Phylogenetic Inference with Weighted Codon Evolutionary Distances

Alexis Criscuolo · Christian J. Michel

Received: 23 July 2008 / Accepted: 16 February 2009 / Published online: 24 March 2009

Abstract We develop a new approach to estimate a matrix of pairwise evolutionary distances from a codon-based alignment based on a codon evolutionary model. The method first computes a standard distance matrix for each of the three codon positions. Then these three distance matrices are weighted according to an estimate of the global evolutionary rate of each codon position and averaged into a unique distance matrix. Using a large set of both real and simulated codon-based alignments of nucleotide sequences, we show that this approach leads to distance matrices that have a significantly better treelikeness compared to those obtained by standard nucleotide evolutionary distances. We also propose an alternative weighting to eliminate the part of the noise often associated with some codon positions, particularly the third position, which is known to induce a fast evolutionary rate. Simulation results show that fast distance-based tree reconstruction algorithms on distance matrices based on this codon position weighting can lead to phylogenetic trees that are at least as accurate as, if not better, than those inferred by maximum likelihood. Finally, a well-known multigene dataset composed of eight yeast species and 106 codon-based alignments is reanalyzed and shows that our codon evolutionary distances allow building a phylogenetic tree which is similar to those obtained by non-distance-based methods (e.g., maximum parsimony and maximum likelihood) and also significantly improved compared to standard nucleotide evolutionary distance estimates.

Keywords Evolutionary distance · Unbiased estimate · Codon model · Evolutionary rate · Phylogeny reconstruction · Phylogenetic signal · Computer simulations

A. Criscuolo · C. J. Michel (✉)
Equipe de Bioinformatique Théorique, LSIIT, FDBT (UMR CNRS-ULP 7005), Université de Strasbourg, Pôle API, Boulevard Sébastien Brant, 67400 Illkirch, France
e-mail: michel@dpt-info.u-strasbg.fr

A. Criscuolo
e-mail: criscuol@pasteur.fr

Introduction

An evolutionary distance between two homologous DNA sequences is defined as the average number of substitutions per nucleotide since their divergence from a common ancestor. These pairwise distances are often used to infer phylogenetic trees under the assumption that if homologous sequences have evolved according to a tree model, then the pairwise distances estimated between each pair of sequences are very close to an additive distance, i.e., equivalent to a valuated phylogenetic tree based on the patristic (i.e., path-length) distance between each pair of taxa (Barthélemy and Guénoche 1991). The distance methods for phylogenetic inference thus proceed in two successive steps: first, an estimate of the evolutionary distance between each pair of sequences is computed; and second, an algorithmic method is applied in order to find the phylogenetic tree associated with the additive distance matrix which is the most representative of the distance matrix estimated in the first step, under either a least-squares (e.g., Cavalli-Sforza and Edwards 1967; Fitch and Margoliash 1967) or a Minimum Evolution (e.g., Rzhetsky and Nei 1992, 1993) criterion. This approach is widely used, particularly because their run time is very fast in practice (e.g., a few minutes for hundreds of taxa). In the 20 last years, much progress has been made in this field, particularly in the development of fast and powerful algorithms for tree inference, e.g., Neighbor-Joining (NJ; Saitou and Nei 1987; Studier and Keppler 1988), BioNJ (Gascuel 1997), WLS-MVR (Gascuel 2000), Weighbor (Bruno et al. 2000), FastME (Desper and Gascuel 2002), and STC (Vinh and von Haeseler 2005). Therefore, distance methods are used today not only in exploratory contexts with datasets including several thousands of taxa, but also to obtain starting trees in non-distance methods in order to improve their computing times. For example, the software PhyML (Guindon and Gascuel 2003) infers an initial BioNJ tree which is then used with a heuristic local search in order to optimize a maximum likelihood (ML) criterion. If the starting tree is reliable (i.e., close to the optimal tree), then PhyML is faster and avoids local optima (Criscuolo et al. 2006). Therefore, the theoretical development of distance methods to improve phylogenetic inference is very important not only as a direct method, but also as an initialization technique for non-distance methods.

If a distance matrix is additive (see above), then all distance-based tree inference algorithms retrieve the correct tree. Therefore, the improvement of performance for distance methods needs the identification of evolutionary distances which are as close as possible to the phylogenetic signal induced by an alignment of homologous sequences. These distance estimates are based on a nucleotide evolutionary model (NEM), i.e., associated with a 4 × 4 nucleotide mutation matrix: the model $\mathrm{NEM}_{1\mathrm{P}}$, with one substitution parameter (Jukes and Cantor 1969); the model $\mathrm{NEM}_{2\mathrm{P}}$, with two substitution parameters (transitions and transversions [Kimura 1980]); the model $\mathrm{NEM}_{3\mathrm{P}}$, with three substitution parameters (transitions and two types of transversions [Kimura 1981]), and their variants based on nonsymmetrical mutation matrices (e.g., Tajima and Nei 1984; Hasegawa et al. 1985; Tamura 1992; Tamura and Nei 1993). However, a great difference in scale is often observed for the distances estimated from the three different codon positions (e.g., Kimura 1981; Tajima and Nei 1984; Reed and Sperling 1999).

Indeed, substitutions in some codon positions do not modify the corresponding amino acid, and these so-called synonymous substitutions generally occur more frequently than the nonsynonymous ones. Therefore, a high substitution rate in the third position and a low substitution rate in the second position are generally observed (e.g., Nei 1987; Mindell and Thacker 1996; Yang 1996; see also below). Consequently, as the codon positions with a slow evolution are poor indicators of the sequence evolutionary history, while those with a fast evolution have a saturation effect (see Fig. 2 of Guindon and Gascuel 2003), both the quantity and the quality of the phylogenetic signal induced by each codon position can be very different (see Figs. 8 and 9 of Cummings et al. 1995) and can strongly affect the inference of a reliable phylogenetic tree based on evolutionary distances estimated by a NEM.

All these NEMs differ by their number of substitution parameters but the size of their associated mutation matrices remains identical (4 × 4). In general, a nucleotide evolutionary distance (NED) can be derived analytically from each model, e.g., a distance $\mathrm{NED}_{1\mathrm{P}}$ ($\mathrm{NED}_{2\mathrm{P}}$ and $\mathrm{NED}_{3\mathrm{P}}$, respectively) at one (two and three, respectively) parameter(s) with the model $\mathrm{NEM}_{1\mathrm{P}}$ ($\mathrm{NEM}_{2\mathrm{P}}$ and $\mathrm{NEM}_{3\mathrm{P}}$, respectively). Several years ago, NEMs have been generalized, on the one hand, to motif evolution models with trinucleotides (Arquès and Michel 1993) and dinucleotides (Arquès and Michel 1995) and, on the other hand, to codon-based evolutionary models (CEMs; e.g., Goldman and Yang 1994; Yang and Nielsen 1998, 2008; Yang et al. 2000; see also the review by Yang and Bielawski 2000). These CEMs consider a sense codon as the unit of evolution and take into account substitution events from one codon to another one if they differ from one position at the maximum (e.g., from the serine codon TCA to the serine codon TCG or to a codon CCA coding a different amino acid). Mainly used in a ML framework (e.g., Ren et al. 2005), these CEMs allow precise likelihood calculation with given trees, but at the present time they are not feasible to search optimal ML phylogenetic trees with large datasets. Moreover, to our knowledge, no simple analytical formula is derived from these CEMs to estimate a pairwise codon evolutionary distance (CED) between two aligned codon sequences.

Recently, Michel (2007) has proposed a CEM based on a 64 × 64 mutation matrix with nine substitution parameters. By assuming that there is one substitution per trinucleotide per time interval, a codon evolutionary distance with nine parameters ($\mathrm{CED}_{9\mathrm{P}}$) can be derived analytically (Sect. 3.4 of Michel 2007). It is also proven that this distance $\mathrm{CED}_{9\mathrm{P}}$ is equal to the sum of the three distances $\mathrm{NED}_{3\mathrm{P}}$ estimated separately from each codon position (property 5 and remark 2 of Michel 2007). Finally, analytical formulae of the particular codon evolutionary distances $\mathrm{CED}_{6\mathrm{P}}$ and $\mathrm{CED}_{3\mathrm{P}}$, with two and one substitution parameters per codon position, respectively, can be easily derived. A CED containing distance information of each three codon positions should allow better recovery of the phylogenetic signal induced by a codon-based alignment of nucleotide sequences compared to a NED. We have developed this approach here with the distance $\mathrm{CED}_{6\mathrm{P}}$ (see Table 1 for a list of abbreviations, formulae, and references related to this particular distance). As with the distance $\mathrm{CED}_{9\mathrm{P}}$, the distance $\mathrm{CED}_{6\mathrm{P}}$ is equal to the sum of the three distances $\mathrm{NED}_{2\mathrm{P}}$ estimated from each codon position (see details below).

After a recall of its analytical formula, we extend the distance $\mathrm{CED}_{6\mathrm{P}}$ (Michel 2007) to a weighted codon evolutionary distance, called $\mathrm{WCED}_{6\mathrm{P}}$, by considering an evolutionary rate μ_p specific to each codon position

Table 1 Summary of abbreviations, formulae, and references related to the nucleotide evolutionary distances NED_{2P} and the codon evolutionary distances CED_{6P}

Evolutionary model	Related distance	Formulae	Original reference
NEM_{2P}	NED_{2P}	(3), (4), (5)	Kimura (1980)
	$UNED_{2P}$	(3), (6), (7)	Tajima (1993)
$GNEM_{2P}$	$GNED_{2P}$	(3), (10), (11)	Jin and Nei (1990)
	$GUNED_{2P}$	(3), (12), (13)	Rzhetsky and Nei (1994)
CEM_{2P}	CED_{6P}	(14), (18)	Michel (2007)
	$WCED_{6P}$	(14), (19)	–
	W^2CED_{6P}	(14), (20)	–

$p = 1, 2, 3$. We demonstrate that this distance $WCED_{6P}$ completes the distance CED_{6P} with a weighting inversely proportional to the three rates μ_p. By using a large set of sequence alignments (both real and simulated), we illustrate the strong heterogeneity of evolutionary rates among the three codon positions, which can lead to a strong bias in the phylogenetic signal produced by estimated NED. Furthermore, we show that the distance $WCED_{6P}$, which naturally compensates for this heterogeneity, leads to distance estimates much closer to additive distance matrices with the same data and, consequently, infer more reliable phylogenetic trees. We also propose another weighting in the distance $WCED_{6P}$, denoted W^2CED_{6P}, for a better discrimination of the codon positions inducing a saturated phylogenetic signal occurring with extreme (high or low) evolutionary rates. Finally, we carry out a distance-based reanalysis of the multigene dataset of Rokas et al. (2003), composed of eight yeast species and 106 codon-based alignments of DNA sequences. We show that these codon distances $WCED_{6P}$ and W^2CED_{6P} allow building the same phylogenetic tree as that inferred by maximum parsimony (MP) or ML, contrary to the trees inferred from NED.

Materials and Methods

The abbreviations used in the following to distinguish the different distances (classical nucleotide distances based on NED_{2P} and new codon distances based on CED_{6P}) are listed in Table 1.

Nucleotide Evolutionary Distance NED_{2P}

Let i and j be two nucleotide sequences having diverged during a time t. Let μ be the global evolutionary rate, and let μ_{ts} and μ_{tv} be the nucleotide substitution rates of transition type (A↔G and C↔T) and transversion type (A↔C, A↔T, C↔G and G↔T), respectively, per site and per time unit. Then the total rate of substitutions is equal to $\mu_{ts} + 2\mu_{tv}$ and the total number $\Delta_{ij}(t)$ of substitutions per site at time t is given by $\Delta_{ij}(t) = 2\mu(\mu_{ts} + 2\mu_{tv})t$. Then the probabilities $P(t)$ and $Q(t)$ of transitions and transversions, respectively, at time t can be easily derived:

$$P(t) = 0.25 + 0.25\exp(-8\mu\mu_{tv}t) - 0.5\exp[-4\mu(\mu_{ts} + \mu_{tv})t] \quad (1)$$

$$Q(t) = 0.5 - 0.5\exp(-8\mu\mu_{tv}t) \quad (2)$$

If the rate μ is not weighted, i.e., $\mu = 1$, and if P_{ij} and Q_{ij} are the observed proportions of transitions and transversions, respectively, between the two sequences i and j, then the nucleotide distance NED_{2P}, noted Δ_{ij}, can be deduced (Kimura 1980) as

$$\Delta_{ij} = \delta_{ij} + \gamma_{ij} \quad (3)$$

where δ_{ij} and γ_{ij}, respectively, are estimated by

$$\delta_{ij} = -0.5\ln\left(1 - 2P_{ij} - Q_{ij}\right) \quad (4)$$

$$\gamma_{ij} = -0.25\ln\left(1 - 2Q_{ij}\right) \quad (5)$$

It should be stressed that the exact estimation of NED_{2P} is $\Delta_{ij} = (\delta_{ij} + \gamma_{ij})/\mu$. However, as the global substitution rate μ has only a homothetic effect on the distance matrix (Δ_{ij}), formula (3) is classically used.

However, the estimators (4) and (5) cannot always be applied. Indeed, the logarithmic term must be positive, implying particularly that $Q_{ij} < 0.5$. Otherwise, formula (3) can underestimate the distance NED_{2P} when the length ℓ of sequences i and j is short (e.g., $\ell \le 100$). Therefore, both Eqs. 4 and 5 should be expressed using the Taylor-series expansion $\ln(1 - x) = \sum_{a=1}^{\infty}(-1)^{a+1}x^a/a$. If s_{ij} and v_{ij} are the observed number of transitions and transversions, respectively, between the two sequences i and j, then s_{ij}, v_{ij}, and $\ell - s_{ij} - v_{ij}$ follow a multinomial distribution with parameters ℓ, P_{ij}, Q_{ij}, and $1 - P_{ij} - Q_{ij}$ leading to unbiased estimates $P_{ij}^{b}Q_{ij}^{a\text{-}b} \approx s_{ij}^{(b)}v_{ij}^{(a\text{-}b)}/\ell^{(a)}$ for $b \le s_{ij}$ and $a - b \le v_{ij}$, and $Q_{ij}^{a} \approx v_{ij}^{(a)}/\ell^{(a)}$ for $a \le v_{ij}$, with $x^{(a)} := x!/(x - a)!$. These different formulae allow unbiased estimates of δ_{ij} and γ_{ij} to be deduced (Tajima 1993):

$$\delta_{ij} = \sum_{a=1}^{s_{ij}+v_{ij}} \frac{1}{a\ell^{(a)}} \sum_{b=\max\{0;a-v_{ij}\}}^{\min\{a;s_{ij}\}} \binom{a}{b} 2^{b-1} s_{ij}^{(b)} v_{ij}^{(a-b)} \tag{6}$$

$$\gamma_{ij} = \sum_{a=1}^{v_{ij}} \frac{2^{a-2} v_{ij}^{(a)}}{a\ell^{(a)}} \tag{7}$$

Formula (3) with estimators (6) and (7) leads to the unbiased nucleotide distance UNED$_{2P}$.

The variance $\mathrm{var}(\Delta_{ij})$ of the distance Δ_{ij} estimated by the delta method (Kimura and Ohta 1972) leads to (Kimura 1980; Tajima 1993)

$$\mathrm{var}(\Delta_{ij}) = \frac{1}{\ell}\left[\lambda_1^2 P_{ij} + \lambda_2^2 Q_{ij} - \left(\lambda_1 P_{ij} + \lambda_2 Q_{ij}\right)^2\right] \tag{8}$$

where $\lambda_1 = \exp(2\delta_{ij})$ and $\lambda_2 = 0.5[\lambda_1 + \exp(4\gamma_{ij})]$, and where δ_{ij} and γ_{ij} are estimated by formulae (4) and (5) or formulae (6) and (7), respectively. This variance $\mathrm{var}(\Delta_{ij})$ can also be estimated by

$$\mathrm{var}(\Delta_{ij}) = \Delta_{ij}^{\rho}\big/\ell \tag{9}$$

with $\rho \approx 2$ (Kuhner and Felsenstein 1994; Criscuolo and Gascuel 2008; see also Sanjuán and Wróbel [2005] for a close bootstrap-based variance estimate).

In order to model the evolutionary rate variation across the sites of the two sequences i and j, the gamma distribution is the most commonly used. The shape of this distribution is related to one parameter, α_Γ. When $\alpha_\Gamma < 1$, this variability is high. When α_Γ increases, e.g., $\alpha_\Gamma \geq 2$, this variability decreases until a common substitution rate for all sites when α_Γ tends toward infinity. Then the gamma nucleotide distance, denoted GNED$_{2P}$, between two sequences i and j is estimated by formula (3), with (Jin and Nei 1990)

$$\delta_{ij} = 0.5\alpha_\Gamma\left[\left(1 - 2P_{ij} - Q_{ij}\right)^{-1/\alpha_\Gamma} - 1\right] \tag{10}$$

$$\gamma_{ij} = 0.25\alpha_\Gamma\left[\left(1 - 2Q_{ij}\right)^{-1/\alpha_\Gamma} - 1\right] \tag{11}$$

An unbiased estimate, denoted GUNED$_{2P}$, of this gamma distance also exists. Based on the Taylor-series expansion $(1-x)^{-m} = 1 + \sum_{a=1}^{\infty} (x^a/a!) \prod_{b=1}^{a} (m+b-1)$, Rzhetsky and Nei (1994) determined the two following unbiased estimates:

$$\delta_{ij} = 0.5\alpha_\Gamma \sum_{a=1}^{s_{ij}+v_{ij}} \frac{1}{\alpha_\Gamma^a \ell^{(a)}} \left[\prod_{b=1}^{a} [(b-1)\alpha_\Gamma + 1]\right] \times \left[\sum_{c=\max\{0;a-v_{ij}\}}^{\min\{a;s_{ij}\}} \frac{2^{c-1} s_{ij}^{(c)} v_{ij}^{(a-c)}}{c!(a-c)!}\right] \tag{12}$$

$$\gamma_{ij} = 0.25\alpha_\Gamma \sum_{a=1}^{v_{ij}} \frac{2^a v_{ij}^{(a)}}{a!\alpha_\Gamma^a \ell^{(a)}} \prod_{b=1}^{a} [(b-1)\alpha_\Gamma + 1] \tag{13}$$

with the same notations as used for formulae (6) and (7). Respective variances of the distances GNED$_{2P}$ and GUNED$_{2P}$ are given by formula (8), with particular estimates of λ_1 and λ_2 (not given; see Tajima 1993 and Rzhetsky and Nei 1994). However, the practical formula (9) can also be used.

Weighted Codon Evolutionary Distance WCED$_{6P}$

Let μ_p be the global evolutionary rate of the codon position $p = 1, 2, 3$, and let $\mu_{\mathrm{ts},p}$ and $\mu_{\mathrm{tv},p}$ be the rates of transitions and transversions, respectively, in the codon position p. The total number $\Delta_{ij}^{\mathrm{cod}}(t)$ of substitutions per codon at time t is given by $\Delta_{ij}^{\mathrm{cod}}(t) = 2t \sum_{p=1,2,3} \mu_p\left(\mu_{\mathrm{ts},p} + 2\mu_{\mathrm{tv},p}\right)$. Then the probabilities $P_p(t)$ and $Q_p(t)$ of transitions and transversions in codon position p, respectively, at time t are obviously obtained by formulae (1) and (2), respectively, by replacing μ, μ_{ts}, and μ_{tv} by μ_p, $\mu_{\mathrm{ts},p}$, and $\mu_{\mathrm{tv},p}$, respectively. Following the same proof as that used to derive the distance CED$_{9P}$ (Michel 2007), the weighted codon evolutionary distance WCED$_{6P}$, denoted $\Delta_{ij}^{\mathrm{cod}}$, is derived:

$$\Delta_{ij}^{\mathrm{cod}} = \sum_{p=1,2,3} w_p \Delta_{ij|p} \tag{14}$$

where $w_p = 1/\mu_p$ and $\Delta_{ij|p}$ is the distance NED$_{2P}$ between the two sequences i and j restricted to the sites corresponding to codon position p. The distances $\Delta_{ij|p}$ can be computed with formula (3), with δ_{ij} and γ_{ij}, respectively, estimated with formulae (4) and (5), (6) and (7), (10) and (11), or (12) and (13). The weighting $w_p = 1/\mu_p$ in formula (14) should allow the important scale differences among the nucleotide distance matrices $(\Delta_{ij|1})$, $(\Delta_{ij|2})$, and $(\Delta_{ij|3})$ to be corrected, for example, in order to avoid the small distances with a slow evolution (e.g., at the second codon position) to be hidden by the large distances with a fast evolution (e.g., at the third codon position). The method SDM* (Criscuolo et al. 2006) will be used to estimate the relative values of the parameters w_p, which will lead to a codon distance matrix $\left(\Delta_{ij}^{\mathrm{cod}}\right)$ satisfactory for tree inference (see below). If each codon position p evolves at the same rate, i.e., $w_p = 1$, then the distance WCED$_{6P}$ is identical to the CED$_{6P}$ one (Michel 2007).

Each codon position p being independent by hypothesis, the variance $\mathrm{var}\left(\Delta_{ij}^{\mathrm{cod}}\right)$ is simply given by

$$\mathrm{var}\left(\Delta_{ij}^{\mathrm{cod}}\right) = \sum_{p=1,2,3} w_p^2 \mathrm{var}\left(\Delta_{ij|p}\right) \tag{15}$$

where $\mathrm{var}\left(\Delta_{ij|p}\right)$ can be estimated by formula (8) or (9).

Real Dataset: The PANDIT Database

This new phylogenetic inference approach based on the weighted codon evolutionary distance WCED$_{6P}$ is evaluated with gene alignments from the database PANDIT 17.0 (Whelan et al. 2006). This database is composed of 7738 families of homologous protein domains. For each family, the alignments of amino acid sequences and their associated codon sequences are available. Furthermore, an optimal phylogenetic tree Φ is also available from each gene alignment (for more details see Whelan et al. 2006).

To avoid a too small or too large number of taxa, only the subset of alignments with at least 10 and at most 100 taxa is considered. There is no restriction with the gene length ℓ, which can be very different in scale, i.e., $27 \leq \ell \leq 5334$ (see below for more details). Therefore, the following results are based on a set of 3211 codon-based alignments.

Exploration Protocol for the Real Dataset

In order to construct a codon distance matrix $\left(\Delta_{ij}^{\text{cod}}\right)$ from each of the 3211 codon-based alignments, the evolutionary distances $\Delta_{ij|p}$ between each pair of taxa i and j are estimated for each codon position p = 1, 2, 3 by formula (3) with unbiased estimates (6) and (7). This choice was preferred compared to estimates (4) and (5) mainly for the following two reasons. On the one hand, the distances $\Delta_{ij|p}$ are impossible to estimate with (4) and (5) when the taxa i and j diverge strongly, a case often observed for $p = 3$, leading to distance matrices $(\Delta_{ij|p})$ containing "infinite" entries which cannot be directly used for phylogenetic inference. On the other hand, estimates (6) and (7) can compute evolutionary distances for genes of short lengths (e.g., $\ell/3 \leq 100$) which are less underestimated than those computed with the estimates (4) and (5) (Tajima 1993; Rzhetsky and Nei 1994).

For each of the 3211 alignments, relative estimates α_p of the inverse $1/\mu_p$ of the evolutionary rates of each codon position $p = 1, 2, 3$ are computed with the method SDM* (Criscuolo et al. 2006; see also the equivalent DistR method of Bevan et al. 2005). With the three codon position distance matrices $(\Delta_{ij|1})$, $(\Delta_{ij|2})$, and $(\Delta_{ij|3})$, SDM* computes the parameters α_1, α_2, and α_3 which minimize the following quadratic criterion:

$$\sum_{i<j}\sum_{p=1,2,3}\ell_p\left(\alpha_p\Delta_{ij|p}-\overline{\Delta_{ij}}\right)^2 \text{ with } \overline{\Delta_{ij}}=\frac{\sum\limits_{p=1,2,3}\ell_p\alpha_p\Delta_{ij|p}}{\sum\limits_{p=1,2,3}\ell_p}, \tag{16}$$

where ℓ_p is the length of each codon position p (note that here $\ell_1 = \ell_2 = \ell_3 = \ell/3$). Criterion (16) is minimized under the linear constraint $\alpha_1 + \alpha_2 + \alpha_3 = 3$. Therefore, the closer α_p is to 3 (0, respectively), the more slowly (quickly, respectively) codon position p evolves compared to the other two. If $\alpha_1 \approx \alpha_2 \approx \alpha_3 \approx 1$, then the three codon positions have similar evolutionary rates. The left part of Fig. 1 illustrates the procedure used to estimate the three values α_1, α_2, and α_3 which correspond to the inverse of the relative evolutionary rates μ_p of the three codon positions p = 1, 2, 3 for each of the 3211 considered alignments.

In order to quantify and evaluate the phylogenetic signal induced by the different estimated distance matrices, the arboricity coefficient *Arb* (Guénoche and Garreta 2000) is used. If the distance matrix (Δ_{ij}) exactly represents a valuated tree T (i.e., if each value Δ_{ij} is equal to the path-length distance between taxa i and j in T), then (Δ_{ij}) is additive and verifies the quadrangular inequality $\Delta_{ij} + \Delta_{xy} \leq \max\{\Delta_{ix} + \Delta_{jy}; \Delta_{iy} + \Delta_{jx}\}$ for each quartet of distinct taxa i,j,x,y (Zaretskii 1966; Buneman 1971). Therefore, if a pair of taxa i,j is separated by another pair of taxa x,y by at least one internal branch in T, then

$$\Delta_{ij} + \Delta_{xy} < \Delta_{ix} + \Delta_{jy} = \Delta_{iy} + \Delta_{jx} \tag{17}$$

In practice, a distance matrix directly estimated from a sequence alignment is rarely additive and the three sums in formula (17) often differ. If these three sums are sorted and denoted according to their increasing values S_{min}, S_{med}, and S_{max}, then the arboricity coefficient *Arb* measures the proportion of quartets of taxa such that the middle sum S_{med} is closer to the highest one, S_{max}, than the lowest one, S_{min}, i.e.,

$$Arb\left(\Delta_{ij}\right) = \left|\left\{\{i,j,x,y\} : S_{max} - S_{med} < S_{med} - S_{min}\right\}\right| \bigg/ \binom{n}{4}$$

This topological criterion being normalized by $\binom{n}{4}$ (i.e., the total number of quartets of distinct taxa induced by an n-taxon tree), *Arb* ranges from 0 to 1 and allows the level of the phylogenetic signal induced by a distance matrix to be quantified. The closer *Arb* is to 1, the closer the distances are to an additive distance, i.e., a phylogenetic tree.

In order to characterize the strong heterogeneity of both the evolutionary rate and the quality of phylogenetic signal induced by the three codon positions, the criterion *Arb* is computed for each of the 3211 gene alignments with the three previously computed codon position distance matrices $(\Delta_{ij|1})$, $(\Delta_{ij|2})$, and $(\Delta_{ij|3})$. These 3 × 3211 values of *Arb* are represented graphically in Fig. 2 as a function of the corresponding parameters α_1, α_2, and α_3, and discussed below.

For each of the 3211 codon-based alignments, the seven distance matrices corresponding to the evolutionary distances summarized in Table 1 are estimated.

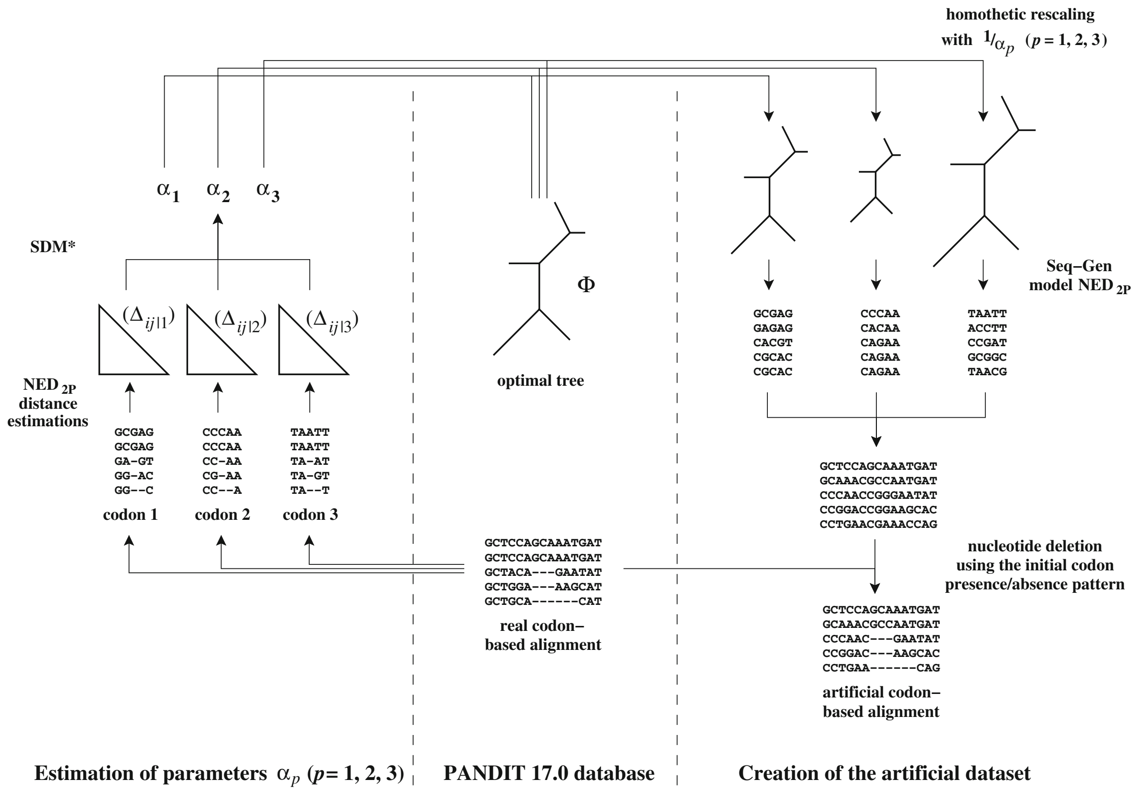
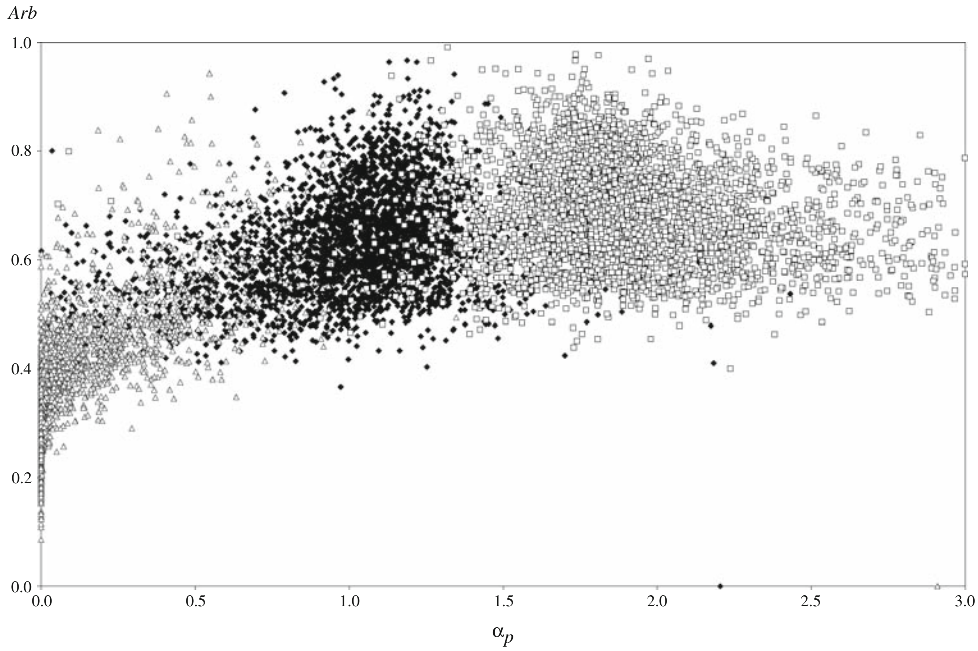


Fig. 1 Flowchart illustrating the process used to estimate the α_p values from each codon position $p = 1, 2, 3$ of the real dataset (left) and to generate the closely related artificial dataset from the α_p values, the model tree Φ, and the codon presence/absence pattern (right)

Fig. 2 Arboricity coefficient *Arb* as a function of the estimate α_p of the inverse of the evolutionary rates of each codon position $p = 1, 2, 3$ for each of the 3211 codon-based alignments of the PANDIT database (real dataset). Codon position $p = 1$ is represented by filled diamonds (◆); $p = 2$, by open squares (□); and $p = 3$, by open triangles (△)



- NED$_{2P}$. The incomputable distances (due to the logarithm of a negative number in 1.8% of the inferred distances with the 3211 alignments) are estimated with an algorithm of completion of incomplete distance matrix (Guénoche and Grandcolas 1999, 2000).

- $UNED_{2P}$. The inferred distances are always computable, whatever the values of P_{ij} and Q_{ij} (see above).
- $GNED_{2P}$. First, ML trees are built with PhyML (model $GNEM_{2P}$), and second, the 3211 estimates of the α_Γ parameter are used with formulae (3), (10), and (11). Incomputable distances are estimated in the same way as NED_{2P}.
- $GUNED_{2P}$. This distance is used with the same α_Γ values as $GNED_{2P}$.
- CED_{6P}. The previously computed distance matrices $(\Delta_{ij|p})$ $(p = 1, 2, 3)$ are used in formula (14) with

  $$w_p = 1 \tag{18}$$

  associated with a common evolutionary rate in the three codon positions.
- $WCED_{6P}$. This procedure is identical to the previous calculation of CED_{6P} but with

  $$w_p = \alpha_p \tag{19}$$

  where α_p are the previously computed estimates of the relative inverse of the evolutionary rates μ_p induced by each of the three codon positions $(p = 1, 2, 3)$. This more general distance definition allows the three codon positions to have different evolutionary rates.
- W^2CED_{6P}. This procedure is identical to the previous calculation but with

  $$w_p = \alpha_p Arb\left(\Delta_{ij|p}\right)/v \tag{20}$$

  where $Arb(\Delta_{ij|p})$ are the previously computed arboricity coefficients $(p = 1, 2, 3)$ and where $v = \sum_{p=1,2,3} \alpha_p Arb\left(\Delta_{ij|p}\right)/3$ normalizes the three weights w_1, w_2, and w_3. This distance definition attributes low weights to the codon positions associated with a too weak phylogenetic signal.

For each of these seven distance estimates, the average arboricity value *Arb* obtained with the 3211 alignments is given in Table 2. A sign test (MacStewart 1941; Dixon and Mood 1946; Hemelrijk 1952) is also performed to assess the statistical significance between the observed average values of the coefficient *Arb*. Two sets of 3211 values of *Arb* are considered significantly different if the *p*-value returned by the sign test is <0.05. Significant best average values *Arb* for the seven observed distance estimates are underscored in Table 2.

Three fast tree inference algorithms are applied to the 3211 codon-based alignments and for each of the seven distance formulae (i.e., 22,477 distance matrices studied).

- FastME (Desper and Gascuel 2002) searches the tree, minimizing the same criterion used in the algorithm NJ but faster and more accurately (Desper and Gascuel 2004; Vinh and von Heaseler 2005; Gascuel and Steel 2006).
- BioNJ (Gascuel 1997) is based on the same algorithmic scheme as NJ but improves the estimation step of tree branch lengths by using a Poisson model of the variances and covariances of the distances.
- WLS-MVR (Gascuel 2000) generalizes BioNJ, as any variance model can be associated with the evolutionary distances. Preliminary tests (results not shown) show that better trees with WLS-MVR are obtained when the variances are estimated with formula (9) with $\rho = 2$ compared to formula (8). Therefore, WLS-MVR is applied to the seven distance matrices with variances estimated by either formula (9) or formula (15) with $\rho = 2$.

These three distance-based tree inference algorithms were chosen among other fast methods because of their significantly accuracy during preliminary tests (results not shown).

In complement to the arboricity coefficient *Arb*, we use the variance accounted for *Vaf* (e.g., Guénoche and Garreta 2000; Hubert et al. 2006). If (Δ_{ij}) is a distance matrix and T its inferred tree, then the variance accounted for *Vaf* is expressed by the following formula:

$$Vaf\left(\Delta_{ij}, T\right) = \max\left\{0; 1 - \frac{\sum_{i<j}\left(\Delta_{ij} - T_{ij}\right)^2}{\sum_{i<j}\left(\Delta_{ij} - \bar{\Delta}\right)^2}\right\}$$

where T_{ij} is the patristic (path-length) distance between taxa i and j, and $\bar{\Delta}$, the average of the values in (Δ_{ij}). The metric criterion *Vaf* corresponds to the quadratic difference between (Δ_{ij}) and (T_{ij}) normalized by the variance of the entries in (Δ_{ij}), and quantifies the level of fitness between the fitted branch lengths of T and the distances in (Δ_{ij}). For each of the seven distance estimates and for each of the three inferred trees, the average criteria *Vaf* with the 3211 alignments are given in Table 2. For each of the three tree reconstruction algorithms, the statistical significance of the best average *Vaf* criterion was assessed by a sign test in a similar way as previously.

Artificial Dataset

We describe here the protocol to generate artificial codon-based alignments associated with the 3211 real codon-based alignments of the PANDIT database. The flowchart in Fig. 1 illustrates the following description.

The PANDIT database associates an optimal tree Φ with each real alignment. Each tree Φ is considered as a model tree to generate an artificial alignment. For each codon position p = 1, 2, 3, the model tree Φ is rescaled such that its total length (i.e., the sum of all its branch lengths) is

Table 2 Accuracy of different estimations of evolutionary distances and tree building methods for real and artificial datasets

	Real datasets				Artificial datasets							
	Arb[a]	*Vaf*[a]			*Arb*[a]	*Vaf*[a]			d_q (proportion of retrieved model trees)[a]			
		FastME	BioNJ	WLS-MVR		FastME	BioNJ	WLS-MVR	FastME	BioNJ	WLS-MVR	PhyML
NED_{2P}	0.6567	0.6534	0.6713	0.6614	0.7016	0.7766	0.8068	0.8029	0.1780 (11.3%)	0.1673 (11.0%)	0.1672 (10.7%)	0.1686 (14.8%)
UNED_{2P}	0.6614	0.6710	0.7116	0.6988	0.7018	0.7766	0.8068	0.8029	0.1780 (11.3%)	0.1673 (11.0%)	0.1672 (10.7%)	–
GNED_{2P}	0.5714	0.4967	0.5007	0.4904	0.6132	0.6590	0.6814	0.6609	0.2065 (8.7%)	0.2018 (8.5%)	0.1921 (8.5%)	$\underline{0.1105\ (24.3\%)}$
GUNED_{2P}	0.5938	0.5502	0.5837	0.5642	0.6216	0.6757	0.6999	0.6796	0.2034 (8.8%)	0.1990 (8.7%)	0.1902 (8.6%)	–
CED_{2P}	0.4358	0.2056	0.2368	0.2141	0.4640	0.3080	0.3329	0.3271	0.4652 (5.1%)	0.4577 (5.5%)	0.4540 (5.0%)	–
WCED_{6P}	0.6794	0.7807	0.8062	0.8019	0.7939	0.9431	0.9454	0.9448	0.1273 (13.5%)	0.1263 (13.4%)	0.1267 (12.4%)	–
$\text{W}^2\text{CED}_{6P}$	$\underline{0.6927}$	$\underline{0.8007}$	$\underline{0.8263}$	$\underline{0.8216}$	$\underline{0.8197}$	$\underline{0.9577}$	$\underline{0.9599}$	$\underline{0.9590}$	$\underline{0.1200\ (16.5\%)}$	$\underline{0.1205\ (16.1\%)}$	$\underline{0.1203\ (15.5\%)}$	–

[a] For each column, the significantly best average criterion value (as assessed by a sign test) is underscored

equal to the global rate $1/\alpha_p$ with the parameter α_p previously estimated by SDM* (see above). The software Seq-Gen (Rambaut and Grassly 1997) simulates evolution of the three codon positions from the model tree Φ rescaled by $1/\alpha_p$. Seq-Gen is used with the model NEM_{2P} with a precomputed transition/transversion ratio κ_p specific to each codon position p. The transition/transversion ratio κ_{ij} between two nucleotide sequences i and j being estimated by $\kappa_{ij} = \delta_{ij}/\gamma_{ij} - 1$ (Jukes 1987; Yang and Yoder 1999), we have defined κ_p here as the average of all pairwise estimates $\kappa_{ij|p}$, i.e., $\kappa_p = 2\sum_{i<j}\kappa_{ij|p}/(n(n-1))$, where $\kappa_{ij|p} = \delta_{ij|p}\big/\gamma_{ij|p} - 1$, with $\delta_{ij|p}$ and $\gamma_{ij|p}$ estimated from the associated real alignment by formulae (6) and (7), respectively (see also Galtier and Gouy [1995] for a similar estimation of the transition/transversion ratio from a nucleotide alignment). The length $\ell/3$ of each of the three generated artificial alignments is obtained from the length ℓ of the associated real alignment. Therefore, three artificial alignments are obtained, each having been generated from the same model tree Φ but with a specific global rate $1/\alpha_p$. Then these three alignments are gathered into a unique artificial one per codon. Finally, all the deletions (i.e., gaps) in the real alignments are introduced into the artificial ones. We thus obtain 3211 artificial codon-based alignments closely associated with the 3211 real ones.

Exploration and Tree Building for the Artificial Dataset

The 3211 artificial codon-based alignments are analyzed according to the same protocol as the 3211 real ones (see above). The average values of arboricity *Arb* and variance accounted for *Vaf* with these 3211 artificial alignments are reported in Table 2.

In addition, from the three distance matrices $(\Delta_{ij|1})$, $(\Delta_{ij|2})$, and $(\Delta_{ij|3})$, the phylogenetic trees T^1, T^2, and T^3, respectively, are inferred with the algorithm BioNJ, and the quartet distance d_q (Estabrook et al. 1985) is computed between the model tree Φ and each tree T^p ($p = 1, 2, 3$). This topological distance measures the number of subtrees with four leaves which exist in only one of both compared trees. Then it is normalized by $2\binom{n}{4}$ in order to restrict it to the interval [0,1]. A distance $d_q = 0$ means that the two compared trees are identical. It was chosen compared to other measures (e.g., Williams and Clifford 1971; Waterman and Smith 1978; Robinson and Foulds 1979; Steel and Penny 1993; Goddard et al. 1994) because it has relative stability, with small differences between pairs of trees (see a simple example in Hartmann and Vision 2008). Moreover, thanks to its large range (i.e., its normalizing factor),

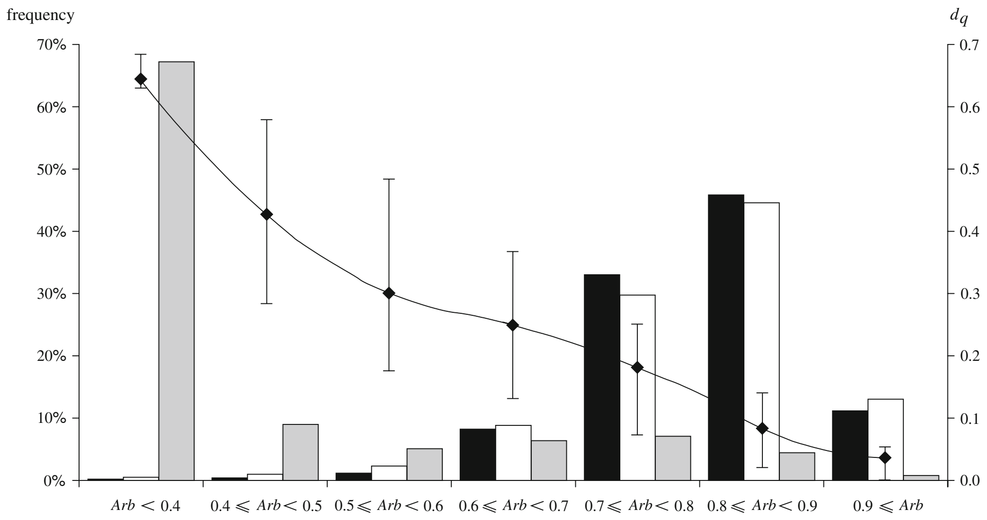


Fig. 3 Relation between the arboricity coefficient *Arb* and the quartet distance d_q observed with the 3211 artificial codon-based alignments. The curve represents the observed average values d_q (right ordinate) as a function of different classes of *Arb* values (abscissa). The error bars represent the two quartiles of the observed d_q values. The histograms represent the proportion of distance matrices (left ordinate) by codon position $p = 1, 2, 3$ associated with each class of *Arb* values (abscissa). Black, white and grey histograms symbolize the codon position $p = 1, 2, 3$, respectively

d_q is more precise than the widely used bipartition distance (Robinson and Foulds 1979). Finally, as it is weakly dependent on the number n of taxa (Steel and Penny 1993; Bryant et al. 2000), it is well suited to the study of the 3211 artificial alignments where n varies between 10 and 100. The $3211 \times 3 = 9633$ average values d_q observed with these 3211 alignments and the three codon positions are represented in Fig. 3 as a function of different value intervals of arboricity *Arb*. Figure 3 also gives the distribution of the proportion of distance matrices per value intervals of *Arb* and for each of the three codon positions.

For each of the 3211 artificial codon-based alignments and for each of the seven types of distance matrices (Table 1), the distance d_q is also computed between the model tree Φ and the three trees FastME, BioNJ, and WLS-MVR previously inferred to evaluate the variance accounted for by *Vaf*. In order to compare the distance-based trees to those inferred with the ML criterion, the software PhyML (Guindon and Gascuel 2003) is used to construct a tree from each artificial alignment according to models $\text{NEM}_{2\text{P}}$ (with the transition/transversion ratio κ left as a free parameter) and $\text{GNEM}_{2\text{P}}$ (with the parameter α_Γ of the gamma distribution also left as a free parameter). The distances d_q between the ML trees and the model trees Φ are also computed. All average values d_q and the percentages of retrieved model trees Φ are reported in Table 2. As previously, a sign test is also performed for each of the four tree reconstruction approaches. As the alignment sizes are extremely heterogeneous, with respect to the number n of taxa (i.e., $n \in [10, 100]$) as well as the number ℓ of sites (i.e., $\ell/3 \in [9, 1778]$), the average values d_q of BioNJ (with the distances $\text{UNED}_{2\text{P}}$, $\text{WCED}_{2\text{P}}$, and $\text{W}^2\text{CED}_{2\text{P}}$) and PhyML (with models $\text{NEM}_{2\text{P}}$ and $\text{GNEM}_{2\text{P}}$) are graphically represented as a function of different classes of sizes n and ℓ in Fig. 4.

Distance Estimations and Tree Building for the Rokas et al. (2003) Dataset

This dataset is composed of 106 codon-based genes sequenced from seven *Saccharomyces* species and *Candida albicans* as the outgroup taxon (for more details see Rokas et al. 2003). As it is a multigene dataset, two levels of gene combination are used to estimate the pairwise evolutionary distances (Schmidt 2003, chap. 7; Criscuolo et al. 2006).

- Low-level combination. The 106 gene alignments are concatenated into a unique supermatrix of characters from which the distances $\text{UNED}_{2\text{P}}$, $\text{WCED}_{6\text{P}}$, and $\text{W}^2\text{CED}_{6\text{P}}$ are directly estimated.
- Medium-level combination. A distance matrix is estimated with the distances $\text{UNED}_{2\text{P}}$, $\text{WCED}_{6\text{P}}$, and $\text{W}^2\text{CED}_{6\text{P}}$ from each alignment. These 106 distance matrices are then combined in a unique distance supermatrix by the method SDM* (Criscuolo et al. 2006). This approach minimizes criterion (16) with $p = 1, 2, \ldots, 106$ where $(\Delta_{ij|p})$ corresponds to the distance matrix estimated from the gene p, and then it considers the distance supermatrix $\overline{\Delta_{ij}}$ to infer a tree (for more details see Criscuolo et al. 2006).

A phylogenetic tree is inferred by BioNJ from these six distance (super)matrices (i.e., low- and medium-level

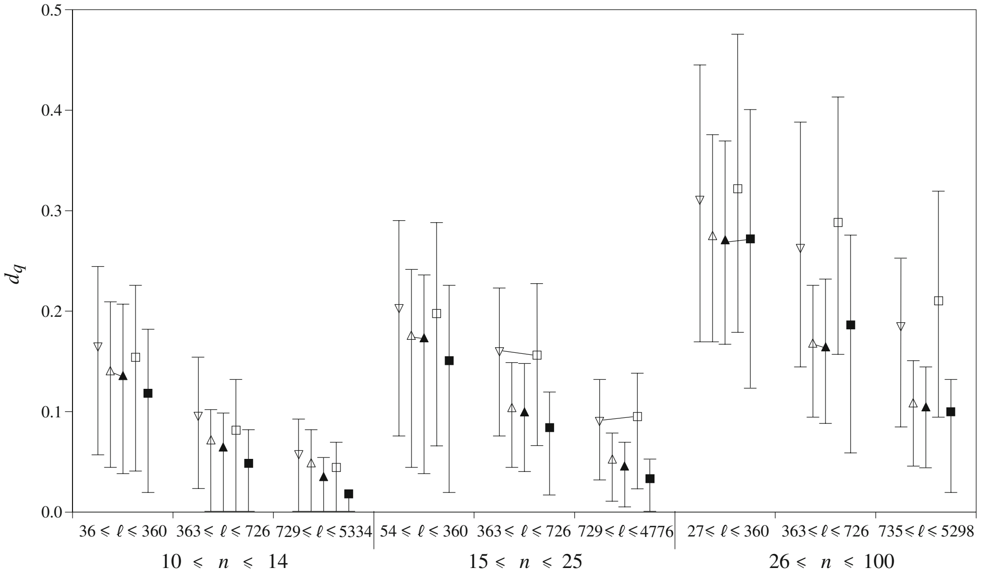


Fig. 4 Average values d_q observed with five phylogenetic tree reconstruction methods as a function of different classes of artificial codon-based alignments depending on the number of taxa n and the number of sites ℓ. The distance-based tree reconstruction method BioNJ is based on the unbiased nucleotide distance UNED$_{2P}$ and on the two weighted codon distances, WCED$_{6P}$ and W^{2}CED$_{6P}$, which are symbolized by open downward triangles (▽), open triangles (△), and filled triangles (▲), respectively. The software PhyML is symbolized by open squares (□) for the model NEM$_{2P}$ and by filled squares (■) for the model GNEM$_{2P}$. Symbols are joined if their average values are not significantly different (as assessed by a sign test with a p-value > 0.05). The error bars represent the two quartiles of the observed d_q values. Note that the 75% quartile is equal to 0 with the third filled square

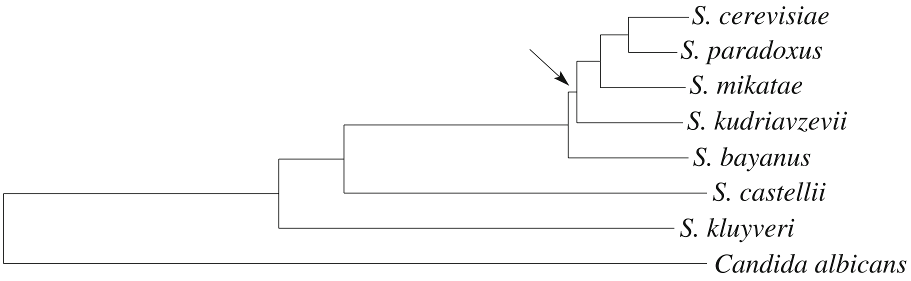


Fig. 5 Phylogenetic tree inferred from the multigene dataset of Rokas et al. (2003) with the distance-based tree reconstruction BioNJ from the distance supermatrix SDM* applied to the 106 distance matrices based on the weighted codon distance W^{2}CED$_{6P}$. All internal branches are 100% bootstrap supported by all distance-based approaches used, except the one marked, whose different bootstrap supports are reported in Table 3

Table 3 Bootstrap support (of 1000 replicates) of the branch marked in Fig. 5 according to distance-based approach used

	Low-level combination	Medium-level combination
UNED$_{2P}$	197	145
WCED$_{6P}$	575	564
W^{2}CED$_{6P}$	645	812

combinations for each of the three distances UNED$_{2P}$, WCED$_{6P}$, and W^{2}CED$_{6P}$). Degrees of support are computed for each internal branch with a codon-based bootstrap process based on 1000 replicates. This procedure samples the codon characters with replacement according to the same procedure as the standard bootstrap with nucleotide characters. In other words, a block-bootstrapping (Künsch 1989; Liu and Singh 1992) is performed, with blocks of size 3. The tree obtained by medium-level combination with the distance W^{2}CED$_{6P}$ is given in Fig. 5. Some observed bootstrap proportions depending on the six distance-based approaches are reported in Table 3 and discussed below. It should be stressed that the distances NED$_{2P}$ and UNED$_{2P}$ give exactly the same tree and bootstrap proportions in both low- and medium-level combinations.

Results and Discussion

Evolutionary Rate and Phylogenetic Signal Heterogeneities Among Codon Positions

The analysis of the 3211 real codon-based alignments of the PANDIT database according to the arboricity coefficient *Arb* is given as a function of the SDM* estimate α_p of the relative inverse of the global evolutionary rate for each codon position $p = 1, 2, 3$ in Fig. 2. An obvious separation of the cloud of points into three clusters associated with each codon position is observed. The position p = 1 (2 and 3, respectively) is centered with an average value α_1 (α_2 and α_3, respectively) of 1.00 (1.89 and 0.10, respectively). As expected, third codon positions have the highest (relative) evolutionary rates, i.e., 1.00/0.10 $\approx$10 times faster than first codon positions and 1.89/0.10 $\approx$19 times faster than second codon positions. Figure 2 also identifies an increasing relation between the parameters α_p and the criterion values *Arb*. The codon positions with fast evolutionary rates (e.g., $\alpha_p < 0.5$) are associated with values of *Arb* generally <0.5. This observation means that more half of the quartets do not verify property (17) of distance matrices defining phylogenetic trees. These codon positions with bad phylogenetic descriptors are globally associated with the third position, as the average *Arb* values are 0.63, 0.68, and 0.37 for codon positions $p = 1$, 2, and 3, respectively. This result is also confirmed by two other statistical parameters which retrieve the same order of *Arb* values. Indeed, their medians are 0.62, 0.68, 0.35 for $p = 1$, 2, 3, respectively, and the two quartiles are 0.55, 0.61, 0.29 and 0.69, 0.75, 0.43 for $p = 1$, 2, 3, respectively. Furthermore, similar results with both the *Arb* criterion and α_p values are obtained with some others nucleotide distances NEDs (e.g., Jukes and Cantor 1969; Kimura 1981; Tamura and Nei 1993; results not shown). Therefore, as the codon positions have very different relative evolutionary rates and as the estimate α_p is inversely proportional to the order of magnitude of the values in the distance matrix ($\Delta_{ij|p}$), a weighting, w_p, is necessary to compute an accurate CED with formula (14). Indeed, Table 2 shows that the worst values of the criteria *Arb* and *Vaf* and the poorest trees (i.e., the highest values d_q) are obtained with the codon distance CED$_{6P}$ (i.e., without weighting), while the codon distances WCED$_{6P}$ and W^{2}CED$_{6P}$ (i.e., with weighting) lead among the (significantly) best ones.

The artificial dataset is generated from the real dataset using the α_p estimated from each codon position $p = 1$, 2, 3 of the real dataset in order to consider the strong evolutionary rate heterogeneity among them (see Materials and Methods and Fig. 1). When the exploration protocol was performed on this artificial dataset, new estimates of α_p, denoted $\hat{\alpha}_p$, were obtained in order to build the distance matrices WCED$_{6P}$ and W^{2}CED$_{6P}$. These estimates $\hat{\alpha}_p$ are globally close to the real α_p (i.e., correlation coefficient r $\approx$0.9), showing that SDM* performs well to estimate the relative inverse of the evolutionary rates induced by each codon position. The 3211 artificial codon-based alignments according to the arboricity coefficient *Arb* as a function of the estimate $\hat{\alpha}_p$ for each codon position p led to a representation similar to that in Fig. 2 (results not shown). Therefore, there still exists an increasing relation between the parameters $\hat{\alpha}_p$ and the criterion values *Arb* with the artificial dataset. Moreover, as there is a relative correlation between the arboricity *Arb* and the distance d_q (Fig. 3), the coefficient *Arb* in the artificial codon-based alignments still represents a good estimate of a distance matrix to analyze a tree reconstruction algorithm which represents evolutionary history as accurately as possible. Graphical representations similar to Fig. 3 are also obtained with FastME and WLS-MVR (results not shown). Figure 3 shows that the saturation effect exists with a great number of codon positions of artificial alignments (mainly the third position, broadly inducing fast evolutionary rates as shown in Fig. 2). More precisely, most of the third codon positions of artificial alignments lead to trees with a distance d_q $\approx$0.66 from the model tree Φ, which corresponds to the expected value of d_q with two random phylogenetic trees in the same leaf set (Steel and Penny 1993).

In general, there is a low value of the coefficient *Arb* with the distance matrices ($\Delta_{ij|3}$), particularly when the third codon position has a very high (relative) evolutionary rate (e.g., $\alpha_3 < 0.5$; Fig. 2). However, this observation is not always verified. Indeed, some matrices ($\Delta_{ij|3}$) have good values of *Arb* (e.g., $Arb > 0.7$) despite a high evolutionary rate (e.g., $\alpha_3 < 0.6$). Moreover, the artificial dataset contains $\approx$14% of third codon positions that induce high values of *Arb* (e.g., $Arb \geq 0.7$; Fig. 3) and, consequently, better d_q values (e.g., $d_q < 0.3$). Therefore, a lack of systematic consideration of the third codon position, or equivalently a phylogeny recontruction based only on the two first codon positions, can lead to a relative lost of phylogenetic signal (Cummings et al. 1995; Yang 1996, 1998; Zardoya and Meyer 1996).

Accuracy of Distance-Based Trees Inferred from Codon-Based Estimation of Evolutionary Distances

In a general way, the weighted codon evolutionary distances WCED$_{6P}$ and W^{2}CED$_{6P}$ lead to distance matrices that represent with accuracy the phylogenetic signal induced by a gene alignment compared to all studied nucleotide evolutionary distances (NED$_{2P}$, UNED$_{2P}$, GNED$_{2P}$, and GUNED$_{2P}$; see Table 2 and Fig. 4). Among these four distances, the unbiased distances (i.e., UNED$_{2P}$ and GUNED$_{2P}$) show performances similar to those of the biased ones (i.e., NED_{2P} and $GNED_{2P}$, respectively), whereas the gamma distances (i.e., $GNED_{2P}$ and $GUNED_{2P}$) lead to phylogenetic signal and trees that are not as good. The values *Arb* and *Vaf* are obviously higher with the artificial dataset, but the hierarchy of performances in Table 2 is very similar to the hierarchy obtained with the real dataset. Finally, among the three distance-based tree reconstruction algorithms, sign tests on both *Vaf* and d_q indicate that BioNJ shows slightly better performances with the seven distance estimations (results not shown).

As reported in Table 2, the codon distances are significantly more efficient than the nucleotide distances for inferring phylogenetic trees (with the already mentioned exception of CED_{6P}). The best result is obtained with the distance W^2CED_{6P} with $d_q = 0.1200$ and $\approx 16\%$ of retrieved model trees Φ with FastME. The distance $WCED_{6P}$ with $d_q = 0.1263$ and $\approx 13\%$ of retrieved model trees Φ with BioNJ also leads to very good performance compared to the (non-gamma) NED with $d_q = 0.1673$ and $\approx 11\%$ of retrieved model trees. Moreover, both distances, $WCED_{6P}$ and W^2CED_{6P}, lead to better trees than the ML approach without the gamma parameter. Indeed, sign tests performed between the d_q values returned by PhyML with the model NEM_{2P} and the d_q values returned by FastME, BioNJ, and WLS-MVR with the distances $WCED_{6P}$ and W^2CED_{6P} show that these six codon distance-based tree reconstruction approaches are significantly better than the ML one (i.e., *p*-values always $<10^{-5}$). However, the use of the gamma correction strongly improves the ML approach. Indeed, the value $d_q = 0.1105$ presented by PhyML with the model $GNEM_{2P}$ is, according to sign tests, significantly the best value among all studied approaches (Table 2). Curiously, the use of such a gamma correction decreases the phylogenetic signal in nucleotide distance-based approaches (e.g., low values of *Arb* and high values of d_q in Table 2). This observation could be explained by a highest variance in the distance estimates $GNED_{2P}$ and $GUNED_{2P}$ or by an incorrect estimate of the α_Γ parameter (Guindon and Gascuel 2002). However, no significantly better trees are obtained using the distance-based estimates α_Γ returned by the software GAME (Guindon and Gascuel 2002; results not shown).

As expected, the approaches that consider the evolutionary rate variation among codon positions lead to more accurate phylogenetic trees. Table 2 shows that the codon distances $WCED_{6P}$ and W^2CED_{6P} give the best results among all distance-based approaches and that the model trees Φ are more often retrieved with an ML-based phylogenetic reconstruction using the gamma correction. In order to better discriminate these different approaches, Fig. 4 illustrates the range of some observed values d_q depending on several intervals of the numbers of taxa n and sites ℓ. These intervals are defined with the 33%- and 66%-deciles of n (i.e., 14 and 26, respectively) and ℓ (i.e., 360 and 726, respectively) in the set of the 3211 observed values. These nine partitions of the artificial dataset have similar sizes, ranging from 322 (i.e., $26 \leq n \leq 100$ and $735 \leq \ell \leq 5298$) to 419 (i.e., $10 \leq n \leq 14$ and $729 \leq \ell \leq 5334$). As expected, the best values d_q are obtained with both low n and high ℓ. As shown in Table 2, approaches taking into account the rate variation among sites (ML with model $GNEM_{2P}$) or codon positions (codon distances $WCED_{6P}$ and W^2CED_{6P}) always retrieve model trees Φ more accurately (especially with large values of n). Surprisingly, for $26 \leq n \leq 100$, BioNJ with the codon distance W^2CED_{6P} shows performances similar to (with $27 \leq \ell \leq 360$) or significantly better than (with $363 \leq \ell \leq 726$) those obtained using PhyML with the model $GNEM_{2P}$. The latter particular cases are obviously not a general conclusion (see the other cases in Fig. 4), but whatever the performance, the tree inference from the codon distances $WCED_{6P}$ and W^2CED_{6P} may become necessary with datasets of large sizes. For example, with a Pentium IV 1.6-GHz (1-Gb RAM) PC, PhyML (with $GNEM_{2P}$) needs a run time of about 25 h (3 days, respectively) to infer the 3211 trees from the artificial (real, respectively) dataset, while all distance-based approaches (e.g., BioNJ with W^2CED_{6P}) complete this whole tree reconstruction process in at most 10 min (see also below the discussion about the dataset of Rokas et al. 2003).

Finally, codon-based likelihoods were estimated using PAML (Yang 2007) in order to compare with the previous computer experiments the most accurate distance-based tree reconstruction method (i.e., the evolutionary distance W^2CED_{6P} and its phylogenetic tree inferred by BioNJ) to the ML-based tree building approach. Since PAML provides numerous parameter-rich codon evolutionary models implying very high computation times, we have selected from the 3211 real codon-based alignments the subset of 270 alignments with $n = 10$ taxa. PhyML was used to infer 270 ML trees from these 270 codon-based alignments using one of the most parameter-rich NEMs, i.e., the General Time Reversible model (GTR; Rodriguez et al. 1990; Lanave et al. 1984; Yang 1994), with corrections for invariant characters and rate variation across sites (GTR + I + G). This 10-parameter NEM was chosen because it is the most selected model in practice to build representative phylogenetic trees from alignments of real nucleotide sequences (Kelchner and Thomas 2006). The likelihood (*lk*) of each of these 270 ML trees as well as their corresponding 270 distance-based trees inferred by BioNJ with the distance matrices W^2CED_{6P} were estimated by PAML with the codon model F3 × 4MG (Muse and Gaut 1994) with all parameters left free (see Ren et al. [2005], Yang and Nielsen [2008], and the PAML documentation for more details). Among the set of 270 pairs of

ML- and distance-based phylogenetic trees, 41 ($\approx$15.2%) present an identical topology (and, consequently, an identical log *lk* value), 131 ($\approx$48.5%) have the ML tree with the best log *lk* value, and 98 ($\approx$36.3%) correspond to more likely distance-based trees. Even if the sign test for these 131 and 98 pairs of trees gives a *p*-value $\approx$0.034 and assesses a significantly better accuracy of one of the most parameter-rich ML tree building approaches (GTR + I + G), our faster distance-based approach based on only six parameters (W^2CED_{6P}) leads to trees with similar or better F3 × 4MG likelihood in $\approx$51.5% of the cases (i.e., $\approx$15.2% + 36.3%).

It should be stressed that several other criteria have been used to characterize the phylogenetic signal induced by a distance matrix, such as δ-plots (Holland et al. 2002), stress or distorsion criteria, and the rate of well-designed quartets (Guénoche and Garreta 2000). All these criteria classify the distance estimates as *Arb* and *Vaf* in Table 2 (results not shown). However, *Arb* in the weighting (20) shows better results among the previous criteria (results not shown). Finally, similar classifications of performances are obtained with other nucleotide distances (e.g., NED_{1P} and NED_{3P}) and their related codon distances (results not shown). It should also be stressed that the average values *Arb* and *Vaf* reported in Table 2 with the real dataset are less accurate than those observed with the amino acid evolutionary distances PAM (Dayhoff 1979) and JTT (Jones et al. 1992) as implemented in the PHYLIP package (Felsenstein 2005), i.e., $Arb = 0.7087$ and $Vaf = 0.8547$ with the PAM distance, and $Arb = 0.7195$ and $Vaf = 0.8786$ with the JTT distance. Indeed, the codon distances $WCED_{6P}$ and W^2CED_{6P} are based on an extension of NEM that does not consider base composition bias. However, our approach can easily be improved by using other distances NED. For example, we obtain $Arb =$ 0.7298 and $Vaf = 0.8788$ when using the formula suggested by Tamura and Nei (1984) to estimate Δ_{ijlp} in Eq. 14 with the weights w_p computed by formula (20). In conclusion, our phylogenetic inference approaches based on weighted codon distances improved the phylogenetic signal in distance matrices, and consequently the accuracy of phylogenetic inference, compared to any nucleotide distances.

Distance-Based Trees from Rokas et al. (2003)

Several phylogenetic analyses of this multigene dataset containing 106 genes have been realized at different levels (nucleotides, codons, and amino acids) and using different criteria (MP, ML, and Bayesian-based [Rokas et al. 2003; Phillips et al. 2004; Taylor and Piel 2004; Ren et al. 2005]). In the low-level combination framework (see Materials and Methods for definition), all these studies infer the same phylogenetic tree (Fig. 5), in which each of the five internal branches is strongly supported (e.g., 100% bootstrap proportion in Rokas et al. [2003] and Ren et al. [2005]). However, with a similar approach (i.e., low-level combination), but using NED on the supermatrix of characters, Phillips et al. (2004) did not produce the same tree. This tree, built from GTR (Rodriguez et al. 1990; Lanave et al. 1984; Yang 1994) and LogDet (Lake 1994; Lockhart et al. 1994; Steel 1994) distance matrices, shows the species *S. kudriavzevii* and *S. bayanus* as monophyletic, supported by a 100% bootstrap proportion, in contradiction to the phylogeny in Fig. 5. However, it should be stressed that Phillips et al. (2004, p. 1455) consider that the tree in Fig. 5 “appears correct” and that it corresponds to the best log *lk* among numerous evolutionary models (Ren et al. 2005).

On the one hand, one can observe that both GTR and LogDet distances are parameter-rich, inducing distance estimates with a large variance and, then, often leading to imprecise distance-based tree reconstruction (e.g., Saitou and Nei 1987; Takahashi and Nei 2000). Nevertheless, our low-level combination using the distance $UNED_{2P}$ (associated with a low variance, thanks to its only two parameters) infers the same tree as reported by Phillips et al. (2004), with the clade *S. kudriavzevii* + *S. bayanus* supported by $\approx$80% bootstrap proportion (and, conversely, 197/1000 $\approx$20% for the clade *S. kudriavzevii* + *S. mikatae* + *S. cerevisiae* + *S. paradoxus*; see Fig. 5 and Table 3). On the other hand, it has been shown that a distance-based analysis of a multigene dataset at the low-level combination leads to incorrect trees (particularly because of the evolutionary rate heterogeneity among genes [Criscuolo et al. 2006]), and thus, the medium-level combination method SDM* was initially designed to compensate this strong bias. However, our medium-level combination using the distance $UNED_{2P}$ leads to similar results compared to the previous low-level combination with the same distance (with the clade *S. kudriavzevii* + *S. bayanus* supported by $\approx$85% bootstrap proportion).

Our approaches using the codon distances $WCED_{6P}$ and W^2CED_{6P} infer the same tree as reported by Rokas et al. (2003) and Ren et al. (2005) in both low- and medium-level combinations. Furthermore, with the distance W^2CED_{6P}, the clade *S. kudriavzevii* + *S. mikatae* + *S. cerevisiae* + *S. paradoxus* is relatively well supported, with $\approx$64% and $\approx$81% bootstrap proportion with the low- and (more appropriately; see above) medium-level combinations, respectively (Table 3). Considering that the tree in Fig. 5 is the correct one, the classifications of the bootstrap-based performances of the three distances $UNED_{2P}$, $WCED_{6P}$, and W^2CED_{6P} are similar to those observed during previous explorations of the real and artificial datasets. Thus, unlike nucleotide distances, codon distances allow better capture of the phylogenetic signal induced by the multigene dataset of Rokas et al. (2003) with similar run times (about 10 min to perform the 1000 distance estimates and tree reconstruction with BioNJ, so <1 s per bootstrap replicate). Even if PhyML is used, which is not possible due to the internal limitation with too large datasets, no similar run times are expected with any accurate ML-based tree reconstruction method on this dataset.

Conclusion

We have proposed here two new codon evolutionary distances generalizing the distance CED_{6P} (Michel 2007): $WCED_{6P}$, which is based on the heterogeneous global evolutionary rates among the three codon positions; and W^2CED_{6P}, based on the treelikeness rate induced by each codon position. Our distances have been compared with the classically used nucleotide evolutionary distances NED_{2P}, $UNED_{2P}$ (unbiased), $GNED_{2P}$ (gamma), and $GUNED_{2P}$ (unbiased gamma; see Table 1) and studied according to different phylogenetic tree reconstruction methods (FastME, BioNJ, WLS-MVR, and PhyML). The codon distances allow fast and reliable phylogenetic trees to be inferred.

From a biological point of view, the distance $WCED_{6P}$ is based on a codon evolutionary model, called CEM_{6P}, assuming that each of the three codon positions evolves independently under the nucleotide evolutionary model NEM_{2P} (Kimura 1980). This model CEM_{6P} allows the derivation of a simple analytical formula to estimate the distance between two codon sequences. Studies with both real and artificial datasets (3211 codon-based alignments) show that this distance $WCED_{6P}$ (and its improved distance W^2CED_{6P}) allows the building of very accurate trees when used with fast distance-based tree reconstruction methods, in some cases even better than the trees obtained by ML approaches. Since these new distances CED are based on an extension of simple nucleotide evolutionary models (e.g., NEM_{1P} [Jukes and Cantor 1969], NEM_{2P} [Kimura 1980], NEM_{3P} [Kimura 1981]), new distances could be defined from more parameter-rich models (e.g., Tajima and Nei 1984; Hasegawa et al. 1985; Tamura 1992; Tamura and Nei 1993) in order to include more information about the DNA sequences and then to refine phylogenetic inference.

From a computational point of view, the approach proposed here allows the estimation of more precise evolutionary distances (in the treelikeness sense) than the distances classically used (nucleotide distances or gamma-based distances). More precisely, by decomposing the sequence alignment into three nucleotide partitions (i.e., the codon positions) and by considering weights inversely proportional to the global evolutionary rate induced by each of these three partitions, the codon evolutionary distance is an accurate alternative to the gamma-based one. Furthermore, this codon evolutionary distance can be generalized to a word of any length (i.e., not necessary a codon [see property 5 in Michel 2007]). Therefore, a general method considering several nucleotide partitions will estimate probably more treelike evolutionary distances and, consequently, build better trees. This alternative to gamma distances is currently under investigation.

The new codon evolutionary distances presented in this paper, as well as classical evolutionary distances and several distance-based tree reconstructions, are implemented in our research software DNAdistree. This software and both the real and the artificial datasets are available at http://lsiit-bioinfo.u-strasbg.fr:8080/DNADISTREE/.

Acknowledgments We thank Simonetta Gribaldo as well as the referees for their helpful comments and suggestions on the manuscript.

References

Arquès DG, Michel CJ (1993) Analytical expression of the purine/pyrimidine codon probability after and before random mutations. Bull Math Biol 55:1025–1038

Arquès DG, Michel CJ (1995) Analytical solutions of the dinucleotide probability after and before random mutations. J Theor Biol 175:533–544

Barthélemy JP, Guénoche A (1991) Trees and proximity relations. Series in discrete mathematics and optimization. Wiley-Interscience, Chichester

Bevan RB, Lang BF, Bryant D (2005) Calculating the evolutionary rates of different genes: a fast, accurate estimator with applications to maximum likelihood phylogenetic analysis. Syst Biol 54:900–915

Bryant D, Tsang J, Kearney P, Li M (2000) Computing the quartet distance between evolutionary trees. In: Proceedings of the 11th annual symposium on discrete algorithms (SODA), pp 285–286

Bruno WJ, Socci ND, Halpern AL (2000) Weighted neighbor joining: a likelihood approach to distance-based phylogeny reconstruction. Mol Biol Evol 17:189–197

Buneman P (1971) The recovery of trees from measures of dissimilarity. In: Hudson F, Kendall D, Tautu P (eds) Mathematics in archaeological and historical sciences. University Press, Edinburgh, pp 387–395

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. Am J Hum Genet 19:223–257

Criscuolo A, Gascuel O (2008) Fast NJ-like algorithms to deal with incomplete distance matrices. BMC Bioinformatics 9:166

Criscuolo A, Berry V, Douzery EJP, Gascuel O (2006) SDM: a fast distance-based approach for (super)tree building in phylogenomics. Syst Biol 55:740–755

Cummings MP, Otto SP, Wakeley J (1995) Sampling properties of DNA sequence data in phylogenetic analysis. Mol Biol Evol 12:814–822

Dayhoff MO (1979) Atlas of protein sequence and structure, vol 5. National Biomedical Research Foundation, Washington, DC Suppl 3

Desper R, Gascuel O (2002) Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. J Comput Biol 9:687–705

Desper R, Gascuel O (2004) Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. Mol Biol Evol 21:587–598

Dixon WJ, Mood AM (1946) The statistical sign test. J Am Statist Assoc 41:557–566

Estabrook GF, McMorris FR, Meacham CA (1985) Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. Syst Zool 34:193–200

Felsenstein J (2005) PHYLIP (Phylogeny Inference Package), version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle

Fitch WM, Margoliash E (1967) The construction of phylogenetic trees—a generally applicable method utilizing estimates of the mutation distance obtained from cytochrome c sequences. Science 155:279–284

Galtier N, Gouy M (1995) Inferring phylogenies from DNA sequences of unequal base compositions. Proc Natl Acad Sci USA 92:11317–11321

Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14:685–695

Gascuel O (2000) Data model and classification by trees: the minimum variance reduction (MVR) method. J Classif 17:67–99

Gascuel O, Steel M (2006) Neighbor-joining revealed. Mol Biol Evol 23:1997–2000

Goddard WE, Kubicka G, Kubicki G, McMorris FR (1994) The agreement metric for labelled binary trees. Math Biosci 123:215–226

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725–736

Guénoche A, Garreta H (2000) Can we have confidence in a tree representation? In: Proceedings of JOBIM00. Lecture notes in computer science. vol 2066, pp 45–56

Guénoche A, Grandcolas S (1999) Approximation par arbre d’une distance partielle. Math Inf Sci Hum 146:51–64 (in French)

Guénoche A, Grandcolas S (2000) Estimating missing values in tree distances. In: Kier HAL et al (eds) Data analyses, classification and related methods. Proceedings of the IFCS’ 2000. Springer, New York, pp 143–148

Guindon S, Gascuel O (2002) Efficient biased estimation of evolutionary distances when substitution rates vary across sites. Mol Biol Evol 19:534–543

Guindon S, Gascuel O (2003) A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood. Syst Biol 52:696–704

Hartmann S, Vision TJ (2008) Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment? BMC Evol Biol 8:95

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

Hemelrijk J (1952) A theorem on the sign test when ties are present. Proc Nederl Akad Weten Ser A 55:322

Holland BR, Huber KT, Dress A, Moulton V (2002) δ plots: a tool for analysing phylogenetic distance data. Mol Biol Evol 19:2051–2059

Hubert L, Arabie P, Meulman J (2006) The structural representation of proximity matrices with MATLAB. SIAM, Philadelphia

Jin L, Nei M (1990) Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol Biol Evol 7:82–102

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275–282

Jukes TH (1987) Transitions, transversions, and the molecular clock. J Mol Evol 26:87–98

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic Press, New York, pp 21–132

Kelchner SA, Thomas MA (2006) Model use in phylogenetics: nine key questions. Trends Ecol Evol 22:87–94

Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Kimura M (1981) Estimation of evolutionary distances between homologous nucleotide sequences. Proc Natl Acad Sci USA 78:454–458

Kimura M, Ohta T (1972) On the stochastic model for estimation of mutational distance between homologous proteins. J Mol Evol 2:87–90

Kuhner MK, Felsenstein J (1994) A simulation comparison of phylogeny algorithms under unequal evolutionary rates. Mol Biol Evol 11:459–468

Künsch HR (1989) The jackknife and the bootstrap for general stationary observations. Ann Stat 17:1217–1241

Lake JA (1994) Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc Natl Acad Sci USA 91:1455–1459

Lanave C, Preparata G, Saccone C, Serio G (1984) A new method for calculating evolutionary substitution rates. J Mol Evol 20:86–93

Liu RY, Singh K (1992) Moving blocks jackknife and bootstrap capture weak dependence. In: LePage R, Billiard L (eds) Exploring the limits of the bootstrap. Wiley, New York, pp 224–248

Lockhart PJ, Steel MA, Hendy MD, Penny D (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol 11:605–612

Michel CJ (2007) Codon phylogenetic distance. J Comput Biol Chem 31:36–43

MacStewart W (1941) A note on the power of the sign test. Ann Math Stat 12:236–239

Mindell DP, Thacker CE (1996) Rates of molecular evolution: phylogenetic issues and applications. Annu Rev Ecol Syst 27:279–303

Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11:715–724

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Phillips MJ, Delsuc F, Penny D (2004) Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol 21:1455–1458

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

Reed RD, Sperling FAH (1999) Interaction of process partitions in phylogenetic analysis: an example from swallowtail butterfly genus Papilio. Mol Biol Evol 16:286–297

Ren F, Tanaka H, Yang Z (2005) An empirical examination of the utility of codon-substitution models in phylogeny reconstruction. Syst Biol 54:808–818

Robinson D, Foulds L (1979) Comparison of weighted labelled trees. In: Lecture Note in Mathematics. Springer-Verlag, Berlin, pp 119–126

Rodriguez R, Oliver JL, Marin A, Medina JR (1990) The general stochastic model of nucleotide substitution. J Theor Biol 142:485–501

Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804

Rzhetsky A, Nei M (1992) A simple method for estimating and testing minimum-evolution trees. Mol Biol Evol 9:945–967

Rzhetsky A, Nei M (1993) Theoretical foundation of the minimum-evolution method of phylogenetic inference. Mol Biol Evol 10:1073–1095

Rzhetsky A, Nei M (1994) Unbiased estimates of the number of nucleotide substitutions when substitution rate varies among different sites. J Mol Evol 38:295–299

Saitou N, Nei M (1987) The neighbor-joining method: a new method to reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Sanjuán R, Wróbel B (2005) Weighted least-squares likelihood ratio test for branch testing in phylogenies reconstructed from distance measures. Syst Biol 54:218–229

Schmidt HA (2003) Phylogenetic trees from large datasets. PhD thesis. University of Dusseldorf

Steel MA (1994) Recovering a tree from the leaf colorations it generates under a Markov model. Appl Math Lett 7:19–23

Steel MA, Penny D (1993) Distribution of tree comparison metrics—some new results. Syst Biol 42:126–141

Studier JA, Keppler KJ (1988) A note on the neighbor-joining method of Saitou and Nei. Mol Biol Evol 4:729–731

Tajima F (1993) Unbiased estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol 10:677–688

Tajima F, Nei M (1984) Estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol 1:269–285

Takahashi K, Nei M (2000) Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. Mol Biol Evol 17:1251–1258

Tamura K (1992) Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C content biases. Mol Biol Evol 9:678–687

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526

Taylor DJ, Piel WH (2004) An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data. Mol Biol Evol 21:1534–1537

le Vinh S, von Haeseler A (2005) Shortest triplet clustering: reconstructing large phylogenies using representative sets. BMC Bioinformatics 8(6):92

Waterman M, Smith T (1978) On the similarity of dendograms. J Theor Biol 73:789–800

Whelan S, de Bakker PIW, Quevillon E, Rodriguez N, Goldman N (2006) PANDIT: an evolution-centric database of protein and associated nucleotide domains with inferred trees. Nucleic Acids Res 34:D327–D331

Williams WT, Clifford HT (1971) On the comparison of two classifications of the same set of elements. Taxon 20:519–522

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

Yang Z (1996) Maximum-likelihood models for combined analyses of multiple sequence data. J Mol Evol 42:587–596

Yang Z (1998) On the best evolutionary rate for phylogenetic analysis. Syst Biol 47:125–133

Yang Z (2007) PAML 4: a program package for phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591

Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15:496–503

Yang Z, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol 46:409–418

Yang Z, Nielsen R (2008) Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol 25:568–579

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

Yang Z, Nielsen R, Goldman N, Pedesen A-MK (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449

Zardoya R, Meyer A (1996) Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. Mol Biol Evol 13:933–942

Zaretskii K (1966) Postroenie dereva po naburo rasstoianii mezhdu visiacimi vershinami. Usp Mat Nauk 20:90–92 (in Russian)