BIOINFORMATICS

Vol. 26 ISMB 2010, pages i115–i123
doi:10.1093/bioinformatics/btq196

SUPERTRIPLETS: a triplet-based supertree approach to phylogenomics

Vincent Ranwez[1,2,*], Alexis Criscuolo[3] and Emmanuel J.P. Douzery[1,2]

[1]Université Montpellier 2, CC064, Place Eugène Bataillon, 34 095 Montpellier Cedex 05, [2]CNRS, Institut des Sciences de l'Evolution (UMR 5554), CC064, Place Eugène Bataillon, 34 095 Montpellier and [3]Institut Pasteur, Département de Microbiologie, Unité de Biologie Moléculaire du Gène chez les Extrêmophiles, 25 rue du Dr Roux, 75 015 Paris, France

ABSTRACT

Motivation: Phylogenetic tree-building methods use molecular data to represent the evolutionary history of genes and taxa. A recurrent problem is to reconcile the various phylogenies built from different genomic sequences into a single one. This task is generally conducted by a two-step approach whereby a binary representation of the initial trees is first inferred and then a maximum parsimony (MP) analysis is performed on it. This binary representation uses a decomposition of all source trees that is usually based on clades, but that can also be based on triplets or quartets. The relative performances of these representations have been discussed but are difficult to assess since both are limited to relatively small datasets.

Results: This article focuses on the triplet-based representation of source trees. We first recall how, using this representation, the parsimony analysis is related to the median tree notion. We then introduce SUPERTRIPLETS, a new algorithm that is specially designed to optimize this alternative formulation of the MP criterion. The method avoids several practical limitations of the triplet-based binary matrix representation, making it useful to deal with large datasets. When the correct resolution of every triplet appears more often than the incorrect ones in source trees, SUPERTRIPLETS warrants to reconstruct the correct phylogeny. Both simulations and a case study on mammalian phylogenomics confirm the advantages of this approach. In both cases, SUPERTRIPLETS tends to propose less resolved but more reliable supertrees than those inferred using MATRIX REPRESENTATION with PARSIMONY.

Availability: Online and JAVA standalone versions of SUPERTRIPLETS are available at http://www.supertriplets.univ-montp2.fr/

Contact: vincent.ranwez@univ-montp2.fr

1 INTRODUCTION

Representing the evolutionary history of a set of taxa is usually performed by using a phylogenetic tree. External nodes (leaves) correspond to operational units, and the root corresponds to its origin. The whole branching order from the root to each leaf provides a representation of the evolutionary relationships among the operational units studied. However, one could readily observe variations among different trees inferred for a given phylum, depending on stochastic errors due to site and taxon sampling, systematic errors induced by model misspecifications during probabilistic inferences, and biological errors due to hidden paralogy, lateral gene transfer and incomplete lineage sorting of ancestral polymorphism (see Swofford *et al*., 1996, for a discussion on the causes of incongruence). Summarizing various trees obtained from independent inferences into a single tree is therefore a recurrent issue in molecular systematics and phylogenomics. Having a single tree at hand is not just a simplification facilitating the management of topological information; it is in most cases a necessity to enable further analyses (see Jeffroy *et al*., 2006).

Pooling a collection of phylogenetic trees when all of them are defined on the same leaf set can be done with a consensus tree (Adams, 1972), such as the widely used strict or majority one (see Bryant, 2003, for an overview). Actually, the consensus tree is a particular case of the more general problem of combining trees defined on partially overlapping leaf sets. The latter case is known as the supertree problem (Gordon, 1986), whereby the aim is to build a tree (the supertree) summarizing the topological information induced by a collection of source (input) trees as well as possible. It has long been proposed that the median tree, i.e. the tree minimizing the sum of distances to the source trees, could be considered as an adequate supertree candidate (Bryant, 1997; Cotton and Wilkinson, 2007; Wilkinson *et al*., 2001). Defining the supertree as a median turns out to be not only an intuitive definition but also a guarantee of the consistency of so-defined supertree methods (Steel and Rodrigo, 2008). Even though this ensures the consistency of any median supertree, these authors stressed the importance of choosing appropriate metrics with regards to the biological setting and expected computing times. The triplet distance, based on rooted, binary, three-leaf topologies, has the advantage of being fine-grained and easy to compute in polynomial time (see Methods section for more formal definitions). We thus worked on designing a supertree method that seeks the triplet median supertree (TMS).

Among the numerous supertree techniques that have been proposed over the past 20 years, MATRIX REPRESENTATION WITH PARSIMONY (MRP; Baum, 1992; Ragan, 1992; see also Doyle, 1992) is the most widely used. In brief, MRP first encodes the whole topological signal induced by the source trees into a matrix of binary pseudocharacters (i.e. the Matrix Representation of the source trees, here noted MR), and secondly considers the optimal supertree(s) as the most parsimonious (MP) tree(s) reconstructed from the MR. If there is more than one MP tree, the standard approach is to consider the supertree as the strict consensus of these MP trees (e.g. Wilkinson *et al*., 2005a; see also Thorley, 2000 for some combinatorial properties). Although alternative consensus methods are available to merge the topological information of all optimal MP trees into a supertree (Wilkinson *et al*., 2007), considering their strict consensus is a standard way for inferring a non-ambiguous supertree from a collection of source trees. For these reasons, all MRP-based supertrees considered in this manuscript

*To whom correspondence should be addressed.

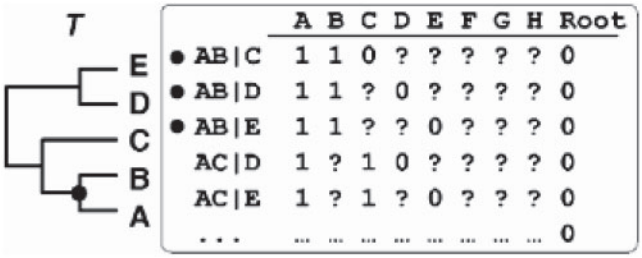

Fig. 1. Matrix representation based on triplets (t-MR). Topology T is decomposed into its set of triplets. The closer phylogenetic affinities of A and B relative to C, or D or E are indicated by the AB|C, AB|D and AB|E triplets. This correspondence is also indicated by black circles. Rooting is ensured by an artificial all-zero taxon.

are strict consensus of MP trees. To build the MR, several binary encodings have been proposed. In the standard encoding, denoted here b-MR, each pseudocharacter represents one bipartition (b) of the leaf set of the corresponding source tree (Baum, 1992; Doyle, 1992; Farris *et al*., 1970; Ragan, 1992). Alternatively, in the triplet-based encoding (Nelson and Ladiges, 1994; Wilkinson *et al*., 2004; Williams and Humphries, 2003), denoted here t-MR, each pseudocharacter represents a triplet (t) of the leaf set of the corresponding source tree (Fig. 1).

In this article, we first recall that, using t-MR, the MP tree(s) could be viewed as a kind of triplet-based median of the source trees. After describing some limits of the corresponding supertree method, i.e. triplet-based MRP (t-MRP), we then propose a new, simple and fast supertree method, named SUPERTRIPLETS, that constructs an initial tree using a polynomial algorithm close to the NEIGHBOR JOINING scheme (NJ; Saitou and Nei, 1987) followed by iterative improvement based on nearest neighbor interchange (NNI) moves. SUPERTRIPLETS is dedicated to triplets and explicitly searches for a triplet-based median supertree. Unlike the TH(SPR/TBR) triplet supertree heuristics recently proposed by Lin *et al*. (2009), SUPERTRIPLETS uses an asymmetric dissimilarity measure and is able to propose supertrees that are not fully resolved. After describing the SUPERTRIPLETS method, we compare its properties to those of b- and t-MRP, and those of another triplet-based supertree method, named TILI (Mosses, 2005), which is the triplet adaptation of the QILI supertree algorithm (Piaggio-Talice *et al*., 2004). Then, to compare these methods, we present a simulation protocol close to those used in previous supertree studies (Criscuolo *et al*., 2006; Eulenstein *et al*., 2004). There are several other triplet-based supertree building heuristics, such as MAXCUT (Moran *et al*., 2005) and TH(SPR/TBR) (Lin *et al*., 2009), but since there is currently no publicly available implementation of these approaches, we could not incorporate them in our comparative analysis. The computer simulation protocol based on larger taxon sets and larger collections of source trees enables both comparison of the relative accuracy and running time performances of SUPERTRIPLETS, t-MRP, b-MRP and TILI, and validation of their previously discussed properties. In particular, we show that SUPERTRIPLETS infers supertrees with much fewer wrong triplets (triplets absent from the initial model tree), and thus represents a pertinent alternative to the b- and t-MRP approaches. Lastly, the utility of SUPERTRIPLETS is illustrated by a phylogenomics inference on 33 mammalian taxa from ∼13 000 source trees collected in the OrthoMaM database (Ranwez *et al*., 2007).

2 METHODS

A phylogenetic tree on a set L of leaves is a rooted tree whose leaves are each bijectively labeled by an element of L (see Semple and Steel, 2003; for simplicity sake, we use in the following the term tree without ambiguity). Given a forest F (i.e. a collection of trees), $L(F)$ denotes its set of leaves; similarly $L(T)$ denotes the leaf set of a tree T. In the following, k denotes the number of tree(s) inside a forest F (i.e. $k = |F|$), and n is the total number of leaves in F (i.e. $n = |L(F)|$). A clade within a phylogenetic tree is the set of all leaves that descend from a given node. For example in Figure 1, the clades of tree T are {A,B}, {A,B,C}, {D,E}, and the trivial ones composed by each of its leaves {A}, {B}, {C}, {D}, {E} and the root-clade $L(T)$. A triplet is a three-leaf tree. Given three leaves X, Y, Z, a triplet is denoted XY|Z when its only non-trivial clade is {X,Y}. With these three leaves, the only possible triplets are then XY|Z, XZ|Y, YZ|X, and the unresolved one, denoted XYZ. Given a tree T, any subset $L' \subseteq L(T)$ induces another tree on leaf set L' denoted as $T_{|L'}$. Informally, this restriction of T is the subtree of T that connects the taxa of L'. In particular, $T_{|\{X,Y,Z\}}$ is a triplet said to be displayed by T (see Fig. 1). A tree T can be completely described by all of its non-trivial clades. It can also be described by the set $tr(T)$ of all the triplets it induces (Bryant, 1997; Grunewald *et al*., 2007). The triplet distance $d_3(T_1,T_2)$ between two trees T_1 and T_2 having the same leaf set is the cardinality of the symmetric difference between $tr(T_1)$ and $tr(T_2)$ (Critchlow *et al*., 1996; Dobson, 1975). This definition is similar to that of the bipartition distance (Bourque, 1978; Robinson and Foulds, 1981), except that the respective sets of triplets of T_1 and T_2 are considered instead of the sets of leaf bipartitions induced by each of their internal branch. When two trees have the same leaf set, the triplet distance is thus the number of resolved triplets present in only one of these two trees. It directly follows from its definition that d_3 is a dissimilarity, but it is also a distance since it satisfies the triangle inequality (Bansal *et al*., 2008).

2.1 Triplet asymmetric median supertree

Several distances have been proposed to compare two trees with the same leaf set (e.g. Bourque, 1978; Critchlow *et al*., 1996; Dobson, 1975; Robinson and Foulds, 1981; see also Steel and Penny, 1993). These distances cannot be directly used to compare a supertree T with a source tree T_i since both do not necessarily have the same leaf set. The similarity between T_i and T is thus generally evaluated through the distance between T_i and the supertree T restricted to $L(T_i)$ (e.g. Cotton and Wilkinson, 2007; Creevey and McInerney, 2005). The resulting measure between T and T_i does not satisfy distance axioms (e.g. trees on different leaf sets can have a distance of 0). Yet in what follows we will, for simplicity sake, refer to this measure as the distance between the supertree T and a source tree T_i. Since a supertree is supposed to summarize a forest, a natural candidate is the median tree minimizing the sum of distances to the source trees (Bryant, 1997; Wilkinson *et al*., 2001; Creevey and McInerney, 2005). Several authors have underlined the equivalence between the majority rule consensus and the median tree based on bipartition metrics, and Cotton and Wilkinson (2007) generalize this approach to the supertree setting. Defining supertree as a median turns out to be not only an intuitive definition but also a guarantee of fundamental theoretical properties. Using a model of phylogenetic error, Steel and Rodrigo (2008) consider supertree in a maximum likelihood (ML) framework, with the so-defined ML-supertree being simply a median tree (with respect to tree metrics). This ML framework allows the authors to demonstrate the statistical consistency of the median supertree approach (i.e. converging to the true tree as more source trees are added). Even though this result ensures the consistency of any median supertree definition, the authors stress the importance of choosing appropriate metrics. The triplet distance has the advantage of being easily computed (unlike the SPR distance; Bordewich and Semple 2004; Hickey *et al*., 2008; but see Lin *et al*., 2009) and more fine-grained than the bipartition distance (e.g. Steel and Penny, 1993). As stated by Williams (2004), it seems that 'three item data are more sensitive to how information accumulates to produce overall summary solutions'.

Considering a source tree T_i and the supertree T, the set of triplets over $L(T)$ can be partitioned into five sets (Page, 2002): $S(T_i,\ T)$ and $D(T_i,\ T)$, the triplets resolved in T_i and $T_{|L(Ti)}$ that have same and different topologies, respectively; $R_1(T_i,\ T)$, the triplets resolved in T_i and not resolved in $T_{|L(Ti)}$; $R_2(T_i,\ T)$, the triplets not resolved in T_i and resolved in $T_{|L(Ti)}$; $X(T_i,\ T)$, the triplets unresolved in both T_i and $T_{|L(Ti)}$. Using this notation, a triplet distance between a source tree and a supertree can be formulated as $d_t(T_i,\ T) = 2|D(T_i,\ T)| + |R_1(T_i,\ T)| + |R_2(T_i,\ T)|$. Minimizing this distance is closely related to maximizing the similarity $|S(T_i,\ T)|$ as done by Lin *et al*. (2009). Our distance formulation has the advantage of describing the way unresolved triplets are handled. Indeed, this distance is increased by the presence of triplet(s) resolved in the supertree T but not in some source trees T_i, i.e. the set $R_2(T_i,\ T)$. Supertrees resolving triplets that are not resolved in some source trees are thus penalized. The use of the (symmetric) distance d_t results in considering irresolution in source trees as a signal to take into account (hard polytomies; Maddison, 1989) rather than as an absence of signal (soft polytomies; Bansal *et al*., 2008). Since polytomies in source trees are generally considered as soft ones, it seems more suitable to consider the asymmetric triplet distance: $\delta_t(T_i,\ T) = 2|D(T_i,\ T)| + |R_1(T_i,\ T)|$. Note that δ_t is no longer a distance between T_i and $T_{|L(Ti)}$ (due to its asymmetry, δ_t is quite a divergence). The corresponding median supertree definition is therefore no longer guaranteed to be consistent overall. However, the consistency still holds in the special case where source trees are all fully resolved, since in this case there is no difference between d_t and δ_t. Using the asymmetric function δ_t weakens the theoretical properties of the corresponding supertree approach but seems to be more in line with the biological acceptation of polytomies. Moreover, the advantages of an asymmetric approach have already been pointed out and exploited in the consensus setting (Phillips and Warnow, 1996).

Given a forest $F = \{T_1,\ T_2 \ldots T_k\}$, the two closely related computational problems of finding a TMS (i.e. according to d_t) and finding a triplet asymmetric median supertree (TAMS, i.e. according to δ_t) can be stated as finding a supertree T with $L(T) = \cup_i L(T_i)$ that minimizes $d_t(F,\ T) = \sum_i d_t(T_i,\ T)$ and $\delta_t(F,\ T) = \sum_i \delta_t(T_i,\ T)$, respectively. As conjectured by Bansal *et al*. (2008), the TMS problem is NP-hard due to the NP-completeness of the maximum triplet compatibility problem (Semple and Steel, 2003). Since both TMS and TAMS problems are equivalent when source trees are fully resolved, we conjecture that the TAMS problem is also NP-hard (nothing indicates that instances of the TMS problem where source trees are fully resolved are easier than others). The TAMS optimization problem can be turned into a classical MP optimization problem by using a t-MR of the source trees (Wilkinson *et al*., 2001, 2005a, 2007). As unresolved triplets of sources trees are not encoded in the matrix, the MP optimization does, indeed, tackle the TAMS problem and not the TMS problem. This approach benefits from long-term efforts to optimize MP local search programs (e.g. Goloboff *et al*., 2008) but is hampered by some limitations due mostly to its inefficient representation of triplet information. We therefore designed SuperTriplets, a heuristic dedicated to the TAMS optimization problem.

2.2 Agglomerative construction of a TAMS

The first step of SuperTriplets aims at converting in $O(kn^3)$ the (rooted) source trees into weighted triplets, where the weight of each triplet is the number of source trees containing it. Then the agglomerative scheme is used to construct an initial supertree. For our triplet-based strategy, this algorithmic scheme is well adapted, since when an agglomeration is performed, all resolved triplets that this agglomeration generates in the final supertree are completely determined (see Fig. 2).

Considering the agglomerative scheme as an iterative improvement of the current supertree T, the aim is to select, during each iteration, the best agglomeration with respect to the TAMS criterion to minimize. When modifying a tree T into T' by proposing a new clade resulting from the joining of two existing clades C_A and C_B, SuperTriplets creates new triplets AB|X for any set of three taxa {A,B,X} such that $A \in C_A$, $B \in C_B$ and $X \notin C_A \cup C_B$ (Fig. 2). The difference between $\delta_t(F,\ T)$ and $\delta_t(F, T')$ depends

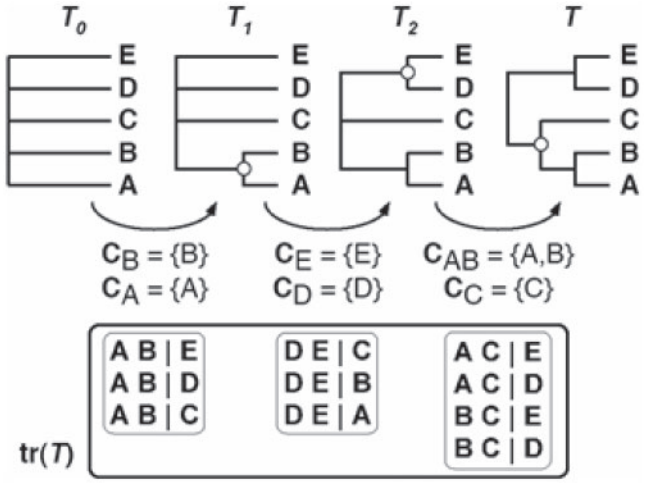

Fig. 2. Detail of the agglomeration process. Each taxon agglomeration step is denoted by an open circle, and creates some new triplets. Their union is $tr(T)$, the set of triplets of the final binary tree.

only on such newly resolved triplets. Consequently, after agglomerating C_A and C_B, the criterion value $\delta_t(F,\ T')$ can be expressed from $\delta_t(F,\ T)$ as $\delta_t(F,\ T') = \delta_t(F,\ T) - N^+(C_A,\ C_B) + N^-(C_A,\ C_B)$, where $N^+(C_A,\ C_B)$ is the number of times the new triplets AB|X appear in F, and $N^-(C_A,\ C_B)$ the number of times an alternative resolution (i.e. AX|B or BX|A) appears in F. Obviously, $\delta_t(F,\ T') < \delta_t(F,\ T)$ if $N^+(C_A,C_B) > N^-(C_A,\ C_B)$. A natural approach is then to agglomerate the two clades such that $N^+(C_A,\ C_B) - N^-(C_A,\ C_B)$ is maximal. Note, however, that all agglomerations do not resolve the same number of triplets, and that the importance of a given triplet is related to the number of times it appears in F, i.e. its weight. During this agglomeration process, SuperTriplets thus does not fully take triplet weights into account, but only uses weights to determine, for each group of three taxa, the most frequent resolution in source trees (i.e. the triplet with the highest weight). If a triplet AB|X is more frequent in source trees than its alternative resolutions (i.e. AX|B and BX|A), it is called a most frequent triplet, denoted *mft*. At each step, SuperTriplets selects the agglomeration leading to the highest proportion of *mft*. This criterion ensures that SuperTriplets will recover the correct phylogeny for every dataset such that, for any three-taxon set {A,B,C}, the *mft* is the correct one. Indeed, in such cases, every creation of a correct clade leads to 100% of *mft*, while this proportion is smaller for incorrect clades. Moreover, the tree obtained during this agglomeration step will contain every *mft* and is therefore the one minimizing the TAMS criterion. More formally, the selected agglomeration is randomly chosen among those maximizing $N^{mft+}(C_A,\ C_B)/[N^{mft+}(C_A,\ C_B) + N^{mft-}(C_A,\ C_B)]$, where $N^{mft+}(C_A,\ C_B)$ is the number of newly resolved triplets that are *mft*, and $N^{mft-}(C_A,\ C_B)$ the number that conflict with *mft*. It follows that to evaluate the agglomeration of any pair C_A and C_B, SuperTriplets only needs to have at hand the corresponding values for $N^{mft+}(C_A,\ C_B)$ and $N^{mft-}(C_A,\ C_B)$.

During the first iteration of the agglomerative process, clades are reduced to a single taxon, so that for any of the $O(n^2)$ pairs of such trivial connected component C_A and C_B, the values $N^{mft+}(C_A,\ C_B)$ and $N^{mft-}(C_A,\ C_B)$ correspond to the creation of new triplets AB|X, where A and B are fixed taxa and where X can take $n-2$ label values. All the $N^{mft+}(C_A,\ C_B)$ and $N^{mft-}(C_A,\ C_B)$ values are therefore initialized in $O(n^3)$. When two clades C_{A1} and C_{A2} are merged in a new clade C_A, this does not change the value of $N^{mft+}(C_X,\ C_B)$ as long as both C_X and C_B differ from C_{A1}, C_{A2} and C_A. Moreover, the new value $N^{mft+}(C_A,\ C_B)$ can easily be obtained from $N^{mft+}(C_{A1},\ C_B)$ and $N^{mft-}(C_{A2},\ C_B)$. Indeed, any triplet AB|X that will be newly resolved by an agglomeration between C_A and C_B is a triplet that would have been newly resolved by agglomerating C_{A1} and C_B, or C_{A2} and

C_B. This is simply due to the fact that if A is inside C_A, it was either in C_{A1} or in C_{A2}. On the other hand, the triplets $A_1B|X$ that would have been newly resolved by the agglomeration of C_{A1} and C_B will also be newly resolved by the agglomeration of C_A and C_B, except if X is within C_{A2}. It follows that $N^{mft+}(C_A, C_B)$ is simply the sum of $N^{mft+}(C_{A1}, C_B)$ and $N^{mft+}(C_{A2}, C_B)$ minus a correction. This correction is equal to twice the number of *mft* $A_1A_2|B$ (with $A_1 \in C_{A1}, A_2 \in C_{A2}$ and $B \in C_B$) since these triplets are counted in both $N^{mft+}(C_{A1}, C_B)$ and $N^{mft+}(C_{A2}, C_B)$, and should not be counted in $N^{mft+}(C_A, C_B)$. The same remark is easily transposed to $N^{mft-}(C_A,C_B)$.

Once the initial supertree is obtained thanks to the above agglomerative heuristics, SUPERTRIPLETS checks whether any of its neighbor trees is closer to the set of source trees F. If so, this particular neighbor is used as the new current supertree to recheck its tree neighbors. This is repeated until a supertree is reached such that none of its neighbor trees is better fitted than itself. The neighborhood used is that defined by the NNI, where a tree T' is a tree neighbor of T if T' can be obtained from T by swapping two subtrees of T, both of which are rooted on the two extremities of a given edge e. Each NNI is therefore related to a specific internal branch. This heuristics is often used in fast phylogenetic reconstruction algorithms (e.g. PHYML; Guindon and Gascuel, 2003). Given an NNI related to an edge e of T, the resolution of a triplet AB|X will be changed after the NNI if and only if e is the sole edge separating a clade containing A and B but not X from the first clade containing the three taxa. To evaluate the whole set of possible NNIs, SUPERTRIPLETS therefore twice considers the occurrences of such triplets as well as the occurrence of their alternative resolutions, so this can be done in $O(n^3)$. Note that this is already faster than consulting all triplets since most of them are not considered at all during this process (no current NNI can change their resolution). Furthermore, for each NNI around an edge e, one can store the value NNI(e) that represents the difference between the actual criterion and the best value for the two trees that differs from T by an NNI around e. Storing those values makes NNI selection very fast, since once an NNI is actually done around an edge e, the value NNI(e') only changes for a handful of edges e' close to e. Since NNI exploration is really fast, SUPERTRIPLETS uses this advantage to better explore the topology space using random NNI modifications that are allowed to temporarily decrease the quality of the current tree.

2.3 Reliability estimation of supertree edges

Collapsing an edge e transforms the supertree T into a less resolved supertree T'. Indeed, a set of resolved triplets AB|X is dependent upon the presence of the edge e and will become unresolved inside T'. The difference between $\delta_t(F, T)$ and $\delta_t(F, T')$ depends only upon such newly unresolved triplets. Following the notation used to describe the agglomerative process, we have $\delta_t(F, T') = \delta_t(F, T) - N^-(e) + N^+(e)$, where $N^+(e)$ represents the number of times the triplets AB|X induced by e appear in F, and $N^-(e)$ the number of appearances of their alternative resolutions (i.e. AX|B or BX|A). A measure of the reliability of an edge e in a tree T can thus be obtained by considering the modification of T by collapsing e. SUPERTRIPLETS estimates the reliability of e as $N^+(e)/[N^+(e)+N^-(e)]$. This criterion value ranges from 0 to 1, and represents the percentage of triplets that supports the edge e. Note that this measure only takes a small subset of the triplets included in the current tree into account. Indeed, some triplets require the collapse of several edges of T to become unresolved. Using this measure, a supertree can contradict some input triplets while having all its edges with maximal support. This can be avoided by considering a global measure involving all triplets. Each edge defines a clade that can be incompatible with some triplets present in the source trees. A global measure can thus be obtained by considering triplets of source trees that disagree (or agree) with this clade. From this standpoint, a triplet AB|X of the source trees will influence the support of every branch of the supertree that defines a clade containing A and B but not X (i.e. those that should be collapsed to obtain an irresolution on A, B and X). The number of such clades varies among triplets. In order to obtain a measure such that all triplets have the same importance, the contribution of each triplet is equally divided among these clades, as suggested by Wilkinson *et al*. (2005b). Hence, all triplets have the same importance, but while some of them focus on a single edge, others influence several edge supports. This provides a global measure of agreement, i.e. $N^{g+}(e)$, and contradiction, i.e. $N^{g-}(e)$, between source trees and a given edge e. These measures are then used to define a global support of edge e defined by $N^{g+}(e)/[N^{g+}(e)+N^{g-}(e)]$. Though each triplet influences all the edges within a path in the supertree and not a single one, the support of each edge can be estimated in $O(n^3)$. SUPERTRIPLETS uses these measures to detect edges that must be collapsed. The modification of the supertree, which involves collapsing edges with local support smaller than 0.5, always reduces the value of $\delta_t(F, T)$. Since these supports are interdependent, SUPERTRIPLETS collapses all such edges simultaneously to avoid arbitrary choice between edges. Then all remaining edges with global support <0.5 are also simultaneously collapsed. Note that these edges may have a local support >0.5, for instance if the taxon sampling in source trees implies that this local support is estimated based on very little information. Collapsing a single one of these branches will not reduce the divergence $\delta_t(F, T)$ between the source trees and the supertree, but collapsing several of them possibly will. Finally, the local supports of the remaining edges are evaluated and displayed in the final supertree as a reliability indication. Using this conservative procedure, the supertree returned by SUPERTRIPLETS is more reliable by displaying strongly supported clades, but is also less resolved as the topological information contained in the source trees is ambiguous.

3 RESULTS

3.1 Simulation Results

Simulations were conducted to compare the behavior of SUPERTRIPLETS to that of the well-known b-MRP and the alternative t-MRP methods when analyzing collections of source trees of varying size and taxon overlap. We also compare these three methods with the more recently implemented TILI algorithm (Mosses, 2005), which is the triplet variant of the quartet-based supertree algorithm QILI (Piaggio-Talice *et al*., 2004).

An ultrametric model topology T^{model} with $n = 100$ ingroup taxa was generated by the *r8s* software (Sanderson, 2003) according to a Yule-Harding branching process (Harding, 1971; Yule, 1925). One outgroup taxon was *a posteriori* added, with a root-to-tip branch length of 1.1 substitution per site. To highlight the substitution rate heterogeneity among taxa and genes, T^{model} was duplicated 50 times, and a global deformation of these identical clocklike trees was achieved by individual branch length deviation followed by 50 independent total tree length rescaling according to Criscuolo *et al*. (2006). We thus obtained 50 rooted 101-taxon trees with the same topology as T^{model} but with different branch lengths and different induced global evolutionary rates. Each of these 50 non-clocklike phylogenetic trees was used by Seq-Gen (Rambaut and Grassly, 1997) to generate 50 nucleotide alignments of 101 taxa under a Kimura (1980) model with transition/transversion ratio of 2.0. The number of aligned sites was uniformly drawn from 200 to 1000 bp. For each alignment, ingroup taxa were randomly deleted with a probability $d = 25$, 50 and 75% following the procedure first suggested by Eulenstein *et al*. (2004).

For each of the three taxon deletion rates ($d = 25$, 50, 75%), 50 ML trees were then inferred from the 50 partially deleted alignments with PHYML (Guindon and Gascuel, 2003) under the Kimura (1980) model with the transition/transversion ratio left as free parameter. The whole procedure of topology generation, phylogram deformation, sequence generation, taxon deletion with three conditions and phylogeny reconstruction by PHYML was repeated 100 times. Each tree returned by PHYML was then rooted

by the outgroup taxon. One thus obtains 100 sets of 50 trees for each of the three values of d.

For each taxon deletion rate d and each of the 100 corresponding sets of trees, the collections of the first $k = 10$, 20, 30, 40, 50 source trees were submitted to SUPERTRIPLETS and TILI. The same forests were analyzed with both b- and t-MRP. Hence, the b-MR and the weighted site t-MR were first built and, secondly, the parsimony analyses were performed by the TNT software (Goloboff *et al*., 2008) using 10 random addition sequences and TBR branch swapping. The ratchet option was used for b-MRP, but not for t-MRP since this approach uses t-MRs of extremely large size (e.g. currently more than 500 000 pseudocharacters with $d = 25\%$ and $k = 50$, despite the site weighting procedure during the t-MR computing step) which require huge running times to be analyzed (e.g. at most 20 h, whereas several days are necessary for one supertree inference with the ratchet option).

To evaluate the degree to which the inferred supertrees agree with their corresponding model tree T^{model}, we used bipartition- and triplet-based distances between trees. Given the five previously described triplet partitions S, D, R_1, R_2, X (see Section 2.1), and the five corresponding triplet rates s, d, r_1, r_2, x, respectively, one can define two error types. The proportion of triplets that are in the supertree but not in the model tree is called the type I error and corresponds to erroneous information displayed in the supertree (sometimes called false positive). Reciprocally, the proportion of triplets that are not in the supertree but in the model tree is called the type II error, and corresponds to topological information that is missing in the supertree (sometimes called false negative). The type I error rate between T^{model} and its related supertree T is then defined as $et_I = (d+r_2)/(d+s+r_2)$; reciprocally, the type II error rate is defined as $et_{II} = (d+r_1)/(d+s+r_1)$. It should be stressed that the type II error rate is closely related to the triplet-fit similarity from Page (2002), which is defined as 1-et_{II}.

Since b-MRP optimizes a bipartition-based criterion, not a triplet-based criterion, both type I and II errors were also computed at a bipartition level. The bipartition-based error rate et_I (et_{II}, respectively) is defined as the proportion of bipartitions of $L(T)$ induced by each of the internal branches of $T^{\text{model}}_{|L(T)}$ (of T, respectively) that are not present in T (in $T^{\text{model}}_{|L(T)}$, respectively).

A low type I error rate means that the supertree T displays few leaf bipartitions (or triplets) that are not present in T^{model}. Respectively, a high type II error means that the supertree T does not represent all the phylogenetic information in T^{model}. For each of the four supertree methods studied (i.e. b- and t-MRP, SUPERTRIPLETS and TILI), each taxon deletion rate (i.e. $d = 25$, 50, 75%) and each forest size (i.e. $k = 10$–50), triplet- and bipartition-based et_I and et_{II} values were computed and averaged over the 100 replicates. These results are graphically represented in Figure 3 for the triplet-based errors and in Figure 4 for the bipartition-based errors.

The simulation results can be alternatively represented by focusing on the *sensitivity* (i.e. the true positive rate) and *specificity* (i.e. the complementary of the false positive rate) of the methods. This representation, unlike the previous one, takes the size of the triplet (or bipartition) space into account. Using notations introduced in Section 2.1, the number of false positive triplets is $\text{FP} = |D| + |R_2|$, the number of false negative triplets is $\text{FN} = |D| + |R_1|$, and the number of true positive triplets is $\text{TP} = |S|$. Finally, since the total number of resolved triplets on n taxa is $n(n-1)(n-2)/2$, the number of true negative triplets is $\text{TN} = n(n-1)(n-2)/2 - (\text{TP} + \text{FP} + \text{FN})$. Using these notations, the true positive and false negative rates are defined as TP/(TP + FN) and FP/(FP + TN), respectively. Figure 5 depicts the plot of these two rates (i.e. the so-called ROC graph; e.g. Fawcett, 2004) for the various methods and simulation conditions.

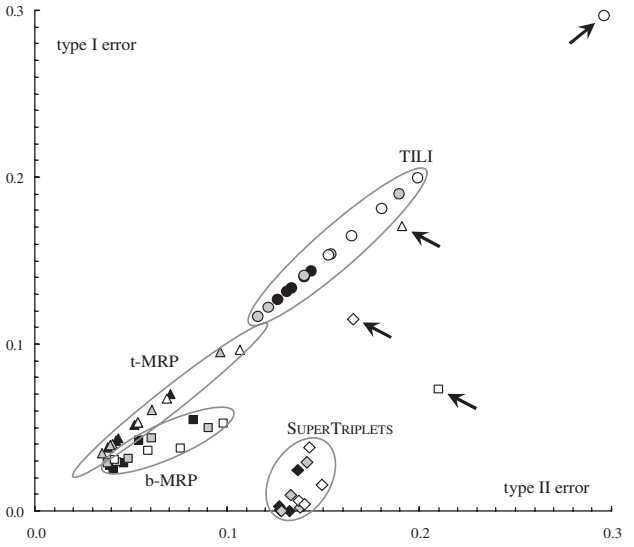

Fig. 3. Type I error rate of supertree methods as a function of type II error rate at the triplet level. Supertrees were inferred from forests by four methods under three different simulation conditions. Points represent the average over 100 replicates of the supertree error rates. Symbols correspond to the methods: b-MRP (open squares), t-MRP (open triangles), SUPERTRIPLETS (open diamonds) and TILI (open circles). For clarity, symbols corresponding to each of the four methods are encircled. Colors correspond to the three taxon deletion rates: $d = 25\%$ (black), 50% (grey) and 75% (white). Arrows point to the difficult case $d = 75\%$ and $k = 10$ (see Section 3.1 for more details).

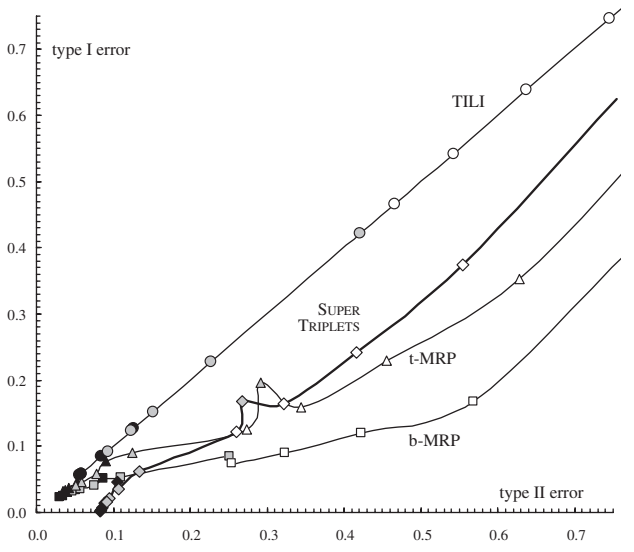

Fig. 4. Type I error rate as a function of type II error rate at the bipartition level. Symbols and colors have the same meaning as in Figure 3. For clarity, symbols that correspond to the same method are connected by a line. The difficult case $d = 75\%$ and $k = 10$ is not represented.

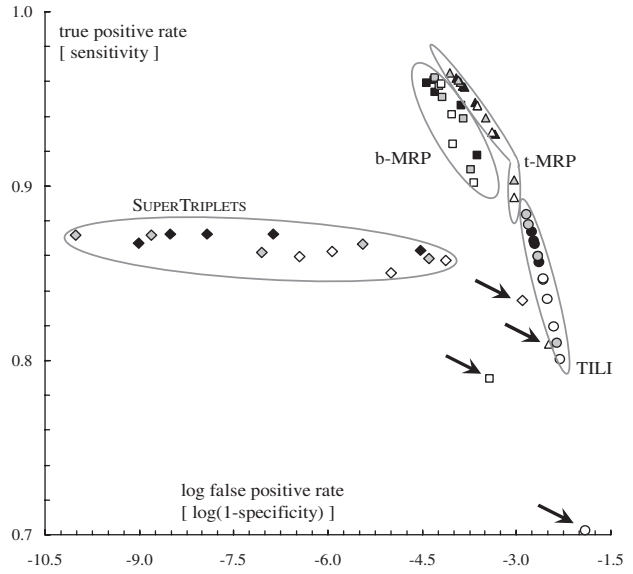

Fig. 5. ROC graph displaying true positive rates as a function of false positive rates at the triplet level. Symbols, colors and arrows have the same meaning as in Figure 3. Note that the *x*-axis is on a logarithmic scale.

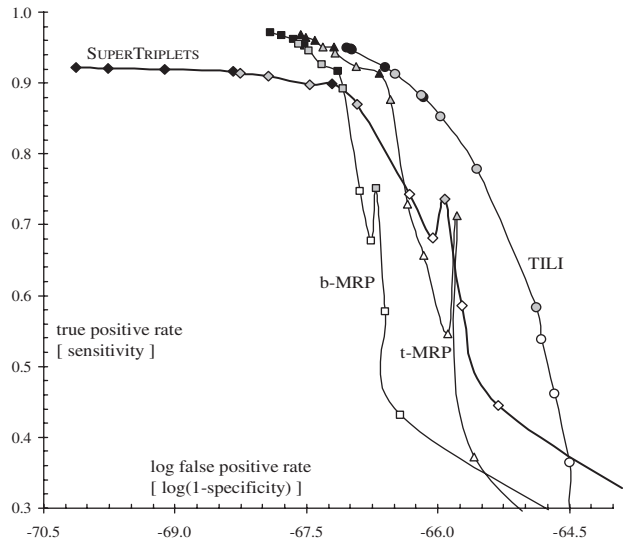

Fig. 6. ROC graph displaying true positive rates as a function of false positive rates at the bipartition level. Symbols that correspond to the same method are connected by a line. Symbols and colors have the same meaning as in Figure 4. Note that the *x*-axis is on a logarithmic scale.

Based on the fact that the total number of non-trivial bipartition on n taxa is $2^{n-1}-n-1$, a similar approach can be used to obtain the bipartition-based ROC graph displayed in Figure 6. It should be stressed that in ROC graphs (Figs 5 and 6) the best results are displayed close to top left corner (i.e. the method is both sensitive and specific; Fawcett, 2005) whereas in the type I/type II plot (Figs 3 and 4) the best results are close to the bottom left corner (i.e. reduced type I and II errors.)

SUPERTRIPLETS displays the lowest triplet-based type I errors (Fig. 3) and false positive rates (Fig. 5) in all cases but the very difficult one ($k = 10$ and $d = 75\%$, i.e. few source trees with low taxon overlap). Indeed, its triplet-based type I errors and false positive rates are often close to 0, and are significantly lower than those obtained for the three other methods (as assessed by a sign test; e.g. Dixon and Mood, 1946). This shows that a supertree inferred by SUPERTRIPLETS displays very few triplets that are not present in the corresponding model tree (i.e. so-called false positive triplets). In addition, SUPERTRIPLETS bipartition-based type I errors (Fig. 4) and false positive rates (Fig. 6) are among the lowest. However, it should be stressed that b-MRP supertrees show better minimization of bipartition-based false positive rates for difficult cases, i.e. when the taxon overlap among source trees is minor (e.g. $d = 50\%$ and $k \leq 20$, or $d = 75\%$). This latter result is still expected since b-MRP optimizes a bipartition-based encoding of the source topologies. Conversely, SUPERTRIPLETS displays higher triplet-based type II errors and lower true positive rates than both b- and t-MRP (Figs 3 and 5) whatever the number of source trees (except for $k = 10$ and $d = 75\%$). Note also that, although based on triplet decomposition, TILI has the worst type I and II errors even when considering triplet error rates (Fig. 3). For this reason, we will not further discuss results obtained with the TILI supertree method.

Overall, SUPERTRIPLETS has a much lower type I error whereas b- and t-MRP have a much lower type II error. Our simulations therefore suggest that SUPERTRIPLETS infers less-resolved but more reliable conservative supertrees, whereas b- and t-MRP infers more resolved but less reliable liberal ones. It is not clear, however, whether or not both error types have the same importance. Suppose that starting from a given supertree T, it is possible to modify it into T' to resolve 10 additional triplets—everything apart from these 10 triplets would be the same in T and T'. Among these 10 triplets, suppose that four of them are wrong whereas the six others are correct. It is thus hard to determine which of these two supertrees is the most informative summary of the source trees. This can be formulated as '*how many true triplets do we require to accept the introduction of a wrong one?*'. The answers certainly vary depending upon the end-users and the analyses they plan to conduct with the supertree. As described previously, b-MRP, t-MRP and SUPERTRIPLETS answer this question in different ways.

3.2 Triplet-based supertree in a phylogenomics context

To stress the benefits of SUPERTRIPLETS for inferring a species tree, we collected the 12 958 trees available in the fifth release of the OrthoMaM database (Ranwez *et al.*, 2007). These ML trees have been inferred from alignments of orthologous gene coding sequences, as identified by EnsEMBL v. 54. The alignment sizes range from 0.1 to 42 kb, and the number of taxa from 6 to 33 mammals. A supertree was obtained by using SUPERTRIPLETS on these 12 958 equally-weighted trees after having rooted them via two outgroup taxa, i.e. one marsupial (*Monodelphis*) and one monotreme (*Ornithorhynchus*). The running time was very fast (30 s on a MacPro computer with 2.66 GHz dual-core Intel Xeon processor).

The topology inferred (see Fig. 7) is in line with current knowledge about mammalian phylogenetics (e.g. Prasad *et al.*, 2008). SUPERTRIPLETS branch support values range from high to moderate for defining several supra-ordinal clades: Cetartiodactyla (0.92), Afrotheria (0.87), Paenungulata (0.82), Euarchontoglires

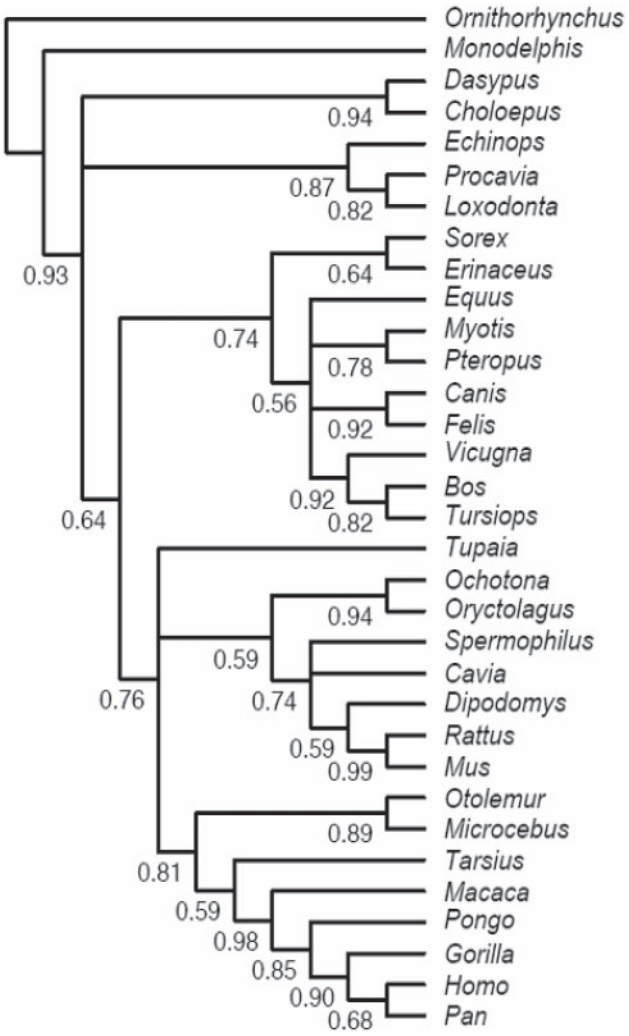

Fig. 7. Phylogenomic case study: reconstruction by SUPERTRIPLETS of the supertree of 33 mammals for which complete genome data are available. The SUPERTRIPLETS supertree has been inferred from the 12 958 source trees of OrthoMaM v. 5, i.e. ML trees estimated from DNA orthologous gene coding sequences (CDS). SUPERTRIPLETS node supports are given.

(0.76), Laurasiatheria (0.74), Boreoeutheria (0.64), Glires (0.59) and Scrotifera (0.56). This branch support hierarchy (showing the level of congruence among source trees) is compatible with those reported by Beck *et al*. (2006). Several multifurcations in the supertree correspond to areas of documented phylogenetic uncertainty: placental tree root (Afrotheria + Xenarthra, versus either Afrotheria or Xenarthra first; Churakov *et al*., 2009), relationships among Laurasiatheria (Prasad *et al*., 2008), *Tupaia* affinities (Janecka *et al*., 2007), and relationships among the three major Rodentia clades (Blanga-Kanfi *et al*., 2009). Rodents and primates are monophyletic (respective supports: 0.74 and 0.81), and the following within-order clades are supported: *Dipodomys* + *Rattus* + *Mus* (0.59), Strepsirrhini (0.89) and Haplorrhini (0.59), Catarrhini (0.98) and Hominoidea (0.85).

Given the huge number of independent multiple sequence alignments that can be harvested from comparative genomics projects, there is a need for efficient tools to quickly summarize the whole phylogenetic signal of the numerous trees inferred from these various loci. In such phylogenomic cases, when using a supermatrix approach (obtained by concatenating all alignments), very large matrices must be dealt with that can exceed the memory capacity of most today computers. There would also be many missing data, which could interfere with the optimization process (Criscuolo *et al*., 2006). Supertree approaches offer a more tractable alternative in this setting. SUPERTRIPLETS is especially well tailored to applications under these conditions since the number of input trees has little impact on the computer resources it uses. Here, for instance, assembling the 12 958 CDSs into a supermatrix would have resulted in a huge matrix made of ∼27-million of sites containing 28% missing character states (some taxa with low-coverage genomes such as *Myotis* or *Otolemur* are not represented for many CDSs). This matrix is so huge that its phylogenetic analysis would be problematic due to memory limitations.

Supertree methods are also used for summarizing previously published phylogenies into a single one that may contain hundreds to thousands of taxa (e.g. Beck *et al*., 2006; Bininda-Emonds *et al*., 2007). We thus compared the computation time of b-MRP, t-MRP and SUPERTRIPLETS on the OrthoMaM CDS dataset, and on the mammalian family-level dataset of Beck *et al*. (2006) that contains a total of 115 taxa and 725 weighted source trees (i.e. 6724 trees to account for the differential weighting of individual source topologies). The computation times were measured on a 3-GHz Intel® Core™2 Duo PC with 1.9 Gb RAM. For t-MRP, the compressed t-MR was obtained in few minutes using the homemade MRtools software (available on the SUPERTRIPLETS webpage). Since no substantial efforts were made to optimize this step, we did not include it when measuring the overall computation time of b-MRP and t-MRP methods. We thus slightly underestimate their running times, especially for t-MRP, since the building of the t-MR is an $O(kn^3)$ step. For the three methods, we used the same parameters as those used for assessing their accuracy (e.g. ratchet for b-MRP but not for t-MRP, see Section 3.1). On the OrthoMaM dataset, b-MRP runs in ∼18 min while t-MRP and SUPERTRIPLETS run in a couple of seconds. On the mammalian family dataset, b-MRP runs in ∼21 min, t-MRP in ∼7 min, and SUPERTRIPLETS in ∼10 s. Moreover, the supertrees obtained by b-MRP and SUPERTRIPLETS are in agreement with current knowledge on the mammalian phylogeny. SUPERTRIPLETS thus seems to be a reliable and fast alternative to the widespread b-MRP method.

4 DISCUSSION

Summarizing various trees obtained from different inferences into a single one is a frequent task in phylogenetic analysis of multigene datasets. Supertree methods are able to achieve this task even when the source trees only have partially overlapping taxon sets. A natural candidate supertree is the tree that is the closest to the source trees in terms of a given dissimilarity measure. SUPERTRIPLETS is a method specially designed to find the asymmetric median supertree according to triplet dissimilarity δ_t, and that has several advantages over the closely related t-MRP method.

In the first step of the t-MRP approach, the t-MR of the source trees depicts them as a set of binary sequences (one sequence per taxon, plus an extra one consisting only of '0' for an artificial outgroup leaf corresponding to the root of all source trees). Given a fully

resolved tree T, the parsimony value for a site encoding XY|Z is one if T entails this triplet resolution and two if T entails an alternative resolution (i.e. XZ|Y or YZ|X). It follows that the parsimony value of a binary supertree T for the t-MR of a forest F is simply the number of pseudocharacters (i.e. the total t-MR length) plus $\delta_t(F, T)/2$ (Moran *et al*., 2005; Wilkinson, 1994). A t-MRP supertree T is thus one of the binary trees that minimizes the triplet dissimilarity $\delta_t(F, T)$ (Wilkinson *et al*., 2001, 2005a, 2007). Nevertheless, there are some practical problems that hamper standard MP analysis performed to optimize this triplet-based criterion.

First, each pseudocharacter of the t-MR is one of the $O(n^3)$ triplets induced by one of the k source trees. For comparison, with the b-MR, one have $O(n)$ pseudocharacters. The number of source trees does not impact the number of pseudocharacters since it is common practice to pre-process them in order to keep only one (weighted) representative for the many identical pseudocharacters.

Second, for each pseudocharacter of the t-MR, there are only four informative character states (Fig. 1). It follows that, for instance, when inferring a supertree on 99 taxa, for each pseudocharacter the t-MR uses 100 states (99 taxa + root), 96 of which are missing character states. The t-MR is then made of 96% uninformative character states and this proportion is constant regardless of whether or not identical sites are compressed using a weighting procedure. If the supertree is defined on 999 taxa, the rate of uninformative character states grows to 99.6%. Each of the $O(n^3)$ triplets, i.e. stored in $O(1)$ memory space by SUPERTRIPLETS, is stored in $O(n)$ by the t-MR. It follows that, when dealing with t-MRP, both the MR size and the proportion of missing character states rapidly grows. Even with weighted pseudocharacters, the size of the t-MR remains large and requires huge memory to compute the t-MRP supertree (see Section 3.1). Moreover, a high proportion of missing character states is known to often slow down any MP tree inference method, since it may increase both the size and number of local parsimony optima. By directly considering all weighted triplets (instead of a t-MR) and by combining them through a polynomial time algorithm, SUPERTRIPLETS is better suited for achieving this task than t-MRP.

Third, the parsimony criterion does not allow comparison of a fully resolved tree T with a tree T' obtained by collapsing edges of T. Indeed, T' will always have a parsimony value greater or equal to that of T. By definition, the supertree returned by t-MRP is a summary (using strict consensus) of all MP (fully) resolved trees. This supertree is not the best according to the MP criterion; it is just a practical summary of the MP trees. On the other hand, the criterion optimized by SUPERTRIPLETS (i.e. asymmetric triplet dissimilarity δ_t) naturally allows comparison of trees with different degrees of resolution.

5 CONCLUSION

We propose SUPERTRIPLETS, a method that first constructs a supertree using a polynomial agglomerative scheme, and, secondly, performs fast local search to find the asymmetric median supertree according to a triplet dissimilarity. We point out that, by using triplet representation (i.e. t-MR), this median supertree is one of the MP binary trees. Computer simulations and a biological case study indicate that SUPERTRIPLETS tends to propose more reliable but less resolved supertrees than MRP methods.

We think that this work provides new prospects for designing supertree methods. On the one hand, SUPERTRIPLETS could certainly be significantly improved. Heuristics searching for the MP tree(s) have strongly benefited from more than 20 years of improvements, and have a considerable advantage with respect to method tuning and implementation tricks (e.g. Goloboff *et al*., 2008) as compared to recently proposed approaches. On the other hand, the confidence values displayed by SUPERTRIPLETS at the supertree edges are well-founded, and having an indication of the supertree branch support is an important feature. Several related confidence values dedicated to supertree-based comparative studies have already been proposed (Bininda-Emonds, 2003; Burleigh *et al*., 2006; Cotton *et al*., 2006; Moore *et al*., 2006; Wilkinson *et al*., 2005b). Comparing their relevance and accuracy is far from easy, as pointed out by Cotton *et al*. (2006), who concluded: '*Our example at least shows that there is no single correct view of support for supertrees*…'. This remark is relevant for supertree approaches but also reflect the intrinsic difficulty of comparing any clade support value in the supertree or supermatrix context (Burleigh *et al*., 2006; Douady *et al*., 2003).

ACKNOWLEDGEMENTS

We thank the three referees for their suggestions on the manuscript. We also thank Guillaume Dugas for the SUPERTRIPLETS webpage and Khalid Belkhir for helpful advices.

Funding: This work was supported by the Research Networks Program in Bioinformatics of the High Council for Scientific and Technological Cooperation between France and Israel; the bioinformatics cluster of ISE-M; and the French Agence Nationale de la Recherche '*Domaines Emergents*' [ANR-08-EMER-011 '*PhylAriane*']. This publication is contribution No. 2010-021 of the Institut des Sciences de l'Evolution de Montpellier (UMR 5554 - CNRS), France.

Conflict of Interest: none declared.

REFERENCES

Adams,E.N. (1972) Consensus techniques and the comparison of taxonomic trees. *Syst. Zool*., **21**, 390–397.

Bansal,M.S. *et al*. (2008) Comparing and aggregating partially resolved trees. *Lect. Notes Comput. Sci*., **4957**, 72–83.

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.

Beck,R.M.D. *et al*. (2006) A higher-level MRP supertree of placental mammals. *BMC Evol. Biol*., **6**, 93.

Bininda-Emonds,O.R.P. (2003) Novel versus unsupported clades: assessing the qualitative support for clades in MRP supertrees. *Syst. Biol*., **52**, 839–848.

Bininda-Emonds,O.R.P. *et al*. (2007) The delayed rise of present-day mammals. *Nature*, **446**, 507–512.

Blanga-Kanfi,S. *et al*. (2009) Rodent phylogeny revised: analysis of six nuclear genes from all major rodent clades. *BMC Evol. Biol*., **9**, 71.

Bordewich,M. and Semple,C. (2004) On the computational complexity of the rooted subtree prune and regraft distance. *Ann. Combinat*., **8**, 409–423.

Bourque,M. (1978) Arbres de Steiner et réseaux dont varie l'emplagement de certains sommets. PhD Thesis, University of Montréal, Montréal, Canada. (In French).

Bryant,D. (1997) Building trees, hunting for trees and comparing trees. PhD Thesis, University of Canterbury, Canterbury, New Zealand.

Bryant,D. (2003) A classification of consensus methods for phylogenies. In Janowitz,M. *et al*. (eds), *Bioconsensus*. DIMACS, AMS, Providence, RI, pp. 163–184.

Burleigh,J.G. *et al*. (2006) Supertree bootstrapping methods for assessing phylogenetic variation among genes in genome scale data. *Syst. Biol*., **55**, 426–440.

Churakov,G. *et al*. (2009) Mosaic retroposon insertion patterns in placental mammals. *Genome Res*., **19**, 868–875.

Cotton,J.A. and Wilkinson,M. (2007) Majority-rule supertrees. *Syst. Biol*., **56**, 445–452.

Cotton,J.A. *et al*. (2006) Discriminating supported and unsupported relationships in supertrees using triplets. *Syst. Biol*., **55**, 345–350.

Creevey,C.J. and McInerney,J.O. (2005) CLANN: investigating phylogenetic information through supertree analyses. *Bioinformatics*, **21**, 390–392.

Criscuolo,A. *et al*. (2006) SDM: a fast distance-based approach for (super)tree building in phylogenomics. *Syst. Biol*., **55**, 740–755.

Critchlow,D.E. *et al*. (1996) The triples distance for rooted bifurcating phylogenetic trees. *Syst. Biol*., **45**, 323–334.

Dixon,W.J. and Mood,A.M. (1946) The statistical sign test. *J. Am. Statist. Assoc*., **41**, 557–566.

Dobson,A.J. (1975) Comparing the shapes of trees. In Street,A.P. and Wallis,W.D. (eds) *Combinatorial Mathematics III*, *LNCS*, vol. 452. Springer, New York, pp. 95–100.

Douady,C.J. *et al*. (2003) Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol. Biol. Evol*., **20**, 248–254.

Doyle,J.J. (1992) Gene trees and species trees: molecular systematics as one-character taxonomy. *Syst. Bot*., **17**, 144–163.

Eulenstein,O. *et al*. (2004) Performance of flip supertree construction with a heuristic algorithm. *Syst. Biol*., **53**, 299–308.

Farris,J.S. *et al*. (1970) A numerical approach to phylogenetic systematics. *Syst. Zool*., **19**, 172–191.

Fawcett,T. (2004) ROC graphs: notes and practical considerations for researchers. Technical Report HPL-2003-4. HP Labs, Palo Alto, CA.

Fawcett,T. (2005) An introduction to ROC analysis. *Pattern Recogn. Lett*., **27**, 86–874.

Goloboff,P.A. *et al*. (2008) TNT, a free program for phylogenetic analysis. *Cladistics*, **24**, 774–786.

Gordon,A.D. (1986) Consensus supertrees: the synthesis of rooted trees containing overlapping sets of labeled leaves. *J. Classif*., **3**, 335–348.

Grunewald,S. *et al*. (2007) Closure operations in phylogenetics. *Math. Biosci*., **208**, 521–537.

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

Harding,E. (1971) The probabilities of rooted tree-shapes generated by random bifurcation. *Adv. Appl. Probab*., **3**, 44–77.

Hickey,G. *et al*. (2008) SPR distance computation of unrooted trees. *Evol. Bioinform. Online*, **4**, 17–27.

Janecka,J.E. *et al*. (2007) Molecular and genomic data identify the closest living relative of primates. *Science*, **318**, 792–794.

Jeffroy,O. *et al*. (2006) Phylogenomics: the beginning of incongruence? *Trends Genet*., **22**, 225–231.

Kimura,M. (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol*., **2**, 87–90.

Lin,H.T. *et al*. (2009) Triplet supertree heuristics for the tree of life. *BMC Bioinformatics*, **10**(Suppl. 1), S8.

Maddison,W.P. (1989) Reconstructing character evolution on polytomous cladograms. *Cladistics*, **5**, 365–377.

Moore,B.R. *et al*. (2006) Increasing data transparency and estimating phylogenetic uncertainty in supertrees: approaches using nonparametric bootstrapping. *Syst. Biol*., **55**, 662–676.

Moran,S. *et al*. (2005) Using semi-definite programming to enhance supertree resolvability. In Istrail,S. *et al*. (eds) *Algorithms in Bioinformatics, Proceedings of WABI 2005*, Vol. 3692 of *LNCS*, Springer, Berlin, ALLEMAGNE, pp. 89–103.

Mosses,C. (2005) Triplet supertrees. PhD Thesis, University of Aarhus, Aarhus, Denmark.

Nelson,G. and Ladiges,P.Y. (1994) Three-item consensus: empirical test of fractional weighting. In Scotland,R.W. *et al*. (eds) *Models in Phylogeny Reconstruction*. Clarendon, Oxford, pp. 193–209.

Page,R.D.M. (2002) Modified MinCut supertrees. In Guigó,R. and Gusfield,D. (eds) Vol. 2452 of *LNCS*, Springer, London, UK, pp. 537–551.

Phillips,C. and Warnow,T.J. (1996) The asymmetric median tree–a new model for building consensus trees. *Discr. Appl. Math*., **71**, 311–355.

Piaggio-Talice,R. *et al*. (2004) Quartet supertrees. In Bininda-Emonds,O.R.P. (ed) *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*. Kluwer Academic Publishers, Dordrecht, pp. 173–191.

Prasad,A.B. *et al*. (2008) Confirming the phylogeny of mammals by use of large comparative sequence data sets. *Mol. Biol. Evol*., **25**, 1795–1808.

Ragan,M.A. (1992) Phylogenetic inference based on matrix representation of trees. *Mol. Phyl. Evol*., **1**, 53–58.

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

Ranwez,V. *et al*. (2007) OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evol. Biol*., **7**, 241.

Robinson,D.F. and Foulds,L.R. (1981) Comparison of phylogenetic trees. *Math. Biosci*., **53**, 131–147.

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

Sanderson,M.J. (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, **19**, 301–302.

Semple,C. and Steel,M. (2003) *Phylogenetics*. Oxford University Press, Oxford.

Steel,M.A. and Penny,D. (1993) Distribution of tree comparison metrices - some new results. *Syst. Biol*., **42**, 126–141.

Steel,M.A. and Rodrigo, A. (2008) Maximum likelihood supertree. *Syst. Biol*., **57**, 243–250.

Swofford,D.L. *et al*. (1996) Phylogenetic inference. In Hillis, D.M. *et al*. (eds) *Molecular Systematics*. Sinauer Associates, Massachussets, pp. 407–514.

Thorley,J.L. (2000) Cladistic information, leaf stability and supertree construction. PhD Thesis. University of Bristol, Bristol, UK.

Wilkinson,M. (1994) Three-taxon statements: when is a parsimony analysis also a clique analysis? *Cladistics*, **10**, 221–223.

Wilkinson,M. *et al*. (2001) Towards a phylogenetic supertree for platyhelminthes? In Littlewood,D.T. and Bray,R.A. (eds) *Interrelationships of the Platyhelminthes*. Taylor and Francis, Chapman Hall, London, pp. 292–301.

Wilkinson,M. *et al*. (2004) The information content of trees and their matrix representations. *Syst. Biol*., **53**, 989–1001.

Wilkinson,M. *et al*. (2005a) The shape of supertrees to come: tree shape related properties of fourteen supertree methods. *Syst. Biol*., **54**, 419–431.

Wilkinson,M. *et al*. (2005b) Measuring support and finding unsupported relationships in supertrees. *Syst. Biol*., **54**, 823–831.

Wilkinson,M. *et al*. (2007) Properties of supertree methods in the consensus setting. *Syst. Biol*., **56**, 330–337.

Williams,D.M. (2004) Supertrees, components and three-item data. In Bininda-Emonds,O.R.P. (ed) *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*. Kluwer Academic Publishers, Dordrecht, pp. 389–408.

Williams,D.M. and Humphries,C.J. (2003) Component coding, three-item coding, and consensus methods. *Syst. Biol*., **52**, 255–259.

Yule,G.U. (1925) A mathematical theory of evolution, based on the conclusions of Dr J.C. Willis. *Philos. Trans. Roy. Soc. B*, **213**, 21–87.