SDM: A Fast Distance-Based Approach for (Super)Tree Building in Phylogenomics
Alexis Criscuolo, Vincent Berry, Emmanuel Douzery, Olivier Gascuel

To cite this version:

HAL Id: lirmm-00136655
https://hal-lirmm.ccsd.cnrs.fr/lirmm-00136655
Submitted on 5 Sep 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Phylogenomics, whereby phylogenies are built from large sets of homologous genes, is currently a popular trend that benefits from the increased quantity of sequenced genes within a huge variety of organisms (Daubin et al., 2002; Gatesy et al., 2002; Eisen and Fraser, 2003; Driskell et al., 2004; Philippe et al., 2004, 2005; Devulder et al., 2005). One of the main difficulties in phylogenomics is that fast methods are required to process the large collections of taxa and genes. Missing data are another difficulty with such data sets, as some genes or species are less represented in databases. Numerous approaches have been proposed to deal with this problem (Bininda-Emonds, 2004); they can be classified into three main categories (Schmidt, 2003; Chap. 7):

- The low-level (or total evidence) methods concatenate all genes to obtain a single alignment, also called supermatrix of characters, which is then analyzed using standard phylogeny reconstruction algorithms. As some genes are missing for some taxa, supermatrices usually contain numerous missing characters (e.g., >90% in Driskell et al., 2004). The various phylogenetic methods used to analyze such supermatrices are more or less vulnerable to missing characters, but the probabilistic ones seem to be not much affected and still provide accurate trees with sparse data (Philippe et al., 2004). Genes evolve under different constraints, and heterogeneity of rates and of evolutionary modes can also be problematic (Yang, 1996; Pupko et al., 2002). Again, probabilistic methods (e.g., MrBayes, Huelsenbeck and Ronquist, 2001) provide ways to circumvent this difficulty, by allowing for different substitution models to be defined among genes (or among codon positions). However, computing time is a main issue, specially with most sophisticated (e.g., Bayesian) approaches.

- The high-level methods arrange in a single tree topological information contained in the set of phylogenies inferred from each gene. Those source phylogenies are inferred independently, possibly using different evolutionary models, and may as well be derived from other (e.g., morphological or transposon-based) data types, which total evidence methods hardly account for. As some genes are missing for some taxa, the different phylogenies are defined on partially overlapping sets of taxa. This generalization of the consensus tree (Bryant, 2001) is called the supertree problem (Bininda-Emonds, 2004). Matrix representation with parsimony (MRP) (Baum, 1992; Ragan, 1992) is the most popular method to deal with this problem. MRP involves coding the topological information of every source tree in a single matrix of partial binary characters, which is then analyzed using parsimony to infer the supertree. This approach has been refined in various ways, such as weighted MRP (Ronquist, 1996) and matrix representation with flipping (Eulenstein et al., 2004). Numerous other combinatorial approaches have been proposed to deal with the supertree problem (Bininda-Emonds, 2004), including the MinCut (Semple and Steel, 2000) and modified MinCut (Page, 2002) algorithms.

- The medium-level methods involve an intermediary gene analysis stage, between simple gene concatenation and complete tree inference. Numerous
solutions do exist to extract information from every single gene, without inferring the complete tree as in high-level methods. The main idea is to extract in a fast way elementary pieces of information from each gene independently, then to combine all these elements for all genes together. The hard combination task is thus performed just once with all genes being accounted for. As the first analysis stage is performed independently for each gene (or information source), these methods offer simple ways to accommodate for genes evolving under different evolutionary constraints or to combine heterogeneous data types. A good example is the quartet approach (Strimmer and von Haeseler, 1996; Schmidt et al., 2002; Piaggio-Talice et al., 2004) whereby every quartet topology is inferred using maximum likelihood from each gene before combining them in a single tree. We shall see in this paper a second example where first analysis stage involves computing for each gene the evolutionary distance between every taxon pair.

The outline of such large categories is blurred and some methods can be seen as intermediary. For example, the (medium-level) quartet approach has been proposed by several authors (Strimmer and von Haeseler, 1996; Schmidt et al., 2002) to deal with the (high-level) supertree problem, and the divide-and-conquer searching methods (e.g., Huson et al., 1999) use a (high-level) tree combination approach to solve the low-level problem. Moreover, the criterion that the method seeks to optimize gives another important point of view. Most practical methods are based on maximum parsimony (MP) and maximum likelihood (ML). However, one aim of phylogenomics is to build large phylogenies from large gene collections. Therefore, it is essential to be able to process huge data sets by low time-consuming methods.

The distance-based approach is the first choice from this standpoint. Using fast algorithms such as NJ (Saitou and Nei, 1987; Studier and Kephner, 1988), BToN (Gascuel, 1997), or FASTME (Desper and Gascuel, 2002), trees with thousands of taxa can be inferred in a few minutes on a standard computer. Moreover, these algorithms are fairly accurate, though not as accurate as likelihood-based approaches. This computational efficiency is why distance-based methods are frequently employed in exploratory studies. They are also used to provide starting trees for procedures aimed at optimizing more time-consuming criteria. The PHYML program (Guindon and Gascuel, 2003) is a good example of this approach with respect to the ML criterion.

Paradoxically, few distance-based approaches have been proposed in phylogenomics. One simple method is to directly estimate pairwise evolutionary distances from the concatenated matrix of characters. For example, PAUP*’s (Swofford, 2002) option MISS=DIST=IGNORE only takes sites that have no missing value in the two sequences into account. This procedure is named distance-based total evidence (DTE) in the following and is obviously limited by large amounts of missing data and severe rate heterogeneity. A second method, the average consensus supertree (ACS) procedure, was proposed by Lapointe and Cucumel (1997) to deal with the supertree problem, where there can be large amounts of missing data. The first step is to compute the path-length distance matrices corresponding to the source trees. Each source matrix is then standardized, and ACS computes the average of the standardized matrices to produce the distance supermatrix that is analyzed using a least-squares method. ACS has been shown to be the same as MRP in the consensus setting with unitary branch lengths, but both are different in the more general supertree context (Lapointe et al., 2003). A similar averaging method was used by Lapointe et al. (1999) and Levasseur and Lapointe (2001) to compare and combine various distance matrices being obtained directly (the medium-level way) from sequences or from DNA hybridization, or corresponding to (high-level) gene trees. The standardization step proposed by Lapointe and Cucumel (1997) involves dividing all distances in each matrix by the maximum distance in that matrix. Other standardization methods have been explored, but they seem to be inaccurate with more than two trees and Lapointe and Levasseur (2004) concluded that “other ways of scaling path-length distance matrices need to be investigated when combining more than two trees of varying size” (p. 100). Recently, Creevey and McInerney (2005) proposed another distance-based method to the supertree problem, named most similar supertree (MSS). The unitary (every branch has length 1) path-length distance matrices corresponding to the source trees are first computed; then, MSS searches for the supertree that best represents these matrices using topological rearrangements.

Here, we propose a novel distance method, which follows the same line as ACS but is based on a much more involved standardization procedure that answers the limitations outlined by Lapointe and Levasseur (2004). This method, named super distance matrix (SDM), first deforms the source matrices without modifying their topological message, so as to bring them as close as possible to each other; then, just as with ACS, so-deformed matrices are averaged and analyzed by usual tree-building algorithms. Simulations show that SDM deals efficiently and accurately with collections containing a large number of source matrices of varying size. SDM was initially designed as a medium-level method (source distance matrices are directly computed from the sequences of each gene), but it is an effective alternative within high-level scenarios (source matrices are obtained from the gene trees, just as with ACS) and within low-level scenarios, where good starting trees are obtained thanks to its speed.

In the following, we first describe the principle and the main features of the SDM algorithm; we then provide comparisons of SDM to other gene combination techniques using simulations; we further compare SDM to other approaches using a phylogenomics data set of placental mammals (Gatesy et al., 2002). Mathematical proofs and equations are provided in the Appendix.
THE SDM METHOD

Notations and Definitions

Let \( \mathcal{C} = \{ (\Delta_{ij}^1), (\Delta_{ij}^2), \ldots, (\Delta_{ij}^k) \} \) be a collection of \( k \) distance matrices (with no missing entries), where \( \Delta_{ij}^p \) is the evolutionary distance between taxa \( i \) and \( j \) for the gene \( p \). \( \mathcal{L}_p \) is the set of taxa covered by the gene \( p \) and defining the entries of \( (\Delta_{ij}^p) \); \( n_p \) is the size of \( \mathcal{L}_p \); \( n \) is the size of \( \mathcal{C} = \bigcup_p \mathcal{L}_p \); and \( k_{ij} \) is the number of occurrences of the taxon pair \( ij \) in the collection \( \mathcal{C} \); i.e., \( k_{ij} = |\{ p : [i,j] \subset \mathcal{L}_p \}| \). We set:

\[
\hat{\mathcal{L}}_p = \{ i \in \mathcal{L}_p : \exists j \in \mathcal{L}_p - \{ i \} , k_{ij} \geq 2 \},
\]

\[
\hat{n}_p = |\hat{\mathcal{L}}_p|,
\]

\[
\hat{\mathcal{L}} = \bigcup_p \hat{\mathcal{L}}_p \quad \text{and}
\]

\[
\hat{n} = |\hat{\mathcal{L}}|.
\]

Our method involves bringing each matrix \( (\Delta_{ij}^p) \) closer (in the least-squares sense) relative to the others. \( \hat{\mathcal{L}} \) is the set of taxa to be used to compare the distances between pairs of taxa appearing in more than one source matrix.

Method

Assume a high-level context and let \( T^p \) be the tree corresponding to gene \( p \). \( (\Delta_{ij}^p) \) is then additive and is obtained from \( T^p \) by computing the path-length for every \( ij \) pair. \( (\Delta_{ij}^p) \) is equivalent to \( T^p \) as \( T^p \) can be unambiguously recovered from \( (\Delta_{ij}^p) \). It is well known (Barthélemy and Guénoche, 1991) that multiplication by a factor \( \alpha_p > 0 \) to obtain a new distance matrix \( (\alpha_p \Delta_{ij}^p) \) does not change the topology of \( T^p \). This operation is equivalent to multiplying every branch length of \( T^p \) by \( \alpha_p \). Similarly, it is easily shown that the addition of a constant \( a_{ip} \geq 0 \) to each of the \( 2(n_p - 1) \) nondiagonal distances corresponding to taxon \( i \) does not change the topology of \( T^p \). This operation is equivalent to elongating, by length \( a_{ip} \), the external branch corresponding to taxon \( i \). This addition can be performed independently for every taxon, and both multiplication and addition operations can be combined to obtain the new matrix \( (\alpha_p \Delta_{ij}^p + a_{ip} + a_{jq}) \) that contains the same topological information as \( (\Delta_{ij}^p) \).

In the medium-level context, distance matrices are estimated from sequences (or other data) and are not additive (i.e., do not exactly correspond to a tree). But the above property still holds in some sense. Indeed, it is easily shown (Gascuel, 1994) that NJ and a number of related algorithms infer the same topology from the original distance matrix \( (\Delta_{ij}^p) \) and from the deformed one \( (\alpha_p \Delta_{ij}^p + a_{ip} + a_{jq}) \). This is still true when using any of the algorithms implemented in the FASTME package (Desper and Gascuel, 2002). Moreover, simulation experiments show that least-squares algorithms, e.g., FITCH from the PHYLIP package (Felsenstein, 1993) and MW (Makarenkov and Leclerc, 1999) from the TREX program (Makarenkov, 2001), are almost insensitive to multiplication and addition operations.

The different ACS standardization methods (Lapointe and Levasseur, 2004) all correspond to the use of the multiplication operation to rescale matrices before averaging them. SDM uses both multiplication and addition, which greatly increases flexibility as multiplication involves one free parameter per source matrix, whereas addition involves one parameter per taxon for each source matrix. The basic principle of SDM is to deform each of the \( k \) distance matrices \( (\Delta_{ij}^p) \) by multiplying them by a positive factor \( \alpha_p \) and adding constants \( a_{ip} \) in order to bring them as close as possible to each other in the least-squares sense. All distances that are shared by at least two matrices of \( \mathcal{C} \) (i.e., such that \( k_{ij} \geq 2 \)) are taken into account in the computation of deformation parameters. Moreover, weights \( (w_p) \) are associated to each of the source matrices to give them a confidence value (see below for more). Thus, for every pair \( ij \) such that \( k_{ij} \geq 2 \), we aim at minimizing the variance term:

\[
V_{ij} = \sum_{p : [i,j] \subset \mathcal{L}_p} w_p (\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp} - \bar{\Delta}_{ij})^2
\]

where \( \bar{\Delta}_{ij} \) is the weighted average of the deformed distances:

\[
\bar{\Delta}_{ij} = \frac{1}{W_{ij}} \sum_{p : [i,j] \subset \mathcal{L}_p} w_p (\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp})
\]

with

\[
W_{ij} = \sum_{p : [i,j] \subset \mathcal{L}_p} w_p.
\]

The \( (\bar{\Delta}_{ij}) \) matrix is the SDM output, once optimal values of the \( \alpha_p \) and \( a_{ip} \) parameters have been computed. By minimizing criterion (1), we bring closer the \( \Delta_{ij}^p \) distances, and we obtain a reliable estimation of their average that defines the SDM superdistance.

As said above, \( w_p \) weights allow to give a confidence value to each of the source matrices. Typically, the variance of any distance estimate is inversely proportional to the length of the sequences used for estimation (Nei and Jin, 1989). Thus, it is coherent to set \( w_p \) to be equal to the sequence length, which is denoted as \( \ell_{ip} \). On the other hand, matrices involving few taxa might have a poor influence in comparison to matrices with many taxa. To compensate for this effect, we can use \( \ell_{ip}/(\bar{n}_p - 1) \), or the intermediate value \( \ell_{ip}/\bar{n}_p \), as the matrix weight. A number of other weightings are possible, and SDM is easily extended to the case where each distance is associated to a confidence value, just as in weighted least-squares tree building methods (Fitch and Margoliash, 1967).

With the minimization of \( V_{ij} \) being applied for every relevant pair of taxa \( ij \), SDM thus involves minimizing
the sum:

$$\sum_{i,j \in L} V_{ij}.$$  (3)

Several linear constraints on the variables are associated with minimization of criterion (3). The $\alpha_p$ factors deal with the various evolutionary rates of each gene in a similar way as the gene-specific rate models first suggested by Yang (1996). Constraint (4), identical to that of the proportional model of Pupko et al. (2002), forces the $\alpha_p$ factors to be equal on average to 1.0:

$$\sum_p \alpha_p = k.$$  (4)

This constraint gives interpretable scaling and is required to avoid the trivial solution $\alpha_p = 0$, $\forall p$.

External branches of a phylogeny are generally longer than the internal branches. Consequently, most of the variance of each pairwise distance is generally supported by the two external branches. Moreover, Lapointe and Levasseur (2004) noticed that high heterogeneity in the branch lengths of source trees deteriorates ACS topological accuracy. The $a_{ip}$ variables thus try to normalize the external branch lengths in the various matrices. Constraint (2) forces, for each taxon $i$, the sum of $a_{ip}$ to be equal to 0 and forbids overelongation (or shortening) of external branch lengths corresponding to taxon $i$:

$$\sum_{p \in \tilde{L}_p} a_{ip} = 0, \forall i \in \tilde{L}.$$  (5)

Constraint (6) forces, for each matrix ($\Delta_i^f$), the sum of $a_{ip}$ to be equal to zero:

$$\sum_{i \in L_p} a_{ip} = 0, \forall p = 1, 2, \ldots, k - 1.$$  (6)

This avoids having some of the matrices deformed into star-like distances by global elongation of the external branches and small $\alpha_p$ values. The topological signal of the original matrix would then be stifled, as was experimentally observed (in the absence of constraint (6)) with matrices having few taxa and low (or contradictory) signal. Note that constraint $\sum_{i \in \tilde{L}_p} a_{ik} = 0$ is useless as it is induced by the other constraints (5) and (6) on $a_{ip}$ values; adding this constraint to the system induces linear dependency and perturbs the resolution.

Minimization of criterion (3) involves calculating its partial derivatives for each of the $k + \tilde{n}$ variables $\alpha_p$ and $a_{ip}$, adding Lagrange multipliers (Luenberger, 1984: Chap. 10) that are associated with each of the $\tilde{n} + k$ linear constraints. We thus obtain a linear system defined by $O(\tilde{n}k)$ variables and equations which has a unique solution (see Appendix). Resolving this system has $O(\tilde{n}^2k^3)$ time complexity, which is theoretically equivalent to the running time of the NJ algorithm with a distance matrix of size $\tilde{n}k$. However, as this linear system is very sparse, the practical time required to solve it is much lower (see below) using an appropriate library (MTJ, available at https://mtj.dev.java.net/, is used in our implementation).

Let $a_{ip}^*$ and $a_{ip}^*$ be the optimal values of the parameters we obtain this way, then the SDM distance supermatrix ($\Delta_{ij}^{SDM}$) is defined by:

$$\Delta_{ij}^{SDM} = \frac{1}{W_{ij}} \sum_{p \ni [ij] \subseteq \mathcal{L}_p} w_p (a_{ip}^* \Delta_i^p + a_{ip}^* + a_{jp}^*)$$

where $W_{ij} = \sum_{p \ni [ij] \subseteq \mathcal{L}_p} w_p$.

Note that this formula applies to all available distances (i.e., $k_{ij} \geq 1$), not only to those used to compute the optimal parameter values (i.e. $k_{ij} \geq 2$). This last step of the SDM approach is fast and requires $O(n^2k)$ running time.

To check the SDM practical running time, we generated 100 collections of distance matrices with $k = 4, 8, 12, 16$ and $n$ from 50 to 500 and then measured the running time $t$ of SDM. Assuming $t = b(nk)^2$, we performed a linear regression on $\log(t)$ as a function of $\log(nk)$ and found that the estimated value of $b$ is below 2.0. Thus, in practice, SDM requires computing time that is at most quadratic in $nk$. For example, it only takes a few minutes to deal with a collection of distance matrices with $n = 500$ and $k = 20$, using a 1.8 GHz Pentium IV PC with 1 Gb RAM.

**Phylogenetic Reconstruction Using SDM**

A distance-based algorithm is applied to the SDM distance supermatrix to obtain a phylogeny. However, just as with ACS, missing entries may occur in this distance supermatrix depending on the extent of taxon overlap within the source matrices. It has been shown (Farrach et al., 1995) that tree reconstruction from distance matrices with missing entries is computationally hard, and heuristics approaches have to be used. Two types of method have been proposed:

- The indirect method involves first estimating missing distances by applying an ultrametric (De Soete, 1984), additive (Landry et al., 1996), decomposition-based (Lapointe and Landry, 2001), or quartet-based (Guénoche and Grandcolas, 1999) completion algorithm. The TREV package (Makarenkov, 2001) provides several implementations of such algorithms to be used before tree building using any standard method with the completed matrix.

- The direct method involves using a weighted least-squares algorithm and associating missing distances with null weight, which means that missing distances are simply discarded from weighted least-squares computations (Swofford et al., 1996:
A combination of both direct and indirect methods is provided by MW* (Makarenkov and Lapointe, 2004) (also available in TReX); this algorithm first applies an ultrametric or additive completion algorithm (depending on the density of missing distances) and then infers a tree using the weighted least-squares algorithm MW (Makarenkov and Leclerc, 1999), where weights are set at 1.0 for known distances, 0.5 for estimated distances, and 0.0 for missing distances (if any remain).

However, missing distances are relatively rare, though the amount of missing characters is usually high in the gene collections that are commonly used in phylogenomics studies. For example, data sets of Gatesy et al. (2002) and Philippe et al. (2004, 2005) have high ratio of missing character states (about 68%, 25%, 35%, respectively) but do not produce any missing distances when using SDM. Indeed, in these data sets some genes (e.g., cytochrome b and ribosomic protein L10) have been sequenced for all taxa; at least one gene distance matrix is then complete, which induces that the SDM supermatrix is also complete. Moreover, it is a simple consequence of randomness that the number of missing distances tends to decrease when the number of genes increases. For example, with the two very large data sets of Driskell et al. (2004), which were collected from Swiss-Prot and GenBank thanks to a computer program (previous collections were collected manually), the ratio of missing distances is ≈19% and ≈1%, whereas the ratio of missing characters is ≈92% and ≈87%, respectively. In the same way, in our simulations study (see below), missing distances are very rarely observed when the number of genes is above 10 and when the ratio of missing characters (equal in expectation to the taxon deletion rate) is of 25%. When the SDM distance supermatrix is complete, fast algorithms (e.g., NJ, BioNJ, or FastME) can be used to infer the corresponding tree.

**Simulation Protocol**

We conducted large-scale simulations to evaluate the topological accuracy of SDM. Our aim was to compare the ability to recover the correct topology and the running times of low-, high-, and medium-level approaches. In the three cases, we present standard methods that are compared to SDM-based scenarios. Moreover, we emphasize distance-based methods as SDM belongs to this category. We first describe the way trees and sequences were generated, then the various methods we tested, and finally the criteria we used in the comparisons.

**Tree Generation**

The procedure was similar to the one used in Guindon and Gascuel (2003), which can be referred to for further details. Random 48-taxon trees were generated using the standard Yule-Harding process, via the r8s program (Sanderson, 2003). This process makes the trees clocklike, so we created a deviation from this model by multiplying every branch length by (1 + X), where X followed an exponential distribution with expectation \( \mu \). The \( \mu \) value represents the extent of deviation and was identical within each tree but different from tree to tree and equal to 0.2/(0.001 + \( U \)), with \( U \) being uniformly drawn from [0, 1]. The smaller the \( U \), the larger the \( \mu \) and the larger the deviation from the molecular clock. Let \( tbl \) be the total branch length of the generated tree. We obtained the nonclocklike tree \( T \) with total length 1.0 by dividing every branch length by \( tbl \). \( T \) was the “correct” tree that the various methods aimed at recovering.

To simulate the evolution of the different genes, we generated \( k \) trees \( T^p \) from \( T \) by multiplying every branch length of \( T \) by 0.4 + 8.6 \( V_p \), where \( V_p \) was uniformly drawn from [0, 1]. The \( V_p \) value was the same within each tree \( T^p \), but different from tree to tree. These \( k \) source trees \( T^p \) thus have the same topology as the tree \( T \). However, they have their own evolutionary rates with relative values ranging from 1.0 to 22.5 (= (0.4 + 8.6)/0.4) in extreme cases; such values are in agreement with real values (see Guindon and Gascuel, 2003, for more and, below, our analysis of Gatesy et al. data, 2002).

**Sequence Generation**

We considered gene collections of size \( k = 2, 4, 6, \ldots, 20 \) and generated 500 data sets per \( k \) value. For each of these data sets, we first generated a correct tree \( T \) and then a collection of \( k \) gene trees \( T^p \), as explained above. For each gene \( p \), we uniformly drew sequence length \( \ell_p \) between 200 and 1000 bp, and then used the SEQ-GEN program (Rambaut and Grassly, 1997) to simulate the sequence evolution along \( T^p \) according to the K2P substitution model (Kimura, 1980). We used a transition/transversion ratio of 2.0 and did not rescale the \( T^p \) trees. To simulate partial overlap that occurs in real data sets, we randomly removed some of the taxa within each gene alignment obtained. Following Eulenstein et al. (2004), two overlap conditions were studied, corresponding to 25% taxon deletion per gene (strong overlap) and 75% deletion (low overlap). However, an overlap of at least four taxa was preserved between each gene pair to maintain a common evolutionary history between genes and avoid meaningless data sets. Note that the expected ratio of missing characters was also equal to 25% and 75%, respectively, due the random processes we used for sequence length generation and taxon deletion.

**Inference Methods**

The (10 gene collection size \times 500 collections \times 2 overlap conditions \approx) 10,000 generated datasets were used to compare a number of tree building approaches. Our aim was (1) to check the properties of SDM when used in various scenarios of low-, medium-, and high level;
(2) to compare these SDM-based scenarios to classical approaches (e.g., MRP), and (3) to compare SDM with other distance-based methods (e.g., ACS). All tested methods and scenarios are described below, grouped according to their combination level. Figure 1 displays a flowchart indicating the way the various scenarios combine several methods to achieve tree construction from gene collections (e.g., PHYML+MRP scenario involves first inferring gene trees using PHYML, then combining these trees using MRP).

Medium-level, Distance-Based Approaches.—For each dataset, we used PAUP* with K2P to estimate k distance matrices from the k sequence alignments. The SDM distance supermatrix was computed from this collection of matrices, with each matrix weighted by the length of the corresponding sequences (i.e. \( w_p = l_p \)) in formula (1).

We then used the FitCh program (with all default options, notably without global rearrangements) to build a phylogeny from the SDM distance supermatrix that (possibly) contains missing entries. This tree-building scenario is called SDM+FitCh (Fig. 1). In order to test for the advantage of using \( a_{ip} \) variables, instead of solely using \( \alpha_p \) variables that deal with gene rate heterogeneity, we ran another similar scenario, where the \( a_{ip} \) variables were forced to be zero. This second scenario is called SDM\( ^* \)+FitCh; it is close to ACS as it only uses multiplication operation to rescale matrices (see also Bevan et al., 2005). We tested other approaches to deal with missing entries, as listed in the Introduction, but they all performed poorer than the FitCh weighted least-squares program (see further results and discussions regarding MW\( ^* \) by Makarenkov and Lapointe, 2004).
Low-Level Combination.—For each dataset, a supermatrix of characters was obtained by concatenating the $k$ partially deleted genes. We computed the K2P distance matrix from this supermatrix of characters using the PAUP*’s MISSDIST = IGNORE option (see the Introduction) and then used BIONJ to infer the DTE (distance-based total evidence) phylogeny.

To obtain an ML-based total evidence phylogeny, we analyzed the supermatrix of characters using PhyML with K2P. This program searches for the optimal tree according to the ML criterion, via topological rearrangements from a starting tree. As these topological rearrangements are local and solely based on nearest neighbor interchange (NNI) (Swofford et al., 1996), the resulting tree depends, to some extent, on the starting tree. We call DTE+PhyML the scenario whereby the PhyML default option is used, which involves using DTE (see above and Fig. 1) to compute the starting tree. As we suspected that DTE would generate poor trees in this phylogenomics context, we also used SDM+FITCH (see above and Fig. 1) to infer the starting tree to be then refined by PhyML; we call this scenario SDM+FITCH+PhyML (Fig. 1). Our aim was to check that using the improved SDM starting tree, we improve the resulting PhyML tree and reduce the number of NNIs and the running time.

High-Level Combination.—A collection of $k$ ML phylogenies was built from the $k$ partially deleted genes using PhyML with K2P. We then combined these trees using the standard MRP technique, which involves first coding the tree topologies in a partial binary matrix, then inferring the supertree by maximum parsimony. To achieve this task, we used TNT (Goloboff et al., 2003), which is well known for its efficiency, and followed the standard approach (Bininda-Emonds and Bryant, 1998) that defines the MRP supertree as the strict consensus of the most parsimonious trees. TNT was run with 25 random addition sequences, TBR branch swapping and ratchet default option. We call this supertree construction scenario PhyML+MRP (Fig. 1).

We also tested three distance-based supertree approaches, using the same PhyML source trees as with MRP. First, we evaluated ACS regarding its pioneer role in the field. Gene trees were transformed into path-length matrices, and we used the standardization procedure of Lapointe and Cucumel (1997), which applies to any number of source matrices, unlike the other standardizations presented by Lapointe and Lévesque (2004). This version of ACS was combined with the recent MW* algorithm (Makarenkov and Lapointe, 2004), which invokes both indirect and direct algorithms to deal with missing distances (see above). This scenario is called PhyML+ACS97+MW* (Fig. 1). We selected MW* to be combined with ACS as it was designed by the same author group, but we also performed experiments substituting MW* by FITCH. We applied SDM to the same path-length matrices as ACS and combined it with FITCH; this scenario is called PhyML+SDM+FITCH (Fig. 1). Finally, we evaluated MSS (Creevey and McInerney, 2005) using default parameters and recommended options: NJ was applied to the MRP matrix using $p$-distances to obtain a starting tree; unitary path-length matrices corresponding to each gene tree were then computed and fed into MSS, which was run with SPR rearrangements. This scenario is called PhyML+MSS (Fig. 1).

Topological Accuracy Measure

We measured the topological accuracy of every scenario using the quartet distance $d_4$ (Estabrook et al., 1985) between the inferred tree $\hat{T}$ and the model tree $T$. $d_4$ counts the number of resolved 4-trees (i.e., four-taxon trees) present in one tree but not in the other. $d_4$ is then the sum of two error types: the type I error corresponding to resolved 4-trees induced by $T$ but not present in $\hat{T}$, and the type II error corresponding to resolved 4-trees in $\hat{T}$ but not induced by $T$. As any fully resolved tree with $n$ taxa induces $\binom{n}{4}$ 4-trees, this measure can take any integer value between 0 and $2\binom{n}{4}$. $d_4$ is then more precise than the widely used bipartition distance (Bourque, 1978; Robinson and Foulds, 1979), which counts the number of internal branches present in one tree but not in the other, and then takes integer values between 0 and $2(n-3)$. Moreover, $d_4$ is less sensitive to slight topological differences; e.g., when just one taxon is misplaced and far away from its correct location, the bipartition distance is high as a number of bipartitions are incorrect, whereas the $d_4$ distance remains moderate as only quartets involving this taxon are modified. Thus, $d_4$ is better suited than bipartition distance to compare remote trees Steel and Penny, 1993, as obtained with 75% deletion rate where tree inference is hard (see below). $d_4$ was normalized by dividing its value by $2\binom{n}{4}$; 0 then corresponds to identical trees, whereas a distance of 1 means that both trees do not share any 4-tree. To avoid giving a topological meaning to very short branches in the inferred trees, every branch length less than 0.0001 was collapsed to make a multifurcation.

Simulation Results

Topological Accuracy

For each of the 20 conditions (10 gene collection sizes × 2 overlap conditions), the collected 500 $d_4$ values were averaged and are graphically represented in Figure 2. These graphs show the average $d_4$ value as a function of the number ($k$) of genes.

As expected, all curves in Figure 2 are decreasing: the correct tree $T$ is better recovered (i.e., the $d_4$ distance between $\hat{T}$ and $T$ decreases) as the number of genes increases. As also expected, the inferred phylogeny is closer to the correct tree as the taxon deletion rate decreases. The more information we have, the easier tree building is, whatever the reconstruction scenario. However, some of the scenarios are clearly more accurate than others.

Among pure distance-based scenarios, SDM+FITCH is best. It outperforms SDM*+FITCH in all conditions, indicating that incorporating $d_4$ variables in criterion (1) gives a significant improvement over using only the $\alpha_p$ multiplication factors. As expected, DTE performance is
very poor and its results are out of the scales used in Figure 1. This is due to weak distance estimation caused by rate heterogeneity among genes and missing sequences. Indeed, when two taxa share slow genes, they are estimated to be close, whereas when their common genes are evolving fast, they are predicted to be distant. Applying BioNJ (or any other algorithm) to such a poor distance matrix inevitably results in a poor tree.

High-level scenarios combine the source trees inferred by PhyML into a supertree. Among the three distance approaches, PhyML+SDM+Fitch is best in all conditions, and PhyML+ACS97+MW* tends to outperform PhyML+MSS when the gene number \( k \) is relatively high. We also tested other combinations, substituting Fitch and MW* to obtain the PhyML+ACS97+Fitch and PhyML+SDM+MW* scenarios. Neither one nor the other is better than PhyML+SDM+Fitch, and PhyML+ACS97+Fitch outperforms PhyML+ACS97+MW*.

This seems to indicate that Fitch (direct method) could be better suited than MW* (combining direct and indirect algorithms) to deal with distance matrices obtained in phylogenomics studies and containing missing entries. This somewhat contradicts findings presented by Makarenkov and Lapointe (2004), but could be explained by differences in the simulation protocols. These authors used a single distance (super)matrix with random deletion of pre-fixed numbers of entries, whereas our protocol is based on the assembly of several gene distance matrices and closer to phylogenomics data. Thus, our supermatrices are likely to be more perturbed than those in Makarenkov and Lapointe (2004), which could penalize indirect methods that only use a few distances to fill each of the missing entries. Note, moreover, that Fitch slightly outperformed MW* in one of the two experiments presented by Makarenkov and Lapointe (2004).
Comparing now SDM and MRP, we see that PhyML+MRP and PhyML+SDM+Fitch show similar accuracy with 25% taxon deletion, whereas the SDM-based scenario outperforms MRP with 75% deletion. In fact, it can be seen that PhyML+SDM+Fitch deals better with missing information (e.g., $k = 2$ with 25% and 75% deletion), whereas PhyML+MRP performs well when information is abundant (e.g., $k = 20$, where the two methods are close in both deletion conditions). This could be explained by the often poor resolution of MRP supertrees, which is due to the use of strict consensus and is higher with low source tree overlap (e.g., for $k = 10$ and 75% deletion rate, MRP supertrees contain 17% unresolved quartets on average). However, in the bootstrap analysis context, we showed that collapsing poorly supported branches improves topological accuracy (Berry and Gascuel, 1996) by decreasing type I error without significantly augmenting type II. A better explanation (Lapointe and Cucumel, 1997) could be that SDM not only uses the topology of the source trees (as MRP) but also their branch lengths. SDM-based trees then have more information than MRP trees. This could also explain the poor results of MSS, which loses information by setting all branches to length 1. Weighted MRP (where branches of the source trees are weighted by their bootstrap support; Ronquist, 1996) performs better than standard MRP (Bininda-Emonds and Sanderson, 2001), but at the expense of huge computing times, as with this approach the initial tree building algorithm (here, PhyML) has to run a number of times (at least 100) on each of the source data sets. However, fast branch support estimates could be used to accelerate these computations (Kishino et al., 1990; Waddell et al., 2002; Anisimova and Gascuel, 2006).

Low-level scenarios analyze the supermatrix of characters using PhyML, which improves, via NNIs, a starting tree built by a fast distance method. SDM+Fitch+PhyML clearly outperforms DTE+PhyML, which is a poor method with 75% taxon deletion. This is due to the extreme weakness of DTE with phylogenomics data, but not true with single gene study or when there are no missing characters. NNIs considerably improve DTE trees (see 25% deletion, where DTE+PhyML shows similar accuracy as MRP), but NNIs are not powerful enough to obtain a satisfactory tree when starting from quasirandom trees, as is the case with 75% deletion.

We then see the advantage of using SDM within the three combination levels. The medium-level SDM+Fitch scenario is even better than standard MRP with 75% deletion, whereas low-level SDM+Fitch+PhyML is clearly the best in all conditions, among the methods we tested. Moreover, this latter scenario could likely be improved by incorporating the specific rate of every gene in likelihood computations, as proposed by Yang (1996), Pupko et al. (2002), and Bevan et al. (2005). Finally, in the high-level context (which greatly simplifies the processing of various data types and evolutionary modes), SDM offers a relevant alternative to MRP as it is nearly equivalent with low (25%) taxon deletion, but significantly better with high (75%) deletion rate. Comparing the three SDM-based scenarios, we see a clear ordering: the low-level approach is best in all conditions, the medium-level method is worst, and the high-level combination is in between. As we shall see (and not surprisingly), this ordering is inverse of that induced by the computing times, the low-level scenario being much heavier than the two other methods. Moreover, the gap between high-level and low-level scenarios is moderate, and their ordering could be inverted with complex data sets showing strong heterogeneity in the evolutionary modes.

**Running Time**

The average running times of the main scenarios used in the simulations are displayed in Table 1, with $k = 10, 20$, and 25% and 75% taxon deletion. We also generated additional data sets with $n = 96$ taxa (10 per condition), using the previously described procedure and the same $k$ values and taxon deletion rates, and reported the running times in Table 1. Note that all these times strongly depend on implementation. For example, the weighted least-squares procedure in PAUP* is clearly faster than Fitch, whereas both follow a closely related scheme. Thus, results in Table 1 illustrate the main tendencies but should not be overinterpreted.

We first see that SDM on its own is a fast algorithm. For example, it only requires 48 s with 96 taxa, $k = 20$, and 25% taxon deletion. It follows that the running times of the SDM-based scenarios mostly depend on the other components of the scenarios, which tend to be (much) slower than SDM itself. The medium-level SDM+Fitch scenario is one of the fastest methods. For example, with 96 taxa, $k = 20$, and 25% taxon deletion, SDM+Fitch requires 539 s. Thanks to the speed of PhyML and TNT, PhyML+MRP is also quite efficient, being slower than SDM+Fitch with 48 taxa, but generally faster with 96. However, most of the computing time required by SDM+Fitch is spent by Fitch (e.g., 491 s as compared to 539 s, in our previous example). Fitch is useful as it copes with missing entries, but, as explained earlier, real data sets often yield full supermatrices of distance. In such cases, much faster inference algorithms do exist. In our simulations, all distance supermatrices are full when $k = 20$, $n = 48, 96$, and for both taxon deletion rates. With this ($k = 20$) data sets, we then used FASTME instead of Fitch. Topological accuracy remains similar (e.g., with 48 taxa and 25% deletion, average $d_2$ topological distances are 0.0242 for SDM+Fitch and 0.0268 for SDM+FASTME) but the tree inference time is less than 1 s in all settings; e.g., with 96 taxa, $k = 20$, and 25% taxon deletion, the SDM+FASTME scenario requires approximately 50 s, as compared to 539 s when using Fitch. In this biologically common case, SDM+FASTME is the fastest inference scenario, by a factor of 10- to 100-fold with 96 taxa, and this factor increases with the number of taxa. With 500 taxa and $k = 20$, SDM+FASTME requires a few minutes, while other scenarios require hours (or days) of computation.
Table 1. Running times. The values correspond to the average running time in seconds using a 1.8-GHz Pentium IV PC with 1.8 Gb RAM. $k$: number of genes used in the reconstruction. 25%, 75%: taxon deletion rates. Sums in parentheses provide the running times required by the different components of the scenarios.

<table>
<thead>
<tr>
<th></th>
<th>SDM$^\ast$+Fitch</th>
<th>SDM+Fitch</th>
<th>SDM+FASTME</th>
<th>DTE+PhyML</th>
<th>SDM+Fitch+PhyML</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k = 10$</td>
<td>$k = 20$</td>
<td>$k = 10$</td>
<td>$k = 20$</td>
<td>$k = 10$</td>
</tr>
<tr>
<td>48 taxa 25%</td>
<td>24</td>
<td>25</td>
<td>10</td>
<td>902</td>
<td>1224</td>
</tr>
<tr>
<td></td>
<td>(1+23)</td>
<td>(1+23)</td>
<td>(2+23)</td>
<td>(9+23)</td>
<td>(9+1)</td>
</tr>
<tr>
<td>75%</td>
<td>18</td>
<td>18</td>
<td>3</td>
<td>2475</td>
<td>4105</td>
</tr>
<tr>
<td></td>
<td>(1+17)</td>
<td>(1+17)</td>
<td>(2+23)</td>
<td>(1+2474)</td>
<td>(1+4104)</td>
</tr>
<tr>
<td>96 taxa 25%</td>
<td>500</td>
<td>507</td>
<td>49</td>
<td>1249</td>
<td>2110</td>
</tr>
<tr>
<td></td>
<td>(3+497)</td>
<td>(10+491)</td>
<td>(48+491)</td>
<td>(1+1248)</td>
<td>(2+2108)</td>
</tr>
<tr>
<td></td>
<td>(1+348)</td>
<td>(1+486)</td>
<td>(4+486)</td>
<td>(1+3162)</td>
<td>(1+6325)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>PhyML+MPR</th>
<th>PhyML+MSS</th>
<th>PhyML+ACS97+MW$^\ast$</th>
<th>PhyML+SDM+Fitch</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k = 10$</td>
<td>$k = 20$</td>
<td>$k = 10$</td>
<td>$k = 20$</td>
</tr>
<tr>
<td>48 taxa 25%</td>
<td>155</td>
<td>246</td>
<td>209</td>
<td>168</td>
</tr>
<tr>
<td></td>
<td>(143+12)</td>
<td>(143+103)</td>
<td>(143+165)</td>
<td>(143+2+23)</td>
</tr>
<tr>
<td>75%</td>
<td>44</td>
<td>387</td>
<td>84</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>(37+7)</td>
<td>(37+350)</td>
<td>(37+1+46)</td>
<td>(37+1+17)</td>
</tr>
<tr>
<td>96 taxa 25%</td>
<td>268</td>
<td>16,778</td>
<td>2218</td>
<td>399</td>
</tr>
<tr>
<td></td>
<td>(230+38)</td>
<td>(451+71)</td>
<td>(230+1+1987)</td>
<td>(451+1+1987)</td>
</tr>
<tr>
<td></td>
<td>53</td>
<td>18,050</td>
<td>1915</td>
<td>591</td>
</tr>
<tr>
<td>75%</td>
<td>50+3</td>
<td>18,101</td>
<td>2066</td>
<td>591</td>
</tr>
<tr>
<td></td>
<td>(50+18,000)</td>
<td>18,101</td>
<td>(50+1+1864)</td>
<td>(101+1+1964)</td>
</tr>
</tbody>
</table>
Comparing high-level scenarios, we see that with \( n = 96 \) PhyML+SDM+Fitch tends to be handicapped in comparison to PhyML+MRP, due to the use of Fitch. Replacing Fitch by FastME in case of full distance supermatrix does not significantly change the topological accuracy (e.g., with \( k = 20, 48 \) taxa, and 25\% deletion rate, average \( d_k \) topological distances are 0.0162 for PhyML+SDM+Fitch and 0.0168 for PhyML+SDM+FastME) but makes the SDM approach faster than PhyML+MRP. The two other high-level scenarios, PhyML+ACSA97+MW* and PhyML+MSS, are slower than MRPP- and SDM-based scenarios. PhyML+ACSA97+MW* is penalized by MW* that is slower than Fitch (used here without global rearrangements, contrary to Makarenkov and Lapointe, 2004). MSS appears as a slow algorithm, likely due to the combination of its complex optimality criterion and of SPR topological rearrangements.

Finally, as trees built with SDM+Fitch are close to the correct tree \( T \), we observe a clear improvement in PhyML running time when using SDM+Fitch as starting tree instead of DTE. Thus, with 96 taxa, \( k = 20, \) and 75\% taxon deletion, SDM+Fitch+PhyML runs 4,286 s, as compared to 6,329 s for DTE+PhyML, i.e., a relative gain of around 50\%.

We see from these comparisons that SDM-based scenarios not only have high topological accuracy but are also efficient relative to the other approaches. Moreover, they become much faster when the distance supermatrix does not contain any missing entry, thanks to the use of a fast distance-based tree building method.

**APPLICATION**

To illustrate the properties of SDM, we analyzed a data set of placental mammals (with focus on Cetartiodactyla), which was used by Gatesy et al. (2002) in a parsimony-based low-level combination framework. This taxonomic group was recently studied using different data and high-level approaches by Mahon (2004) and Price et al. (2005). We first describe Gatesy et al.’s data set and the various tree-building scenarios we tested, then provide the results, both in terms of running time and likelihood of the inferred trees. As we shall see, these results confirm our findings with simulated data sets.

**Data and Tree Building Scenarios**

The original Gatesy et al. data set comprises 57 character sources: 3 morphological data sets, 5 protein sequences, 1 tranposon, 33 nuclear genes, and 15 mitochondrial genes. As the current version of PhyML does no allow for separate analysis of various data types, we only retained the DNA coding sequences. We then considered a data set of 48 genes, 36,639 sites, and 75 placental mammals, from which 7 Afrotherians were used to root the inferred trees. As shown in Gatesy et al., this gene collection has high taxonomic sampling heterogeneity, and 68\% of the characters are missing. To obtain a fair comparison between Gatesy et al.’s tree-building approach and the other scenarios, we run TNT on the 48 concatenated genes, with 25 random taxon additions, TBR branch swapping, and ratchet default option. The corresponding tree is called Gatesy-TNT in the following.

All scenarios described in our simulations were also applied to this gene collection. The distance matrices were estimated using the GTR model (Rodriguez et al., 1990). To weight source matrices in SDM, we used \( w_p = \ell_p/\ell_p (\lambda_p - 1)] \) in Equation (1), which compensates for taxon number heterogeneity among genes (e.g., 10 taxa for \( \alpha \)-lactalbumin and 75 for cytochrome b). As the cytochrome b gene is present for all of the 75 taxa, the SDM distance supermatrix does not contain any missing entry and we used FastME instead of Fitch. Likelihood computations were performed using PhyML with the GTR+\( \Gamma \) model; we used eight rate categories and the gamma distribution parameter was estimated from the data. Invariant sites were not used as their proportion was estimated to be zero in preliminary studies. Moreover, to estimate the likelihood of all the topologies from the various scenarios, we fitted branch lengths and parameters to the original supermatrix of characters, using PhyML with the same GTR+\( \Gamma \) substitution model.

**Results**

The most likely tree is built by SDM+FastME+PhyML. This phylogeny is shown in Figure 3 and its log-likelihood is equal to \(-330,354\). This tree is relatively close to Gatesy et al.’s original tree, which has a log-likelihood of \(-330,492\). Quartet distance between both is of 0.028. Although Gatesy et al. found that Camelidae+Tayassuidae+Suidae were monophyletic, our tree displays a basal position of Camelidae among Cetartiodactyla and a sister-group relationship between Suina and Hippopotamidae+Cetacea+Ruminantia. Such basal position of Camelidae has already been proposed and discussed by Madsen et al. (2001) and Waddell et al. (2003) in low-level combination studies, and by Price et al. (2005) in a high-level combination framework. We also found that Pholidota+Carnivora is the nearest parent of Perissodactyla, which is another different branching relative to the topology found by Gatesy et al. As the corresponding branching has low bootstrap support in the tree of Gatesy et al., the tree in Figure 3 represents a likely alternative (biologically and mathematically). Gatesy-TNT tree is not much different from Gatesy et al.’s original tree; its log-likelihood is of \(-330,428\) (instead of \(-330,492\)) and quartet distance between both is equal to 0.007.

Results of all the scenarios are summarized in Table 2. We measured (1) the log-likelihood (as explained above), (2) the running time (using a 1.8-GHz Pentium IV PC with 1 Gb RAM), and (3) the topological distance between the corresponding tree and the best (SDM+FastME+PhyML) tree. As all trees are relatively close, we used the bipartition distance instead of the quartet distance to augment the contrast (see above
The SDM+FASTME tree is inferred in a running time of approximately 30 s, considerably faster than any other scenario (except SDM*+FASTME and DTE). This confirms the benefits of this medium-level approach at an exploratory stage, or for building a starting tree. This speed should also be useful to perform bootstrap analysis, which is hardly achievable with other scenarios. The $\alpha_p$ values estimated by SDM range from 0.26 to 2.80, with a median value of 1.17. As the $\alpha_p$ parameters are inversely proportional to the evolutionary rates, these results show that the data set of Gatesy et al. is composed of genes with quite heterogeneous rates. For example, $1/\alpha_p = 0.356$ corresponds to the slowest gene, the nuclear ZFX, and $1/\alpha_p = 3.755$ corresponds to the fastest one, the mitochondrial ATP8; the rate ratio between both is about 11.0. SDM medium-level based scenario can then be used to obtain the evolutionary rates of the studied genes in a quick way (i.e., much faster than any ML-based method). The advantage of such approach was already discussed by Bevan et al. (2005), who used it to account for gene rate heterogeneity in ML-based tree inference with very low computational cost.

**Conclusion**

We have presented a novel method, named SDM, to combine a collection of source distance matrices into a single distance supermatrix. SDM can be used in tree-building scenarios of various levels and computational costs. Using large-scale simulations and a real phylogenomics case study, we have shown that SDM, used together with FITCH or FASTME tree building programs, has topological accuracy similar to that of the popular MRPI method. With low taxon overlap, SDM tends to outperform MRPI, notably when it is used in a high-level way to combine gene trees. Moreover, in a low-level context, SDM can be used to quickly construct a starting tree to be refined by a maximum likelihood method. According to our simulations, this latter approach seems to be the most accurate gene combination method to date. This result could be affected by strong heterogeneity in the evolutionary modes of the studied genes, which was not incorporated in our simulations but may occur with real phylogenomic data. However, likelihood-based separate analysis (e.g., MrBayes; Huelsenbeck and Ronquist, 2001) provides a way to deal with such schemes in the low-level context, and SDM-based medium and high-level scenarios should not be affected as different models can be used to estimate their input (i.e., distances and trees, respectively).

The SDM algorithm is very fast. The computing time required by the SDM approach (i.e., first running SDM, then inferring the tree from the SDM supermatrix using a distance algorithm) greatly depends on the taxon overlap among genes. When the SDM supermatrix is complete (which occurs frequently, as some genes have been sequenced for a large number of species), the SDM approach is very efficient thanks to the use of fast algorithms such as NJ, BioNJ, or FASTME; in this case, huge data sets can be dealt with in a few minutes on a standard computer.

---

### Table 2. Results of the various tree building scenarios with Gatesy et al.’s (2002) data set.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Log-likelihood</th>
<th>Running time</th>
<th>$d_{df}$</th>
<th>$P$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDM+FASTME+PHYML</td>
<td>−330,354</td>
<td>4.6 h</td>
<td>0.000</td>
<td>0.986</td>
</tr>
<tr>
<td>DTE+PHYML</td>
<td>−330,394</td>
<td>4.7 h</td>
<td>0.097</td>
<td>0.014</td>
</tr>
<tr>
<td>PhyML+MRP</td>
<td>−330,427</td>
<td>2.0 h</td>
<td>0.160</td>
<td>10$^{-4}$</td>
</tr>
<tr>
<td>Gatesy-TNT</td>
<td>−330,428</td>
<td>48 min</td>
<td>0.188</td>
<td>&lt;10$^{-4}$</td>
</tr>
<tr>
<td>PhyML+SDM+FASTME</td>
<td>−330,577</td>
<td>2.0 h</td>
<td>0.236</td>
<td>&lt;10$^{-4}$</td>
</tr>
<tr>
<td>SDM+FASTME</td>
<td>−331,532</td>
<td>30 s</td>
<td>0.458</td>
<td>&lt;10$^{-4}$</td>
</tr>
<tr>
<td>PhyML+MSS</td>
<td>−332,212</td>
<td>4.1 h</td>
<td>0.458</td>
<td>&lt;10$^{-4}$</td>
</tr>
<tr>
<td>PhyML+ACS97+MW*</td>
<td>−332,261</td>
<td>2.2 h</td>
<td>0.444</td>
<td>&lt;10$^{-4}$</td>
</tr>
<tr>
<td>DTE</td>
<td>−333,692</td>
<td>2 s</td>
<td>0.590</td>
<td>&lt;10$^{-4}$</td>
</tr>
</tbody>
</table>

Comparison between both measures. This distance, also called Robinson and Foulds ($d_{df}$) distance, was normalized; 0 corresponds to identical trees, whereas 1 means that both trees do not share any bipartition (clade). Finally (4), we checked for the significance of our findings using Shimodaira asymptotically unbiased test (2002), as implemented in CONSEL software.

The SDM+FASTME+PHYML tree is significantly better than the other trees ($P = 0.986$). Overall, the results are in good accordance with simulations, even though the ranking criteria are not the same (likelihood versus topological distance with the model tree). Low-level methods tend to be the best ones, including Gatesy-TNT parsimony-based approach (but excluding DTE). Moreover, using SDM+FASTME (instead of DTE) to build a starting tree increases the likelihood of the resulting PHYML tree. Among high-level scenarios, PHYML+MRP performs best (~70 log-likelihood units below the best tree), PHYML+SDM+FASTME is also efficient (~220 log-likelihood units below the best tree), whereas PhyML+MSS and PhyML+ACS+MW* are outperformed (~1900 log-likelihood units below the best tree). SDM+FASTME medium-level scenario (~1200 log-likelihood units below the best tree) is behind the best high-level methods, but performs better than PhyML+MSS and PhyML+ACS+MW*. Finally, DTE is the worst of all methods, just as in simulations (~3,000 log-likelihood units below the best tree). Topological distances (measured by $d_{df}$) between the best and the other trees are also significant; e.g., DTE and the best trees share only 41% of clades, whereas the PhyML+MRP value is of 84%. Results with quartet distance are much less contrasted as those measures become 88.5% and 98.5%, respectively.

This ordering of the scenarios is very similar to that of Figure 2, with 25% taxon deletion and large number of genes. Even though the ratio of missing characters in the Gatesy et al. data set is closer to 75% than to 25%, the gene number in this data set is large (48) and some genes are sequenced for all taxa (e.g., *cytochrome b*); information is then abundant, which explains the closeness with 25% (rather than 75%) taxon deletion.
computer. When the SDM supermatrix contains missing entries, as is the case for some recent very sparse data sets selected by computer programs (Driskell et al., 2004), slower algorithms such as FITC or MW* must be used; then the SDM approach is not as efficient with running times similar to those of MRP.

A key direction for further research is to develop fast algorithms, as fast as NJ or FASTME, to accurately reconstruct trees from distance matrices with missing entries. Other directions include exploring new weighting schemes within the SDM optimality criterion (1), or new linear constraints on the parameters.

Our implementation of the SDM method, in JAVA 1.4 for better portability, is available at http://www.lirmm.fr/~criscuol/soft/sdm. All simulated data sets can be downloaded from http://www.lirmm.fr/~criscuol/soft/sdm/data sets.

ACKNOWLEDGMENTS

We thank Nicolas Galtier, Fabienne Thomarat, and Maria Anisimova for helpful comments and suggestions. We also thank Olaf Bininda-Emonds, Roderic Page, François-Joseph Lapointe and an anonymous reviewer for their help during the reviewing process. This work has been supported by ACI Informatique, Mathématique et Physique en Biologie Moléculaire (ACI IMP-Bio), and by IFR119 Biodiversité Continentale Méditerranéenne et Tropicale (Montpellier) computing facilities. This publication is the contribution number 2005-054 of the Institut des Sciences de l’Évolution de Montpellier (UMR 5554, CNRS).

REFERENCES


(N. El-Mabrouk, T. Lengauer, D. Sankoff, eds.). Publications CRM, Montreal.


APPENDIX

The goal is to minimize criterion (3), which can be written as

$$f(v) = \sum_{i,j} \sum_{p \in \mathcal{L}_p} w_p(\alpha_i \Delta_{ij}^p + a_i + a_j - \hat{\Delta}_{ij})^2$$

where $v = (a_1, \ldots, a_p, \ldots, a_s, \ldots)$ and

$$\hat{\Delta}_{ij} = \frac{1}{n_{ij}} \sum_{p \in \mathcal{L}_p} w_p(\alpha_i \Delta_{ij}^p + a_i + a_j)$$

with $W_p = \sum_{i \neq j} w_p$.

subject to linear constraints (4), (5), and (6):

$$h^{(1)}(p) = \sum_{i \in \mathcal{L}_p} a_i = k,$$

$$h^{(2)}(p) = \sum_{i \in \mathcal{L}_p} a_i = 0, \forall i \in \mathcal{L}_p,$$

$$h^{(3)}(p) = \sum_{i \in \mathcal{L}_p} a_i = 0, \forall p \neq k.$$
monodimensional optimization) that any minimizer must satisfy is (Luenberger, 1984: 300):

\[
\begin{align*}
\frac{\partial}{\partial a_m} f(v) + \lambda \frac{\partial}{\partial a_m} h_m(v) &= 0, \quad \forall m = 1, 2, \ldots, k, \\
\frac{\partial}{\partial a_m} f(v) + \mu_i \frac{\partial}{\partial a_m} h_i(v) + \eta_m \frac{\partial}{\partial a_m} h_{i+}(v) &= 0, \quad \forall m < k, \forall i \in \mathcal{L}_m, \\
\frac{\partial}{\partial a_m} f(v) + \mu_i \frac{\partial}{\partial a_m} h_i(v) &= 0, \quad \forall i \in \mathcal{L}_i, \\
h_i(v) &= k, \quad \forall i \in \mathcal{L}_i, \\
h_i(v) &= 0, \quad \forall i \in \mathcal{L}_i, \\
h_i(v) &= 0, \quad \forall p \neq k.
\end{align*}
\]

where \(\lambda, \mu_i,\) and \(\eta_m\) are the Lagrange multipliers induced by linear constraints \(h_m(v), h_i(v),\) and \(h_i(v)\) respectively. We have:

\[
\begin{align*}
\frac{\partial}{\partial a_m} f(v) &= 2 \sum_{i,j:p \in \mathcal{L}_p} w_{ij} \left( \Delta_{ij}^n - \frac{w_{ij}}{W_{ij}} \Delta_{ij}^p \right) \left( a_m \Delta_{ij}^n + a_{mn} + a_{jm} - \Delta_{ij} \right) \\
&\quad + \sum_{i,j:p \in \mathcal{L}_p} w_{ij} \left( - \frac{w_{ij}}{W_{ij}} \Delta_{ij}^p \left( a_p \Delta_{ij}^p + a_{jp} + a_{ip} - \Delta_{ij} \right) \right) \\
&= 2w_m \sum_{i,j:p \in \mathcal{L}_p} \Delta_{ij}^n \left( a_m \Delta_{ij}^n + a_{mn} + a_{jm} - \Delta_{ij} \right) \\
&\quad - \frac{1}{W_{ij}} \sum_{p \not\in \mathcal{L}_p} w_{ij} \left( a_p \Delta_{ij}^p + a_{jp} + a_{ip} - \Delta_{ij} \right) \left( \alpha_i \Delta_{ij}^n + a_{jm} + a_{pn} - \Delta_{ij} \right) \\
&= 2w_m \sum_{i,j:p \in \mathcal{L}_p} \Delta_{ij}^n \left( a_m \Delta_{ij}^n + a_{mn} + a_{jm} - \Delta_{ij} \right)
\end{align*}
\]

and, with similar arithmetic,

\[
\frac{\partial}{\partial a_m} f(v) = 4w_m \sum_{j \in \mathcal{L}_{m-1}} \left( a_m \Delta_{ij}^n + a_{mn} + a_{jm} - \Delta_{ij} \right).
\]

Linear system \(S\) can then be written as:

\[
\begin{align*}
\sum_{i,j:p \in \mathcal{L}_p} \Delta_{ij}^n \left( a_p \Delta_{ij}^p + a_{ip} + a_{jp} - \Delta_{ij} \right) + \lambda &= 0, \quad \forall p = 1, 2, \ldots, k, \\
\sum_{j \in \mathcal{L}_{j-1}(p)} (a_p \Delta_{ij}^n + a_{ip} + a_{jp} - \Delta_{ij}) + \mu_i + \eta_p &= 0, \quad \forall p < k, \forall i \in \mathcal{L}_p, \\
\sum_{j \in \mathcal{L}_{j-1}(i)} (a_i \Delta_{ij}^n + a_{ji} + a_{jk} - \Delta_{ij}) + \mu_i &= 0, \quad \forall i \in \mathcal{L}_i, \\
a_p &= k, \quad \forall p \not\in \mathcal{L}_p, \\
a_p &= 0, \quad \forall i \in \mathcal{L}_i, \\
a_p &= 0, \quad \forall p \not\in \mathcal{L}_p.
\end{align*}
\]

\(S\) is a square linear system, with \(n + 2k + \sum_p n_p\) equations and parameters (including Lagrange multipliers), and \(S\) has at least one solution. For \(S\) to define the unique global optimum of \(f(v)\) subject to the constraints, the second-order necessary condition (Luenberger, 1984: 306) must be fulfilled (equivalent to positivity of the second derivative in unconstrained mono dimensional minimization). In our quadratic programming problem, where \(f(v)\) is non-negative, this condition becomes:

\[
\begin{align*}
f(v) &= 0 \\
h_i(v) &= 0, \quad \forall i \in \mathcal{L}_i \\
h_{i+}(v) &= 0, \quad \forall p \neq k.
\end{align*}
\]

\(f(v)\) is a sum of squares. \(f(v) = 0\) implies that all the squares are null, which means that for any \(i, j\) pair \((k_i \geq 2)\) we have \(a_i \Delta_{ij}^n + a_{ip} + a_{jp} = a_i \Delta_{ij}^p + a_{ip} + a_{jp}, \forall p, p' \not\in \mathcal{L}_p, \mathcal{L}_{p'}\). \(f(v) = 0\) then forces \(\sum_{i,j:p \not\in \mathcal{L}_p} \Delta_{ij}^p (k_i - 1)\) independent linear equations. Combining these equations with the \(k + n\) constraints, we obtain a linear system with \(k + \sum_p n_p\) parameters (i.e., the size of \(v\)). The second-order sufficiency condition is then equivalent to testing the linear independence of the set of column vectors defining this linear system. This can easily be achieved numerically. However, except in very special cases (corresponding to equalities or redundancies, see example below), this vector set is linearly independent as soon as the number of vectors is less than, or equal to, the vector dimension, that is:

\[
k + \sum_p n_p \leq k + n + \sum_{i,j:p \not\in \mathcal{L}_p} (k_i - 1),
\]

which simplifies into:

\[
\sum_{i,j:p \not\in \mathcal{L}_p} (k_i - 1) + n - \sum_p n_p \geq 0 \quad (7)
\]

When (7) is fulfilled, the linear system \(S\) should then define the unique global optimum of our constrained optimization problem. The only exception we were able to find in all our simulations and experiments involved flawed data sets, where one of the source matrices was duplicated.

The left-hand side term in (7) measures the matrix overlap. For example, in the extreme case where we only have two source matrices that only share two taxa, the three components in this term equal 1, 2, and 4, respectively, and (7) is violated; in other words, a single (1) distance comparison plus 4 \((= 2 + k)\) constraints is not enough to estimate 6 \((= 4 + k)\) parameters. Assuming now that the two matrices share 3 taxa, the sum in (7) becomes \(3 + 3 - 6 = 0\), i.e., 3 comparisons are enough to estimate 8 parameters subject to 5 constraints, and \(S\) defines a unique global optimum.

Unicity of the global optimum yields the consistency of the SDM approach in estimating the relative rates of the genes. Assume that all source matrices are issued from a single \((\Delta_a)\) matrix through multiplication by \(\theta\) factors, each representing the evolutionary rate of gene \(p\). We have \((\Delta_a^p = \theta \Delta_a)\), just as in the proportional model of Yang (1996). Moreover, without loss of generality, assume that the \(\theta\) factors are rescaled to obtain \(\sum_i 1/\theta_p = k\). SDM then consistently estimates the \(\theta\) values, as soon as condition (7) is fulfilled. Let \(v^*\) be defined by \(a_i^* = 1/\theta_i\) and \(a_{ip} = 0, \forall i, p\). It is easily seen that \(f(v^*) = 0\) and that all the constraints are satisfied by \(v^*\). As (7) is fulfilled, \(v^*\) is the unique solution of \(S\), and \(\theta_p = 1/\alpha_p^*\) is a consistent estimator of \(\theta_p\), which finishes the consistency proof.