

# SDM: a Fast Distance-based Approach for (Super)Tree Building in Phylogenomics

Alexis Criscuolo<sup>1,2</sup>, Vincent Berry<sup>2</sup>, Emmanuel J. P. Douzery<sup>1</sup> and Olivier Gascuel<sup>2,\*</sup>

<sup>1</sup> *Groupe Phylogénie Moléculaire. ISEM, Université Montpellier 2, CC 064, 34095 Montpellier Cedex 05, FRANCE.*

<sup>2</sup> *Equipe Méthodes et Algorithmes pour la Bioinformatique. LIRMM (CNRS, Université Montpellier 2), 161 rue Ada, 34392 Montpellier Cedex 05, FRANCE.*

\* Corresponding author:

<http://www.lirmm.fr/~gascuel>

[gascuel@lirmm.fr](mailto:gascuel@lirmm.fr)

Tel. (33 or 0 from France) 4 67 41 85 47

Fax. (33 or 0 from France) 4 67 41 85 00

**Keywords:** phylogenomics, evolutionary distances, distance method, supertree, supermatrix, MRP, total evidence

## Abstract

Phylogenomic studies aim to build phylogenies from large sets of homologous genes. Such “genome-sized” data require fast methods, because of the typically large numbers of taxa examined. In this framework, distance-based methods are useful for exploratory studies and building a starting tree to be refined by a more powerful maximum likelihood (ML) approach. However, estimating evolutionary distances directly from concatenated genes gives poor topological signal as genes evolve at different rates. We propose a novel method, named *Super Distance Matrix* (SDM), which follows the same line as *Average Consensus Supertree* (ACS, Lapointe and Cucumel, 1997) and combines the evolutionary distances obtained from each gene into a single distance supermatrix to be analyzed using a standard distance-based algorithm. SDM deforms the source matrices, without modifying their topological message, to bring them as close as possible to each other; these deformed matrices are then averaged to obtain the distance supermatrix. We show that this problem is equivalent to the minimization of a least-squares criterion subject to linear constraints. This problem has a unique solution which is obtained by resolving a linear system. As this system is sparse, its practical resolution requires  $O(n^a k^a)$  time, where  $n$  is the number of taxa,  $k$  the number of matrices and  $a < 2$ , which allows the distance supermatrix to be quickly obtained. Several uses of SDM are proposed, from fast exploratory studies to more accurate approaches requiring heavier computing time. Using simulations, we show that SDM is a relevant alternative to the standard *Matrix Representation with Parsimony* (MRP) method, notably when the taxa sets of the different genes have low overlap. We also show that SDM can be used to build an excellent starting tree for an ML approach, which both reduces the computing time and increases the topological accuracy. We use SDM to analyze the dataset of Gatesy et al. (2002) that involves 48 genes of 75 placental mammals. The results indicate that these genes have strong rate heterogeneity and confirm the simulation conclusions.

## INTRODUCTION

Phylogenomics, whereby phylogenies are built from large sets of 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; Devulder et al., 2005; Philippe 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 is another difficulty with such datasets, 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, ch. 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 seems 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 the *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 datasets 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 Keppler, 1988), BIONJ (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 con-

catenated 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 dataset of placental mammals (Gatesy et al., 2002). Mathematical proofs and equations are provided in an appendix.

## THE SDM METHOD

### *Notations and Definitions*

Let  $\mathcal{C} = \{(\Delta_{ij}^1), (\Delta_{ij}^2), \dots, (\Delta_{ij}^p), \dots, (\Delta_{ij}^k)\}$  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  $\mathcal{L} = \bigcup_p \mathcal{L}_p$ ; 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:

$$\tilde{\mathcal{L}}_p = \{i \in \mathcal{L}_p : \exists j \in \mathcal{L}_p - \{i\}, k_{ij} \geq 2\}, \quad \tilde{n}_p = |\tilde{\mathcal{L}}_p|,$$

$$\tilde{\mathcal{L}} = \bigcup_p \tilde{\mathcal{L}}_p \quad \text{and} \quad \tilde{n} = |\tilde{\mathcal{L}}|.$$

Our method involves bringing each matrix  $(\Delta_{ij}^p)$  closer (in the least-squares sense) relative to the others.  $\tilde{\mathcal{L}}$  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  $i, j$  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  $2(n_p - 1)$  non-diagonal 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  $(\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp})$  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 ( $\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp}$ ). 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, while 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:

$$V_{ij} = \sum_{p:\{ij\} \subset \mathcal{L}_p} w_p (\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp} - \bar{\Delta}_{ij})^2 \quad (1)$$

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

$$\bar{\Delta}_{ij} = \frac{1}{W_{ij}} \sum_{p:\{ij\} \subset \mathcal{L}_p} w_p (\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp}) \quad \text{with} \quad W_{ij} = \sum_{p:\{ij\} \subset \mathcal{L}_p} w_p. \quad (2)$$

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  $\ell_p$ . 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  $\ell_p/(\tilde{n}_p(\tilde{n}_p-1))$ , or the intermediate value  $\ell_p/\tilde{n}_p$  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:

$$\sum_{\substack{i,j:i \neq j \\ k_{ij} \geq 2}} V_{ij}. \quad (3)$$

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:

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

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 (5) forces, for each taxon  $i$ , the sum of  $a_{ip}$  to be equal to zero and forbids over-elongation (or shortening) of external branch lengths corresponding to taxon  $i$ :

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

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

$$\sum_{i \in \tilde{\mathcal{L}}_p} a_{ip} = 0, \quad \forall p = 1, 2, \dots, k-1. \quad (6)$$

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  $\sum_{i \in \tilde{\mathcal{L}}_k} a_{ik} = 0$  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  $k + \sum_p \tilde{n}_p$  variables  $\alpha_p$  and  $a_{ip}$ , adding Lagrange multipliers (Luenberger, 1984, ch.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  $O(\tilde{n}^3 k^3)$  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  $\alpha_p^*$  and  $a_{ip}^*$  be the optimal values of the parameters we obtain this way, then the SDM distance supermatrix  $(\Delta_{ij}^{\text{SDM}})$  is defined by:

$$\Delta_{ij}^{\text{SDM}} = \frac{1}{W_{ij}} \sum_{p:\{ij\} \subset \mathcal{L}_p} w_p (\alpha_p^* \Delta_{ij}^p + a_{ip}^* + a_{jp}^*) \quad \text{where} \quad W_{ij} = \sum_{p:\{ij\} \subset \mathcal{L}_p} w_p.$$

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  $t=b(nk)^a$ , we performed a linear regression on  $\log(t)$  as a function of  $\log(nk)$  and found that the estimated value of  $a$  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 (Farach 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 TREX 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, p.449). The MWMODIF algorithm (Makarenkov and Leclerc, 1999) from TREX and the FITCH algorithm (Felsenstein, 1997) from the PHYLIP package (Felsenstein, 1993) implement this technique.

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, datasets of Gatesy et al. (2002), Philippe et al. (2004) and Philippe et al. (2005) have high ratio of missing characters (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 *ribosomal 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 datasets 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 of  $\approx 19\%$  and  $\approx 1\%$ , while the ratio of missing characters is of  $\approx 92\%$  and  $\approx 87\%$ , 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 non-clocklike 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  $0.4 + 8.6V_p$ , 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  $k = 2, 4, 6, \dots, 20$  and generated 500 datasets per  $k$  value. For each of these datasets, 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 re-scale the  $T^p$  trees. To simulate partial overlap that occurs in real datasets, 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 4 taxa was preserved between each gene pair to maintain a common evolutionary history between genes and avoid meaningless datasets. 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 sizes  $\times$  500 collections  $\times$  2 overlap conditions =) 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-levels, (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.  $w_p = \ell_p$ ) 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 (Figure 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 data, 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 Introduction), and then used BIONJ to infer the DTE (*Distance-based Total Evidence*) phylogeny.

To obtain a ML-based total evidence phylogeny, we analysed 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 Figure 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 Figure 1) to infer the starting tree to be then refined by PHYML; we call this scenario SDM+FITCH+PHYML (Figure 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 ad-

dition sequences, TBR branch swapping and ratchet default option. We call this supertree construction scenario PHYML+MRP (Figure 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 Levasseur (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\* (Figure 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 (Figure 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 (Figure 1).

### *Topological Accuracy Measure*

We measured the topological accuracy of every scenario using the quartet distance  $d_q$  (Estabrook et al., 1985) between the inferred tree  $\hat{T}$  and the model tree  $T$ .  $d_q$  counts the number of resolved 4-trees (i.e. four taxon trees) present in one tree but not in the other.  $d_q$  is then the sum of two error types: the type I error corresponding to resolved 4-trees induced by  $\hat{T}$  but not present in  $T$ , the type II error corresponding to resolved

4-trees in  $T$  but not induced by  $\widehat{T}$ . As any fully resolved tree with  $n$  taxa induces  $\binom{n}{4}$  4-trees, this measure can take any integer value between 0 and  $2\binom{n}{4}$ .  $d_q$  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,  $d_q$  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, while the  $d_q$  distance remains moderate as only quartets involving this taxon are modified. Thus,  $d_q$  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).  $d_q$  was normalized by dividing its value by  $2\binom{n}{4}$ ; 0 then corresponds to identical trees, while 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  $\times$  2 overlap conditions), the collected 500  $d_q$  values were averaged and are graphically represented in Figure 2. These graphs show the average  $d_q$  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  $d_q$  distance between  $\widehat{T}$  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  $a_{ip}$  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, while 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, while 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), while 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 datasets. 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 (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 quasi-random 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, while 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 datasets 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 datasets 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, while both follow a closely related scheme. Thus, results in Table 1 illustrate the main tendencies but should not be over-interpreted.

We first see that SDM on its own is a fast algorithm. For example, it only requires 48 seconds 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 seconds. 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 seconds as compared to 539 seconds, in our previous example). FITCH is useful as it copes with missing entries, but, as explained earlier, real datasets 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 ( $k = 20$ ) datasets, we then used FASTME instead of FITCH. Topological accuracy remains similar (e.g. with 48 taxa and 25% deletion, average  $d_q$  topological distances are 0.0242 for SDM+FITCH and 0.0268 for SDM+FASTME) but the tree inference time is less than 1 second in all settings; e.g., with

96 taxa,  $k = 20$  and 25% taxon deletion, the SDM+FASTME scenario requires approximately 50 seconds, as compared to 539 seconds 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.

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  $d_q$  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+ACS97+MW\* and PHYML+MSS, are slower than MRP and SDM-based scenarios. PHYML+ACS97+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 seconds, as compared to 6,329 seconds 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 dataset 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 high-level approaches by Mahon (2004) and Price et al. (2005) and different data. We first describe Gatesy et al. dataset 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 datasets.

### *Data and Tree Building Scenarios*

The original Gatesy et al. dataset comprises 57 character sources: 3 morphological datasets, 5 protein sequences, 1 transposon, 33 nuclear genes, and 15 mitochondrial genes. As the current version of PHYML does not allow for separate analysis of various data types, we only retained the DNA coding sequences. We then considered a dataset 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. (2002), this gene collection has high taxonomic sampling heterogeneity, and 68% of the characters are missing. To obtain a fair comparison between Gatesy's et al. 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  $w_p = \ell_p / (\tilde{n}_p(\tilde{n}_p - 1))$  in equation (1), which compensates for taxon number heterogeneity among genes (e.g. 10 taxa for *α-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. TNT was run with 500 random taxon additions, TBR branch swapping and ratchet default option. Likelihood computations were performed using PHYML with the GTR+ $\Gamma$  model; we used 8 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+ $\Gamma_8$  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. original tree, which has a log-likelihood of  $-330,492$ . Quartet distance between both is of 0.028. While 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 the 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). Gagesy-TNT tree is not much different from the 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 comparison between both measures). This distance, also called Robinson and Foulds ( $d_{RF}$ ) distance, was normalized; 0 corresponds to identical trees, while 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.

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 Gagesy-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 ( $\sim 70$  log-likelihood units below the best tree). PHYML+SDM+FASTME is also efficient ( $\sim 220$  log-likelihood units below the best tree), while PHYML+MSS and PHYML+ACS+MW\* are outperformed ( $\sim 1,900$  log-likelihood units below the best tree). SDM+FASTME medium-level scenario ( $\sim 1,200$  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 ( $\sim 3,000$  log-likelihood units below the best tree). Topological distances

(measured by  $d_{BE}$ ) between the best and the other trees are also significant; e.g. DTE and the best trees share only 41% of clades, while 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 Gatesy et al. dataset 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.

The SDM+FASTME tree is inferred in a running time of approximately 30 seconds, considerably faster than any other scenario (except SDM\* 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 dataset 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 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 MRP method. With low taxon overlap, SDM tends to outperform MRP, 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 scheme 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 datasets can be dealt with in a few minutes on a standard computer. When the SDM supermatrix contains missing entries, as is the case for some recent very sparse datasets selected by computer programs

(Driskell et al., 2004), slower algorithms such as FITCH 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 datasets can be downloaded from <http://www.lirmm.fr/~criscuol/soft/sdm/datasets> !

## ACKNOWLEDGEMENTS

We thank Nicolas Galtier, Fabienne Thomarat and Maria Anisimova for helpful comments and suggestions. 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 N° 2005-YYY of the Institut des Sciences de l'Evolution de Montpellier (UMR 5554 - CNRS).

## REFERENCES

- Anisimova, M. and O. Gascuel. 2006. Approximate likelihood ratio test for branches: a fast, accurate and powerful alternative. *Syst. Biol.*, in press.
- Barthélemy, J. P. and A. Guénoche. 1991. Trees and proximity relations. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, Ltd., Chichester.
- Baum, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon*, 41:3–10.
- Berry, V. and O. Gascuel. 1996. Interpretation of bootstrap trees: threshold of clade selection and induced gain. *Mol. Biol. Evol.*, 13(7):999–1011.
- Bevan, R. B., B. F. Lang and D. Bryant. 2005. Calculating the evolutionary rates of different genes: a fast, accurate estimator with applications to maximum likelihood phylogenetic analysis. *Syst. Biol.*, 54:900–915.
- Bininda-Emonds, O. R. P., editor. 2004. Phylogenetic supertrees: Combining information to reveal the tree of life. Kluwer Academic Publishers, Dordrecht.
- Bininda-Emonds, O. R. P. and H. N. Bryant. 1998. Properties of matrix representation with parsimony analyses. *Syst. Biol.*, 47:497–508.
- Bininda-Emonds, O. R. P. and M. J. Sanderson. 2001. Assessment of the accuracy of matrix representation with parsimony supertree construction. *Syst. Biol.*, 50:565–579.

- Bourque, M. 1978. Arbres de Steiner et réseaux dont varie l'emplacement de certains sommets. PhD thesis, Département d'informatique et de recherche opérationnelle, Université de Montréal.
- Bryant, D. 2001. A classification of consensus methods for phylogenetics. Pages 163–184 *in* Bioconsensus, Janowitz, M., F.-J. Lapointe, F.R. McMorris, B. Mirkin, F.S. Roberts, DIMACS-AMS edition. Providence.
- Creevey, C. J. and J. O. McInerney. 2005. CLANN: investigating phylogenetic information through supertree analyses. *Bioinformatics*, 21:390–392.
- Daubin, V., M. Gouy, and G. Perrière. 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res.*, 12(7):1080–1090.
- De Soete, G. 1984. Additive tree representations of incomplete dissimilarity data. *Quality and Quantity*, 18:387–393.
- Desper, R. and O. Gascuel. 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J. Comp. Biol.*, 19(5):687–705.
- Devulder, G., M. Pérouse de Montclos, and J. P. Flandrois. 2005. A multigene approach to phylogenetic analysis using the genus *Mycobacterium* as a model. *Int. J. Syst. Evol. Microbiol.*, 55:293–302.
- Driskell, A. C., C. Ané, J. G. Burleigh, M. M. McMahon, B. C. O'Meara and M. J. Sanderson. 2004. Prospects for building the tree of life from large sequence databases. *Science*, 306:1172–1174.
- Eisen, J. A. and C. M. Fraser. 2003. Phylogenomics: Intersection of evolution and genomics. *Science*, 300:1706–1707.

- Estabrook, G. F., F. R. McMorris, and C. A. Meacham. 1985. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Syst. Zool.*, 34:193–200.
- Eulenstein, O., D. Chen, J. G. Burleigh, D. Fernandez-Baca, and M. J. Sanderson. 2004. Performance of flip supertree construction with a heuristic algorithm. *Syst. Biol.*, 53:299–308.
- Farach, M., S. Kannan and T. Warnow. 1995. A robust model for finding optimal evolutionary trees. *Algorithmica*, 13:155–179.
- Felsenstein, J. 1993. PHYLIP: Phylogeny inference package, version 3.6b. Distributed by the author. University of Washington, Seattle.
- Felsenstein, J. 1997. An alternating least-squares approach to inferring phylogenies. *Syst. Biol.*, 46:101–111.
- Fitch, W. M. and E. Margoliash. 1967. Construction of phylogenetic trees. *Science*, 155:279–284.
- Gascuel, O. 1994. A note on Sattath and Tversky's, Saitou and Nei's and Studier and Keppler's algorithms for inferring phylogenies from evolutionary distances. *Mol. Biol. Evol.*, 11(6):961–963.
- 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.
- Gatesy, J., C. Matthee, R. DeSalle, and C. Hayashi. 2002. Resolution of a supertree/supermatrix paradox. *Syst. Biol.*, 51:652–664.

- Goloboff, P., J. Farris, and K. Nixon. 2003. TNT: Tree analysis using new technology. Distributed by the authors.
- Guénoche, A. and S. Grandcolas. 1999. Approximations par arbre d'une distance partielle. *Math. Inf. Sci. Humaines*, 146:51–64.
- Guindon, S. and O. Gascuel. 2003. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, 52:696–704.
- Huelsenbeck, J. P. 2001. When are fossils better than extant taxa in phylogenetic analysis? *Syst. Zool.*, 40(4):458–469.
- Huelsenbeck, J. P. and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755.
- Huson, D. H., S. M. Nettles and T. J. Warnow. 1999. Disk-covering, a fast-converging method for phylogenetic tree reconstruction. *J. Comp. Biol.*, 6(3-4):369–386.
- Kearney, M. 2002. Fragmentary taxa, missing data, and ambiguity: mistaken assumptions and conclusions. *Syst. Biol.*, 51:369–381.
- 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.
- Kishino, H., T. Miyata and M. Hasegawa. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.*, 31:151–160.
- Landry, P.-A., F.-J. Lapointe and J. A. W. Kirsch. 1996. Estimating phylogenies from lacunose distance matrices: Additive is superior to ultrametric estimation. *Mol. Biol. Evol.*, 13:818–823.

- Lapointe, F.-J. and G. Cucumel. 1997. The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa. *Syst. Biol.*, 46:306–312.
- Lapointe, F.-J. and P.-A. Landry. 2001. A fast procedure for estimating missing distances in incomplete matrices prior to phylogenetic analysis. Pages 189–190 *in* N. El-Mabrouk, T. Lengauer, D. S., editor, *Currents Computational Molecular Biology*. Publications CRM, Montréal.
- Lapointe, F.-J., J. A. W. Kirsch and J. M. Hutcheon. 1999. Total evidence, consensus and bat phylogeny: A distance-based approach. *Mol. Phylogenet. Evol.*, 11:55–66.
- Lapointe, F.-J. and C. Levasseur. 2004. Everything you always wanted to know about the average consensus, and more. Pages 87–105 *in* O.R.P. Bininda-Emonds, editor, *Phylogenetic supertrees: Combining information to reveal the tree of life*. Kluwer Academic, Dordrecht.
- Lapointe, F.-J., M. Wilkinson and D. Bryant . 2003. Matrix representation with parsimony or with distances: two sides of the same coin? *Syst. Biol.*, 52:865–868.
- Levasseur, C. and F.-J. Lapointe. 2001. War and peace in phylogenetics: A rejoinder on total evidence and consensus. *Syst. Biol.*, 50:881–891.
- Luenberger, D. G. 1984. *Linear and Nonlinear programming*. Addison-Wesley.
- Madsen, O., M. Scally, C. J. Douady, D. J. Kao, R. W. DeBry, R. Adkins, H. M. Amrine, M. J. Stanhope, W. W. de Jong, and M. S. Springer. 2001. Parallel adaptive radiations in two major clades of placental mammals. *Nature*, 409(6820):610–614.
- Mahon, S. A. 2004. A molecular supertree of the Artiodactyla. Pages 411–437 *in* O.R.P. Bininda-Emonds, editor, *Phylogenetic supertrees: Combining information to reveal the tree of life*. Kluwer Academic, Dordrecht.

- Makarenkov, V. 2001. TREX: reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics*, 17:664–668.
- Makarenkov, V. and F.-J. Lapointe. 2004. A weighted least-squares approach for inferring phylogenies from incomplete distance matrices. *Bioinformatics*, 20:2113–2121.
- Makarenkov, V. and B. Leclerc. 1999. An algorithm for the fitting of a phylogenetic tree according to a weighted least-squares criterion. *J. classif.*, 16(1):3–26.
- Nei, M. and L. Jin. 1989. Variances of the average numbers of nucleotides substitutions within and between populations. *Mol. Biol. Evol.*, 6:290–300.
- Page, R. 2002. Modified mincut supertrees. In Guigo, R. and Gusfield, D., editors, Vol. 2452 of *Lecture Notes In Computer Science*, 537–552.
- Philippe, H., E. A. Snell, E. Baptiste, P. Lopez, P. W. H. Holland, and D. Casane. 2004. Phylogenomics of eukaryotes: Impact of missing data on large alignments. *Mol. Biol. Evol.*, 21(9):1740–1752.
- Philippe, H., N. Lartillot and H. Brinkmann. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of ecdysozoa, lophotrochozoa, and protostomia. *Mol. Biol. Evol.*, 22(5):1246–1253.
- Piaggio-Talice, R., G. Burleigh and O. Eulenstein. 2004. Quartet supertrees. Pages 173–191 *in* O.R.P. Bininda-Emonds, editor, *Phylogenetic supertrees: Combining information to reveal the tree of life*. Kluwer Academic, Dordrecht.
- Price, S. A., O. R. P. Bininda-Emonds and J. L. Gittleman. 2005. A complete phylogeny of the whales, dolphins and even-toed hoofed mammals (Cetartiodactyla). *Biol. Rev. Comb. Philos. Soc.*, 80(3):445–473.

- Pupko, T., D. Huchon, Y. Cao, N. Okada, and M. Hasegawa. 2002. Combining multiple data sets in a likelihood analysis: which models are the best? *Mol. Biol. Evol.*, 19(12):2294–2307.
- Ragan, M. A. 1992. Phylogenetic inference based on matrix representation of trees. *Mol. Phylogenet. Evol.*, 1:53–58.
- Rambaut, A. and N. C. Grassly. 1997. SEQ-GEN: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comp. Appl. Biosc.*, 13:235–238.
- Robinson, D. and L. Foulds. 1979. Comparison of weighted labeled trees. *Lect. Notes Math.*, 748:119–126.
- Rodriguez, R., J. L. Oliver, A. Marin, and J. R. Medina. 1990. The general stochastic model of nucleotide substitution. *J. Theo. Biol.*, 142:485–501.
- Ronquist, F. 1996. Matrix representation of trees, redundancy, and weighting. *Syst. Biol.*, 45:247–253.
- Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4:406–425.
- Sanderson, M. J. 2003. Inferring absolute rates of molecular evolution and divergence times in the absence of molecular clock. *Bioinformatics*, 19:301–302.
- Schmidt, H. A. 2003. *Phylogenetic Trees from Large Datasets*. PhD thesis, Düsseldorf, Germany.
- Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler. 2002. TREE-PUZZLE:

- Maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 18:502–504.
- Semple, C. and M. Steel. 2000. A supertree method for rooted trees. *Disc. Appl. Math.*, 105:147–158.
- Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.*, 51:492–508.
- Steel, M. A. and D. Penny. 1993. Distribution of tree comparison metrics – Some new results. *Syst. Biol.*, 42:126–141.
- Strimmer, K. and A. von Haeseler. 1996. Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.*, 13:964–969.
- Studier, J. A. and K. J. Keppler. 1988. A note on the neighbor-joining method of Saitou and Nei. *Mol. Biol. Evol.*, 5:729–731.
- Swofford, D. L. 2002. PAUP\*: Phylogenetic Analysis using Parsimony (\*and other methods), version 10, Sinauer Associates, Sunderland, Massachusetts.
- Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. In D.M. Hillis, Moritz, C. and B.K. Mable, editors, *Molecular Systematics*. Sinauer, Sunderland, Massachusetts.
- Waddell, P. J., H. Kishino and R. Ota. 2002. Very fast algorithms for evaluating the stability of ML and Bayesian phylogenetic trees from sequence data. In R. Lathrop, K. Nakai, S. Miyano, T. Takagi and M. Kanehisa, editor, *Genome Informatics*. Universal Academy Press, Tokyo.

- Waddell, P. J. and S. Shelley. 2003. Evaluating placental inter-ordinal phylogenies with novel sequences including RAG1,  $\gamma$ -fibrinogen, ND6, and mt-tRNA, plus MCMC-driven nucleotide, amino acid, and codon models. *Mol. Phylogenet. Evol.*, 28:197–224.
- Wiens, J. J. 1998. Does adding characters with missing data increase or decrease phylogenetic accuracy? *Syst. Biol.*, 47:625–640.
- Wiens, J. J. and T. W. Reeder. 1995. Combining data sets with different numbers of taxa for phylogenetic analysis. *Syst. Biol.*, 44:548–558.
- Yang, Z. 1996. Maximum-likelihood models for combined analysis of multiple sequence data. *J. Mol. Evol.*, 42:587–596.

# Appendix

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

$$f(v) = \sum_{\substack{i,j:i \neq j \\ k_{ij} \geq 2}} \sum_{p:\{ij\} \subset \tilde{\mathcal{L}}_p} w_p (\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp} - \bar{\Delta}_{ij})^2$$

where  $v = (\alpha_1, \dots, \alpha_p, \dots, \alpha_k, \dots, a_{ip}, \dots)$  and

$$\bar{\Delta}_{ij} = \frac{1}{W_{ij}} \sum_{p:\{ij\} \subset \tilde{\mathcal{L}}_p} w_p (\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp}) \quad \text{with} \quad W_{ij} = \sum_{p:\{ij\} \subset \tilde{\mathcal{L}}_p} w_p,$$

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

$$h^{(1)}(v) = \sum_p \alpha_p = k, \quad h_i^{(2)}(v) = \sum_{p:i \in \tilde{\mathcal{L}}_p} a_{ip} = 0, \forall i \in \tilde{\mathcal{L}}, \quad h_p^{(3)}(v) = \sum_{i \in \tilde{\mathcal{L}}_p} a_{ip} = 0, \forall p \neq k.$$

This is a quadratic programming problem with equality constraints. In principle, inequalities  $\alpha_p \geq 0$  should be added, but in practice we have never found negative  $\alpha_p$  values, neither with simulated sequences nor with biological datasets. The necessary first-order condition (equivalent to nullity of the first derivative in unconstrained mono-dimensional optimization) that any minimizer must satisfy is (Luenberger, 1984, p. 300):

$$S \left\{ \begin{array}{ll} \frac{\partial}{\partial \alpha_m} f(v) + \lambda \frac{\partial}{\partial \alpha_m} h^{(1)}(v) = 0, & \forall m = 1, 2, \dots, k, \\ \frac{\partial}{\partial a_{im}} f(v) + \mu_i \frac{\partial}{\partial a_{im}} h_i^{(2)}(v) + \eta_m \frac{\partial}{\partial a_{im}} h_m^{(3)}(v) = 0, & \forall m < k, \quad \forall i \in \tilde{\mathcal{L}}_m, \\ \frac{\partial}{\partial a_{ik}} f(v) + \mu_i \frac{\partial}{\partial a_{ik}} h_i^{(2)}(v) = 0, & \forall i \in \tilde{\mathcal{L}}_k, \\ h^{(1)}(v) = k, & \\ h_i^{(2)}(v) = 0, & \forall i \in \tilde{\mathcal{L}}, \\ h_p^{(3)}(v) = 0, & \forall p \neq k. \end{array} \right.$$

where  $\lambda$ ,  $\mu_i$  and  $\eta_p$  are the Lagrange multipliers induced by linear constraints  $h^{(1)}$ ,  $h_i^{(2)}$  and  $h_p^{(3)}$ , respectively.

We have:

$$\begin{aligned}
\frac{\partial}{\partial \alpha_m} f(v) &= 2 \sum_{\substack{i,j:i \neq j \\ \{ij\} \subset \tilde{\mathcal{L}}_m}} \left[ w_m \left( \Delta_{ij}^m - \frac{w_m}{W_{ij}} \Delta_{ij}^m \right) (\alpha_m \Delta_{ij}^m + a_{im} + a_{jm} - \bar{\Delta}_{ij}) \right. \\
&\quad \left. + \sum_{\substack{p:\{ij\} \subset \tilde{\mathcal{L}}_p \\ p \neq m}} w_p \left( -\frac{w_m}{W_{ij}} \Delta_{ij}^m \right) (\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp} - \bar{\Delta}_{ij}) \right] \\
&= 2w_m \sum_{\substack{i,j:i \neq j \\ \{ij\} \subset \tilde{\mathcal{L}}_m}} \Delta_{ij}^m \left[ \alpha_m \Delta_{ij}^m + a_{im} + a_{jm} - \bar{\Delta}_{ij} \right. \\
&\quad \left. - \frac{1}{W_{ij}} \sum_{p:\{ij\} \subset \tilde{\mathcal{L}}_p} w_p (\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp} - \bar{\Delta}_{ij}) \right] \\
&= 2w_m \sum_{\substack{i,j:i \neq j \\ \{ij\} \subset \tilde{\mathcal{L}}_m}} \Delta_{ij}^m (\alpha_m \Delta_{ij}^m + a_{im} + a_{jm} - \bar{\Delta}_{ij})
\end{aligned}$$

and, with similar arithmetic,

$$\frac{\partial}{\partial a_{im}} f(v) = 4w_m \sum_{j \in \tilde{\mathcal{L}}_m - \{i\}} (\alpha_m \Delta_{ij}^m + a_{im} + a_{jm} - \bar{\Delta}_{ij}).$$

Linear system  $S$  can then be written as:

$$S \left\{ \begin{array}{ll}
\sum_{\substack{i,j:i \neq j \\ \{ij\} \subset \tilde{\mathcal{L}}_p}} \Delta_{ij}^p (\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp} - \bar{\Delta}_{ij}) + \lambda = 0, & \forall p = 1, 2, \dots, k, \\
w_p \sum_{j \in \tilde{\mathcal{L}}_p - \{i\}} (\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp} - \bar{\Delta}_{ij}) + \mu_i + \eta_p = 0, & \forall p < k, \quad \forall i \in \tilde{\mathcal{L}}_p, \\
\sum_{j \in \tilde{\mathcal{L}}_k - \{i\}} (\alpha_k \Delta_{ij}^k + a_{ik} + a_{jk} - \bar{\Delta}_{ij}) + \mu_i = 0, & \forall i \in \tilde{\mathcal{L}}_k, \\
\sum_p \alpha_p = k, & \\
\sum_{p:i \in \tilde{\mathcal{L}}_p} a_{ip} = 0, & \forall i \in \tilde{\mathcal{L}}, \\
\sum_{i \in \tilde{\mathcal{L}}_p} a_{ip} = 0, & \forall p \neq k.
\end{array} \right.$$

$S$  is a square linear system, with  $\tilde{n} + 2k + \sum_p \tilde{n}_p$  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, p. 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:

$$\left. \begin{array}{l} f(v) = 0 \\ h^{(1)}(v) = 0 \\ h_i^{(2)}(v) = 0, \forall i \in \tilde{\mathcal{L}} \\ h_p^{(3)}(v) = 0, \forall p \neq k \end{array} \right\} \Rightarrow v = 0.$$

$f(v)$  is a sum of squares.  $f(v) = 0$  implies that all the squares are null, which means that for any  $i, j$  pair ( $k_{ij} \geq 2$ ) we have  $\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp} = \alpha_{p'} \Delta_{ij}^{p'} + a_{ip'} + a_{jp'}, \forall p, p' : \{i, j\} \subset \tilde{\mathcal{L}}_p, \tilde{\mathcal{L}}_{p'}$ .  $f(v) = 0$  then induces  $\sum_{i,j:i < j, k_{ij} \geq 2} (k_{ij} - 1)$  independent linear equations. Combining these equations with the  $k + \tilde{n}$  constraints, we obtain a linear system with  $k + \sum_p \tilde{n}_p$  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:

$$k + \sum_p \tilde{n}_p \leq k + \tilde{n} + \sum_{\substack{i,j:i < j \\ k_{ij} \geq 2}} (k_{ij} - 1),$$

which simplifies into:

$$\sum_{\substack{i,j:i < j \\ k_{ij} \geq 2}} (k_{ij} - 1) + \tilde{n} - \sum_p \tilde{n}_p \geq 0 \quad (7)$$

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 datasets, 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  $(\Delta_{ij})$  matrix through multiplication by  $\theta_p$  factors, each representing the evolutionary rate of gene  $p$ . We have  $(\Delta_{ij}^p) = (\theta_p \Delta_{ij})$ , just as in the proportional model of Yang (1996). Moreover, without loss of generality, assume that the  $\theta_p$ s are rescaled to obtain  $\sum_p 1/\theta_p = k$ . SDM then consistently estimates the  $\theta_p$  values, as soon as condition (7) is fulfilled. Let  $v^*$  be defined by  $\alpha_p^* = 1/\theta_p$  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  $\hat{\theta}_p = 1/\alpha_p^*$  is a consistent estimator of  $\theta_p$ , which finishes the consistency proof.

**Figure 1: Flowchart of the reconstruction scenarios.**

Starting from the data comprising a collection of  $k$  genes, the various scenarios combine several methods, as indicated by the successive arrows. Triangles represent distance matrices, and hatched areas indicate missing data (characters or distances).

**Figure 2: Accuracy of the eight reconstruction scenarios for 25% and 75% taxon deletion rates.**

$k$ : number of genes used in the reconstruction.  $d_q$ : quartet distance between the correct tree and the inferred tree. Triangles: medium-level, distance-based methods; circles and diamonds: high-level scenarios; squares: low-level scenarios. DTE does not appear in the graphics due to its poor accuracy ( $d_q > 0.06$  and  $> 0.3$  with 25% and 75% deletion rates, respectively). Note the difference between the two  $d_q$  scales in the two graphics.

**Figure 3: Phylogeny inferred by the SDM+FASTME+PHYML scenario on the 48-gene dataset of Gatesy et al. (2002).**

This dataset comprises 75 taxa and 36,639 sites. The tree log-likelihood is -330,354. The Afrotheria root the topology. Suina = Tayassuidae + Suidae. Cetartiodactyla = Camelidae + Tayassuidae + Suidae + Hippopotamidae + Cetacea + Ruminantia.

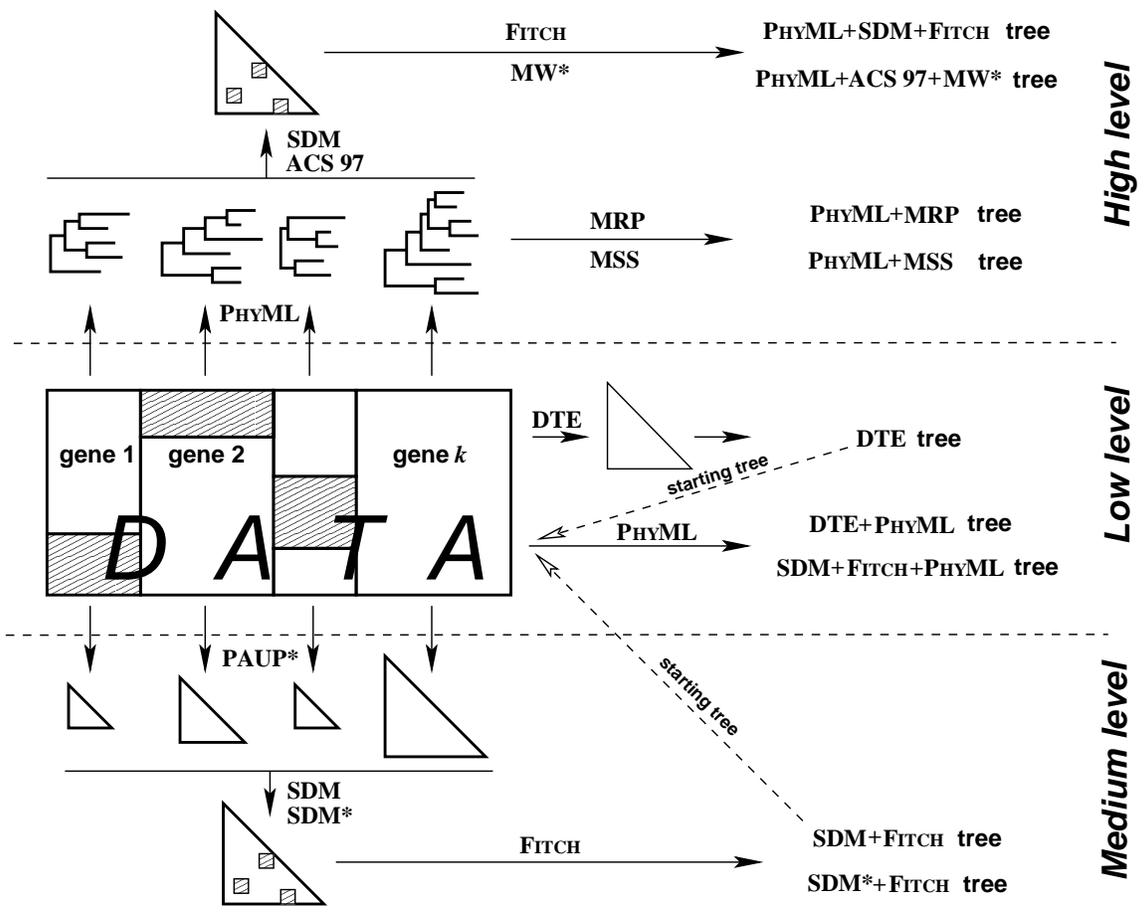


Figure 1:

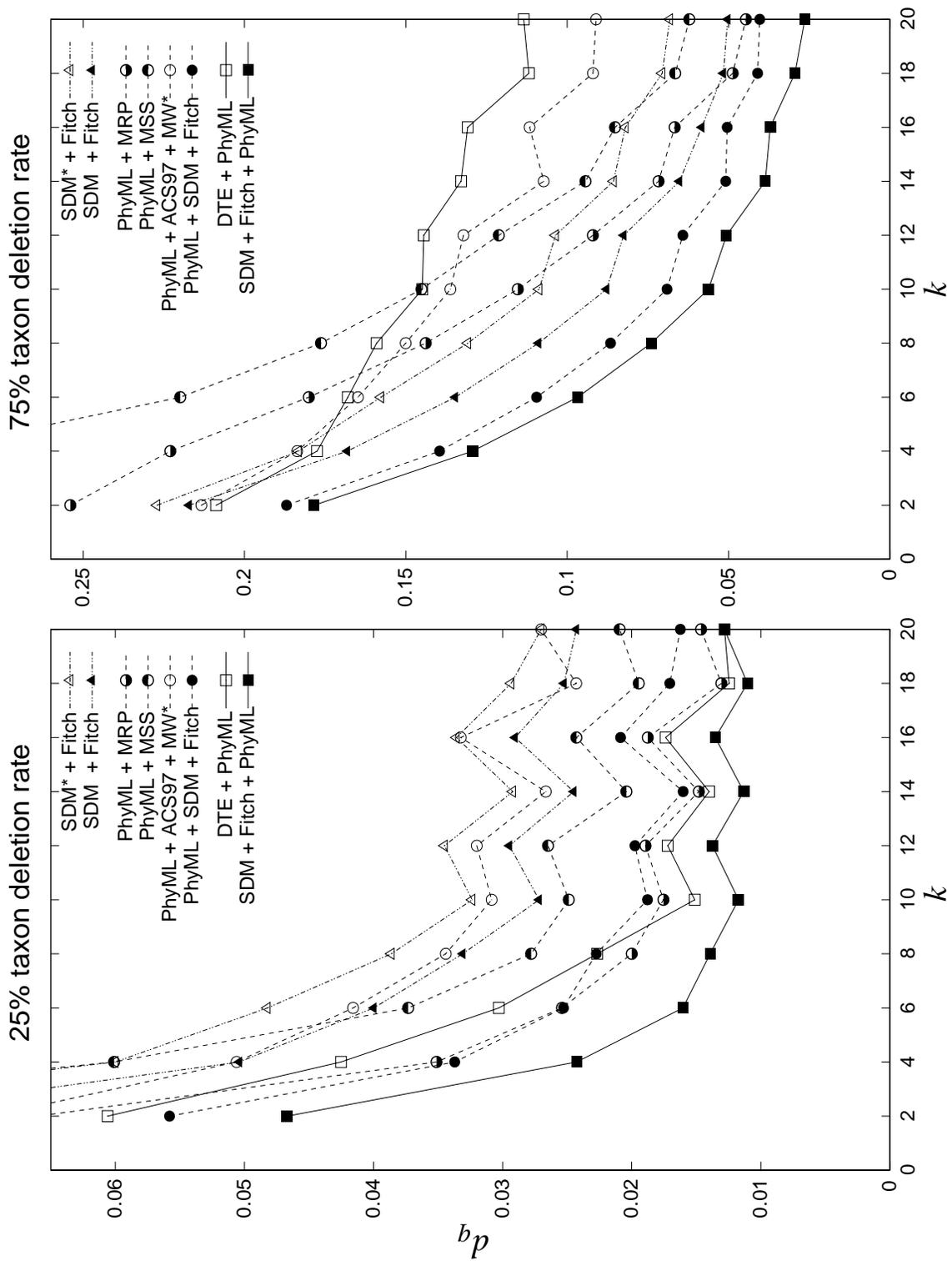


Figure 2:

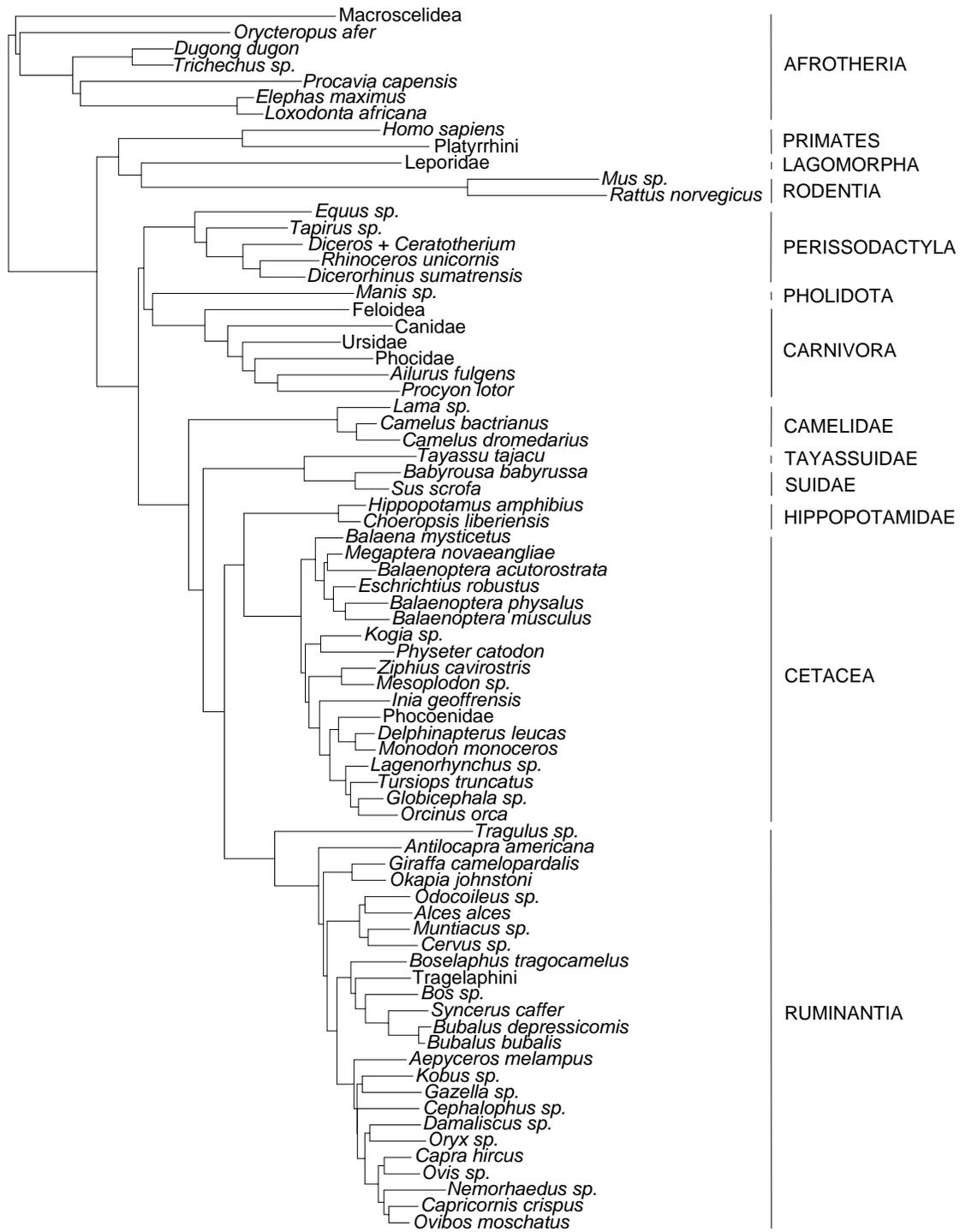


Figure 3:

**Table 1: Results of the various tree buiding scenarios with Gatesy et al. (2002) dataset.**

Gatesy-TNT stands for the MP tree that is inferred by TNT from the 48 concatenated genes;  $d_{RF}$  denotes the bipartition (Robinson and Foulds) distance between the best and the inferred tree.

**Table 2: 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.