Construction of evolutionary distance trees with TREECON for Windows: accounting for variation in nucleotide substitution rate among sites

Yves Van de Peer and Rupert De Wachter

Abstract

Motivation: To improve the estimation of evolutionary distances between nucleotide sequences by considering the differences in substitution rates among sites.

Results: TREECON for Windows (Van de Peer, Y. and De Wachter, R. Comput. Applic. Biosci., 9, 569–570, 1994) is a software package for the construction and drawing of phylogenetic trees based on distance data computed from nucleic acid and amino acid sequences. For nucleic acids, we here describe the implementation of a recently developed method for estimating evolutionary distances taking into account the substitution rate of individual sites in a sequence alignment.

Availability: TREECON for Windows is available on request from the authors. A small fee is asked in order to support the work and to reinvest in new computer hard- and software. More information about the program and substitution rate calibration can be found at URL http://bioc-www.uia.ac.be/yvdp/treeconw.html.

Contact: E-mail: yvdp@uia.ua.ac.be

Introduction

The construction of evolutionary trees on the basis of distance methods such as neighbor-joining (Saitou and Nei, 1987) depends greatly on the accurate estimation of the evolutionary distances from the observed sequence dissimilarities. Conversion of dissimilarity into distance is usually based on a hypothetical substitution model. However, most available models are unrealistic since they either do not take into account differences in substitution rate among the different sites of a molecule (e.g. Jukes and Cantor, 1969; Kimura, 1980) or assume that these rates are distributed according to some mathematical function (Olsen, 1987; Jin and Nei, 1990). We have recently developed a new method to measure the substitution rate or variability of every nucleotide position in a sequence alignment. From the specific spectrum of nucleotide substitution rates, a new equation can be derived that describes better the relationship between sequence dissimilarity and evolutionary distance than equations previously available. The algorithmic details and methodology are discussed in Van de Peer et al. (1993, 1996a).

Application of this new method, called 'substitution rate calibration', to eukaryotic small ribosomal subunit rRNA sequences has already yielded some important improvements in tree topology (Van de Peer et al., 1996a,b). The overall conclusion of applying rate calibration is that evolutionary trees are more reliable and suffer less from anomalies such as those caused by the presence of long branches.

System requirements

TREECON is written in C for Windows (Borland C + + 4.5; Borland International, 1995) and runs on IBM-compatible computers (80486 or higher is recommended) under MS-Windows 3.1, 3.11 or MS-Windows 95. It is assumed that users are familiar with the basic principles of the Microsoft Windows environment.

Implementation

In practice, estimation of relative nucleotide substitution rates and computation of the evolutionary distances are carried out as follows (see also the outline in Figure 1). First, for an alignment of $N$ sequences, TREECON computes $N(N-1)/2$ pairwise evolutionary distances $d$ according to the equation of Jukes and Cantor (1969)

$$d = \frac{3}{4} \ln \left( 1 - \frac{4}{3} f \right)$$

(1)

where $f$ is the dissimilarity or fraction of observed substitutions between two sequences. Next, nucleotide variabilities are computed by observing the frequencies with which sequence pairs differ at homologous positions as a function of the evolutionary distance between the sequence pair. This goes as follows. When all pairwise distances have been computed, they are classified into a number of distance intervals, e.g. distances from 0 to 0.005, distances from 0.005 to 0.010, and so on. For each alignment position and for each distance interval, the fraction of sequence pairs possessing a different nucleotide is computed. For each position, the fraction of sequences showing a difference is plotted as a function of the distance between them (Van de Peer et al.,...
Y. Van de Peer and R. De Wachter

Nucleotide substitution rate calibration

- Compute distances according to Jukes & Cantor (1969)
- Estimate nucleotide substitution rates for each site
- Derive new function to convert dissimilarity into distance
- Compute distances based on new function
- Estimate nucleotide substitution rates for each site
- Distribution of nucleotide substitution rates
- Adjust parameters of new function
- Repeat until nucleotide substitution rates are stable

Fig. 1. Schematic outline of the 'substitution rate calibration' method. See the text for details.

A curve obeying the equation

$$ p_i = \frac{3}{4} \left[ 1 - \exp \left( -\frac{4}{3} v_i d \right) \right] $$

is then fitted to these points by non-linear regression. Equation (2) expresses the probability $p_i$ that an alignment position $i$ contains a different nucleotide in two sequences, as a function of the evolutionary distance $d$ separating these sequences. The slope of the curve in the origin yields the specific nucleotide substitution rate $v_i$ for the position under consideration (Van de Peer et al., 1993). The main advantages of this approach are that nucleotide variabilities do not depend upon a given tree topology, contrary to variability estimation based on maximum parsimony (see, for example, Kumar and Rzhetsky, 1996; Sullivan et al., 1996), and that variabilities can be estimated on the basis of several hundreds of sequences. This is important since the more sequences taken into consideration, the more accurate the estimate of nucleotide variabilities, and the new equation derived. Relative nucleotide variabilities of a section of 18S rRNA are plotted in Figure 2, middle window.

Actually, the estimated nucleotide substitution rates are not yet optimal, because they are derived on the basis of a distance matrix computed by means of equation (1). This equation only gives a first approximation of the relationship between dissimilarity and distance, since it starts from the unrealistic assumption that all nucleotides have the same substitution rate. Therefore, after estimation of all $v_i$ values, alignment positions are grouped into a number of sets of similar variability. A spectrum of relative nucleotide substitution rates is thus obtained, such as the one shown in Figure 2, bottom right. Once the shape of the spectrum is known, it is possible to derive the following equation for the dissimilarity, $f$, as a function of the evolutionary distance $d$ (see Van de Peer et al., 1996a):

$$ f = \frac{3}{4L} \sum_{i=-j}^{k} l_i \left[ 1 - \exp \left( -\frac{4}{3}(1+a)Rd \right) \right] $$

where

$$ R = \frac{L}{\sum_{i=-j}^{k} l_i (1+a)^j} $$

and where $L$ is the total number of nucleotides, $l_i$ the number of nucleotides in set $i$, $j$ the number of sets with a rate lower than the average rate of the complete molecule, $k$ the number of sets with a higher rate, and $(1+a)$ is the ratio of the relative evolutionary rate in set $i$ to this rate in set $i-1$.

A disadvantage of equation (3) is that no expression can be derived for its inverse, needed to convert measured sequence dissimilarities into evolutionary distances. In TREECON, the latter conversion is obtained during the iterative approximation of the evolutionary rate spectrum by listing numerical values of the function for a large number of values of the argument and by interpolating linearly. On the basis of the new evolutionary distances thus obtained, the relative substitution rate of each alignment position is estimated again, as described above, and a new spectrum of evolutionary rates is derived (see the outline in Figure 1). This iterative process is repeated several times until the nucleotide substitution rates $v_i$ do not change anymore. At this moment, the rate spectrum acquires its definitive shape and the final function (3) is known (Van de Peer et al., 1996a). In general, no more than three iterations are needed for the changes to become imperceptible (see Figure 2, bottom left), but it is up to the user to decide how many iterations should be performed.

In the actual process of distance estimation, the use of equation (3) and its inverse is not very practical for repeated computations of distance matrices required for the
Evolutionary trees accounting for nt substitution rates

Fig. 2. Main window: Phylogenetic tree constructed on the basis of 107 small ribosomal subunit RNA sequences computed by substitution rate calibration and displayed by the TREECON drawing module. Bootstrap values (Felsenstein, 1985) above 50% are shown at the internodes. Small window middle: Individual nucleotide variabilities of part of the small ribosomal subunit RNA alignment (nucleotides 300–600) relative to the average variability of the entire section. Small window bottom right: Final substitution rate spectrum of the nucleotides of small ribosomal subunit RNA sequences obtained after three iterations. Small window bottom left: Black curve: graphic representation of dissimilarity as a function of distance (inverse of equation (1)). Colored curves: graphic representations of equation (3) after one, two and three iterations.
construction of evolutionary trees, especially when bootstrap analysis (Felsenstein, 1985) is performed. Fortunately, it is possible to find the following explicit expression that closely matches function (3), provided that an appropriate value is chosen for parameter $p$:

$$f = \frac{3}{4} \left\{ 1 - \exp \left[ -\frac{4}{3} p \ln \left( 1 + \frac{d}{p} \right) \right] \right\}$$

(4)

The value of parameter $p$ depends on the shape of the substitution rate spectrum and is computed automatically by the program. The inverse of equation (4)

$$d = p \left[ \left( 1 - \frac{4}{3} d \right)^{-\frac{1}{4}} - 1 \right]$$

(5)

gives a more accurate conversion of dissimilarity into distance than equation (1) and serves to compute the final evolutionary distances used for tree construction.

Applying nucleotide substitution rate calibration is rather time consuming. Rate calibration of a sequence alignment of 100 sequences of 1000 nucleotides and derivation of the final value for parameter $p$ (after three iterations) took about 15 min on a Pentium computer running at 90 MHz. However, for a specific alignment, $p$ has to be estimated only once.

Acknowledgements

Our research was supported by the Programme on Interuniversity Poles of Attraction (contract 23) of the Federal Office for Scientific, Cultural and Technical Affairs of the Belgian State, and by the National Fund for Scientific Research (NFWO). Y.V.d.P. is Research Assistant of the NFWO.

References


