BIOINFORMATICS

Vol. 16 no. 4 2000
Pages 395–399

Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies

Andrew Rambaut

Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

Received on November 29, 1999; revised and accepted on March 7, 2000

Abstract

Motivation: *TipDate is a program that will use sequences that have been isolated at different dates to estimate their rate of molecular evolution. The program provides a maximum likelihood estimate of the rate and also the associated date of the most recent common ancestor of the sequences, under a model which assumes a constant rate of substitution (molecular clock) but which accommodates the dates of isolation. Confidence intervals for these parameters are also estimated.*

Results: *The approach was applied to a sample of 17 dengue virus serotype 4 sequences, isolated at dates ranging from 1956 to 1994. The rate of substitution for this serotype was estimated to be* 7.91×10^{-4} *substitutions per site per year (95% confidence intervals of* 6.07×10^{-4}, 9.86×10^{-4}*). This is compatible with a date of 1922 (95% confidence intervals of 1900–1936) for the most recent common ancestor of these sequences.*

Availability: *TipDate can be obtained by WWW from {http://evolve.zoo.ox.ac.uk/software}. The package includes the source code, manual and example files. Both UNIX and Apple Macintosh versions are available from the same site.*

Contact: *andrew.rambaut@zoo.ox.ac.uk*

Introduction

Concomitant with the explosion in the amount of data being deposited in sequence databases is the fact that these sequences have been isolated at different dates. In most cases, the date of the isolation of the sequence is inconsequential compared with the evolutionary history of the organisms from which the sequences came. However, for fast-evolving organisms such as RNA viruses, this variation in date of isolation may represent a significant proportion of the time since they last shared a common ancestor. Furthermore, in a number of cases, viruses have been sequenced from stored tissue samples taken during the first half of this century (e.g. Fitch *et al.*, 1991; Lanciotti *et al.*, 1997; Taubenberger *et al.*, 1997). If all we are interested in is the phylogenetic relationships between these sequences then different isolation times pose no great problem. For viral sequences, however, if we wish to perform analyses based on the assumption of a constant rate of molecular evolution, as is required to obtain estimates of the date of divergence between sequences, the date of isolation must be accounted for. More importantly, the differences in isolation dates, under the assumption of rate constancy, provide a source of information about the rate of molecular evolution. That is, the amount of evolutionary change that has accumulated in each sequence would be expected to be correlated with the date of isolation. Here, I describe a method and computer application for estimating this rate from a set of sequences that have been isolated at different dates.

A previous approach to this problem has been to compare pairs of sequences from different dates with an outgroup sequence to calculate the mean substitution rate (Li *et al.*, 1988). These pairs must be chosen to be independent of each other by ensuring that no two pairs share any evolutionary history (branches on a tree) since their respective common ancestors (Figure 1). However, this approach is problematic because it is perfectly reasonable to expect that, due to the stochastic nature of the substitution process, the sequence sampled earlier will exhibit more divergence from the outgroup than that sampled later (Bollyky and Holmes, 1999). Such cases cannot be easily included in the analysis (they erroneously suggest a negative rate), yet excluding them will bias the analysis towards higher rates.

In comparison, the method presented here incorporates the sequence dates into the maximum likelihood (ML) tree reconstruction method. This not only allows these sequences to be fitted to a constant rate model, enabling us to infer the relative times of lineage splitting, but also provides a rate of evolution that calibrates the tree into absolute time.

System and methods

TipDate is a command-line controlled program written in ANSI C. It should be compiled easily and run on any

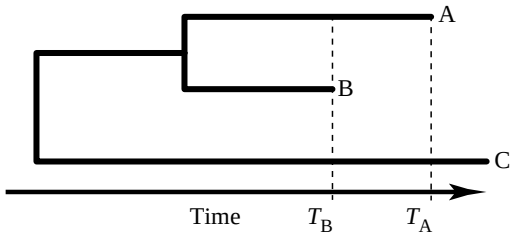

Fig. 1. Calculating the rate of evolution between two sequences, A and B, isolated at different points in time, T_A and T_B, by comparing them with an outgroup sequence, C. Assuming the rate of evolution is the same in lineage A and B, the difference in genetic distance between A and C and between B and C is the amount of evolution that has occurred on lineage A in the time between T_B and T_A. Therefore, the rate of evolution is $(AC - BC)/(T_A - T_B)$.

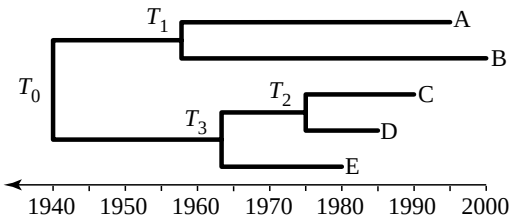

Fig. 2. Constructing a phylogeny of dated sequences. Each of the samples, A–E, was isolated at different points in time, ranging from 1980 to 2000, a span of 20 years. The phylogenetic relationships are known and are shown above. The four internal nodes are at times T_0–T_3, which are to be inferred by maximum likelihood (see text).

UNIX system or workstation. The code will also compile on the Apple Macintosh using the Metrowerks Codewarrior compiler. A separate package is available that includes compiled executables, source and instructions for compiling and running the program on these machines. This paper describes the use of TipDate on a UNIX machine. The application requires an amount of memory proportional to the number and length of sequences being analysed.

Algorithm

Consider a set of five virus sequences isolated on a range of dates from 1980 to 2000 (I shall assume they were all isolated on the same day of the year). Furthermore, assume that the phylogenetic relationships between these sequences have been estimated without error. On a horizontal scale representing time, we position the tips of the tree at points on this scale that correspond to their isolation date (Figure 2). We now have four internal nodes that are at times T_0–T_3; but to begin with, these times are unknown and are given arbitrary values. In effect, the tree has two scales; the time-scale measured in years and the branch-length scale measured in expected number of substitutions per site. We now assume a single rate of evolution for the entire tree, μ, which gives the linear relationship between these two scales. The information that allows us to estimate μ, is provided by the differences in tip dates. Thus, for any given values of our parameters, T_0, T_1, T_2, T_3 and μ, we can calculate the branch lengths for the whole tree. With these branch lengths and our assumed topology, we can calculate the likelihood of the tree for these parameters using the procedure of Felsenstein (1981). These parameters are estimated in the program by finding the values that provide the maximum likelihood.

A range of substitution models are implemented in TipDate, including HKY (Hasegawa *et al.*, 1985), F84 (Felsenstein, 1993) and REV (e.g. Yang, 1994a). These can be combined with two approaches to accommodate rate heterogeneity between sites: the discrete gamma distribution (Yang, 1994b) and relative rates for each codon position. All or any of the parameters of these models may be estimated as part of the maximum likelihood optimization.

Alternative models

Two alternative models are implemented to provide a comparison for the model described above. The first is the single rate (molecular clock) model that has been implemented in a number of phylogenetics packages (PHYLIP, PAML, PAUP etc.). This model assumes that all the sequences have been isolated at the same point in time or that the range of isolation dates is insignificant in the time-scale of the tree. Following the terminology of Goldman (1993) I shall refer to this model as the single rate (SR) model and to the model described above, which relaxes the assumption of contemporaneous sequences, as the single rate dated tips (SRDT) model.

The second model is the different rate (DR) model, originally described by Felsenstein (1981). This model assumes each branch of the tree has a different rate of substitution. However, due to the fact that we do not know what period of time each branch represents, we have no way of estimating these rates. The DR model does provide a suitable general model against which to test the assumption of the SR and SRDT models that the rates are constant across all lineages using a likelihood for ratio test. The test statistic is the difference in the log likelihood (Δ) between our single rate model (SR or SRDT model) and the unconstrained branch length model (DR model). Twice the difference of the log likelihoods (2Δ) is expected to be χ^2 distributed with degrees of freedom equal to the difference in the number of free parameters between the models we are comparing (Wilks, 1938). For a tree of n tips, the DR model has $2n-3$ free parameters (one for each branch length) and the

SR model has $n-1$ (one for each internal node). The SRDT model has one additional free parameter over the SR model (i.e. the rate of evolution). If the SR model is rejected in favour of the DR but the SRDT is not, then the latter can be accepted as no worse a description of the evolution of the data than the DR model.

We could also use Monte Carlo simulation (Goldman, 1993) to obtain the null distribution against which to test Δ. To perform this, simulated data sets of nucleotide sequences are generated along the ML tree inferred under the SRDT model (the null hypothesis). We simulate the data using the ML parameters of the substitution model inferred from the real sequences using a program such as Seq-Gen (Rambaut and Grassly, 1997). For each of these simulated sets of sequences, we estimate the ML hypothesis under both the SRDT and the DR models to produce a distribution of likelihood ratios. If Δ falls outside the 95 percentile of this distribution then the SRDT hypothesis can be rejected. This approach is not discussed further here.

Estimating confidence intervals

An estimate of the confidence intervals (CIs) of the rate of evolution can be obtained using a standard approximation (e.g. Kalbfleisch, 1985). We find the values of μ, either side of the ML value, that give a likelihood that is $\frac{1}{2}\chi^2$ with 1 degree of freedom less than the maximum likelihood. This is the equivalent of asking what is the range of μ, which we would be unable to reject under a likelihood for ratio test. This procedure has been shown by simulation to provide realistic confidence intervals for a method related to that presented here (Rambaut and Bromham, 1998).

Table 1. The likelihood ratio test of fit of the SR and SRDT models

Model	ln L	parameters	Δ	d.f.	χ^2
DR	−3681.64	32	—	—	—
SR	−3726.07	16	44.43	16	$P \ll 0.01$
SRDT	−3692.14	17	10.50	15	$P = 0.14$

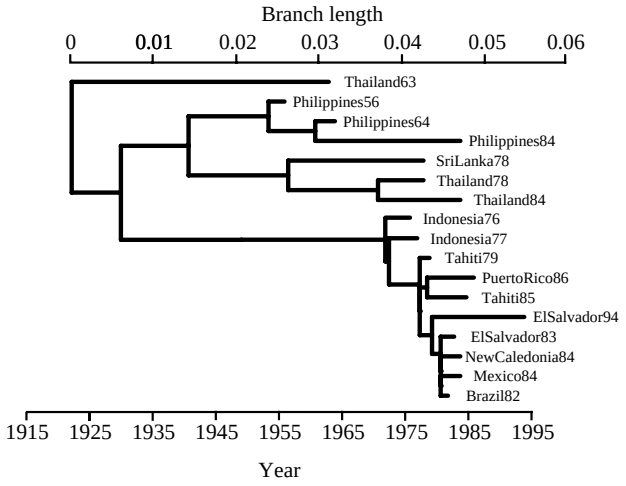

Fig. 3. Dengue-4 phylogeny constructed under the SRDT model. The phylogeny of Dengue-4 isolates constructed under the SRDT model. Under the assumptions of this model the isolates are positioned relative to each other with respect to their dates of isolation. The root of the tree is placed at 1922 but the 95% confidence intervals range from 1900 to 1936.

Results

TipDate was used to analyse a dataset consisting of E (envelope) gene sequences from dengue virus serotype 4 (Dengue-4) (Lanciotti *et al.*, 1997) which have been obtained from samples isolated between 1956 and 1994. One sequence that has been identified as having evidence of recombination (Worobey *et al.*, 1999) was omitted leaving 17 (of length 1485 bp). These sequences exhibit no insertions or deletions with respect to each other and thus alignment was trivial. A maximum likelihood tree was obtained using a heuristic search under the DR model within PAUP* (version 4.0d65: Swofford, 1999). This was done using the HKY model of substitution, allowing different rates of evolution at each codon position. Once obtained, this tree was used in all subsequent analyses. The alignment and tree are provided as an example in the TipDate package.

Using the topology estimated above, TipDate was used to compare the DR model with the SR and SRDT models using the likelihood ratio test. For both the SR and SRDT model, the root of the tree that gave the maximum likelihood was used. The results are summarized in Table 1. As we might expect, the SR model is rejected as an inadequate description of the evolution of these viruses ($P \ll 0.01$), presumably because this model makes no accommodation for the temporal sampling of the isolates. However, the SRDT model provides an adequate fit to the data. The estimated rate of molecular evolution is 7.91×10^{-4} substitutions per site per year (95% confidence intervals of 6.07×10^{-4} to 9.86×10^{-4}) resulting in a date estimate for the root of the tree of 1922 (95% confidence intervals of 1900–1936). Under the SRDT model, the tree has an explicit relationship between branch length and time, such that every node can be dated both relative to each other and on an absolute time scale (Figure 3).

The assumptions in this method

In constructing this tree, a number of assumptions have been made that must be justified. First, it has been assumed that the phylogenetic relationships (topology) between the sequences is known. Actually, we need not make this as-

Table 2. The results of the simulation study

ML tree[a]	Estimates of rate of substitution		
	Mean ($\times 10^{-4}$ subst. site^{-1} year^{-1})	95% Range[b] ($\times 10^{-4}$ subst. site^{-1} year^{-1})	Error[c]
All	7.88	5.87, 9.93	0.044
Correct	7.89	5.92, 9.94	0.039
Wrong	7.77	5.76, 9.65	0.056

[a] In 36 of the 500 replicate simulations, the ML tree was not the 'true' (generating) tree.
[b] The central range that contains 95% of the estimates.
[c] The error is the frequency of simulations in which the 'true' rate of substitution (7.91×10^{-4} subst. site^{-1} year^{-1}) lay outside the estimated 95% confidence intervals.

sumption, as it would be possible to use standard topology searching techniques to obtain the maximum likelihood topology in a manner similar to the ML tree reconstruction methods commonly in use. However, this may make the technique computationally intractable for anything other than small data sets. The practical alternative used here is to estimate the ML topology using a DR model in one of the standard tree reconstruction programs such as PAUP* (Swofford, 1999) or DNAML from the PHYLIP package (Felsenstein, 1993). This makes no assumptions about the constancy of rates and will be unlikely to reconstruct the wrong tree simply as the result of the rate being approximately constant amongst branches. It should be pointed out, however, that the tree estimated for the DR model may not be the maximum likelihood tree for the SR or SRDT models. If this is the case then the SR or SRDT models will have a worse likelihood and thus be more likely to be rejected. Thus, this simplification makes the test of fit of the model conservative.

Simulations

To investigate the properties and assumptions of the TipDate method and to test the reliability of the inferences made for Dengue-4, a set of simulations were performed. These followed the protocol:

(1) A 'true' tree was created by arbitrarily removing seven tips from the inferred SRDT tree for the Dengue-4 sequences without disturbing the branch lengths. This was done to allow a reasonable amount of replicate simulations to be performed.

(2) Sequences of the same length as the Dengue-4 were simulated along this tree 500 times using Seq-Gen (Rambaut and Grassly, 1997). The model of substitution inferred for Dengue-4 under the SRDT model was used except that the rate was assumed to be homogeneous between sites.

(3) Maximum likelihood trees were constructed under the DR model using PAUP* (Swofford, 1999).

(4) TipDate was used to estimate the rate of substitution and its confidence intervals for each simulated tree using the procedure outlined above.

The results of this simulation study are presented in Table 2. The trees produced in step (3) were compared with the generating 'true' tree and in 36 of the 500 replicates the ML tree was 'wrong'. The mean and 95 percentiles of the rate of substitution are given for the entire set of 500 estimates and for those estimates based on the 'correct' and 'wrong' trees separately. There is very little difference amongst these treatments. For the entire set, the type I error for the estimate of 95% confidence intervals was found by obtaining the proportion of estimates for which the 'true' rate (that with which the simulations were generated) lay outside the confidence intervals. This is expected to be binomially distributed with parameters $\alpha = 0.05$, $N = 500$. The type I error of 0.044 lies within the confidence intervals of this distribution (0.033, 0.073).

From this simulation study, the following conclusions can be made: (1) The DR model is a reasonable way of obtaining the tree topology for use in TipDate. (2) The estimation of rate of substitution using TipDate is robust to errors in the estimation of the tree. (3) The approximation for estimating confidence intervals using a likelihood ratio test performs as expected.

Implementation

On a UNIX workstation, TipDate is run by typing its name at the command line. The input file is redirected to the standard input and the resulting phylogeny is written to the standard output, which may be redirected to a file. The switches and parameters that control the program are supplied on the command line. The manual, included with the package, describes how to run the program in detail. The example Dengue-4 alignment described here and its results are also included.

Input file format

The input to TipDate is a text file containing one or more nucleotide sequence alignments, each accompanied by a tree, in the format used by Felsenstein's (1993) PHYLIP package. The date of isolation of each sequence is appended to the end of the sequence's name. Full details of the input file format are supplied in the accompanying manual.

Discussion

The analysis of dengue fever virus serotype 4 using the SRDT model gave a good fit to the data. The estimate of the divergence date of these sequences was in a similar range to that estimated for this serotype by Zanotto *et al*. (1996). However, the latter estimate was based on two pair-wise comparisons (in a manner similar to that described in the introduction) which were not phylogenetically independent, thereby erroneously reducing the variance of the estimate. Furthermore, the error inherent in the substitution process was not accommodated. This error results from the fact that a given rate of substitution can produce a large variation in the number of substitutions that occur along lineages in the same amount of time. By modelling the substitution process explicitly, the approach described here produces confidence intervals that realistically express this uncertainty.

Acknowledgements

This work was supported by grant 50275 from the Wellcome Trust and by the BBSRC. I would like to thank Eddie Holmes, Paul Harvey and Nick Grassly for their assistance and Ziheng Yang for allowing me to use some invaluable code from PAML.

References

Bollyky,P. and Holmes,E. (1999) Reconstructing the complex evolutionary history of hepatitis B virus. *J. Mol. Evol.*, **49**, 130–141.

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

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

Fitch,W.M., Leiter,J.M. E., Li,X. and Palese,P. (1991) Positive Darwinian evolution in human influenza A viruses. *Proc. Natl Acad. Sci. USA*, **88**, 4270–4274.

Goldman,N. (1993) Statistical tests of models of DNA substitution. *J. Mol. Evol.*, **36**, 182–198.

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

Kalbfleisch,J.G. (1985) *Probability and Statistical Inference*. Springer-Verlag, New York.

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

Li,W.-H., Tanimura,M. and Sharp,P.M. (1988) Rates and dates of divergence between AIDS virus nucleotide sequences. *Mol. Biol. Evol.*, **5**, 313–330.

Rambaut,A. and Bromham,L. (1998) Estimating divergence dates from molecular sequences. *Mol. Biol. Evol.*, **15**, 442–448.

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

Swofford,D.L. (1999) *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Sinauer Associates, Sunderland, MA.

Taubenberger,J.K., Reid,A.H., Krafft,A.E., Bijwaard,K.E. and Fanning,T.G. (1997) Initial genetic characterization of the 1918 'Spanish' influenza virus. *Science*, **275**, 1793–1796.

Wilks,S.S. (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.*, **9**, 60–62.

Worobey,M., Rambaut,A. and Holmes,E.C. (1999) Widespread intra-serotype recombination in natural populations of dengue virus. *Proc. Natl Acad. Sci. USA*, **96**, 7352–7357.

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

Yang,Z. (1994b) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, **39**, 306–314.

Zanotto,P.M. d. A., Gould,E.A., Gao,G.F., Harvey,P.H. and Holmes,E.C. (1996) Population dynamics of flaviviruses revealed by molecular phylogenies. *Proc. Natl Acad. Sci. USA*, **93**, 548–553.