BIOINFORMATICS ORIGINAL PAPER
Vol. 23 no. 2 2007, pages 169–176
doi:10.1093/bioinformatics/btl577

Phylogenetics

Tree and rate estimation by local evaluation of heterochronous nucleotide data

Zhu Yang[1,†], John D. O'Brien[2,†], Xiaobin Zheng[1], Huai-Qiu Zhu[1] and Zhen-Su She[1,3,*]

[1]State Key Lab for Turbulence and Complex Systems and Center for Theoretical Biology, Peking University, Beijing 100871, China, [2]Department of Biomathematics and [3]Department of Mathematics, University of California, Los Angeles, Los Angeles, CA 90095, USA

Received on April 1, 2006; revised on November 8, 2006; accepted on November 10, 2006
Advance Access publication November 16, 2006
Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Heterochronous gene sequence data is important for characterizing the evolutionary processes of fast-evolving organisms such as RNA viruses. A limited set of algorithms exists for estimating the rate of nucleotide substitution and inferring phylogenetic trees from such data. The authors here present a new method, Tree and Rate Estimation by Local Evaluation (TREBLE) that robustly calculates the rate of nucleotide substitution and phylogeny with several orders of magnitude improvement in computational time.

Methods: For the basis of its rate estimation TREBLE novelly utilizes a geometric interpretation of the molecular clock assumption to deduce a local estimate of the rate of nucleotide substitution for triplets of dated sequences. Averaging the triplet estimates via a variance weighting yields a global estimate of the rate. From this value, an iterative refinement procedure relying on statistical properties of the triplets then generates a final estimate of the global rate of nucleotide substitution. The estimated global rate is then utilized to find the tree from the pairwise distance matrix via an UPGMA-like algorithm.

Results: Simulation studies show that TREBLE estimates the rate of nucleotide substitution with point estimates comparable with the best of available methods. Confidence intervals are comparable with that of BEAST. TREBLE's phylogenetic reconstruction is significantly improved over the other distance matrix method but not as accurate as the Bayesian algorithm. Compared with three other algorithms, TREBLE reduces computational time by a minimum factor of 3000. Relative to the algorithm with the most accurate estimates for the rate of nucleotide substitution (i.e. BEAST), TREBLE is over 10 000 times more computationally efficient.

Availability: jdobrien.bol.ucla.edu/TREBLE.html

Contact: jdobrien@ucla.edu

1 INTRODUCTION

Until recently, the study of evolution via phylogenetics has been limited to a static perspective, attempting to infer the distant past by observations of the present. In this context, the common ancestor of a set of sequences is so remote in time that the difference in sampling times provides little or no additional statistical information. The recent advent of inexpensive and rapid genome sequencing now makes it possible to monitor evolution on a time scale concomitant with the rate of such change (Drummond *et al*., 2003). In the important case of RNA virus evolution, samples made days or weeks apart may well have evolved substantial genomic differences. In this situation, sampling time provides information on the rate of the nucleotide substitution and, consequently, the phylogenetic relationships among taxa.

We hence distinguish between isochronous nucleotide sequences, where either all samples have the same date or the difference between dates is irrelevant to the evolutionary situation under consideration, and heterochronous nucleotide sequences, where sampling time provides information on the rate of nucleotide substitution. The development of techniques to infer phylogenies and estimates of the rate of nucleotide substitution from heterochronous data sets promises to substantially extend knowledge of inter- and intra-host RNA virus evolution, human archaeology, and the process of speciation (Drummond *et al*., 2003). In the case of RNA viruses, comparisons of the rates of evolution across different hosts may provide insight into the dynamics of periodic epidemics (Ferguson *et al*., 2003).

Methods for analyzing heterochronous sequence data are only beginning to become available. Thus far, two programs have been introduced for simultaneously estimating the rate of nucleotide substitution per site and constructing a phylogenetic tree from such data: PEBBLE [V1.0; Drummond *et al*., 2002] and BEAST [V1.0; Drummond and Rodrigo, 2000)] Several additional methods can provide estimates of the rate of nucleotide substitution per site given a user-defined tree topology (Rambaut, 2004; Sanderson, 2003).

In order to make any calculation of the rate of nucleotide substitution mathematically tractable, all methods rely on some version of the molecular clock assumption (MCA). Generally, the MCA provides an explicit functional or statistical relationship between time and the frequency of nucleotide alterations (Felsenstein, 2004). Most commonly the association is assumed to be constant, as it is in *T*ree and *R*ate *E*stimation *B*y *L*ocal *E*valutation (TREBLE). Recently several programs including BEAST [V1.3, (Drummond *et al*., 2002) and r8s (Sanderson, 2003)] have implemented capacities to estimate the rate of nucleotide substitution under a relaxed molecular clock model via the assumption of a prior distribution of rates along branches. In either its relaxed or strict formulation, the MCA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

induces a zero-th order approximation of the dynamic evolutionary process, taking the rate or rate process to be constant over the tree.

Here the authors present a new method for analyzing heterochronous sequence data, TREBLE, an extremely computationally efficient procedure that utilizes an inherent geometry arising from the application of the MCA to heterochronous sequences. Using triplets of sequences, TREBLE calculates the appropriate topology for the set. As the sampling times are known, the topology enforces specific ratios between the pairwise distances among the sequences that are then used to calculate the rate of nucleotide substitution for that set. Using this solution, the probability of a misspecified topology is calculated and only those triplets above a given threshold are included for further calculation. A variance weighting is then used to integrate these local calculations into a global estimate of the rate of substitution. Finally, a bootstrap is applied to produce confidence intervals on the estimate. The computational efficiency of the algorithm results from the use of rate estimates derived from heterochronous triplets of sequences and the corresponding absence of other costly parameter estimations.

Simulation studies show that TREBLE's estimations of the rate of nucleotide substitution and phylogeny are comparable with those provided by the best available methods. The phylogenetic trees produced are significantly improved over the other distance matrix method (PEBBLE) but still substantially less accurate than the Bayesian algorithm (BEAST). In terms of computational time, TREBLE represents a substantial improvement of at least three orders of magnitude over the next fastest method. Applied to real data sets of dated Dengue, Severe Acute Respiratory Syndrome (SARS) and influenza A gene sequences, TREBLE provides estimates for the rate of nucleotide substitution nearly identical with those provided by other methods and estimates of the time of most recent common ancestors consistent with observation.

In Section 2, we will review the mathematical framework and assumptions used in the algorithm. Section 3 presents the structure and implementation of the algorithm itself. Section 4 lays out the selection of real and simulated data, comparison of results of the simulations studies with three other algorithms and results of the real sequence data examples. Finally, Section 5 summarizes the conclusions, and details possible extensions of the algorithm and future directions for research.

2 NOTATION AND MATHEMATICAL BACKGROUND

Consider a set of n aligned nucleotide sequences, $S = \{s_i{:}i = 1, \cdots, n\}$, each with m aligned base pairs, sampled at possibly distinct times $\{t_i{:}i = 1, \cdots, n\}$. In order to determine an appropriate Markovian model of evolution, we apply the statistics developed by Rzhetsky and Nei to the ensemble of S, yielding an infinitesimal transition matrix of one of the standard models (Rzhetsky and Nei, 1995) such as Jukes-Cantor (1969). This matrix together with any pair of sequences (s_i, s_j) allows an estimate of the number of substitutions between them, $\hat{K}_{ij}$, via the equations developed by Li and Gu (1995). For the Jukes-Cantor model, parameterized by D^s_{ij} (the observed proportion of substitutions between sequences s_i and s_j) the branch length between them is given by

$$\hat{K}_{ij} = -\frac{3}{4}\log\left(1 - \frac{4}{3}D^s_{ij}\right)$$

These values are referred to as the *distance* between sequences, although they can lack metric properties.

The rate of nucleotide substitution, p, is taken to be the expected number of base pair changes per site per unit of time. For brevity, p is referred to as the rate. For an evolutionary situation modeled by a Markov process of nucleotide substitution, necessarily $p \geq 0$. Further, every pair of sequences, s_i and s_j, is assumed to have a most recent common ancestor (MRCA). The MRCA for a set of sequences is also referred to as the root. The time when the sequences s_i and s_j diverged from their MRCA is labeled τ_{ij}. If the sequences are sampled at t_i and t_j, respectively, then necessarily $\tau_{ij} \leq t_i$ and $\tau_{ij} \leq t_j$. These relationships are diagrammed in Figure 1a. Here the MCA is taken to mean that p is constant over a given set of sequences. This statement of the MCA is common, although it is more restrictive than some recent work (Sanderson, 2003; Drummond *et al*., 2003, 2006). Applied to a pair of sequences, this formulation enforces a geometric relationship between the number of base pair substitutions per site between two sequences and the total amount of time of their independent evolution. This is neatly summarized in the equation:

$$\hat{K}_{ij} = p_{ij}(t_i + t_j - 2\tau_{ij}) + \epsilon_{ij}, \tag{1}$$

where ϵ_{ij} is the error associated with the estimate $\hat{K}_{ij}$. For most of the equations below this value is treated as negligible. p_{ij} is subscripted by the sequence numbers to indicate that the MCA only applies to the pair (s_i, s_j).

As τ_{ij} and p_{ij} are unknown, Equation (1) has no definite solution. In the general case, with no knowledge of the topology of sequence divergence, such a system has n equations and $n + 1$ unknowns and so is underdetermined. In the case where $n = 3$, the introduction of topological information makes the system complete. Considering a triplet of sequences (s_i, s_j, s_k) sampled at corresponding times (t_i, t_j, t_k) with topology as given in Figure 1b yields the solution:

$$\begin{aligned}
\hat{\tau}_{ij} &= \frac{1}{2}\left[(t_i + t_j) - \frac{\hat{K}_{ij}}{(\hat{K}_{ik} - \hat{K}_{jk})}(t_i - t_j)\right] \\
\hat{\tau}_{jk} = \hat{\tau}_{ik} &= \frac{1}{2}\left[\frac{\hat{K}_{ik}t_j - \hat{K}_{jk}t_i}{(\hat{K}_{ik} - \hat{K}_{jk})} + t_k\right] \\
\hat{p}^k_{ij} &= \frac{\hat{K}_{ik} - \hat{K}_{jk}}{t_i - t_j},
\end{aligned} \tag{2}$$

where $\hat{\tau}$ denotes an estimate of τ.

The topology necessary for such a solution is not known a priori and so must be constructed, the technique for which will be detailed below. It is also important to note that considering a larger number of taxa (i.e. $n > 3$) together with topological information will make a system analogous to Equation (2) overdetermined and as such will require approximate methods. The non-normality of the errors and the involved correlation structure make such approximations mathematically complex (Fox, 1997).

As can be seen in Equation (2), the estimate of the rate of nucleotide substitution depends only on the time values of sequences i and j. This pair of sequences is thus known as the informative pair. The remaining sequence k is called the outgroup, as shown in Figure 1b. As the calculation yields an estimate of some true p, we write $\hat{p}$. Given the local application of the MCA, for an informative pair (i, j) and outgroup k we write $\hat{p}^k_{ij}$. Such an estimate is valid if p is non-negative and $\hat{\tau}_{ij} \leq t_i$, $\hat{\tau}_{ij} \leq t_j$, $\hat{\tau}_{ik} \leq t_k$, $\hat{\tau}_{ik} \leq \hat{\tau}_{ij}$.

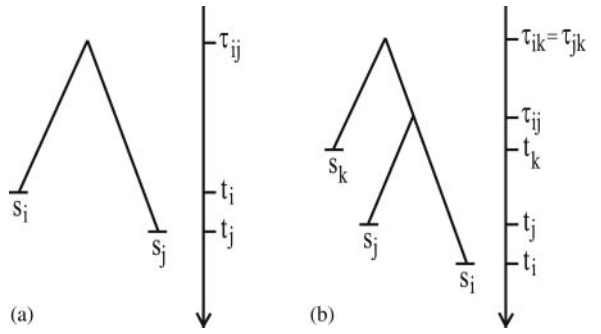

Fig. 1. Diagrams of sequence notation. (**a**) Two sequences, s_i and s_j, having been sampled at times t_i and t_j, diverged from their most recent common ancestor (MRCA) at τ_{ij}. (**b**) Three sequences, s_i, s_j and s_k, sampled at t_i, t_j and t_k, diagrammed with their respective most recent common ancestors located at τ_{ij} and τ_{ik}.

There are several identified sources of error that cause deviation of the estimates $\hat{p}_{ij}^{k}$ from the global rate even in the presence of the molecular clock: natural (i.e. statistical) fluctuations; misspecification of the triplet topology; misspecification of the number of substitutions arising from poor alignment; and variance in the estimate itself. The first sort of error is inherent to the process under observation. Alignment uncertainty is left untreated for the purposes of this paper. TREBLE attempts to increase the accuracy of the estimates by limiting the remaining forms of error. Variance in the estimates is minimized by a weighting scheme similar to iterative reweighted least squares (Fox, 1997). Errors from misspecification of topology are reduced by selecting triplets that have a set probability of having been correctly assigned a topology, as will be shown below.

Again using Equation (2), the variance of the rate estimate can be calculated. According to (Rzhetsky and Nei, 1995), the covariance between the distances for sequences i and k and sequences j and k' is given by

$$\mathrm{cov}(\hat{K}_{ik}, \hat{K}_{jk'}) = \mathrm{var}(\hat{K'}_{ik,jk'}) \tag{3}$$

where $\hat{K}'_{ik,jk'}$ is the length of branches that are shared by the path connecting sequences i and k and the path connecting sequences j and k'. For the triplet shown in Figure 1b, this becomes

$$\mathrm{cov}(\hat{K}_{ik}, \hat{K}_{jk}) = \mathrm{var}(\hat{K'}_{ik,jk}) = \mathrm{var}(\hat{K}_{lk}) \tag{4}$$

where l (i.e. the inferred node corresponding to τ_{ij} in Figure 1b) is the divergence point of sequences i and j. Further note that, following Rzhetsky and Nei (1995), $\mathrm{var}(\hat{K}_{ij})$ is known analytically for most common models of evolution.

$$\begin{aligned}\mathrm{var}(\hat{p}_{ij}^{k}) &= \mathrm{var}\left(\frac{\hat{K}_{ik} - \hat{K}_{jk}}{t_i - t_j}\right) \\ &\approx \frac{\mathrm{var}(\hat{K}_{ik}) + \mathrm{var}(\hat{K}_{jk}) - 2\mathrm{cov}(\hat{K}_{ik}, \hat{K}_{jk})}{(t_i - t_j)^2} \\ &\approx \frac{\mathrm{var}(\hat{K}_{ij})}{(t_i - t_j)^2}\end{aligned} \tag{5}$$

Note that the resultant variance calculation is independent of the outgroup and hence is identical for all triplets with the same informative pair.

In order to limit the error resulting from misspecified topologies, only those triplets with a unique solution to Equation (2) are considered. To further limit such error, we impose the constraint on $\hat{K}_{jk}$ that

$$\hat{p}t_i + \hat{K}_{jk} - \hat{p}t_k - \hat{K}_{ij} > \epsilon_{jk} - \epsilon_{ij} \tag{6}$$

and a corresponding constraint on $\epsilon_{ik} - \epsilon_{ij}$. Here ϵ denotes, as noted above, the normally distributed (Nei and Kumar, 2000) error of the observation $\hat{K}$ from the true number of substitutions. This constraint follows from Equation (1) after substituting the globally calculated $\hat{p}$ for the random variable p. Geometrically, this is equivalent to assuming that half the distance between the two τ's is >0.

The errors have zero expectation and, consequently, so does their difference. The variance of their difference is given, as above, by

$$\mathrm{var}(\epsilon_{ik} - \epsilon_{ij}) \approx \mathrm{var}(\hat{K}_{ij}).$$

A similar result is given for $\mathrm{var}(\epsilon_{jk} - \epsilon_{ij})$. Utilizing the user-defined confidence level α and passing to the asymptotic normal approximation, there is a $1 - \alpha$ probability that the assumption stated in Equation (6) is true if

$$\hat{p}t_i + \hat{K}_{jk} - \hat{p}t_k - \hat{K}_{ij} > Z_\alpha \sqrt{\mathrm{var}(\hat{K}_{ij})}. \tag{7}$$

where Z_α is the corresponding one-tailed normal statistic. A corresponding constraint on $\hat{K}_{jk}$ is also provided. The right-hand side of Equation (7) will be referred to as the noise threshold.

3 ALGORITHM

File input and data preparation. TREBLE reads the set of sequences from an aligned interleaved Phylip file with dates appended to each sequence. Dates can be appended in any time units. Output time is in the same time units as the input. Further information on file preparation is available in the Supplementary information.

Content. The primary content of the TREBLE algorithm has five steps:

(1) For each informative pair, reliable outgroups are selected. A rate calculation and corresponding weighting are created for each informative pair. A global rate is calculated via a weighted average of all informative pairs.

(2) Using the global rate, the noise threshold is applied to each triplet. Those outgroups that do not exceed the noise threshold are removed. The rate for each informative pair is then recalculated.

(3) A new global rate is calculated and compared with the rate obtained at the previous step. If they differ beyond a certain threshold, steps (2) and (3) are repeated until the rate is fixed.

(4) A bootstrap calculation is performed to obtain the confidence interval for the estimated rate.

(5) The rooted phylogenetic tree is reconstructed via an UPGMA-like algorithm.

In the first step, the same procedure is performed for each of the $\binom{n}{2}$ informative pairs of sequences. For a given pair (s_i, s_j), each of the $(n-2)$ possible outgroup sequences is considered. An additional criterion that $\hat{K}_{ik} \geq \hat{K}_{ij}$ and $\hat{K}_{jk} \geq \hat{K}_{ij}$ is used to determine valid

outgroup sequences to be retained. Let A_{ij} be the set of remaining outgroups for the informative pair (s_i, s_j). The averaged rate for the informative pair is calculated as

$$\hat{p}_{ij} = \frac{1}{|A|} \sum_{k \in A_{ij}} \hat{p}_{ij}^{k}.$$

Each $\hat{p}_{ij}$ is associated with the variance calculation in Equation (2) and weights are set such that $w_{ij} = 1/[\mathrm{var}(\hat{p}_{ij})]$ for each distinct pair (i, j). With the weighting, TREBLE then calculates the first global estimate of the rate of nucleotide substitution per site:

$$\hat{p} = \frac{1}{W} \sum_{i,j=1,\cdots,n} w_{ij}\hat{p}_{ij},$$

where $W = \sum_{i,j} w_{ij}$.

In the second step, the global estimate $\hat{p}$ is used as p in Equation (7). Only those outgroups with statistics exceeding the noise threshold are retained, and the rate for each informative pair is recalculated. In the third step, a new global rate is calculated. If it differs from the rate previously obtained by more than the user-defined threshold, a new calculation of the noise threshold is made. The process repeats until the global rate is fixed or begins to oscillate among a finite number of values. For the latter case, the global rate is set as the average of these values.

Then confidence intervals are placed on $\hat{p}$ (and, consequently, on the date of the MRCA) by use of a bootstrap algorithm (Felsenstein, 1985) with a user-defined number of replications in which nucleotides are sampled from the observed data set with replacement. A sample $\hat{p}$ is calculated from this set and confidence intervals are constructed from the distribution of these samples.

In the final step, $\hat{p}$ is then used to construct the phylogenetic relationships among the n sequences. As there are n tips, the construction of $n-1$ interval nodes is necessary for the construction of the full phylogeny. For each pair of sequences s_i and s_j, $\hat{p}$ is used to establish their divergence time, $\hat{\tau}_{ij}$. Given Equation (1) and $\hat{p}$, this calculation is direct. A new node is then placed at their divergence time. Taking the node at $\hat{\tau}_{ij}$ to be a tip replacing s_i and s_j, there are then $n-1$ remaining tips. The algorithm then proceeds to establish the divergence time between this node and the sequence nearest in distance. This process is repeated until only one tip remains. This algorithm is identical to UPGMA (Sneath and Sokal, 1973) except that it starts with the latest divergence node instead of beginning with the smallest pairwise distance. We observed no significant difference between this algorithm and other distance matrix methods, provided distances were converted to time units.

Implementation: TREBLE was implemented in C++ as an executable file. All simulations and tests were run on a Linux operating system (kernel 2.4.20) with an Intel Pentium IV 2.4 GHz processor.

Output: TREBLE returns the final estimate $\hat{p}$ as well as the time necessary for its calculation in seconds. The constructed phylogenetic tree is stored in a user-specified file in Newick format. The bootstrap distribution is saved as a text file. Users can also specify file output for distance and variance matrices.

4 RESULTS

4.1 Real sequence data selection

The real data sets analyzed in this study were taken from the literature or selected from sequence databases. The SARS and Dengue data were downloaded from the GenBank. The first, previously analyzed by the Chinese SARS Molecular Epidemiology Consortium (Lu *et al*., 2001), contains six sequences of the SARS genome. The second viral data set consists of 17 envelope gene sequences from Dengue virus serotype 4 (Dengue-4) (Lanciotti *et al*., 1997), and was also downloaded from the GenBank. This data was previously used in the paper introducing TipDate. The final example comprises 197 sequences of influenza A H3N2 hemagglutinin genes downloaded from the Los Alamos Influenza Sequence Database (Machen *et al*., 2001). For all these sets, α was set to 0.5.

4.2 Generation of simulated sequences

For the MCA simulated data, a set of heterochronous trees was generated by reassorting randomly the branch lengths of the trees created from a random tree generating program (Joyce, 1996, http://alephO.clarku.edu/~djoyce/java/phyltree/intro.html) and then imposing a uniform time scale. The nucleotide sequences were then generated by SeqGen (V1.25; Rambaut, 2000) and 15 trees of 15 taxa were generated and 5 trees each of 30, 60 and 100 taxa. Each sequence contained 2000 nt. For the test of the bootstrap error rate, 1000 trees of 50 taxa were generated. Again, sequences were 2000 nt in length. For all trees, a strict molecular clock was assumed and the rate of substitution was set to 3.0×10^{-4} bp changes per site per unit time. For all simulated tests with TREBLE, α was set to 0.2.

Two data sets were generated to simulate situations where the MCA fails. For the first set, two randomly generated 100 taxa were simulated under the MCA and then partitioned in the following way: each 2000 nt sequence was divided into 200 nt sequences, creating 10 trees with identical topology. Following Equation (8) below, the statistical fluctuations are now on the same order as the rate estimates. In the second data set, 9 trees of 40 taxa were randomly generated with approximately half of the tree simulated under one rate value and the remaining half under another value. One value was always 3.0×10^{-4} substitutions per site per unit time (s/s/ut). The other value ranged discretely over 3.3×10^{-4}, 3.6×10^{-4} and 3.9×10^{-4} s/s/ut. No tree was evenly split between the two rates. Three simulated data sets were generated for each rate partition, for a total of 27 simulated trees. Here also, α was set to 0.2.

4.3 Comparison of TREBLE with BEAST, PEBBLE and TipDate

Of the several programs in use for rate and tree estimation on heterochronous sequence data, we have chosen three that we feel represent the primary methodologies: TipDate, employing a maximum likelihood method; BEAST, a Bayesian approach employing a MCMC algorithm; and PEBBLE, which utilizes a distance-matrix method coupled with a parameterized estimation of branch lengths. PEBBLE and BEAST produce estimates both of the rate and of the tree. TipDate requires a user-defined tree and does not produce an independent phylogenetic construction.

All methods were applied to all trees. TipDate, a maximum likelihood method, did not converge for the 100 taxa examples and provided spurious calculations of its running time for sets of 60 taxa.

As can be seen in Table 1 and Figure 2, TREBLE performs similarly to the best of currently available methods in its point estimation of the rate of nucleotide substitution. In all cases TREBLE produces point estimates extremely close to the simulated

Table 1. The average estimated rate of nucleotide substitution p ($\times 10^{-4}$) and its standard deviation σ_p ($\times 10^{-4}$) by TipDate, PEBBLE, BEAST and TREBLE

Taxa	TipDate		PEBBLE		BEAST		TREBLE	
	p	σ_p	p	σ_p	p	σ_p	p	σ_p
15	2.85	0.73	2.17	2.43	3.35	0.75	3.28	0.80
30	2.83	0.55	1.72	1.67	3.03	0.58	2.94	0.21
60	3.08	0.21	2.12	1.05	3.06	0.22	2.93	0.22
100	—	—	1.44	0.97	3.15	0.20	2.91	0.23

Results cover all data sets simulated under the MCA. The simulated rate is 3.0×10^{-4} s/s/ut.

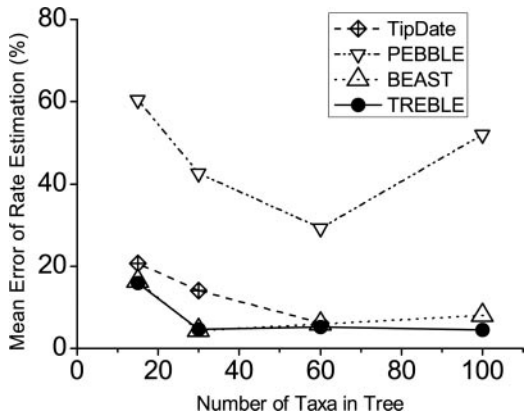

Fig. 2. Mean error of methods given as a percentage from the true rate. TipDate did not yield results for >60 taxa and gave spurious time values for $\geq$60 taxa.

values of the rates and the second most consistent topological constructions after BEAST. TREBLE also produces confidence intervals that are comparable with those from either BEAST or TipDate.

TREBLE performance in topological reconstruction is improved over the other distance method, although it is less accurate than the Bayesian method. In Table 2, the average symmetric distance index (SDI) is provided for each number of taxa and each method. For each test, the SDI was taken between the true tree and that estimated by PEBBLE, BEAST and TREBLE, respectively. The table lists the mean value with standard deviation for each number of taxa. BEAST produces the lowest and hence most precise scores for small numbers of taxa, although TREBLE performs well given its use of a distance method for constructing the tree topology.

Figure 3 compares the average run-time for the four applied methods for each number of taxa. In terms of computational efficiency, TREBLE exceeds BEAST by more than four orders of magnitude. In the 100 taxa case, total computational time does not exceed 2 s, which is over 30 000 times faster than BEAST. For comparison, a list which includes the effective sample size of each run of BEAST is provided in the Supplementary information.

The Jukes-Cantor model is overly simple, relative to the evolutionary processes under consideration. However, with simulations under the F84 model (Felsenstein, 1985) of nucleotide substitution TREBLE yielded similar levels of accuracy. For 5 trees of 100 taxa

Table 2. The error, given by the symmetric distance index and its standard deviation σ_{SDI}, of topology estimation by PEBBLE, BEAST and TREBLE

No. of taxa	PEBBLE		BEAST		TREBLE	
	SDI	σ_{SDI}	SDI	σ_{SDI}	SDI	σ_{SDI}
15	3.7	3.7	0.3	0.7	2.5	2.9
30	10.4	5.5	2.0	2.0	6.8	2.8
60	24.8	7.1	5.2	3.0	19.0	8.1
100	56.8	8.3	9.6	4.1	36.9	6.8

Results cover the same data sets as Table 1.

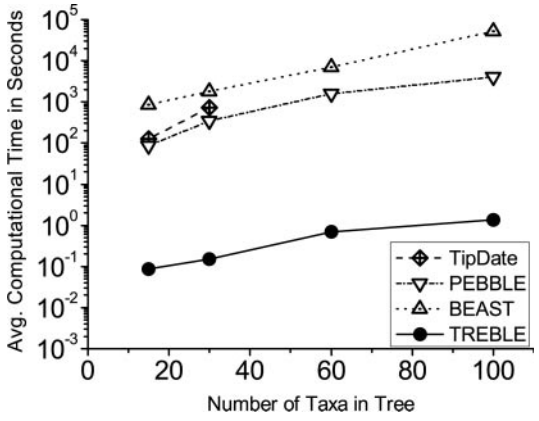

Fig. 3. Mean computational time for each method given in seconds by the number of taxa in the calculated rate and tree. The dependent axis is given in logarithmic scale.

simulated under F84, the mean error in rate estimation was 8.34 ± 4.96% with a mean symmetric distance index of 38 ± 10.5.

To ascertain the rate of error from the bootstrap confidence intervals, we ran TREBLE on 1000 simulated trees, each with 50 taxa. For each tree, 200 bootstrap iterations were performed. For a 95% confidence interval (CI), this resulted in a 7.25% rate of false positives, where the true (simulated) value fell outside the calculated interval.

4.4 Application to viral sequence data

TREBLE was applied to three real viral data sets. For all runs, α was set to 0.5. The first data set consists of six sequences of SARS, obtained from the GenBank. The sampling dates are known from Lu *et al*. (2004).

Kimura's 2-parameter model was used to calculate pairwise distances for the data set. Applying TREBLE to these sequences yielded a global rate of substitution of 8.55×10^{-6} substitutions per site per year (s/s/year) with a 95% CI of (2.06×10^{-6}, 1.54×10^{-5}) s/s/year. The resultant phylogenetic tree is included in Figure 4. This indicates the MRCA divergence at October 23, 2002 with a corresponding CI from September 2, 2002 to February 20, 2003. BEAST gave the estimated rate as 8.14×10^{-6} s/s/year with a 95% credible interval from 1.34×10^{-6} to 1.54×10^{-5} s/s/year. This implies divergence from the MRCA on about October 8, 2002 with a 95% CI from April 28, 2002 to January 21, 2003. Although the calculations of the Chinese SARS Molecular

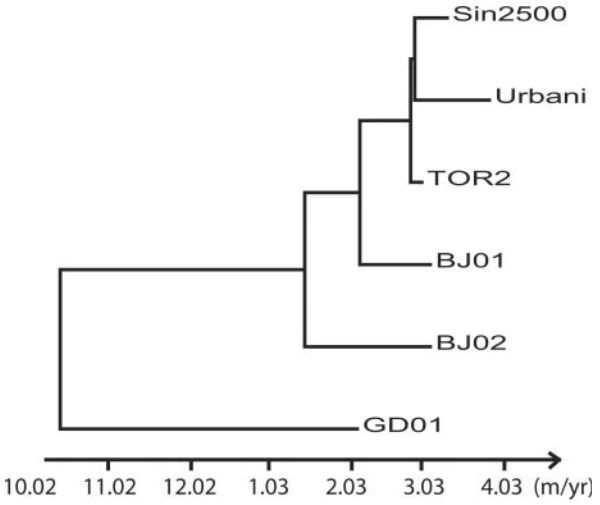

Fig. 4. Six taxa from the 2002-2003 outbreak of SARS, in phylogenies as constructed by TREBLE. The root time closely matches the observed emergence of SARS into human populations.

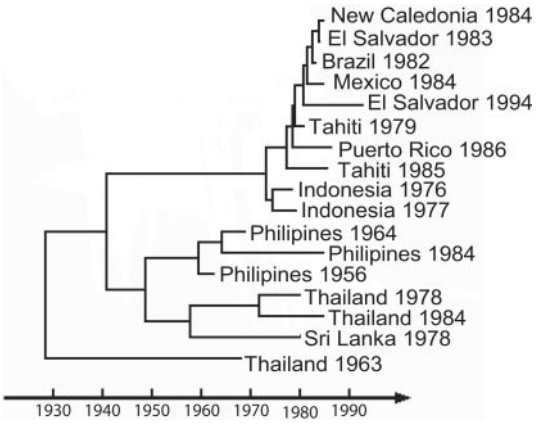

Fig. 5. The phylogenetic tree of 17 samples of the envelope protein of Dengue subtype-4 constructed via TREBLE. The tree is very close to the input tree of TipDate, which was reconstructed by maximum likelihood method, giving evidence for the validity of the molecular clock for these sequences.

Epidemiology Consortium provided the rate of synonymous s/s/year and hence are not commensurate with the results from TREBLE or BEAST, the inferred MRCA is similar at mid-November 2002 with a 95% CI from early June, 2002, to late December, 2002. The phylogenetic tree constructed by the Consortium agrees nearly identically with that of TREBLE (Chinese SARS Molecular Epidemology Consortium, 2004).

The second viral data set considered consists of 17 gene sequences coding for the envelope protein of the Dengue virus serotype 4 that have been obtained from samples isolated between 1956 and 1994 (Lanciotti *et al*., 1997). This data has been previously used in the paper introducing TipDate (Rambaut, 2000).

We applied TREBLE to the Dengue subtype-4 data and, to establish a correspondence with TipDate, set the evolutionary model to HKY (Hasegawa *et al*., 1985). TREBLE estimated the rate as 8.55×10^{-4} s/s/year with a 95% CI of $(6.61 \times 10^{-4}, 1.13 \times 10^{-4})$ s/s/year. The phylogenetic tree constructed is presented in Figure 5. The MRCA is estimated to be 1931 with a confidence interval of (1910, 1941). Similarly, BEAST yielded a rate of substitution of 8.14×10^{-4} s/s/year with a 95% credible interval of $(6.17 \times 10^{-4}, 1.01 \times 10^{-3})$ s/s/year and a MRCA at 1926 (1913, 1937). The phylogenetic tree constructed is presented in Figure 5.

TREBLE's estimation of substitution rate is in close agreement with that provided by TipDate. TipDate gives the substitution rate as 7.91×10^{-4} s/s/year with an 95% CI of $(6.07 \times 10^{-4}, 9.86 \times 10^{-4})$ s/s/year.

For the third set of 197 sequences of influenza A H3N2 hemagglutinin gene, BEAST and TREBLE were both applied using the HKY model of nucleotide substitution (Figure 6). Due to alignment considerations, only the bp 78–1061 were used for calculation. TREBLE's rate calculations varied little using either full or restricted alignment; BEAST had difficulty converging in the former case. TREBLE yields an estimate of the rate of nucleotide substitution as 4.014×10^{-3} s/s/year with a confidence interval of $(3.10 \times 10^{-3}, 4.453 \times 10^{-3})$ s/s/year. The inferred MRCA is set at November, 1965, with CI from January, 1965 to May, 1966, which is highly consistent with the observed emergence of the

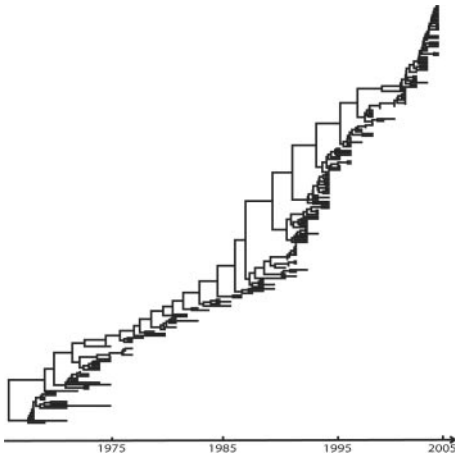

Fig. 6. The phylogenetic tree in time of 197 influenza A H3N2 hemagglutinin gene sequences as constructed by TREBLE. The dendritic structure of the phylogeny is consistent with previous observations of the monophyletic nature of influenza A strain evolution. The root of the tree, located at August, 1966, closely matches the emergence of the H3N2 influenza A strain.

H3N2 influenza A strain in the 'Hong Kong Flu' pandemic of 1966–1967 (Brown, 2000). BEAST's calculations lead to an estimated rate of 4.1043×10^{-3} s/s/year with a credible interval of $(3.775 \times 10^{-3}, 4.475 \times 10^{-3})$ s/s/year. Hence, BEAST and TREBLE's rate calculations for this tree, while not identical, are commensurate.

4.5 Performance in the absence of a molecular clock

In order to examine the performance of TREBLE in a non-clock environment, we ran two simulation studies on the two non-MCA data sets described in Section 4.2. For all non-MCA simulations $\alpha = 0.2$.

For the first data set we found that the introduction of additional noise substantially raises the sampling variance of the estimates but does not increase bias. For example, in the study under the MCA

presented above for trees with 100 taxa the mean percentage deviation was ~6.5%, while in this study the value was ~14.8%. This is a relatively small increase given a 10-fold reduction in observed data. The estimates also appear to be distributed evenly around the simulated value for substitution rate. This gives evidence that under the molecular clock, even in the presence of overwhelming noise, TREBLE makes unbiased (although increasingly variable) estimates of the rate of nucleotide substitution.

When TREBLE is applied to the second data set the bias increases with both the difference in size of the partitions and the difference between the simulated rates. With $\alpha = 0.2$, we observe that TREBLE tends to bias its estimate of the rate of nucleotide substitution towards the rate of the larger set of taxa (see Supplementary information). As the difference between the rate of the two partitions increases, so too does this bias. This is a product of noise threshold filter.

TREBLE allows for setting α to a range of values. The imposition of the noise threshold and the corresponding bias in non-clock-like evolutionary situations can be alleviated by setting $\alpha = 0.5$. In this situation, TREBLE does produce a true zeroth-order approximation of the system, averaging the sampled triplets evenly over the tree without bias. However, the variance on these estimates increases noticeably. It is difficult for us to quantify this increase, as it varies for different distributions of the substitution rate across the tree. For comparison, in the case of the strict molecular clock simulations for trees with 100 taxa the sampled variance grows from 6.5 to 12%. In the case of the Dengue data presented above, varying the values of α had <5% effect on the rate point estimate.

5 DISCUSSION

TREBLE provides a new methodology for estimating the rate of nucleotide substitution and phylogenies that exploits a geometric relationship imposed by the MCA applied to serially sampled sequences. TREBLE's ability to compute the rate of substitution and the phylogenetic tree black is comparable with that of the best available algorithms. In all cases, TREBLE presents a vast improvement in computational efficiency and hence also provides a substantial improvement for calculations on large phylogenies. The maximum number of sequences that can be handled by TREBLE is not known but is likely in the order of thousands. TREBLE provides rooted trees naturally produced from the time sampling of the data. It also provides high accuracy estimates of the rate of nucleotide substitution for both small and large number of taxa, as exemplified above.

As outlined in the algorithm section, TREBLE is only appropriate for heterochronous sequence data sets, that is, only data in which the sampling times vary. The relationship between the accuracy of the TREBLE estimate, the number of nucleotide sites, and the sampling interval, $\bar{T}$ (i.e. the total time of evolution on a particular phylogeny) can be provided via Equation (2), showing that

$$m \cdot p \cdot \bar{T} \sim O(1) \tag{8}$$

Currently the TREBLE algorithm only provides confidence intervals on the estimated rate of nucleotide substitution, giving no indication of posterior support of the computed phylogeny. However, users can quantize such variation by using the outputted trees from the rate bootstrap and applying available tree distance measure programs. TREBLE's current use of a UPGMA-like algorithm can be extended to a more flexible and robust tree-reconstruction algorithm.

The accuracy and computational efficiency of the algorithm result from the novel utilization of a geometric interpretation of the MCA applied to triplets of sequences that are sampled at distinct points in time. The simultaneous ascertainment of the topology and the rate of nucleotide substitution results from the enforcement via the staggered time sampling of a set of relationships between the pairwise distance among the sequences. Hence, it is the heterogeneity of time sampling coupled with the MCA that allows this increase in accuracy and computational speed by trading robust local calculations for costly and inefficient parameter estimations. This previously unnoticed aspect of heterochronous sequence sets suggests that isochronous rate estimation is a special case of heterochronous estimations.

In order to improve the accuracy of TREBLE, greater understanding of the statistical structure of heterochronous trees is necessary. Most saliently, the effect of misspecification of topology on the estimates $\hat{p}_{ij}^{k}$ appears to limit the accuracy of global calculation. In a related problem, the distribution of estimates $\hat{p}_{ij}^{k}$ also appears to reflect the topological structure of the phylogenetic relationships. Establishing more precise understanding of these statistical and mathematical foundational elements will likely yield even more powerful and robust methods of rate and tree estimation.

TREBLE's power to make precise evolutionary inferences will be further maximized by coupling the algorithm with Bayesian methodologies for hierarchical clustering and with hidden Markov models of the state/condition of internal nodes (Suchard *et al*., 2003). The former provides for the rigorous construction of 'super' trees consistent across multipartite data (e.g. multiple genes). The latter allows the assignment to each internal node posterior probabilities of different states, as for example, treatment status or host physiology.

By subdividing genes into minimal windows as determined by Equation (8) and repeatedly applying TREBLE, a 'sliding-window' profile of the rate of substitution within each gene can be constructed. Coupled with an evolutionary model measuring the rate of synonymous and nonsynonymous mutations, this profile provides a quantitative picture of the genomic location of evolutionary pressure. Such profiles may present important distinctions in evolutionary pressure in zoonotic pathogens such as influenza that cycle through several host species.

ACKNOWLEDGEMENTS

We thank Xinqiu Yao, Gangqing Hu, Jian-Kuan Liu, Bonnie Foote and Vladimir Minin for beneficial discussions. The work received partial support from the National Natural Science Foundation and MOST (10225210, 30300071, 90208021; 973 Project grant 2003CB715905) of China. J.D.O. was supported by NIGMS SIB training grant 5T32GM008185. Funding to pay the Open Access publication charges for this article was provided by the NIGMS SIB training grant.

Conflict of Interest: None declared.

REFERENCES

Brown,E. (2000) Influenza virus genetics. *Biomed. Pharma.*, **54**, 196–209.

Chinese SARS Molecular Epidemology Consortium. (2004), Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. *Science*, **303**, 1666–1669.

Drummond,A. and Rodrigo,A. (2000) Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-Sample UPGMA. *Mol. Biol. Evol.*, **17**, 1807–1815.

Drummond,A. *et al.* (2003) Measurably evolving populations. *Trends. Ecol. Evol.*, **19**, 481–488.

Drummond,A. *et al.* (2006) Relaxed Phylogenetics and dating with confidence. *PLoS*, **4**, e88.

Drummond,A.J. *et al.* (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, **161**, 1307–1320.

Felsenstein,J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

Felsenstein,J. (2004) *Inferring phylogenies*. Sinauer Associates, Sunderland, MA.

Ferguson,N.M. *et al.* (2003) Ecological and immunologial determinants of influenza evolution. *Nature*, **422**, 428–433.

Fox,J. (1997) *Applied Regression Analysis, Linear Models, and Related Methods*. Sage Publications, Thousand Oaks, CA.

Hall,P. (1988) Rate of convergence in bootstrap approximations. *Ann. Probab.*, **16**, 1665–1684.

Hasegawa,M. *et al.* (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, **22**, 160–174.

Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. In Munro,H.N. (ed.), *Mammalian Protein Metabolism*. Academic Press, NY, pp. 21–32.

Lanciotti,R.S. *et al.* (1997) Molecular evolution and phylogeny of dengue-4 viruses. *J. Gen. Virol.*, **78**, 2279–2286.

Li,W.-H. and Gu,X. (1995) Statistical models for studying DNA sequence evolution. *Physica A*, **221**, 159–167.

Lu,H.C. *et al.* (2004) Date of origin of the SARS coronavirus strains. *BMC Infect. Dis.*, **4**, Art. No. 3.

Macken,C. *et al.* (2001) The value of a database in surveillance and vaccine selection. In Osterhaus,A.D.M.E., Cox,N. and Hampson,A.W. (eds), *Options for the Control of Influenza IV*. Elsevier Science, Amsterdam, Netherlands, pp. 103–106.

Nei,M. and Kumar,S. (2000) *Molecular Evolution and Phylogenetics*. Oxofrd University Press, Oxford, UK.

Rambaut,A. (2000) Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics*, **16**, 395–399.

Rambaut,A. and Grassly,N.C. (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Com. App. Bios.*, **13**, 235–238.

Rzhetsky,A. and Nei,M. (1995) Tests of applicability of several substitution models for DNA sequence data. *Mol. Biol. Evol.*, **12**, 131–151.

Sanderson,M.J. (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, **19**, 301–302.

Sneath,P.H.A. and Sokal,R.R. (1973) *Numerical Taxonomy*. Freeman, San Francisco, CA.

Suchard,M.A. *et al.* (2003) Hierarchical phylogenetic models for analyzing multipartite sequence data. *Syst. Biol.*, **52**, 649–664.

Thompson,J.D. *et al.* (1997) The Clustal_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **24**, 4876–4882.

Twiddy,S.S. *et al.* (2001) Inferring the rate and time-scale of dengue virus evolution. *Mol. Biol. Evol.*, **20**, 122–129.