SUPERTRIPLETS: a triplet-based supertree approach to phylogenomics

Vincent Ranwez1,2,*, Alexis Criscuolo3 and Emmanuel J.P. Douzery1,2

1Université Montpellier 2, CC064, Place Eugène Bataillon, 34 095 Montpellier Cedex 05, 2CNRS, Institut des Sciences de l’Evolution (UMR 5554), CC064, Place Eugène Bataillon, 34 095 Montpellier and 3Institut Pasteur, Département de Microbiologie, Unité de Biologie Moléculaire du Gêne 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 cladades, 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 some 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 phylogeny, 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 Swoford 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 pseuodcharaters (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...
Lastly, the utility of collected in the OrthoMaM database (Ranwez et al. in previous supertree studies (Criscuolo et al. 2006; Eulenstein et al., 2007), TH(SPR/TBR) 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'.

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, the clades of tree T are (A,B), (A,B,C), (D,E), and the trivial ones comprised 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 XYZ (e.g. Baum, 1992; Critchlow and Fiscus, 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 Ti since both do not necessarily have the same leaf set. The similarity between Ti and T is thus generally evaluated through the distance between Ti and the supertree T restricted to L(Ti) (e.g. Cotton and Wilkinson, 2007; Creevey and McInerney, 2005). The resulting measure between T and Ti 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 Ti 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'.

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 A(B,C)AB and A(B,E) triplets. This correspondence is also indicated by black circles. Rooting is ensured by an artificial all-zero taxon.

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 Ti since both do not necessarily have the same leaf set. The similarity between Ti and T is thus generally evaluated through the distance between Ti and the supertree T restricted to L(Ti) (e.g. Cotton and Wilkinson, 2007; Creevey and McInerney, 2005). The resulting measure between T and Ti 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 Ti 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): $X(T_i, T)$ and $D(T_i, T)$, the triplets resolved in $T_i$ and $T$, respectively; $R(T_i, T)$, the triplets resolved in $T_i$ and not resolved in $T$; $R(T, T_i)$, the triplets not resolved in $T_i$ and resolved in $T$; and $X(T_i, T)$, the triplets unresolved in both $T_i$ and $T$. Using this notation, a triplet distance between a source tree and a supertree can be formulated as $d(T_i, T) = 2|X(T_i, T)| + |R(T_i, T)| + |R(T, T_i)|$. 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 triplets resolved in the supertree but not in some source trees $T_i$, i.e. the set $R(T_i, T)$. Supertrusses resolving triplets that are not resolved in some source trees are thus penalized. The use of the (symmetric) distance $d$ results in considering introduction 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: $d_i(T_i, T) = 2|X(T_i, T)| + |R(T_i, T)|$.

2.2 Agglomerative construction of a TAMS

Agglomerative construction of a TAMS (SuperTriplets) is related to the process of merging two existing clades $T_i$ and $T_j$ in the current supertree $T$ into a new clade $T_{i\cup j}$, where $T_{i\cup j}$ is used to construct an initial supertree. For our triplet-based strategy, we transform each source tree into weighted triplets, where the weight of each triplet is $|X(T_i, T)|$ for any set of three taxa $\{A,B,X\}$ such that $A < B$, and $X \in C_{A\cap B}$. The weight $|X(T_i, T)|$ of a triplet $T$ is obtained as the number of times the triplet appears in source trees. The weight of a triplet $T$ that appears in $n$ source trees is attributed a weight $|X(T_i, T)| = n$. As a natural approach is then to agglomerate the two clades such that $N^\ast(C_{A\cap B})$ is maximal. Moreover, note that in such cases, every creation of a correct clade leads to $100\%$ of $nft$, while this proportion is smaller for incorrect clades. Moreover, the tree obtained during this agglomeration step will contain every $nft$ and is therefore the one minimizing the TAMS criterion. More formally, the selected agglomeration is randomly chosen among those maximizing $\sum_{C_{A\cap B}} \frac{n_{\text{mft}}(C_{A\cap B})}{\sum_{C_{A\cap B}}} (|C_{A\cap B}|-1)|X(C_{A\cap B})|$, where $n_{\text{mft}}(C_{A\cap B})$ is the number of newly resolved triplets that are $nft$, and $N^\ast(C_{A\cap B})$ the number of conflict with $nft$. It follows that to evaluate the agglomeration of any pair $C_{A\cap B}$, $\text{SuperTriplets}$ only needs to have at hand the corresponding values for $N^\ast(C_{A\cap B})$ and $N^\ast(C_{A\cap B}).$

During the first step of the agglomeration process, clades are reduced to a single taxon, so that for any of the $O(n^3)$ pairs of such trivial connected component $C_{A\cap B}$, the values $N^\ast(C_{A\cap B})$ and $n_{\text{mft}}(C_{A\cap B})$ correspond to the creation of new triplets $ABX$, where $A$ and $B$ are fixed taxa and where $X$ can take $n-2$ label values. All the $N^\ast(C_{A\cap B})$ and $n_{\text{mft}}(C_{A\cap B})$ values are therefore initialized in $O(n^4)$.
**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 content. We also compared 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 $T_{\text{model}}$ 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 $a \text{ posteriori}$ added, with a root-to-tip branch length of 1.1 substitution per site. To highlight the substitution rate heterogeneity among taxa and genes, $T_{\text{model}}$ 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 $T_{\text{model}}$ 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 PrivML (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 PrivML was repeated 100 times. Each tree returned by PrivML 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 pseudocaracters 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 \( T_{\text{model}}^d \), 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 \( x, d, r_1, r_2, x \) respectively, one can define two error types. The proportion of triplets that are in the supertree but not in the model tree is defined as 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 defined as 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 \( T_{\text{model}}^d \) and its related supertree \( T \) is then defined as \( et = (d + r_2)/(d + r_1) \); reciprocally, the type II error rate is defined as \( et = (d + r_1)/(d + r_2) \). 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 \( 1 - et \).

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 \( et = (et_B, \text{respectively}) \) is defined as the proportion of bipartitions of \( L(T) \) induced by each of the internal branches of \( T_{\text{model}}^d \) (of \( T \), respectively) that are not present in \( T \) (in \( T_{\text{model}}^d \), respectively).

A low type I error rate means that the supertree \( T \) displays few leaf bipartitions (or triplets) that are not present in \( T_{\text{model}}^d \). Respectively, a high type II error means that the supertree \( T \) does not represent all the phylogenetic information in \( T_{\text{model}}^d \). For each of the four supertree methods studied (i.e. b- and t-MRP, SuperTriplets and TILI), each taxon deletion rate \( d = 25, 50, 75% \) and each forest size \( k = 10-50 \), triplet- and bipartition-based \( et \) and \( et_B \) 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 negative triplets is \( FN = |D| - |R_1| \), and the number of true positive triplets is \( TP = |S| \). Finally, since the total number of resolved triplets on nx taxa is \( n(n-1)(n-2)/2 \), the number of true negative triplets is \( TN = n(n-1)(n-2)/2 - (TP + FP + FN) \). 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.
Based on the fact that the total number of non-trivial bipartition as in Figure 4. Note that the method are connected by a line. Symbols and colors have the same meaning as in Figure 3. Note that the x-axis is on a logarithmic scale.

**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, colors and arrows have the same meaning as in Figure 4. Note that the x-axis is on a logarithmic scale.

**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 ≤ 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 supraordinal clades: Cetartiodactyla (0.92), Afrotheria (0.87), Paenungulata (0.82), Eutherianarchontogires
Laurasiatheria (Prasad placental tree root (Afrotheria + Xenarthra, versus either Afrotheria correspond to areas of documented phylogenetic uncertainty: by Beck (0.76), Laurasiatheria (0.74), Boreoeutheria (0.64), Glires (0.59) and Faills (0.76), Xenarthra (0.74), Eulipotyphla (0.66) and Insectivora (0.58). The supertree method SuperTriplets 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.

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, 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

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.
resolved tree $T$, the parsimony value for a site encoding XYZ is one if $T$ entails this triplet resolution and two if $T$ entails an alternative resolution (i.e. XYZ or YZX). It follows that the parsimony value of a binary supertree $T$ for the t-MR of a forest $F$ is simply the number of pseudocaracters (i.e. the total t-MR length) plus $b_r(F, T)/2$ (Moran et al., 2005; Wilkinson, 1994). A t-MRP supertree $T$ is thus one of the binary trees that minimizes the triplet dissimilarity $b_r(F, T)$ (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 pseudocaracter 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 $O(n)$ pseudocaracters. The number of source trees does not impact the number of pseudocaracters since it is common practice to pre-process them in order to keep only one (weighted) representative for the many identical pseudocaracters.

Second, for each pseudocaracter 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 pseudocaracter 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 and the proportion of missing character states rapidly grows.

Even with weighted pseudocaracters, 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 $b_r$) naturally allows comparison of trees with different degrees of resolution.

5 CONCLUSION

We propose SuperTriplets, a method that first constructs a supertree using a polygonal 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; Dowdy 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 ‘PhyLiAriane’]. 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

Adam,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, 4977, 72–85.
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, 8–10.
Bork,P. et al. (2006) A higher-level MR 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.
Bliangs-Kank,S. et al. (2009) Redest 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. Comb., 8, 409–423.
Bourque,M. (1976) Arbres de Scénario et réseau dont varie l’emplacement 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 Janson,W. et al. (eds), Bioinformatics, 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, 426–446.
Charuvkar,G. et al. (2009) Mosaic retroposition 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.
