ABOUT THE LARGEST SUBTREE COMMON TO SEVERAL X-TREES

Alain GUENOCHE[1]
Henri GARRETA[2]
Laurent TICHIT[3]

RÉSUMÉ – *Etant donnés plusieurs X-arbres, ou arbres phylogénétiques, sur le même ensemble X, nous cherchons à construire un plus grand sous-ensemble $Y \subset X$ tel que les arbres partiels induits sur Y soient identiques d'un point de vue topologique, c'est-à-dire indépendemment des longueurs des arêtes. Ce problème, connu sous le nom de MAST (Maximum Agreement SubTree), est NP-Difficile, dans le cas général, dès que le nombre de X-arbres est supérieur à 2. Nous présentons un algorithme approché qui construit un arbre partiel commun maximal. Il est facilement programmable et suffisamment efficace sur une centaine de X-arbres connectant une centaine d'éléments pour évaluer la taille moyenne d'un sous-arbre commun à des X-arbres indépendants. La distribution observée permet d'estimer la taille critique d'un sous-arbre commun et de mesurer la congruence de plusieurs arbres évolutifs.*

MOTS CLÉS – X-arbres, Arbres phylogénétiques, Sous-arbre commun, MAST, MCT

SUMMARY – *Given several X-trees or unrooted phylogenetic trees on the same set of taxa X, we look for a largest subset $Y \subset X$ such that all the partial trees reduced by Y are topologically identical. This common subtree is called a MAST for Maximum Agreement SubTree. The problem has polynomial complexity when there are only two trees but generally it is NP-hard for more than two. We introduce a polynomial approximation algorithm for the multiple case, which is easy to implement, very efficient and which produces a maximal common subtree. It begins with the computation of an upper bound for its size and designates elements in X that cannot belong to a common subtree of a given size. Simulations on random and real data have shown that this heuristic often provides an optimal solution as soon as the number of trees is larger than 5. Then, we develop a statistical study to evaluate the average size of a MAST corresponding to independent trees. The computed distribution allows to estimate the critical size of a MAST to reveal some congruence between trees.*

KEYWORDS – X-trees, Phylogenetic trees, Partial common tree, MAST, MCT

[1] Institut de Mathématiques de Luminy, 163 Av. de Luminy, 13009 Marseille, e-mail : guenoche@iml.univ-mrs.fr

[2] Laboratoire d'Informatique Fondamentale, 163 Av. de Luminy, 13009 Marseille, e-mail : garreta@univmed.fr

[3] Institut de Mathématiques de Luminy, 163 Av. de Luminy, 13009 Marseille, e-mail : tichit@iml.univ-mrs.fr

1. INTRODUCTION

Let us recall that an X-tree is a partially labeled tree such that (i) X is the set of labeled leaves, (ii) all the unlabeled nodes have degree at least 3 and (iii) the edges have positive or null length. X-trees are unrooted, and when a root is placed on one edge, they become *phylogenetic trees*, adapted to represent the evolution of a set X of several species or *taxa*. When all the nodes have degree 3, X-trees are denoted as binary trees. Given several X-trees on the same set of taxa X, the problem tackled in this paper is to look for a largest subset $Y \subset X$, such that all the partial trees reduced by Y are topologically identical, that is without considering edge length. This common subtree is called a MAST for Maximum Agreement SubTree. This problem appears when comparing several X-trees connecting the same set of taxa X.

In phylogenetic studies, these X-trees are computed from $n = |X|$ aligned sequences with a bootstrap strategy or when comparing the trees obtained from several genes. In the latter case, the comparison of the orthologous gene sequences in each taxon, determined with any reconstruction method (maximum likelihood, parsimony, distance method, etc.), gives an X-tree or a phylogenetic tree. Generally, different genes lead to different trees because of biological reasons such as the nucleotide composition, the evolution speed along the branches or the horizontal gene transfers. When considering p genes, one gets a set $\mathcal{T} = \{T_1, T_2, \ldots T_p\}$ of X-trees. In other domains, such as Cognitive Sciences for category definition (Dubois, 2000) or for comparison of tree reconstruction methods (Guénoche and Garreta, 2001), the congruence of several X-trees can be measured. In any case, the question is to study the compatibility of these T_i trees.

Aside the Robinson-Foulds metric (1981), counting the number of common internal edges between two trees, one edge corresponding to a bipartition or split, this compatibility can be measured by the size of a largest subset $Y \subset X$ for which the trees agree ; this means that the edge lists become identical, independently of their lengths. When trees depend on genes, this identity reveals the same evolutive history.

The mathematical and computational study of X-trees has been established all along the last forty years, beginning with the seminal article of Buneman (1971), working on a completely different application domain, the manuscript filiation. Over numerous articles, one can refer to the books of Barthélemy & Guénoche (1991) & Semple and Steel (2003). In the latter, the question considered here is tackled in Chapter 6 as the Maximum Agreement SubTree (MAST) problem, initially formulated by Finden & Gordon (1985). When there are only two rooted trees, Steel & Warnow (1993) defined a dynamic programming scheme (hence a polynomial algorithm) which can be extended to the unrooted case. But when $p > 2$ the problem becomes NP-hard, except if the degree of nodes is bounded in at least one of the trees. Since that time, Cole et al. (1996) and Amir & Keselman (1997) proposed $O(nlog\ n)$ algorithms for two rooted binary trees and Berry & Nicolas (2004) give a $O(3^k pnlog\ n)$ algorithm for p binary rooted trees, parameter k being the number of elements to eliminate from the input trees to make a MAST. Some recent results have been obtained by Bryant (2007), Guillemot (2008) and Berry et al. (2009).

In this article, we describe a heuristic method to establish a common subtree as large as possible, which can be applied to more than two binary or not binary unrooted trees, which is easy to program and very efficient on real problems up to $n = p = 100$. Simulations on random and real data have shown that this heuristic always provides an optimal solution as soon as the number of trees is at least 5 and the number of taxa is lower than 100. Then, we develop a statistical study to determine the critical size of a MAST corresponding to independent X-trees, to be significant. This permits to evaluate the minimum size of a common subtree to reveal some congruence of the trees.

2. METHODOLOGY

Here, we are only interested in the X-tree shape, defined by the edge list, which is pompously called its *topology*, whatever the length of the edges are. In that case, these lengths can all be set to 1, and the path length metric in the tree, becomes the *unitary* distance, receiving integer values. It is a tree distance D, satisfying the Four Point Condition :

$$\forall \{x, y, z, t\}, D(x, y) + D(z, t) \leq \max\{D(x, z) + D(y, t), D(x, t) + D(y, z)\}.$$

The distance values indicate any quartet topology ; if, for $\{x, y, z, t\}$ $D(x, y)+D(z, t)$ is the smallest of the three sums and is unique, this quartet is said to be *resolved* and has topology $xy|zt$. In the support tree of D, at least one edge separates these two pairs. If the three sums are equal, the topology is said to be *non resolved* and there is no separating edge between any two pairs.

To decide if two X-trees A and B have the same topology, it is sufficient to compare (i) their distances D_A and D_B, or (ii) the splits (or bipartitions) corresponding to the edges in A and B, or (iii) the quartet topologies. The distances, the split sets or the quartet sets must be identical. For these three cases, the corresponding procedures have polynomial time complexity.

2.1. Score function, upper bound and elimination procedure

A quartet is said to be *compatible* with an X-tree set $\mathcal{T}$, if all its topologies in the different X-trees are either non resolved or identical. Consequently, a quartet is incompatible if it possesses at least two different resolved topologies. Clearly, compatible quartets can be assembled in a common tree structure. This is not exactly the MAST problem, since subtrees are not identical, and it is referred as the MCT (for Maximum Compatible Tree) problem. When all the trees are resolved, as for our simulations, these are the same.

Let the score function $Sc : X \longrightarrow \mathbb{N}$ be defined as the number of quartets containing x that are compatible with $\mathcal{T}$, and $ScMax$ be its maximum value over the X set of n elements :

$$Sc(x) \leq ScMax(n) = \binom{n-1}{3} = ScMax(n-1) \times \frac{n-1}{n-4}.$$

The first $ScMax$ values are given in the following table :

n	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$ScMax(n)$	4	15	20	35	56	84	120	165	220	286	364	455	560	680

This allows to calculate an upper bound for the maximum number of elements admitting a common topology.

proposition 1. *The MAST size is lower than or equal to the highest value m such that*

$$\left|\{x \textit{ such that } Sc(x) \geq ScMax(m)\}\right| \geq m.$$

Proof : There cannot exist a partial common tree with at least m leaves if there are less than m elements in X with a score larger than or equal to $ScMax(m)$.

The elimination of the elements having a score lower than $ScMax(m)$ is not a safe strategy even if they cannot belong to a common subtree with m leaves. If finally the largest common tree has $m' < m$ leaves, some elements x having a score $ScMax(m') \leq Sc(x) < ScMax(m)$ would have been eliminated. However, they can belong to a common subtree with $m'+1$ leaves or more. Nevertheless, we eliminate them for the moment, and we only deal with elements having a score larger than or equal to $ScMax(m)$. We denote X' the remaining elements, and set $n' = |X'|$.

2.2. One by one elimination

For each $x \in X'$, let $Nq(x)$ be the number of incompatible quartets in X' containing x and NbQ the total number of incompatible quartets on X'.

$$\sum_{x \in X'} Nq(x) = \sum_{x \in X'} \big(ScMax(n') - Sc(x)\big)$$

If $NbQ > 0$, at least one element must be deleted. This is a classical problem that is to cover a set with a minimum number of given subsets (see Figure 1). The whole set contains all the incompatible quartets, and the n' subsets correspond to the incompatible quartets containing one given element. Erasing one of them eliminates the corresponding quartets, reduces the total number of incompatible quartets and sets n' to $n'-1$.

This covering problem is well known to be NP-difficult. The proposed method consists in first deleting at each step one element covering the largest number of incompatible quartets, that is the one having the largest Nq value. Clearly it is a greedy algorithm, very close to the Chvatal heuristic (1979), since it never comes back on previous eliminations. When only compatible elements remain, a supplementary procedure is performed. It tries to reintroduce one after another the eliminated elements, in case their incompatibility was due to elements that have been erased later. Doing so, the selected set of taxa Y is maximal, since it cannot be extended. This strategy has also been used by Berry et al. (2009).

3. ALGORITHM

The X-trees are given in the standard *newick*[4] format, although it is poorly adapted to computation because it imposes an artificial root and, for a non resolved

[4] http ://evolution.genetics.washington.edu/phylip/newicktree.html

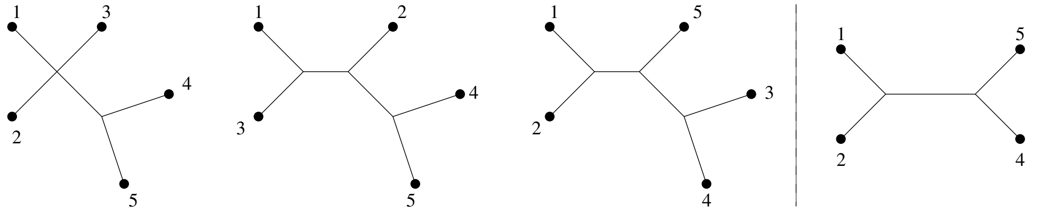

FIGURE 1. Three X-trees (left) and their largest common subtree (right) ; the Nq values are respectively equal to $(3, 3, 4, 3, 3)$, leading to the elimination of element 3.

tree, it could contain arbitrary edges with null length. The initial procedure consists in transforming each tree into two data structures : an unitary distance array and a table of all the splits corresponding to the edges with positive length. Notice that it is out of question to memorize the covering relation, since the usual values of n (≈ 100) makes such a task unrealistic.

3.1. The LAST (large agreement sub-tree) algorithm

```
/* Input */
A set T of p X-trees

/* Score computing */
For all quartets (x<y<z<t)
   if all quartet topologies on T are compatible
       Sc[x]++, Sc[y]++, Sc[z]++, Sc[t]++
End of For All

Determine the maximum number m of compatible elements.
Eliminate from X the elements x such that Sc[x]<ScMax[m]

/* Recursive elimination */
While (NbQuad>0)
  NbQuad:=0
  For each quartet (x<y<z<t) of remaining elements
      If all the topologies are not compatible among T
         NbQuad++
         Nq[x]++, Nq[y]++, Nq[z]++, Nq[t]++
  End For all
  Eliminate one element with maximum Nq ;
End of While

Let Y be the remaining set

/* Insertion of eliminated elements */
For any eliminated element x in X\Y
  For all triples (y<z<t) of elements in Y
      If all the {x,y,z,t} topologies are compatible
         Y <- x
  End of For all
End of For all
```

Let Y be the final subset of X given by the LAST algorithm. It is clear that all the quartets in Y are compatible and that Y is maximal in X. Generally, Y is not unique and nothing proves that Y has the MAST property. This could be partially tested by temporarily erasing an element in Y and looking for other compatible subsets as in the insertion procedure. If this truncated Y can be extended with another element, an equivalent solution will be found and, if it can be extended again, a better solution could be detected.

3.2. Establishing the most resolved common tree

When comparing bacterial strains, giving unresolved gene trees because genes are identical, it is often necessary to build the common tree for which any node which is resolved at least once is resolved. In that case, all the X-trees restricted to Y are not identical, but they are compatible. Starting from a single initial X-tree involves too much edge processing. The simplest way is to add all the unitary distances restricted to Y, using a well known property of tree distances :

proposition 2. *Let A and B be two X-trees and D_A and D_B their associated tree distances (unitary or path length metrics). The sum $D_A + D_B$ is a tree distance if and only if their topologies are compatible.*

Thus, if a quartet is resolved in one way, even in a single tree, it will necessarily be compatible and the most resolved common tree will appear. To get it from the sum of distances, any consistent algorithm can be applied ; the optimal ones are in $O(n^2 log\ n)$ for a tree distance, which is the case here. Several methods can be used as Hein (1989) or Guénoche & Leclerc (2001).

3.3. Complexity

The initial step establishing p unitary tree distances is in $O(pn^3)$ and it is run just once. The score computation, including the elimination step of elements having a minimal score is in $O(pn^4)$ at each iteration. Their number being bounded by n, this heuristic is in $O(pn^5)$. The reintegration of the at most n elements is in $O(pn^4)$.

Nevertheless the program in C is fast, since it takes for instance, less than 10 seconds for two trees having 100 leaves, and 100 seconds for 100 trees with 30 leaves, on a ordinary desktop computer. It uses a limited memory space, pn^2 integer values for distances, and a few arrays with n or $2n$ positions [5].

4. VALIDATIONS

The LAST method does not guarantee anything about the maximality of the common tree size. To tackle this question we performed simulations. We generate problems, that are sets of trees having a common subtree and we apply the LAST procedure to recover it, or a common part. We expect that it is not smaller than

[5] This software is freely available at http ://www.bioinformatics.lif.univ-mrs.fr

p	2	3	4	5	10	20	50
$k = 10$	18.50	14.52	13.06	12.80	11.60	11.16	10.88
$k = 20$	27.18	23.96	22.79	22.41	21.22	20.81	20.58
$k = 30$	35.85	33.34	32.48	32.05	31.12	30.65	30.37

Tab. 1. Average number of m returned by the LAST algorithm. These elements have the same topology in p X-trees with $n = 50$ taxa, k of them beeing fixed and the others randomly permutated.

the common subtree. All the generated X-trees have positive edge length, so the compatibility of topologies, which is the only thing tested, is an identity.

Let Y be the selected set of taxa, $m = |Y|$ and $m^* \geq m$ be the cardinality of a maximum agreement subtree. Even if the scores are bounded (using a branch-and-bound strategy), it is not possible to enumerate subsets with $m + 1$ elements. However, it is possible to build problems for which the value m^*, or a value close to it, is known.

It suffices to take an X-tree as input (randomly generated or coming from a real data set) and to generate as many trees as desired by permuting some of the labels but not all of them. Let k be the number of unchanged labels (labels that stay at the same place in each tree). Considering only these unchanged vertices, the generated X-trees are topologically identical, and one can assert that $m^* \geq k$. One can have $m^* > k$ when permuted labels go to the same location, which certainly occurs when p is small. As soon as the number p of X-trees is large enough, the variability of permutations implies that no quartet on the $n - k$ other elements will have the same position in the X-trees and, if the algorithm is efficient, we tend to have $m = m^* = k$. On the contrary, if $m < k$, then the algorithm failed, finding less than the k compatible elements.

We tested our procedure with $n = 50$. For each tested value of p (varying between 2 and 50), and each value of k (successively, 10, 20 and 30 elements), we generated 100 sets of p trees. Then, we computed the average number of elements selected by the LAST algorithm. These values are reported in Table 1.

When a subset with k elements is conserved, we observe that the average values of m monotonously decrease to k when p increases. We have also observed something more remarkable : none of the 2100 problems treated in these simulations produces a solution with less than k elements ! This does not prove that our algorithm is optimal, but that it nonetheless never *failed*.

Another simulation process has been tested using the Yule-Harding model (Harding, 1971). Starting with a tree with two leaves (a simple edge), the other elements are inserted one by one by selecting uniformly at random a leaf v and grafting the new leaf on the external edge leading to v, simulating a bifurcation. At each step a new node and two edges are placed in the tree. To get a set of trees having a common subtree with at least k elements, we generate first a common tree with k leaves, and we continue adding $n - k$ elements applying the previous procedure for each generated tree ; these $n - k$ elements are not (rarely) placed the same way.

For $n = 50$ and $p \geq 5$, the LAST algorithm returns systematically the k first elements as a common part, whatever is the k value. But when $p = 2, 3$ or 4 some of them may be missing, replaced by others forming a common subtree with less than

p	2	3	4	5	10	20
$n = 10$	5.42	4.32	3.99 (99)	3.35 (83)	3.06	3.
	8	6	5	5	4	4
$n = 20$	7.63	5.31	4.38	4.11	3.19 (19)	3.
	10	7	6	6	5	4
$n = 30$	9.12	5.97	4.70	4.23	3.58 (64)	3.
	12	8	7	6	5	4
$n = 50$	11.58	6.90	5.21	4.42	3.95 (95)	3.
	15	9	7	6	5	4
$n = 75$	13.85	7.68	5.64	4.67	3.92 (91)	3.
	17	10	8	7	5	4

TAB. 2. Average number of elements having the same topology in p random X-trees with n leaves. The values between parenthesis indicate the percentage of problems with a single conserved quartet ; the other problems have none. The critical values are shown in the second line.

k leaves. This occurs mainly for $k = 10$.

5. SIGNIFICANCE OF THE COMMON TREE SIZE

There have been many investigations about the average size of a MAST common to binary trees, and Bryant et al. (2003) obtain analytical bounds for two trees generated according to random models. Our purpose is much more simple : When, for an (n, p) problem, a largest common subtree size m is obtained, we would like to know how far it is from the expected value under the null hypothesis, claiming that these p trees are *independent*. Does this value occur frequently or is it large enough to suggest that the taxa share a common part of evolutive history ?

The same question has been posed by Lapointe & Rissler (2005) for a phylogeographic study. They compared trees having different sets of leaves, dividing the MAST size by the number of common leaves. Two trees are declared congruent when this normalized MAST score is greater than the value expected by chance. We extend their test in the same way for sets of more than two X-trees.

We generated, for each (n, p) values, 500 random sets of p X-trees. Then, we computed the average of the m values and the critical value at 5%, denoted μ, so that the proportion of trees giving a value $m \geq \mu$ is not larger than 5%. These figures are printed in Table 2. Thus, each time the computed value $m(n, p)$ is greater than or equal to $\mu(n, p)$ we can reject the null hypothesis of independence to conclude that these trees are not independent and have some congruence.

Let's comment the row of Table 2 corresponding to $n = 50$. For $p = 2$, both trees have on the average 11.58 compatible elements. This value results from the distribution detailed in Table 3 (rescaled in percentage). It thus appears that the critical value at 5% is equal to 15, since it is necessary to include the 6 cases giving $m = 14$ to reach 95%. The other critical values are obtained in the same way. For $n = 50$ and $p \geq 10$, there are less than 4 compatible elements. The value 3.95 corresponds to 95 trees that share a single compatible quartet, and 5 trees having none ; in this case, there are always at least 3 compatible elements, since no topology

m	9	10	11	12	13	14	15
% of problems	5	16	28	28	15	6	2

Tab. 3. Percentage of problems for 2 X-trees ($n = 50$) sharing a set of m compatible taxa.

is required. For $p \geq 20$, an average value equal to 3 means that there's no preserved quartet (4 vertices with a score equal to 1) and thus 4 is the critical value.

One can conclude that these critical values are very low, as soon as $p \geq 5$, and the independence hypothesis is easy to reject. For instance, a single compatible quartet is sufficient to ensure that more than 20 phylogenetic trees share a common evolutive history part, whatever the number of taxa is.

Acknowledgements

This work has been partly granted by the french GDR Operational Research and a PEPS-STIC Program from CNRS. We would like to thank an anonymous referee for some improvements and many accurate references.

REFERENCES

AMIR A., KESELMAN D., Maximum agreement subtree in a set of evolutionary trees : Metrics and efficient algorithms. *SIAM Journal on Computing*, 26, 1997, p. 758-769.

BARTHELEMY J.P., GUENOCHE A., *Trees and Proximity Representations.* Chichester, Wiley, 1991.

BERRY V., NICOLAS F., Maximum agreement and compatible supertrees. In *in Proceedings of CPM*, 2004, p. 205-219.

BERRY V., GUILLEMOT S., NICOLAS F., PAUL C., On the approximability of the MAST and MCT problems, *Discrete Applied Mathematics*, 157(7), 2009, p. 1555-1570.

BRYANT D., *Building trees, hunting for trees and comparing trees : theory and method in phylogenetic analysis*, PhD thesis, University of Canterbury, 1997.

BRYANT, D., MCKENZIE, A. , STEEL, M., The size of a maximum agreement subtree for random binary trees, in M. Janowitz (Ed.) *Bioconsensus*, DIMACS Ser. Discrete Math. Theoret. Sci., 61, 2003, p. 55-65.

BUNEMAN P., The Recovery of Trees from Measures of Dissimilarity. In D.G. Kendall and P. Tautu, editors, *Mathematics the the Archeological and Historical Sciences*, Edinburgh University Press, 1971, p. 387-395.

CHVATAL V., A greedy heuristic for the set covering problem, *Mathematics and Operations Research*, 4, 1979, p. 233-235.

COLE R., FARACH M., HARIHARAN R., PRZYTYCKA T., THORUP M., An $o(nlog\ n)$ algorithm for the maximum agreement subtree problem for binary trees. *SIAM Journal on Computing*, 30, 1996, p. 1385-1404.

DUBOIS, D., Categories as acts of meaning : the case of categories in olfaction and audition, *Cognitive Science Quartely*, 1, 2000, p. 35-68.

FINDEN C.R., GORDON A.D., Obtaining common pruned trees, *Journal of Classification*, 2, 1985, p. 255-276.

GUENOCHE A., GARRETA, H., Can we have confidence in a tree representation ? In O. Gascuel, M.-F. Sagot (eds.), *JOBIM 2000, Lecture Notes on Computer Science, 2066*, 2001, p. 45-56.

GUENOCHE A., LECLERC B., The triangles method to build X-trees from incomplete distance matrices, *RAIRO Operations Research*, 35, 2001, p. 283-300.

GUILLEMOT S., *Approches combinatoires pour le consensus d'arbres et de séquences*, Thèse de l'Université de Montpellier II, 2008.

HARDING E.F., The probabilities of rooted-tree shapes generated by random bifurcation, *Advances in Applied Probability*, 3, 1971, p. 44-77.

LAPOINTE, F-J., RISSLER L.J., Congruence, consensus, and the comparative phylogeography of codistributed species in California, *American. Naturalist*, 166, 2005, p. 290-299.

HEIN J.J., An optimal algorithm to reconstruct trees from additive distance data, *Bulletin of Mathematical Biology*, 51 (5), 1989, p. 597-603.

ROBINSON D.F., FOULDS L.R., Comparison of phylogenetic trees. *Mathematical Biosciences*, 53, 1981, p. 131-147.

SEMPLE Ch., STEEL M., *Phylogenetics*. Oxford University Press, 2003.

STEEL M., WARNOW T., Kaikoura tree theorems : computing the maximum agreement subtree. *Inf. Process. Lett.*, 48(2), 1993, p. 77-82.