity may aid in our understanding of the patterns of molecular evolution, reliable methods of estimating the transition bias are needed. Proper estimation is also important because the ratio of the rates of transitional to transversional changes (often called the TI:TV ratio) plays a role in evolutionary distance correction methods and is used in several common evolutionary models (e.g., the F84 model) (Wakeley, 1996).

Several methods have been proposed for estimating the TI:TV ratio from a collection of N aligned DNA sequences. These can be broadly grouped into three categories: distance-based methods that use only pairwise distance measures in forming the estimate, parsimony methods that use a most parsimonious (MP) reconstruction of the phylogenetic tree to aid in forming the estimate, and methods that make use of the likelihood function for estimation. We give a brief overview of several of the available methods here.

Among the available distance methods are the pairwise method (Wakeley, 1996) and a method proposed by Ina (1998, Eq. (11)) (here referred to as Ina's method). In the pairwise method, the overall numbers of transitions and transversions are obtained for each pair of sequences in the data set under consideration, and a ratio is taken for each pair. Since there are $N(N-1)/2$ possible pairwise comparisons, the final number of estimates is $N(N-1)/2$. Generally, the average of these pairwise estimates is used to provide a single estimate for the entire data set, although this need not be the case. Pollock and Goldstein (1995) proposed a modification of the pairwise method in which a particular weighted average of the pairwise estimates is taken in an attempt to correct for multiple substitutions. Ina's method is similar to the general pairwise method, except that the

* Corresponding author. Fax: 1-505-277-5505.
E-mail addresses: kickilin@stat.unm.edu (A.K.K. Strandberg), salter@stat.unm.edu (L.A. Salter).

1055-7903/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2004.01.013

numbers of transitions and transversions are added for the entire data set first, and a single ratio of the number of transitions to transversions is obtained.

The first step in estimating the TI:TV ratio in the parsimony method (Wakeley, 1996) is to find the MP reconstruction of ancestral states at the internal nodes of a given (or estimated) phylogenetic tree. The numbers of transitional and transversional changes are then counted along each branch of the tree and an estimate of the TI:TV ratio is obtained by taking the ratio of the total number of transitions to the total number of transversions.

Because estimates of the TI:TV ratio may depend on the time since divergence between the sequences (Purvis and Bromham, 1997; Wakeley, 1996) developed a method that utilizes time since divergence. Their method is based on phylogenetically independent comparisons of sequences (see, e.g., Burt (1989) for a method of selecting phylogenetically independent pairs of sequences), and requires a known phylogenetic tree with known divergence times at the internal nodes. For each independent pair of sequences, the number of transversional changes, P, is plotted against time since divergence, t. A non-linear model of the form

$$P = a + (s - a)(1 - e^{-kt}) \tag{1}$$

is then fitted to the data. Under this model, the proportion of transversional changes will increase from an initial value, a, and finally reach an asymptote, s, where s depends on the base frequencies in the sequence through the expression

$$s = \frac{(\pi_T + \pi_C)(\pi_A + \pi_G)}{\pi_A * \pi_G + \pi_C * \pi_G + (\pi_T + \pi_C)(\pi_A + \pi_G)}, \tag{2}$$

where π_x is the frequency of nucleotide x for $x = A, C, G, T$. The parameter k is a measure of the rate at which the proportion of transversions increases from its initial value to the point of saturation. Since $\hat{a}$ is an estimate of the instantaneous rate of transversional changes, an estimate of the TI:TV ratio is given by $(1 - \hat{a})/\hat{a}$.

Maximum likelihood (ML) can also be used to estimate the TI:TV ratio. For a fixed tree and set of branch lengths, many programs (e.g., *PAUP** (Swofford, 1998) and PAML (Yang, 2000)) can provide ML estimates of the parameters in a specified evolutionary model. Using a model that includes a parameter for the TI:TV ratio, such as the F84 model (Felsenstein, 1984) or the HKY85 model (Hasegawa et al., 1985), allows an estimate of the TI:TV ratio to be obtained. Bayesian techniques (Huelsenbeck and Ronquist, 2001; Li et al., 2000; Mau et al., 1999; Yang and Rannala, 1997) also allow for the estimation of parameters in a specified evolutionary model, and hence an estimate of the TI:TV ratio can be obtained through Bayesian inference as well.

Several authors have compared the performance of various estimators of the TI:TV ratio (Pollock and Goldstein, 1995; Purvis and Bromham, 1997; Wakeley, 1994, 1996; Yang and Yoder, 1999). Wakeley (1994) examined the effect of rate variation among sites analytically, and showed that the presence of rate variation causes underestimation of the TI:TV ratio. Pollock and Goldstein (1995) showed that their estimate based on a weighted average of the pairwise ratios performed well on simulated data in terms of mean squared error. Wakeley (1996) reviewed several of the commonly used methods for estimating the TI:TV ratio, and compared the performance of "classical" estimates (e.g., the pairwise and parsimony methods) with the best available estimate for four real data sets. Purvis and Bromham (1997) compared their independent pairs method (described above) to the pairwise method, a modified pairwise method that corrected for multiple substitutions at a site, and the maximum likelihood method for two real data sets. They found that their method gave larger values than ML estimates, and that the pairwise methods both tended to be lower than these estimates. Yang and Yoder (1999) compared the corrected pairwise method and the ML method for both real and simulated data. For their simulated data, they found that both the pairwise and ML methods overestimated the TI:TV ratio, most significantly when sequence divergence was low. They attributed this to a tendency to over-correct for multiple substitutions.

Our goal is to evaluate and compare several of the available methods for estimating the TI:TV ratio. In particular, we examine the (uncorrected) pairwise method, Ina's modification of the pairwise method, the parsimony method, the independent pairs method of Purvis and Bromham, the ML method, and a Bayesian technique. For each of these methods, we consider both simulated and real data for varying numbers of taxa, and examine the effect of varying levels of rate heterogeneity among sites. We anticipate that methods that do not account for either multiple substitutions or time since divergence (e.g., the pairwise methods and the parsimony method) will underestimate the TI:TV ratio; however, these methods are included because they are computationally more tractable than methods which do account for such phenomena, and it is thus interesting to quantify the extent of their bias. To our knowledge, this is the first study to compare a wide range of methods for estimating the TI:TV ratio using data simulated so that the actual TI:TV ratio is known, and it is the first to assess the performance of Bayesian methods for estimating this parameter.

2. Methods

Data sets were simulated using Seq-Gen Version 1.2.5 (Rambaut and Grassly, 1997). Seq-Gen is a program that generates DNA sequences of a specified

length according to a Markov model for a given phylogeny with fixed branch lengths. In this study, we consider three fixed trees, with $N = 14$, 30, and 57 taxa, with specified parameter values (described in detail below) to generate the sequences. All trees were determined by obtaining ML estimates of the topology and branch lengths for a corresponding real data set: the 14-sequence tree is one estimated for a commonly used mammal data set consisting of mtDNA (Hayasaka et al., 1988), the 30-sequence tree is the estimate obtained using the L1 gene for papillomaviruses (Chan et al., 1992, 1995; Ong et al., 1997), and the 57-sequence tree is the estimate obtained using the 12S mitochondrial ribosomal gene for several species of damselfishes (Jang-Liaw et al., 2002). The trees are shown in Figs. 1–3, respectively.

For each model tree, eight groups of data sets were generated. The 14-sequence data sets were specified to a length of 231 sites, the 30-sequence data sets were specified to a length of 1382 sites, and the 57-sequence data sets were specified to a length of 1053 sites (each chosen to reflect the length of the corresponding real data). The sequences were generated under the F84 model (Felsenstein, 1984) with varying parameter settings. For each model tree, the TI:TV ratio was set to be either low or high. For the 14-sequence data sets, the ratios were 1.31 and 29.21, where the lower value is an estimate achieved for a subset of the data, and the higher value is the maximum likelihood estimate (MLE) for the real data. For the 30-sequence and 57-sequence data sets, the ratios were set to 1.12 and 10.0, and 2.07 and 10.0, respectively. The lower value in each case is the MLE for the real data, while the higher value is a reasonable upper bound. Base frequencies were fixed at

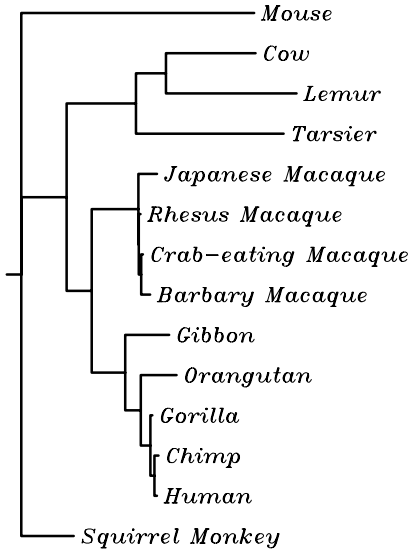

Fig. 1. Model phylogenetic tree used to generate the 14-sequence simulated data sets.

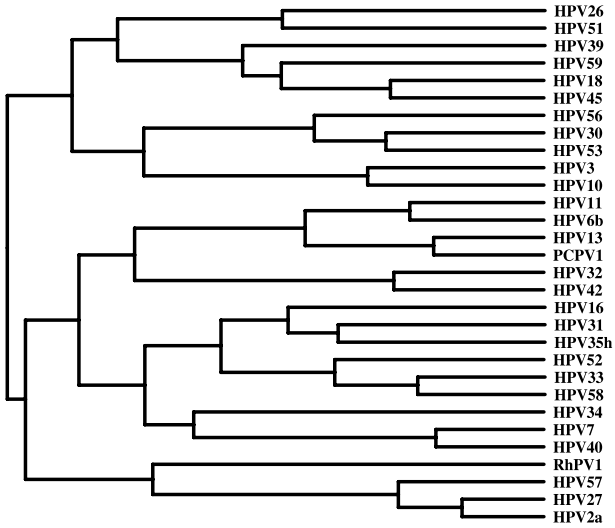

Fig. 2. Model phylogenetic tree used to generate the 30-sequence simulated data sets.

different values for each group of data sets, corresponding to empirical estimates for the real data. The frequencies were: $\pi_A = 0.3760$, $\pi_C = 0.4007$, $\pi_G = 0.0393$, and $\pi_T = 0.1840$ for the 14-sequence data sets; $\pi_A = 0.2856$, $\pi_C = 0.1926$, $\pi_G = 0.2187$, $\pi_T = 0.3032$ for the 30-sequence data sets; and $\pi_A = 0.3062$, $\pi_C = 0.2591$, $\pi_G = 0.2214$, $\pi_T = 0.2133$ for the 57-sequence data sets. The nucleotide substitution rates were either fixed or allowed to vary between sites. When rates varied, they were selected from a gamma distribution with mean 1 and shape parameter α. Three values of α were selected to allow for different types of rate variation. The lowest α value selected was 0.2335, which corresponds to a situation in which the majority of sites have very low rates of evolution while a few sites experience high rates of evolution (this value was selected so that approximately 25% of sites have a rate >1). Our second setting, $\alpha = 1.0$, represents moderate rate variation so that many sites have low rates of evolution, but many have high rates as well (in this case, approximately 37% of sites have a rate >1). Our final setting, $\alpha = 20.0$, represents a scenario in which the majority of sites have intermediate rates of evolution, while a few sites evolve more quickly and a few evolve more slowly (here, approximately 47% of sites have a rate >1). We note that $\alpha = \infty$ represents a constant rate of evolution for all sites, while $\alpha < 1.0$ represents a highly skewed distribution of rates across sites (Yang and Kumar, 1996).

For each combination of TI:TV ratio and rate variation, 100 sub-data sets were generated. An estimate of the TI:TV ratio was obtained for all data sets using each of the methods described below. For each method (with the exception of the Bayesian technique), all parameters except for the TI:TV ratio were assumed to be fixed and known. This was done so that the performance of each

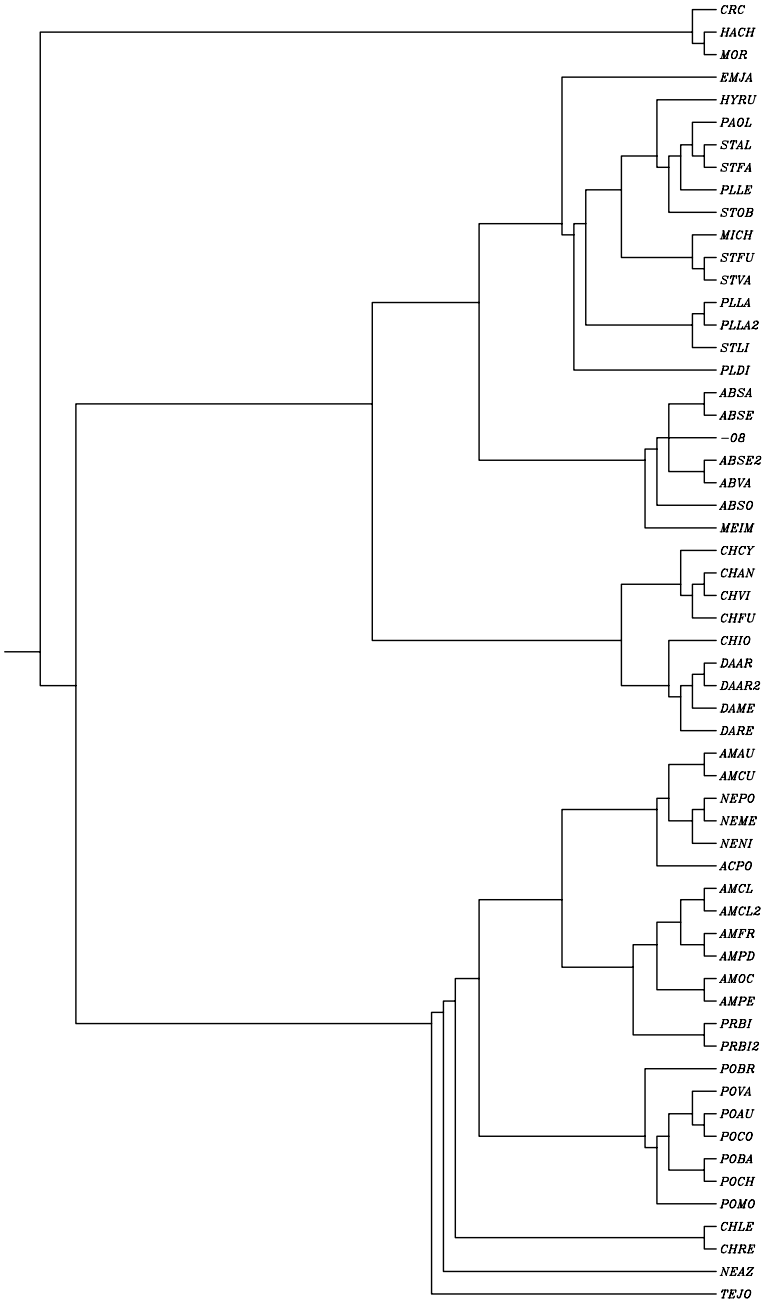

Fig. 3. Model phylogenetic tree used to generate the 57-sequence simulated data sets.

estimator could be assessed under the most favorable conditions. We address the more realistic case where other parameters, including the tree itself, are unknown in the Discussion. To examine the bias and variability of each of the estimators, we report the mean and the variance of these 100 estimates. To get an overall summary, the mean squared error (MSE) is reported (MSE = bias^2 + variance).

For the pairwise method and Ina's method, the number of transitional and transversional changes were tallied using *PAUP** (Swofford, 1998). Ratios between them could then easily be obtained. A problem with the pairwise method is that when the transition bias is high, it is possible that no transversions will be observed in some pairs of sequences. This would lead to division by zero in forming the pairwise estimate. Fortunately, no pairwise comparisons resulted in zero transversions in this simulation, and hence a valid estimate could always be obtained. Ina's method almost surely takes care of this, since it is very unlikely that there are no transversions in the whole data set.

For the parsimony method, the phylogenetic tree that generated the sequences was assumed. A MP reconstruction was obtained using a program written by one

of the authors (L.A.S.). The transitional and transversional changes were then added over the entire tree, and an estimate of the TI:TV ratio was obtained for each of the sub-data sets.

The divergence times used for the 14-sequence data sets when applying Purvis and Bromham's method were derived by Hayasaka and colleagues while working with this common set of species (Hayasaka et al., 1988). For the 30- and 57-sequence data sets, we used the fact that a tree satisfying the molecular clock assumption was used when generating the data. This can be done since the final model does not depend on an exact measure of time, only on the proportional differences in divergence times between the species. Using the generating phylogeny, pairs of species that could be considered independent were derived using Burt's method (Burt, 1989). For the 14-sequence data sets, seven independent pairs were derived; for the 30-sequence data sets, 14 pairs were derived; and for the 57-sequence data sets, 28 pairs were derived. The NLIN procedure in SAS Version 8 (SAS Institute Inc, 2002) was then used to fit the regression model.

The maximum likelihood analyses were carried out in *PAUP** (Swofford, 1998). MLEs of the TI:TV ratio were obtained by using the generating tree and branch lengths. The K2P and F84 models were each used for estimation. We note that the data were generated under the same model (F84) as was used to obtain the maximum likelihood method estimates. Of course this gives an advantage to the ML estimates in the evaluation of the methods, which is why the K2P model was also used in the estimation process. Additionally, our simulations allow examination of the data sets with varying substitution rates over sites when estimated under constant rate models, which might be useful in identifying how sensitive ML estimates are to correct specification of the underlying model.

Bayesian analysis was carried out using the program MrBayes (Huelsenbeck and Ronquist, 2001). Because an option for specifying a fixed tree on which to estimate parameters is unavailable, simultaneous estimates of the tree and substitution model parameters were obtained. Four chains were run for 100,000 generations, and samples were collected each 100 generations. Estimates for each sub-data set were calculated according to the HKY85 model using the posterior means of the κ parameter and the nucleotide frequency parameters.

3. Results

The results for the simulation are found in Tables 1–6. The value reported for each data set is the mean of the 100 sub-data sets, and the variance is the sample variance of these 100 estimates. The MSE is computed for an easy overall comparison of the methods.

Looking at the results in the tables, we see that the distance-based methods and the parsimony method significantly underestimate the true value of the TI:TV ratio under all simulation conditions. Each of these methods additionally shows a general trend for increased underestimation in cases where rate variation among sites is more pronounced (i.e., $\alpha = 0.2335$). This trend is consistent for the varying tree sizes examined here.

Purvis and Bromham's regression-based method (labeled as "Indep Pairs" in the tables) performs well for all data sets, which supports the theory that an adjustment for time since divergence is needed. The estimates are close to the true values, and the MSEs are lower than for the distance-based methods for $N = 14$ and $N = 30$. For $N = 57$, the method shows little bias, but the variability in the estimates is large, resulting in large MSEs. We examine possible causes for this in the Discussion. As for the distance and parsimony methods, there does appear to be an indication of larger underestimation for data sets with varying substitution rates, though this effect is less pronounced for the larger data sets.

Table 1
Simulation results for the data sets with $N = 14$ and TI:TV = 1.31[a]

Method	No rate variation			$\alpha = 0.2335$			$\alpha = 1.0$			$\alpha = 20.0$		
	Mean	Variance	MSE	Mean	Variance	MSE	Mean	Variance	MSE	Mean	Variance	MSE
Pairwise	0.4185	0.0005	0.7953	0.4823	0.0012	0.6862	0.4484	0.0005	0.7429	0.4177	0.0003	0.7965
Ina	0.3872	0.0004	0.8520	0.4608	0.0011	0.7222	0.4189	0.0005	0.7945	0.3879	0.0002	0.8505
Parsimony	0.4165	0.0007	0.7991	0.4044	0.0012	0.8213	0.3916	0.0011	0.8446	0.3769	0.0012	0.8718
Indep Pairs	0.8145	0.0390	0.2846	0.6356	0.0148	0.4697	0.7535	0.0172	0.3269	0.8009	0.0266	0.2858
MLE K2P	0.4301	0.0043	0.7785	0.4636	0.0033	0.7196	0.4606	0.0037	0.7252	0.4243	0.0037	0.7880
MLE F84	1.3292	0.0495	0.0498	0.5956	0.0042	0.5146	0.9199	0.0194	0.1715	1.2741	0.0514	0.0527
Bayes	0.7854	0.0056	0.2808	0.6146	0.0039	0.4876	0.7315	0.0050	0.3400	0.7634	0.2988	0.3034

[a] The mean, variance, and MSE of the 100 replicates are given. "Pairwise" refers to the pairwise method, "Indep Pairs" refers to the method of Purvis and Bromham, "MLE K2P" is the MLE of the TI:TV ratio under the K2P model, "MLE F84" is the MLE of the TI:TV ratio under the F84 model, and "Bayes" is the mean of the 100 estimates obtained under the HKY85 model (see text for details) using the program MrBayes (Huelsenbeck and Ronquist, 2001). "No rate variation" corresponds to the assumption of equal rates of evolution among sites, while the α values indicate that the rate of evolution among sites was allowed to vary according to a γ distribution with mean 1 and shape parameter α.

Table 2
Simulation results for the data sets with $N = 14$ and TI:TV = 29.21[a]

Method	No rate variation			$\alpha = 0.2335$			$\alpha = 1.0$			$\alpha = 20.0$		
	Mean	Variance	MSE	Mean	Variance	MSE	Mean	Variance	MSE	Mean	Variance	MSE
Pairwise	2.5634	0.2887	710.3292	2.0841	0.1082	735.9206	2.4570	0.2227	715.9458	2.5513	0.2048	710.8903
Ina	1.0167	0.0066	794.8670	1.3724	0.0264	774.9606	1.2623	0.0133	781.0855	1.0272	0.0073	794.2778
Parsimony	1.5130	0.0153	767.1419	0.3754	0.0010	831.4371	0.8855	0.0518	802.3308	0.9520	0.0785	798.5932
Indep Pairs	12.1128	29.3114	321.6267	5.6164	3.2376	559.8959	8.5692	13.7572	439.8006	12.5471	41.0737	318.7249
MLE K2P	2.2738	0.0851	725.6414	1.9812	0.0630	741.4707	2.3570	0.0777	721.1635	2.3218	0.0902	723.0655
MLE F84	34.0502	152.8578	176.2855	2.6755	0.1608	704.2431	15.3863	25.6288	216.7229	31.9493	99.4364	106.9400
Bayes	4.7922	0.1399	596.3680	2.7818	0.1174	698.5647	4.5160	0.1379	609.9337	4.7515	0.0836	598.3035

[a] The mean, variance, and MSE of the 100 replicates are given. "Pairwise" refers to the pairwise method, "Indep Pairs" refers to the method of Purvis and Bromham, "MLE K2P" is the MLE of the TI:TV ratio under the F81 model, "MLE F84" is the MLE of the TI:TV ratio under the F84 model, and "Bayes" is the mean of the 100 estimates obtained under the HKY85 model (see text for details) using the program MrBayes (Huelsenbeck and Ronquist, 2001). "No rate variation" corresponds to the assumption of equal rates of evolution among sites, while the α values indicate that the rate of evolution among sites was allowed to vary according to a γ distribution with mean 1 and shape parameter α.

Table 3
Simulation results for the data sets with $N = 30$ and TI:TV = 1.12[a]

Method	No rate variation			$\alpha = 0.2335$			$\alpha = 1.0$			$\alpha = 20.0$		
	Mean	Variance	MSE	Mean	Variance	MSE	Mean	Variance	MSE	Mean	Variance	MSE
Pairwise	0.8863	0.0006	0.0551	0.7377	0.0005	0.1466	0.8246	0.0004	0.0877	0.8833	0.0006	0.0566
Ina	0.8773	0.0006	0.0595	0.7298	0.0005	0.1528	0.8161	0.0004	0.0928	0.8741	0.0006	0.0611
Parsimony	0.9885	0.0006	0.0245	0.6774	0.0022	0.1981	0.7672	0.0032	0.1277	0.8224	0.0035	0.0921
Indep Pairs	1.1355	0.0117	0.0120	0.9957	0.0149	0.0303	1.0798	0.0139	0.0155	1.1316	0.0160	0.0161
MLE K2P	1.1105	0.0009	0 0010	0.8678	0.0008	0.0644	1.0174	0.0010	0.0115	1.1080	0.0011	0.0013
MLE F84	1.1189	0.0009	0.0009	0.8775	0.0008	0.0596	1.0265	0.0010	0.0097	1.1162	0.0011	0.0011
Bayes	1.1195	0.0009	0.0009	0.8803	0.0009	0.0583	1.0274	0.0010	0.0096	1.1164	0.0010	0.0011

[a] The mean, variance, and MSE of the 100 replicates are given. "Pairwise" refers to the pairwise method, "Indep Pairs" refers to the method of Purvis and Bromham, "MLE K2P" is the MLE of the TI:TV ratio under the K2P model, "MLE F84" is the MLE of the TI:TV ratio under the F84 model, and "Bayes" is the mean of the 100 estimates obtained under the HKY85 model (see text for details) using the program MrBayes (Huelsenbeck and Ronquist, 2001). "No rate variation" corresponds to the assumption of equal rates of evolution among sites, while the α values indicate that the rate of evolution among sites was allowed to vary according to a γ distribution with mean 1 and shape parameter α.

Table 4
Simulation results for the data sets with $N = 30$ and TI:TV = 10.0[a]

Method	No rate variation			$\alpha = 0.2335$			$\alpha = 1.0$			$\alpha = 20.0$		
	Mean	Variance	MSE	Mean	Variance	MSE	Mean	Variance	MSE	Mean	Variance	MSE
Pairwise	6.1875	0.0950	14.6304	3.5346	0.0412	41.7429	5.0129	0.0778	24.9489	6.1251	0.1009	15.1159
Ina	5.9891	0.0965	16.1839	3.3893	0.0390	43.7409	4.8340	0.0783	26.7662	5.9333	0.0987	16.6365
Parsimony	7.6600	0.1181	5.5935	1.9150	0.4965	65.8642	2.6926	1.1280	54.5258	3.1856	1.6348	48.0710
Indep Pairs	10.2347	3.2114	3.2665	9.2839	6.8453	7.3581	10.2104	3.1070	3.1513	10.3137	3.5768	3.6752
MLE K2P	9.9219	0.1857	0.1918	5.4216	0.0856	21.0475	7.9666	0.1278	4.2625	9.7980	0.2222	0.2630
MLE F84	10.0442	0.1888	0.1908	5.4756	0.0866	20.5572	8.0665	0.1285	3.8668	9.9214	0.2279	0.2341
Bayes	10.1745	0.2341	0.2646	5.5247	0.0917	20.1195	8.1425	0.1429	3.5932	10.0227	0.2606	0.2611

[a] The mean, variance, and MSE of the 100 replicates are given. "Pairwise" refers to the pairwise method, "Indep Pairs" refers to the method of Purvis and Bromham, "MLE K2P" is the MLE of the TI:TV ratio under the K2P model, "MLE F84" is the MLE of the TI:TV ratio under the F84 model, and "Bayes" is the mean of the 100 estimates obtained under the HKY85 model (see text for details) using the program MrBayes (Huelsenbeck and Ronquist, 2001). "No rate variation" corresponds to the assumption of equal rates of evolution among sites, while the α values indicate that the rate of evolution among sites was allowed to vary according to a γ distribution with mean 1 and shape parameter α.

The results for the maximum likelihood and Bayesian methods are uniformly the best. The estimates are close to the true values and the MSEs are generally lower than for the other methods. As with the other methods, the results indicate that varying substitution rates lead to larger underestimation. For the ML method, it appears that misspecification of the model impacts the estimation of the TI:TV ratio. The estimates obtained under the K2P model show larger bias than those obtained using the F84 model, with the most substantial differences being observed for the $N = 14$ data sets.

The results from the analysis of the real data, shown in Table 7, are consistent with the results for the simulation. As seen in the table, the distance-based

Table 5
Simulation results for the data sets with $N = 57$ and TI:TV = 2.07[a]

Method	No rate variation			$\alpha = 0.2335$			$\alpha = 1.0$			$\alpha = 20.0$		
	Mean	Variance	MSE	Mean	Variance	MSE	Mean	Variance	MSE	Mean	Variance	MSE
Pairwise	1.8677	0.0172	0.0582	1.4847	0.0121	0.3547	1.7350	0.0116	0.1238	1.8636	0.0167	0.0593
Ina	1.8056	0.0174	0.0872	1.4063	0.0116	0.4521	1.6651	0.0121	0.1761	1.8010	0.0176	0.0899
Parsimony	0.8114	0.0784	1.6624	0.6961	0.0392	1.9267	0.7665	0.0590	1.7582	0.8041	0.0773	1.6798
Indep Pairs	2.0807	0.3134	0.3135	2.2229	0.4272	0.4506	2.2484	0.4196	0.4515	2.1771	0.4927	0.5041
MLE K2P	2.0682	0.0102	0.0102	1.7750	0.0087	0.0958	1.9793	0.0082	0.0165	2.0698	0.0089	0.0089
MLE F84	2.0705	0.0102	0.0102	1.7767	0.0085	0.0946	1.9816	0.0082	0.0161	2.0722	0.00878	0.0088
Bayes	2.0856	0.0104	0.0107	1.7847	0.0091	0.0905	1.9931	0.0085	0.0144	2.0881	0.0093	0.0096

[a] The mean, variance, and MSE of the 100 replicates are given. "Pairwise" refers to the pairwise method, "Indep Pairs" refers to the method of Purvis and Bromham, "MLE K2P" is the MLE of the TI:TV ratio under the K2P model, "MLE F84" is the MLE of the TI:TV ratio under the F84 model, and "Bayes" is the mean of the 100 estimates obtained under the HKY85 model (see text for details) using the program MrBayes (Huelsenbeck and Ronquist, 2001). "No rate variation" corresponds to the assumption of equal rates of evolution among sites, while the α values indicate that the rate of evolution among sites was allowed to vary according to a γ distribution with mean 1 and shape parameter α.

Table 6
Simulation results for the data sets with $N = 57$ and TI:TV = 10.0[a]

Method	No rate variation			$\alpha = 0.2335$			$\alpha = 1.0$			$\alpha = 20.0$		
	Mean	Variance	MSE	Mean	Variance	MSE	Mean	Variance	MSE	Mean	Variance	MSE
Pairwise	0.9012	0.7960	1.7723	6.1537	0.4592	15.2531	8.0655	0.8246	4.5670	9.0859	0.9794	1.8150
Ina	8.3146	0.6327	3.4731	5.4617	0.4017	20.9983	7.3608	0.7237	7.6892	8.3511	0.8660	3.5849
Parsimony	1.3721	1.9624	76.4033	1.0483	0.6946	80.8267	1.2582	1.5622	77.9812	1.3652	0.0017	74.5612
Indep Pairs	11.3401	19.8301	21.6261	10.7522	27.1681	27.7339	10.1484	15.8265	15.8485	12.0617	32.4128	36.6632
MLE K2P	9.9946	0.4234	0.4234	7.6477	0.3476	5.8811	9.3108	0.4831	0.9581	9.9558	0.3946	0.3965
MLE F84	10.0128	0.4241	0.4243	7.6644	0.3481	5.8031	9.3284	0.4838	0.9348	9.9621	0.3994	0.4009
Bayes	11.7482	104.7358	107.7920	7.8629	0.4495	5.0169	9.6356	0.5245	0.6573	10.4514	1.5692	1.7730

[a] The mean, variance, and MSE of the 100 replicates are given. "Pairwise" refers to the pairwise method, "Indep Pairs" refers to the method of Purvis and Bromham, "MLE K2P" is the MLE of the TI:TV ratio under the K2P model, "MLE F84" is the MLE of the TI:TV ratio under the F84 model, and "Bayes" is the mean of the 100 estimates obtained under the HKY85 model (see text for details) using the program MrBayes (Huelsenbeck and Ronquist, 2001)."No rate variation" corresponds to the assumption of equal rates of evolution among sites, while the α values indicate that the rate of evolution among sites was allowed to vary according to a γ distribution with mean 1 and shape parameter α.

Table 7
Results from the analysis of the real data sets[a]

Method	Mammal	Papillomaviruses	Damselfishes
Pairwise	2.72	0.81	1.84
Ina	1.02	0.80	1.68
Parsimony	1.37	0.98	0.87
Indep Pairs	11.14	1.31	3.31
MLE	29.21	1.12	2.07
Bayes	3.73	1.10	2.14

[a] "Pairwise" refers to the pairwise method, "Indep Pairs" refers to the method of Purvis and Bromham, "MLE" is the MLE of the TI:TV under the F84 model, and "Bayes" is the estimate computed under the HKY model using the program MrBayes (Huelsenbeck and Ronquist, 2001).

methods and the parsimony method yield estimates much lower than the values obtained by the remaining methods.

4. Discussion

The distance-based methods and the parsimony method significantly underestimate the true TI:TV ratio for all data sets. This is expected since these methods merely count the number of differences between sequences, and thus fail to take into account multiple substitutions at a site over time. Overall, the estimates for the $N = 14$ data sets are more biased than the estimates for the $N = 30$ and $N = 57$ data sets. This is likely because these data sets are larger, both in the number of sites and in the number of sequences, and thus potentially contain more information. The estimates for the $N = 14$ data set with a true TI:TV ratio of 29.21 are highly variable. It should be pointed out that a TI:TV ratio as high as 29.21 is rarely seen in real data sets. When the ratio is that high, few transversions may be observed between some sequences in the simulation, while between others there may be several transversions. This results in substantial differences between the estimates from data set to data set.

The methods that do account for divergence times perform significantly better than the methods that do not. Purvis and Bromham's independent pairs method gives very accurate estimates when the number of sequences is sufficiently large. As noted in the Results, the variance associated with this method can be large. This

is likely due to the fact that, even with a large data set, the number of observations (i.e., number of independent pairs) used to fit the non-linear regression model is rather small (e.g., for $N = 57$, the number of pairs is 28). Thus, even one observation that doesn't fit the general form of the model in Eq. (1) could result in an unusual estimate of the TI:TV ratio. The choice of which pairs of taxa to use is also somewhat subjective, and thus some selections might result in more robust estimates than others. In addition, the dependency on a known phylogenetic tree and good estimates of divergence times is a drawback to this method, since this information may not always be readily available.

The maximum likelihood and Bayesian methods generally perform the best, with the least bias and smallest variability. When the K2P model is assumed in the likelihood framework, the estimates are generally biased downward. When Purvis and Bromham's independent pairs method was applied assuming equal base frequencies, the same problems were encountered (results not shown), which leads us to believe that an estimate of the base frequencies is important for proper estimation of the TI:TV ratio. It is interesting that the results for the Bayesian method are not very good for the $N = 14$ data sets, but are very accurate for the $N = 30$ and $N = 57$ data sets. Since the pattern is the same for Purvis and Bromham's independent pairs method, we find it likely that this is due to more information in the larger data sets. We also note that the variance for several of the Bayesian analyses is larger than for the corresponding ML analysis, even though the Bayesian estimate of the TI:TV ratio is similar. This is in part expected, since the Bayesian analyses involved simultaneous estimation of the TI:TV ratio with the tree and other substitution model parameters. This introduces additional variability into the estimates of the TI:TV ratio.

However, two of the Bayesian analyses show variances that appear to be larger than expected based on other similar runs. This occurs for $N = 57$ with TI:TV ratio 10.0 in the case of no rate variation and in the case where $\alpha = 20.0$. Examination of the individual results for the 100 sub-data sets shows that in these cases, one or two of the 100 replicates had an extreme value estimated for the parameter κ in the HKY85 model. If these replicates are removed and the remaining sub-data sets are summarized as before we obtain the following results: mean $= 10.4388$, variance $= 1.2012$, and MSE $= 1.3937$ for the case of no rate variation; and mean $= 10.3503$, variance $= 0.5525$, and MSE $= 0.6752$ for the case in which $\alpha = 20.0$. We are unsure about the cause of the high values for the estimate of κ. Repeating the Bayesian analysis for these three sub-data sets resulted in values more similar to the other replicates, suggesting that the MCMC algorithm showed some undesirable behavior for these cases in our first analysis.

The results from this study support the findings of Wakeley (1996). Distance-based methods and the parsimony methods do substantially underestimate the true TI:TV ratio. The fact that the methods that account for divergence times perform so well supports the theory that the underestimation is due to multiple substitutions. Additionally, our simulations support Wakeley's finding that substitution rate variation among sites leads to further underestimation. Across all estimation methods and all data sets, the most biased estimates are those for which rate variation is the most extreme ($\alpha = 0.2335$). As the extent of rate variation decreases (i.e., as α increases), the bias decreases, with estimates in the case of $\alpha = 20.0$ fairly similar to those for which there is no rate variation.

It is important to remember that two of the methods that perform the best, the ML method under the F84 model and the Bayesian method, incorporate the model used to generate the data (or one very similar to it) as part of the analysis. This, of course, provides an unfair advantage to these methods, and makes a comparison to the other methods biased. However, the results for Purvis and Bromham's independent pairs method, which performs almost as well as the ML method, are heavily dependent on correct estimates of the tree and the divergence times. Because of this limitation, the ML or Bayesian methods are preferred, since they require only the choice of an evolutionary model that incorporates a parameter for the TI:TV ratio.

Although our application of several of the available methods (namely, the parsimony, independent pairs, and ML methods) here assumed that the phylogenetic tree was known, in practice this tree would need to be estimated along with the TI:TV ratio. Fortunately, it appears that the estimate of the TI:TV ratio is rather robust to misspecification of the tree, provided that a reasonably realistic evolutionary model is used (Yang et al., 1994), and thus the results as described here are likely to be fairly generally applicable. We do point out, however, that estimation of the tree may be dependent on correct specification of the TI:TV ratio (Wang et al., 2002), which underscores the importance of obtaining an accurate estimate of this parameter.

Finally, we note that the observed estimates for all three real data sets follow the same pattern as for the simulated data. We thus conclude that the results described here are likely to be generally applicable to many currently-used data sets.

Acknowledgments

L.A.S. was supported by NSF Grant DMS 0104290. A.K.K.S. was supported by an athletic scholarship from the University of New Mexico Ski Team.

References

Burt, A., 1989. Comparative methods using phylogenetically independent contrasts. Oxford Surv. Evol. Biol., 33–53.

Chan, S.-Y., Bernard, H.-U., Ong, C.-K., Chan, S.-P., Hoffman, B., Delius, H., 1992. Phylogenetic analysis of 48 papillomavirus types and 28 subtypes and variants: a showcase for the molecular evolution of DNA viruses. J. Virol. 66, 5714–5725.

Chan, S.-Y., Delius, H., Halpern, A., Bernard, H.-U., 1995. Analysis of genomic sequences of 95 papillomavirus types: uniting typing, phylogeny, and taxonomy. J. Virol. 69, 3074–3082.

Felsenstein, J., 1984. Distance methods for inferring phylogenies: a justification. Evolution 38, 16–24.

Hasegawa, M., Kishino, H., Yano, T.-A., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 21, 160–174.

Hayasaka, K., Gojobori, T., Horai, S., 1988. Molecular phylogeny and evolution of primate mitochondrial DNA. Mol. Biol. Evol. 5, 626–644.

Huelsenbeck, J., Ronquist, F., 2001. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754–755.

Ina, Y., 1998. Estimation of the transition/transversion ratio. J. Mol. Evol. 46, 521–528.

Jang-Liaw, N.-H., Tang, K., Hui, C.-F., Shao, K.-T., 2002. Molecular phylogeny of 48 species of damselfishes (Perciformes: Pomacentridae) using 12S mtDNA sequences. Mol. Phyl. Evol. 25, 445–454.

Li, S., Pearl, D., Doss, H., 2000. Phylogenetic tree construction using Markov Chain Monte Carlo. J. Am. Stat. Assoc. 95, 493–508.

Mau, B., Newton, M., Larget, B., 1999. Bayesian phylogenetic inference via Markov Chain Monte Carlo methods. Biometrics 55, 1–12.

Ong, C.-K., Nee, S., Rambaut, A., Bernard, H.-U., Harvey, P.H., 1997. Elucidating the population histories and transmission dynamics of papillomaviruses using phylogenetic trees. J. Mol. Evol. 44, 199–206.

Pollock, D.B., Goldstein, D.D., 1995. A comparison of two methods for reconstructing evolutionary distances from a weighted contribution of transition and transversion differences. Mol. Biol. Evol. 12, 713–717.

Purvis, A., Bromham, L., 1997. Estimating the transition/transversion ratio from independent comparisons with an assumed phylogeny. J. Mol. Evol. 44, 112–119.

Rambaut, A., Grassly, N., 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13, 235–238.

SAS Institute Inc. 2002. SAS Version 8, Cary, NC.

Swofford, D., 1998. PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sinauer Associates, Sunderland, MA.

Wakeley, J., 1994. Substitution-rate variation among sites and the estimation of transition bias. Mol. Biol. Evol. 11, 436–442.

Wakeley, J., 1996. The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. TREE 11, 158–163.

Wang, Q., Salter, L.A., Pearl, D.K., 2002. Estimation of evolutionary parameters with phylogenetic trees. J. Mol. Evol. 55, 684–695.

Yang, Z., 2000. Phylogenetic analysis by maximum likelihood (PAML), Version 3.0. University College, London, England.

Yang, Z., Goldman, N., Friday, A., 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. 11, 316–324.

Yang, Z., Kumar, S., 1996. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Mol. Biol. Evol. 13, 650–659.

Yang, Z., Rannala, B., 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. Mol. Biol. Evol. 14, 717–724.

Yang, Z., Yoder, A., 1999. Estimation of the transition/transversion rate bias and species sampling. J. Mol. Evol. 48, 274–283.