DISTREE: A tool for estimating genetic distances between aligned DNA sequences

Johannes Schäfer and Michael Schöniger

Abstract

Motivation: Substitution rates estimated from aligned DNA data can be used as genetic distances to investigate the phylogenetic relationship of those sequences. For this purpose, a Markov model of nucleotide substitution has to be assumed that describes this process most adequately.

Results: A program is presented that estimates substitution rates and their standard errors for a variety of Markov models. The model introduced by Hasegawa et al. (J. Mol. Evol., 22, 160–174, 1985) is the only one for which distances and standard deviations need to be calculated numerically, since analytical formulae cannot be derived. Each model is implemented in two different variants: (i) assuming rate homogeneity or (ii) starting from Gamma-distributed substitution rates across sequence sites. The estimation of heterogeneous substitution rates is based on a method suggested by Tamura and Nei (Mol. Biol. Evol., 10, 512–526, 1993). All required parameters are estimated from sequence data, hence the user is not asked to supply any additional input. One goal of the program is to support the user when choosing a particular model that describes most adequately the evolution of the given data set. For this purpose, a more detailed analysis of this model fit is provided. Phylogenetic trees reconstructed from the inferred distances using the neighbor-joining algorithm are also available.

Availability: http://www.ebi.ac.uk, http://evol10.theochem.tu-muenchen.de/pub

Contact: E-mail: schaefer@theochem.tu-muenchen.de, schoeniger@theochem.tu-muenchen.de

Introduction

Estimating distances from aligned DNA sequence data plays an important role among the methods used for phylogenetic inference. Distances are used, for example, to reconstruct trees or to calibrate divergence dates. During the last few years, several software packages have been published for this purpose. Among these is the widespread PHYLIP package by Felsenstein (1995) that includes a maximum likelihood approach to infer substitution rates. MEGA (Kumar et al., 1993) is another well-known package allowing for the estimation of genetic distances without maximum likelihood algorithms. MEGA was developed for DOS environments.

Here, we present our program, DISTREE, which estimates substitution rates without maximum likelihood routines. Owing to the simplicity of the implemented calculations, it is very fast even for large data sets. Moreover, as we provide source code written in C, our program will run on a variety of systems, including UNIX environments.

DISTREE utilizes the model of nucleotide substitution suggested by Tamura and Nei (1993) and some of those that can be derived from it as special cases. These are the well-known models set up by Jukes and Cantor (1969), and the generalizations introduced by Kimura (1980), Felsenstein (1981), Hasegawa et al. (1985), and Kishino and Hasegawa (1989). The latter, among others, being implemented in the programs DNAML and DNADIST of the PHYLIP package by Joseph Felsenstein since 1984.

All these models assume that each sequence site evolves at the same rate. From the biological point of view, this is known to be an unrealistic restriction. For example, the different positions of a codon are expected to evolve at different rates. Recently, the stochastic description of evolutionary models including rate heterogeneity has been refined considerably. Several strategies exist to model the case of varying substitution rates across sites (Felsenstein and Churchill, 1996; Yang, 1996). In addition to the six homogeneous rate models mentioned above, DISTREE employs the six corresponding augmented models including rate heterogeneity using the ‘method of momenta’ (Yang, 1996). Thereby, we follow the approach of Tamura and Nei (1993) and assume that substitution rates are Gamma-distributed across sequence sites.

At least for the models introduced by Jukes and Cantor (1969) and Kimura (1980), it is proven (Saitou, 1990) that our method and the maximum likelihood method are equivalent in the case of two sequences (see also Zharkikh, 1994). The distances and other parameters of interest are calculated analytically for all models, except for the one proposed by Hasegawa et al. (1985). For this particular model, a formula to calculate the distance analytically cannot be derived. However, this is not prohibitively restrictive, since all required parameters may be computed numerically.

Neither of the software packages mentioned above,
PHYLIP and MEGA, has implemented this particular model. The DNADIST program of the PHYLIP package uses a maximum likelihood approach to estimate pairwise distances and requires the user to supply an estimate for the transition/transversion ratio. Our program infers this parameter from the given sequence data and presents it for use in other software packages. If MEGA is run assuming Gamma-distributed substitution rates, this program asks the user to specify a value for the shape parameter of the Gamma distribution. Our program estimates this parameter automatically.

DISTREE calculates pairwise distances and their standard deviations under each of the 12 models. In addition, it calculates $\chi^2$ values to measure the agreement between observed and predicted dissimilarities. Those quantities are intended as support when a user has to choose a particular model which describes the evolution of the given data adequately under the condition that the corresponding method is not hampered by too large variability.

**Models of nucleotide substitution**

Nucleotide substitution is described by a stationary, continuous-time Markov process which is defined by a matrix $R$ of substitution rates $r_{ij}$ $(i,j \in \{A,G,C,T\})$. Nucleotide $i$ is replaced instantaneously by nucleotide $j$ with rate $r_{ij}$. For the model introduced by Tamura and Nei (1993), the rate matrix is given by:

$$R = (r_{ij}) = \begin{pmatrix} A & G & C & T \\ A & - \alpha_A \pi_G & \beta \pi_C & \beta \pi_T \\ G & \alpha_A \pi_A & - \beta \pi_C & \beta \pi_T \\ C & \beta \pi_A & \beta \pi_G & - \alpha_C \pi_T \\ T & \beta \pi_A & \beta \pi_G & \alpha_C \pi_T \end{pmatrix}$$

Here, $\pi$ is the stationary nucleotide composition, $\alpha_1$ is the transition rate parameter for purines, $\alpha_2$ is the transition rate parameter for pyrimidines, and $\beta$ is the transversion rate parameter. The diagonal elements, abbreviated by dashes, are determined by the condition $\sum r_{ij} = 0$. This ensures that the Markov process preserves the stationary composition $\pi$. DISTREE estimates these parameters directly by counting the frequencies of the four nucleotides in the sequence pair. This is done independently for each pair of sequences in the data set.

The special cases of this model include those introduced by Jukes and Cantor (1969), with $\alpha_1 = \alpha_2 = \beta$, $\pi_i = \frac{1}{4}$; Kimura (1985), with $\alpha_1 = \alpha_2$, $\pi_i = \frac{1}{4}$; Felsenstein (1981), with $\alpha_1 = \alpha_2 = \beta$; Hasegawa et al. (1985), with $\alpha_1 = \alpha_2$; and Kimisho and Hasegawa (1989), which uses a parameter $k$ instead of $\alpha_1$ and $\alpha_2$ with $\alpha_1 = \beta(\frac{k}{k+1})$ and $\alpha_2 = \beta(\frac{k}{k+1})$, where $\pi_A = \pi_A + \pi_G$ and $\pi_T = \pi_C + \pi_T$. The Jukes–Cantor Model is the simplest: it assumes no difference between transitions and transversions, and all base frequencies are equal. The Kimura model allows transition bias. The Felsenstein model includes unequal base frequencies, whereas the Hasegawa et al. and the Kimisho–Hasegawa models allow transition bias and unequal base frequencies. The expected number of substitutions per site $d$ during time $t$ can then be expressed as:

$$d = t \sum_{i,j} \pi_ir_{ij} = -t \sum_i \pi_ir_{ii} \quad (2)$$

In more detail, the expected numbers of transition $(d_s)$ and transversion $(d_v)$ substitutions are given by:

$$d_s = 2\pi_A\pi_G\alpha_1t + 2\pi_C\pi_T\alpha_2t \quad (3)$$

$$d_v = 2\pi_R\pi_Y\beta t \quad (4)$$

The probability $p_{ij}$ of replacing nucleotide $i$ by nucleotide $j$ during time $t$ can be calculated assuming rate homogeneity or Gamma-distributed substitution rates across sequence sites:

$$P = (p_{ij}) = \begin{pmatrix} \exp(R) & \text{rate homogeneity} \\ (1 - \frac{R}{a})^a & \text{Gamma distribution} \end{pmatrix} \quad (5)$$

Here, $a$ denotes the shape parameter of the Gamma distribution and $\frac{R}{a}$ is the expectation of $R$ with respect to the Gamma distribution. The representation of $P$ in terms of the rate matrix for Gamma-distributed substitution rates is obtained by calculations which are similar to those described in Yang (1996). However, explicit use is made of the fact that the eigenvectors of $R$ do not depend on the rate parameters, which is true for each of the models implemented in DISTREE. Equations (2–4) also apply for Gamma-distributed substitution rates if the entries of $R$ and the parameters to be estimated $(d, d_s$ and $d_v$) are replaced by their expected values.

**Adequacy of models**

To judge the adequacy of a model, it is necessary to quantify the uncertainty of estimated parameters (e.g. the distance $d$), as well as the agreement between predicted and observed values (e.g. the composition $\pi$). As usual, we provide $d \pm 1.96\sigma(d)$ to quantify the statistical error of the various estimates of $d$, where $\sigma(d)$ denotes the standard deviation of $d$. The coefficient of variation $\frac{\sigma(d)}{d}$ is used to express the relative error.

Let $n_{ij}$ denote the number of sites in an aligned sequence pair of length $L$, in which nucleotide $i$ and nucleotide $j$ are observed at the same site in the first and in the second sequence, respectively. The Markov models predict these numbers to be $c_{ij} = L\pi_ip_{ij}$. They are referred to as the elements of a matrix $C = (c_{ij})$. The agreement between these two sets of values is measured by a $\chi^2$ for match/mismatch probabilities:

$$\chi^2_{df=15} = \sum_{ij} \frac{(n_{ij} - c_{ij})^2}{c_{ij}} \quad (6)$$
The index \( df = 15 \) indicates the number of degrees of freedom. Since \( \sum_{ij} c_{ij} = \sum_{ij} n_{ij} = L_i \), there are only 15 independent values to be estimated. A similar formula can be derived for the nucleotide composition:

\[
\chi^2_{df=3} = \sum_i \frac{(n_i - c_i)^2}{c_i}
\]

Here \( n_i = \sum_j (n_{ij} + n_{ji}) \) denotes the observed nucleotide composition of the sequence pair and \( c_i = 2L_i x_i \) are the corresponding predicted values. By definition, the \( \chi^2 \) for nucleotide frequencies \( \chi^2_{df=3} = 0 \) for the Felsenstein, Hasegawa et al., Kishino-Hasegawa and Tamura-Nei models.

For the transition and transversion differences, the following formula applies:

\[
\chi^2_{df=2} = \frac{(n_x - c_x)^2}{c_x} + \frac{(n_y - c_y)^2}{c_y}
\]

where \( n_x(n_y) \) is the sum of elements of the matrix \( (n_{ij}) \) that specify transition (transversion) differences. The estimates \( c_x \) and \( c_y \) are the corresponding values obtained from the matrix \( C \). By definition, the \( \chi^2 \) for transition/transversion probabilities \( \chi^2_{df=2} = 0 \) for the Kimura, Hasegawa et al., Kishino-Hasegawa and Tamura-Nei models.

It is important to note that for a particular Markov model, the \( \chi^2 \) values are the same, no matter whether identically or Gamma-distributed substitution rates are assumed. This is due to the fact that the same entries of the probability matrix \( P \) are estimated from the observed sequence data in both cases [cf. equations (5) and (A4–A6)]. For this particular question, tests like the ones described by Goldman (1993) or Yang et al. (1994) should be performed.

The \( \chi^2 \) values are only meaningful if the sequences are not too short. Moreover, for a small number of degrees of freedom, each observed or predicted value must be at least of order five. For the \( \chi^2_{df=15} \) value for match/mismatch probabilities, the latter restriction need not be fulfilled for each of the numbers \( n_{ij}, c_{ij} \) (e.g. Sachs, 1984).

In general, the more complicated a model is, i.e. the more parameters it comprises, the better will be its fit to the data measured by equations (6–8). However, estimating more parameters for a given sample size will increase the variability quantified by \( \sigma(d) \). Therefore, the careful user of DISTREE should try to find a compromise between model adequacy and low variability.

Algorithms

Yang (1994) summarizes conditions that are required to derive analytical formulae for the distance \( d \) and its standard deviation \( \sigma(d) \). They are fulfilled for all homogeneous rate models implemented in DISTREE but the Hasegawa et al. (HKY) model. In the case of the HKY model, DISTREE computes the rate parameter \( \alpha := \alpha_1 = \alpha_2 \) numerically by a Newton algorithm (e.g. Burden and Faires, 1985). Details of this calculation are given in the Appendix. If the desired accuracy of \( \alpha \) cannot be achieved within a maximum number of recursive steps, the HKY model is decided to be not applicable for the sequence pair under consideration. The standard deviation \( \sigma(d) \) is also calculated numerically (see Appendix).

Rate heterogeneity is implemented in DISTREE under the assumption that substitution rates are Gamma distributed across sites. The Gamma distribution is defined by a shape parameter \( a \), which is inferred from the sequence data by DISTREE. We utilize the 'method of momenta' to estimate this parameter (Tamura and Nei, 1993; Yang, 1996). It is defined by the following recursive, self-consistent algorithm:

1. An arbitrary, initial value for the shape parameters \( a \) is chosen.
2. A distance matrix is inferred from the sequence data, using the chosen parameter \( a \). From this matrix, a tree \( A \) is reconstructed utilizing the neighbor-joining algorithm (Saitou and Nei, 1987).
3. Based on tree \( A \), the minimum number of substitutions for each site of the alignment is inferred using maximum parsimony (Farris et al., 1970; Fitch, 1971). A negative binomial distribution is fit to these numbers by estimating its first two momenta and \( a \) can be calculated from them.
4. Using this \( a \), a new distance matrix and neighbor-joining tree \( B \) is calculated.
5. The topologies of the trees \( A \) and \( B \) are compared (Colonius and Schulze, 1981). If they differ, tree \( B \) is renamed \( A \) and step 3 is performed, otherwise the algorithm stops.

Whenever we applied DISTREE, the recursion of steps 3–5 converged fast (after two or three cycles). Once the parameter \( a \) is known, the distance \( d \) can be estimated using the entries of the rate matrix. For all models but HKY, we employ analytical formulae for the distance \( d \) and its standard deviation \( \sigma(d) \) (Tamura and Nei, 1993; Rzhetsky and Nei, 1994) for Gamma-distributed substitution rates. In the case of the HKY model, these estimates are again calculated numerically (see Appendix).

Models for Gamma-distributed substitution rates are excluded from the calculations if the data set contains less than four sequences. In such cases, the statistics obtained from the parsimony analysis are not detailed enough to estimate \( a \) with any confidence.

Implementation

DISTREE is written in C and reads its arguments from the command line. If no argument is specified, a list of all options available is printed to the screen. These options control
running time and the creation of files. In general, output files will be provided only on users’ demand. The only input needed to run DISTREE is a file containing the aligned sequence data.

The program calculates distances, their variances and the \( \chi^2 \) values of model adequacy for each pair of sequences and for all models mentioned above. At the end of each run, a short table is printed to the screen, summarizing the information on the adequacy of the various models. More detailed results for each pair can optionally be printed to a separate output file called detail.log.

The name of the input file is read from the command line. This file must contain a set of aligned sequences which may be supplied in EMBL, GenBank or PHYLIP interleaved (Felsenstein, 1995) format. The file format is detected automatically and need not be specified. If an inconsistency in the alignment is detected, e.g. mismatches in sequence lengths, the program aborts with an error message. No further restrictions are made on the number or the length of the sequences. DISTREE performs all calculations separately for each pair of sequences in the data set. Positions that exhibit gaps are ignored in those calculations. The program offers distance matrices to be written to files in a format compatible with PHYLIP. A distance matrix can be produced for each of the 12 Markov models and for uncorrected distances (observed dissimilarities).

Furthermore, the program offers to reconstruct trees including estimated branch lengths. Based on the distance matrices, the trees are created by making use of the neighboring algorithm (Saitou and Nei, 1987). They can be printed to files in two different formats, intended to be read by the user or to be passed to other programs. A simple graphic that consists entirely of standard ASCII characters followed by a list of branch lengths is available, as well as the standard bracket notation, which is also PHYLIP compatible.

As already mentioned, DISTREE is very fast, even for large data sets (Table I). The most time-consuming part of the calculation is the algorithm used to compare the topologies of the trees (Colonius and Schulze, 1981) while determining the shape parameter \( a \) of the Gamma distribution (cf. step 5 in the description of the algorithm above). This is particularly true for many sequences since the algorithm has to compare \( \binom{N}{4} \) quartets of sequences in the worst case, where \( N \) is the number of sequences in the data set. In the example of Table I, where we analyzed 100 sequences of length 1137, it took \( \approx70\% \) of the computing time to compare the topologies. Therefore, the models using Gamma-distributed rates may optionally be skipped in the calculations.

DISTREE has been successfully compiled and run on various Unix operating systems (including HP-UX, AIX, SunOS4.1, SunOS5.4, OSF1 and FreeBSD) and on a personal computer using a DOS environment. Table I shows the performance of DISTREE on a FreeBSD operating system for six different data sets. As expected, the running time is roughly proportional to \( N^2 \), since DISTREE has to perform \( \binom{N}{2} \) pairwise calculations. However, for large \( N \), the speed is mainly determined by the topology comparison, which may require time of order \( N^3 \) in the worst case. On the other hand, speed is only marginally affected by the length, \( L \), of the sequences for large \( N \). Memory usage is approximately linear in \( N \) and clearly sublinear in \( L \).

Results obtained using different evolutionary models may critically depend on the underlying assumptions. One ability of DISTREE is to provide information on the appropriateness of the various Markov models. This is intended as support for the user to select the particular model that fits best to a given set of sequence data. Table II is printed to the screen if the sample file supplied with DISTREE is used as input. The alignment (M. Cummings, personal communication) consists of 10 cytochrome \( b \) sequences which are 1149 nucleotides long.

The first column displays the rate heterogeneity parameter \( a \) estimated by the ‘method of momenta’. The column titled \( <L_{ST}/L_{TV}> \) shows the transition/transversion ratio \( d_t/d_v \) averaged over all sequence pairs. The next two columns display the number of sequence pairs for which the coefficient of variation falls below the given values. The last column contains the number of pairs for which the corresponding model has the lowest \( \chi^2 \) value for match/mismatch probabilities. As explained above, this value is the same for both variants of a particular model (rate homogeneity or Gamma-distributed rates). The displayed values of this table consider only those sequence pairs for which the parameter in question is available. If, for example, a distance is not defined since the argument of a logarithm is non-positive, the entry displays n/a.

The uncertainty of the distance \( d \) depends on the number of parameters to be estimated from the data set. The relative and absolute errors of the distance \( d \) are thus expected to grow with the complexity of the model. This tendency becomes obvious from Table III which contains detailed results for one of the 45 sequence pairs in this data set. On the other hand, the more general models provide better fits to the data,
## Table II. Summary output

**Summary:**
The data set consisted of 10 sequences, each 1149 nucleotides long.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jukes-Cantor G</td>
<td>n/a</td>
</tr>
<tr>
<td>Kimura G</td>
<td>n/a</td>
</tr>
<tr>
<td>Felsenstein G</td>
<td>n/a</td>
</tr>
<tr>
<td>Hasegawa G</td>
<td>n/a</td>
</tr>
<tr>
<td>Kishino-Hasegawa G</td>
<td>n/a</td>
</tr>
<tr>
<td>Tamura-Nei G</td>
<td>n/a</td>
</tr>
<tr>
<td>Jukes-Cantor</td>
<td>1.56</td>
</tr>
<tr>
<td>Kimura</td>
<td>1.56</td>
</tr>
<tr>
<td>Felsenstein</td>
<td>1.56</td>
</tr>
<tr>
<td>Hasegawa</td>
<td>1.56</td>
</tr>
<tr>
<td>Kishino-Hasegawa</td>
<td>1.56</td>
</tr>
<tr>
<td>Tamura-Nei</td>
<td>1.56</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>$&lt;ts/tv&gt;^c$</th>
<th>#s/d &lt; 5% $^d$</th>
<th>#s/d &lt; 10% $^d$</th>
<th>#min $X^2$ $^e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jukes-Cantor G</td>
<td>0.50</td>
<td>0</td>
<td>45</td>
<td>0</td>
</tr>
<tr>
<td>Kimura</td>
<td>0.96</td>
<td>0</td>
<td>45</td>
<td>0</td>
</tr>
<tr>
<td>Felsenstein G</td>
<td>0.49</td>
<td>0</td>
<td>45</td>
<td>1</td>
</tr>
<tr>
<td>Hasegawa G</td>
<td>0.97</td>
<td>0</td>
<td>45</td>
<td>6</td>
</tr>
<tr>
<td>Kishino-Hasegawa G</td>
<td>1.05</td>
<td>0</td>
<td>45</td>
<td>3</td>
</tr>
<tr>
<td>Tamura-Nei G</td>
<td>0.98</td>
<td>0</td>
<td>45</td>
<td>35</td>
</tr>
</tbody>
</table>

$^a$These results are obtained from the sample input file that is supplied with DISTREE. It consists of an alignment of cytochrome b sequences (M. Cummings, personal communication). The letter G is appended to names of models assuming Gamma-distributed substitution rates.

$^b$The estimated value of the shape parameter of the Gamma distribution.

$^c$Transition/transversion ratio averaged over all sequence pairs.

$^d$The number of sequence pairs for which the coefficient of variation falls below 5 or 10%, respectively.

$^e$The number of sequence pairs for which these models have the smallest $X^2$ for match/mismatch probabilities.

## Table III. Detailed output

sequence pair ‘Loach - Bovine’:

- observed dissimilarities:
  - p = 0.2770
  - pv = 0.1530
  - ps = 0.1240
  - ps[r] = 0.0413
  - ps[y] = 0.0827

- nucleotide frequencies:
  - pi[A] = 0.2823
  - pi[G] = 0.1486
  - pi[C] = 0.3047
  - pi[T] = 0.2643

<table>
<thead>
<tr>
<th>d</th>
<th>$d^a$</th>
<th>$ts/tv^c$</th>
<th>$X^2, 2df^g$</th>
<th>$X^2, 3df^g$</th>
<th>$X^2, 15df^g$</th>
</tr>
</thead>
<tbody>
<tr>
<td>juc</td>
<td>0.3458 ± 0.0413</td>
<td>0.5000 (131.06/262.12)</td>
<td>18.51</td>
<td>132.08</td>
<td>170.18</td>
</tr>
<tr>
<td>jcg</td>
<td>0.4022 ± 0.0554</td>
<td>0.5000 (152.43/304.87)</td>
<td>0.0</td>
<td>132.08</td>
<td>156.68</td>
</tr>
<tr>
<td>kim</td>
<td>0.3476 ± 0.0418</td>
<td>0.9029 (187.55/207.72)</td>
<td>0.0</td>
<td>0.0</td>
<td>70.23</td>
</tr>
<tr>
<td>kmg</td>
<td>0.4062 ± 0.0568</td>
<td>0.9737 (227.86/234.00)</td>
<td>18.58</td>
<td>0.0</td>
<td>70.23</td>
</tr>
<tr>
<td>fel</td>
<td>0.3477 ± 0.0417</td>
<td>0.4995 (131.69/263.61)</td>
<td>0.0</td>
<td>0.0</td>
<td>50.99</td>
</tr>
<tr>
<td>hky</td>
<td>0.3498 ± 0.0424</td>
<td>0.9072 (189.21/208.56)</td>
<td>0.0</td>
<td>0.0</td>
<td>50.85</td>
</tr>
<tr>
<td>hkg</td>
<td>0.4107 ± 0.0582</td>
<td>0.9820 (231.38/235.62)</td>
<td>0.0</td>
<td>0.0</td>
<td>51.38</td>
</tr>
<tr>
<td>kih</td>
<td>0.3497 ± 0.0424</td>
<td>0.9067 (189.09/208.56)</td>
<td>0.0</td>
<td>0.0</td>
<td>50.85</td>
</tr>
<tr>
<td>khg</td>
<td>0.4105 ± 0.0581</td>
<td>0.9809 (231.13/235.62)</td>
<td>0.0</td>
<td>0.0</td>
<td>51.38</td>
</tr>
<tr>
<td>tan</td>
<td>0.3500 ± 0.0425</td>
<td>0.9079 (189.35/208.56)</td>
<td>0.0</td>
<td>0.0</td>
<td>50.85</td>
</tr>
<tr>
<td>tng</td>
<td>0.4110 ± 0.0584</td>
<td>0.9834 (231.71/235.62)</td>
<td>0.0</td>
<td>0.0</td>
<td>51.38</td>
</tr>
</tbody>
</table>

$^a$Results for one sequence pair from the detailed output which was obtained from the supplied sample input file (10 cytochrome b sequences, 1149 nucleotides long).

$^b$Observed long p, transversion (pv) and transition (ps) differences per site. ps is further split into contributions made by purines (ps[r]) and pyrimidines (ps[y]).

$^c$The nucleotide composition counted from the sequence pair.

$^d$Abbreviations are as follows: three-letter encodings with g denote Gamma-distributed substitution rates: juc/jcg, Jukes-Cantor; kim/kmg, Kimura; fel/fg, Felsenstein; hky/hkg, Hasegawa et al.; kih/khg, Kishino-Hasegawa; tan/tng, Tamura-Nei.

$^e$The distances and statistical errors.

$^f$The ratio and the total numbers of the expected transition and transversion substitutions.

$^g$The ratio and the total numbers of the expected transition and transversion substitutions.
as expressed by the $\chi^2$ values. Table II shows that this is true for most, but not all, sequence pairs.

The detailed output (Table III) also contains transition/transversion ratios for each sequence pair. PHYLIP uses 2.0 as the default value for this ratio, but for our cytochrome $b$ example this might not be a good choice (Tables II and III).

Some notes on the estimation of the shape parameter $a$ for variable rates among sites are required. Since only the minimum number of substitutions can be obtained from a parsimony analysis, both the first and second momentum of the assumed underlying negative binomial distribution are considered to be underestimated and thus the shape parameter $a$ is overestimated (Wakeley, 1993). Yang and Kumar (1996) suggested a new method that is also based on a parsimony analysis to reduce this bias. They used a maximum likelihood calculation and the Jukes–Cantor model of evolution to estimate the probability that a given number of different nucleotides is observed at a site in an alignment.

In this paper, the emphasis was put on distance estimation for sequence data. If one is interested in inferring phylogenies, many authors (Kishino and Hasegawa, 1989; Huelsenbeck et al., 1994; Olsen et al., 1994; Tateno et al., 1994; Yang, 1994, 1996; Schöniger and von Haeseler, 1995) agree that maximum likelihood methods for tree reconstruction (Felsenstein, 1981) are, in general, superior to distance matrix methods like neighbor joining, but in cases of large data sets, DISTREE offers a fast estimation of distances and their standard errors which may be used by any distance matrix method.

One should recall that any model discussed in this study is far from biological reality. However, this does not mean that more realistic scenarios like the Tamura–Nei model should be rejected. On the contrary, simulation studies (e.g. Schöniger and von Haeseler, 1995) have shown that using an oversimplified model of base substitution can lead to an inconsistent estimation of evolutionary parameters.

Acknowledgements

We thank the European Bioinformatics Institute at Cambridge for distributing DISTREE. G.L. Hofacker for helpful discussions. the Deutsche Forschungsgemeinschaft for financial support and an anonymous referee for valuable suggestions.

References


Received on January 30, 1997; revised on March 17, 1997; accepted on March 17, 1997.
Appendix

In the following, we indicate how the distance \( d \) and its standard deviation \( \sigma(d) \) are calculated for the HKY model. Let \( \xi = (\alpha - \beta t) \) denote the difference of the rate parameters. With the following abbreviations

\[
q_s = \begin{cases} 
\exp(-\beta t) & \text{rate homogeneity} \\
\left(1 + \frac{\alpha}{\beta} \right)^{\alpha} & \text{Gamma distribution}
\end{cases} \quad (\text{A1})
\]

\[
q_r = \begin{cases} 
q_s \exp(-\beta \xi t) & \text{rate homogeneity} \\
\left(\frac{\alpha}{\alpha} + \frac{\beta}{\beta} \right)^{\alpha} & \text{Gamma distribution}
\end{cases} \quad (\text{A2})
\]

\[
q_t = \begin{cases} 
q_s \exp(-\beta \xi t) & \text{rate homogeneity} \\
\left(\frac{\alpha}{\alpha} + \frac{\beta}{\beta} \right)^{\alpha} & \text{Gamma distribution}
\end{cases} \quad (\text{A3})
\]

the matrix \( P = (p_{ij}) \) can be expressed as:

\[
p_{ij} = \begin{cases} 
\pi_r \left(1 + \frac{\pi_y}{\pi_r} q_i - 1 \right) + \delta_{ij} q_i, i, j \in \{A, C\} \\
\pi_r \left(1 + \frac{\pi_y}{\pi_r} q_i - 1 \right) + \delta_{ij} q_i, i, j \in \{C, T\} \\
\pi_r (1 - q_s), \text{else}
\end{cases} \quad (\text{A4})
\]

where \( \delta_{ij} \) is the Kronecker delta (\( \delta_{ij} = 1 \) if \( i = j \), 0 otherwise).

Let \( p_v, (p_o) \) denote the number of observed transition (transversion) differences per site. The parameter \( \beta t \) (\( \beta t \), respectively) can be estimated from:

\[
p_v = 2 \pi_y (1 - q_v) \quad (\text{A5})
\]

whereas \( \beta t \) (\( \beta t \)) is defined by:

\[
p_v = 2 \pi_A \pi_C \pi_T \pi_r (1 - q_v) + \pi_A \pi_C \pi_T \pi_r (q_v - q_v) + \frac{2 \pi_A \pi_C \pi_T \pi_r (q_v - q_v)}{\pi_r} \quad (\text{A6})
\]

\( \beta t \) (\( \beta t \)) can be calculated numerically from equation (A6). Then the distance \( d \) is obtained from \( \beta t \) and \( \xi t \).

The variance \( \sigma^2(d) \) can be expressed as:

\[
\sigma^2 = \frac{1}{L} \left[ c_v^2 p_v + c_r^2 p_v - (c_v p_v + c_r p_r) \right] \quad (\text{A7})
\]

where \( c_v \) and \( c_r \) are given by:

\[
c_v = \frac{\partial}{\partial p_v} d = 2 \pi_A \pi_C + \pi_C \pi_T \frac{\partial}{\partial p_v} \xi t + \left(1 - \sum \pi_i^2\right) \frac{\partial}{\partial p_v} \beta t
\]

\[
c_r = \frac{\partial}{\partial p_r} d = 2 \pi_A \pi_C + \pi_C \pi_T \frac{\partial}{\partial p_r} \xi t + \left(1 - \sum \pi_i^2\right) \frac{\partial}{\partial p_r} \beta t \quad (\text{A8})
\]

Once the parameter \( \xi t \) is known, the derivatives can be calculated directly from equations (A1–A3), (A5) and (A6), except for the special case of \( \frac{\partial}{\partial p_v} \beta t \), assuming Gamma-distributed substitution rates across sites. This particular derivative is obtained as follows. Using equations (A1–A3), (A5) and (A6), a function \( f(p_v, \beta t) \) is introduced by:

\[
f(p_v, \beta t) = 0 = -p_v + 2 \pi_A \pi_C + \pi_C \pi_T (1 - q_v) + \frac{2 \pi_A \pi_C \pi_T (q_v - q_v)}{\pi_r} \quad (\text{A9})
\]

Since this implicitly defines a function \( p_v(\beta t) \), it holds that (e.g. Königsberger, 1984):

\[
\frac{\partial}{\partial p_v} \beta t = -\frac{\partial}{\partial p_v} f \left( \frac{\partial}{\partial \beta t} f \right) \quad (\text{A10})
\]