BMC
Bioinformatics

METHODOLOGY ARTICLE   **Open Access**

Multiple consensus trees: a method to separate divergent genes

Alain Guénoche

Abstract

Background: It is generally admitted that the species tree cannot be inferred from the genetic sequences of a single gene because the evolution of different genes, and thus the gene tree topologies, may vary substantially. Gene trees can differ, for example, because of horizontal transfer events or because some of them correspond to paralogous instead of orthologous sequences. A variety of methods has been proposed to tackle the problem of the reconciliation of gene trees in order to reconstruct a species tree. When the taxa in all the trees are identical, the problem can be stated as a consensus tree problem.

Results: In this paper we define a new method for deciding whether a unique consensus tree or multiple consensus trees can best represent a set of given phylogenetic trees. If the given trees are all congruent, they should be compatible into a single consensus tree. Otherwise, several consensus trees corresponding to divergent genetic patterns can be identified. We introduce a method optimizing the *generalized score*, over a set of tree partitions in order to decide whether the given set of gene trees is homogeneous or not.

Conclusions: The proposed method has been validated with simulated data (random trees organized in three topological groups) as well as with real data (bootstrap trees, homogeneous set of trees, and a set of non homogeneous gene trees of 30 *E. Coli* strains; it is worth noting that some of the latter genes underwent horizontal gene transfers). A computer program, MCT - Multiple Consensus Trees, written in C was made freely available for the research community (it can be downloaded from http://bioinformatics.lif.univ-mrs.fr/Consensus/index.html). It handles trees in a standard Newick format, builds three hierarchies corresponding to RF and QS similarities between trees and the greedy ascending algorithm. The generalized score values of all tree partitions are computed.

Background

The comparison of gene trees and their assembling in a unique tree representing the species tree is a fundamental problem in phylogeny. A large variety of methods have been proposed to tackle the problem of the reconciliation of gene trees in order to reconstruct a species tree. A panel of methodological approaches can be found in [1] preceding the authors' own method. When the taxa in all the trees are identical, the problem can be stated as a consensus tree problem.

Given a set of phylogenetic trees, summarizing them into a consensus tree implies the homogeneity of this tree set. According to an alignment of sequences from a set X of n taxa, each gene gives an unrooted X-tree [2]; X is the set of leaves, the internal nodes have degree 3 and the

edges have a positive or null length. These trees can be established with several kinds of methods, distance, likelihood, or parsimony [3]. A bootstrapping approach [4] is used to evaluate the robustness of each gene tree, using a consensus methodology. At this level the homogeneity of the tree set is guaranteed. Then comes the second consensus tree problem. It consists in computing, from the gene tree set, denoted in the following as a *profile* of m X-trees, a global consensus tree summarizing them, i.e. producing one species X-tree. For this consensus tree, the homogeneity of the profile is questionable because of horizontal transfer events or because some genes can correspond to paralogous instead of orthologous alleles.

Here, we only deal with unrooted X-trees. Several consensus strategies can be used [5]. We focus on those proceeding the same way : an X-tree is considered as a set of bipartitions, each one corresponding to an internal edge of the tree, the external ones connecting the leaves to

Correspondence: guenoche@iml.univ-mrs.fr
Institut de Mathématiques de Luminy, 163 Av. de Luminy, 13009 Marseille, France

the tree. Any internal edge clearly separates two subsets of X having more than one element. Removing this edge creates a *split*, inducing a bipartition of X. The weight of each bipartition $P_i = X_i \cup X'_i$ is the number m_i of X-trees in the profile containing this bipartition.

Given a set of X-trees,

- the strict consensus tree is only made of bipartitions common to all the trees ($m_i = m$). This is a theoretical consensus leading to very unresolved trees (with very few internal edges).
- the majority consensus tree contains all the bipartitions that are present in a majority of trees ($m_i > m/2$). The bipartitions are compatible with a tree structure that is, generally, incompletely resolved.
- the extended majority consensus tree contains all the majority bipartitions, as well as all those that are compatible with the previous ones, the edges being selected according to decreasing values of m_i. This greedy consensus is the usual one, since it leads to the most resolved consensus trees.
- the Nelson consensus tree is made with the heaviest set of compatible bipartitions ($\sum_i m_i$ is maximum). It corresponds to finding a clique of maximum weight in a compatibility graph of the whole bipartition set, which is NP-hard [6].

Example 1. Let Π be the profile containing the 5 trees of Figure 1 : The computation of the consensus tree will first establish the complete set of bipartitions present in these trees, which is given in Table 1. The only majority bipartitions are the second and third, which are present in 3 and 4 trees respectively. They make the majority consensus tree
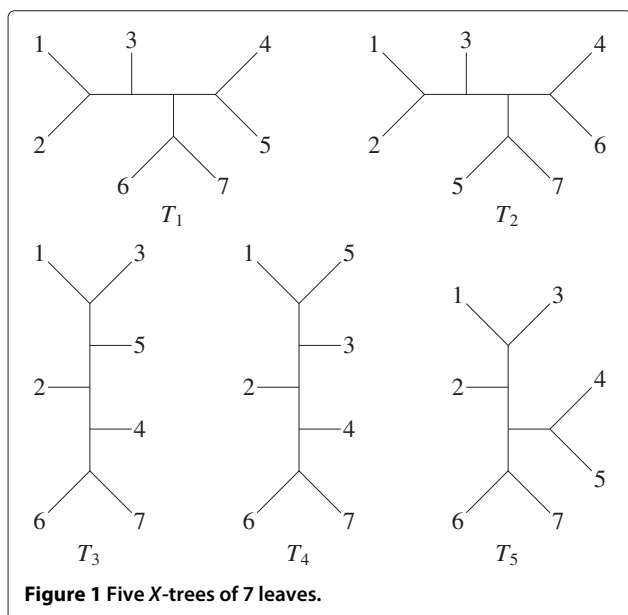
Table 1 The whole set of bipartitions in the trees of Figure 1

	T_1	T_2	T_3	T_4	T_5	Bipartitions
1	x	x				1 2 | 3 4 5 6 7
2	x	x			x	1 2 3 | 4 5 6 7
3	x		x	x	x	1 2 3 4 5 | 6 7
4	x				x	1 2 3 6 7 | 4 5
5		x				1 2 3 4 6 | 5 7
6		x				1 2 3 5 7 | 4 6
7			x		x	1 3 | 2 4 5 6 7
8			x	x		1 3 5 | 2 4 6 7
9			x	x		1 2 3 5 | 4 6 7
10				x		1 5 | 2 3 4 6 7

C which can be extended with the 7-th and 9-th bipartitions to give the extended majority tree C_E; both are drawn in Figure 2. C can also be extended with the 1-st and 9-th bipartitions to give a tree similar to C_E exchanging leaves 2 and 3.

How to assign a weight to the consensus tree?

In bootstrapping a gene sequence alignment, phylogeneticists are mainly interested in strong majority edges, i.e. edges that are in a great number of trees (generally at least 90%) indicating subsets of taxa derived from a common ancestor. This permits both analyzing the tree, to decide if it is a strong consensus, and also comparing it to another tree[a]. For these comparisons, and also for the consensus tree, we will focus only on the majority edges, those that are present in more than half the number of trees, because (i) differences based on minority edges would be less convincing, (ii) the majority consensus tree is a median of the profile, for the *Robinson-Foulds* distance [7] and (iii) it is unique.

The weight of an X-tree A, relative to profile Π, is equal to the sum of the weights of its internal majority edges (those present in $m_i > m/2$) corresponding to bipartitions P_i. Let A_m be the tree A restricted to its majority edges :

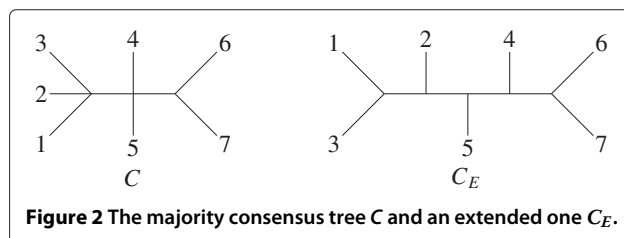$$W_\Pi(A) = \sum_{P_i \in A_m} w(P_i) = \sum_{P_i \in A_m} m_i.$$

Figure 1 Five *X*-trees of 7 leaves.

Figure 2 The majority consensus tree *C* and an extended one *C$_E$*.

In the following, we only consider the majority consensus tree of Π, i.e. the X-tree maximizing this weight function and containing only majority internal edges.

Single or multiple consensus trees?

Now comes the main question : why summarize a set of X-trees with a single tree? When the profile is provided by a bootstrap procedure a single tree is expected ; adding noise to an alignment which is not certain, is to consider other potential alignments to test if they produce the same or a similar tree. But when several X-trees corresponding to different genes are examined, several consensus trees can be expected. If these genes are not congruent, i.e. reflecting the same evolutionary history, the unique consensus tree could contain few majority edges, with low weights, and even no internal edges at all, a star tree. Consequently, given a set of X-trees corresponding to several genes, one can ask if there is one consensus tree or several trees associated to subsets of genes having evolved differently and each having their own strong consensus.

This idea was first formulated by Maddison [8] with the Phylogenetic Islands concept based on pruning and regrafting of just one subtree. He observed that the consensus trees of the islands are different and with a better resolution than for the whole set. This principle has been extended by [9] who investigate several clustering procedures of a given tree set to compare only strict consensus, without indicating how to fix the number of clusters. More recently, [10] give a method to build a minimum number of trees and display all the splits whose support is above a predefined threshold. When the threshold is lower than .5, such a tree set has no consensus meaning.

To decide if a single consensus tree is acceptable of if several trees make a better representation, we introduce the *generalized score* criterion of an X-tree profile Π :

Definition 1. Let P_{Π} be a partition of Π in k classes ($\Pi = \{\Pi_1, \ldots, \Pi_k\}$), containing respectively $\{p_1, \ldots, p_k\}$ X-trees and let $\{C_1, \ldots, C_k\}$ be the majority consensus trees corresponding to these sub-profiles. The generalized score of P_{Π}, denoted $\mathcal{W}^k(P_{\Pi})$ (or $\mathcal{W}^k$ when there is no ambiguity) is the sum over the classes, of the weight of the consensus tree multiplied by the number of trees in each class.

$$\mathcal{W}^k(P_{\Pi}) = \sum_{i=1,\ldots,k} p_i \times W_{\Pi_i}(C_i).$$

This index can be viewed as a voting process ; the p_i trees of Π_i designate C_i as their representative tree ; its weight is equal to $W_{\Pi_i}(C_i)$. Each one of the p_i trees has this weight and the Π_i class counts for $p_i \times W_{\Pi_i}(C_i)$. So, the generalized score is the sum of the class weights. It

is a quadratic criterion, since majority edges count for $W_{\Pi_i}(C_i)$ and for p_i.

Counted so, the generalized score of the profile considered as homogeneous –that is with a single class–, is $\mathcal{W}^1 = m \times W_{\Pi}(C_1)$ which is a reference value. If a partition of the profile in $k > 1$ classes which gives a greater generalized score ($\mathcal{W}^k > \mathcal{W}^1$) exists, then we conclude that the profile is not homogeneous and also that the consensus tree is not unique since there are k groups of genes with their own consensus, leading to a partition of the profile with a set of consensus trees of greater score.

When $\mathcal{W}^m$ is the largest value, it means that each gene has its own evolutionary history and so, the generalized score corresponds to the atomic partition P_0 on Π. This case is denoted as the *atomic* consensus. Each class (singleton) has its own X-tree as consensus and each edge receives a majority weight equal to 1. Each tree gets its number of internal edges as weight, which is bounded by $n - 3$.

$$\mathcal{W}^m = \sum_{i=1}^{m} n_i \leq m \times (n - 3).$$

So, maximizing the generalized score over the set of all the partitions of Π (denoted $\mathcal{P}(\Pi)$) one can decide if Π admits a single, a multiple or an atomic consensus which is certainly justified if there is no majority edge.

Example 2. Consider profile Π of Example 1, and the decomposition $\Pi_1 = \{T_1, T_2, T_5\}$ and $\Pi_2 = \{T_3, T_4\}$. Π_1 has one common bipartition $(1, 2, 3 | 4, 5, 6, 7)$ and three majority ones, $(1, 2 | 3, 4, 5, 6, 7)$, $(1, 2, 3, 4, 5 | 6, 7)$ and $(1, 2, 3, 6, 7 | 4, 5)$, which provide a consensus tree with weight 9. Trees in Π_2 also share three majority (thus common) bipartitions, $(1, 3, 5 | 2, 4, 6, 7)$, $(1, 2, 3, 5 | 4, 6, 7)$ and $(1, 2, 3, 4, 5 | 6, 7)$, making a consensus tree with weight 6. Both consensus trees are depicted in Figure 3.

So, the generalized score of profile Π with its single consensus tree C is $\mathcal{W}^1 = 5 \times W_{\Pi}(C) = 5 \times 7 = 35$. But the score resulting from decomposition $\Pi_1 | \Pi_2$ is : $\mathcal{W}^2 = 3 \times W_{\Pi_1}(C_1) + 2 \times W_{\Pi_2}(C_2) = 27 + 12 = 39$.

Figure 3 Consensus trees C_1 and C_2 of classes Π_1 and Π_2.
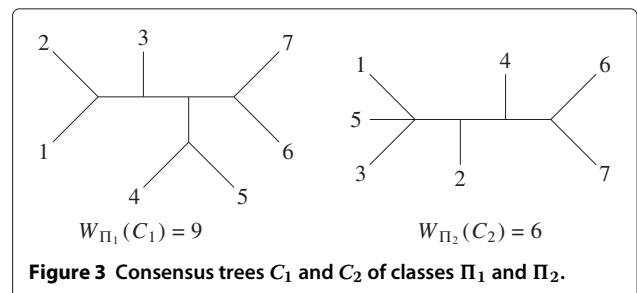
It is also greater than the generalized score corresponding to the atomic partition of Π, each tree T_k having a weight of 4, giving $\mathcal{W}^5 = 5 \times 4 = 20$.

Proposition 1. Two X-trees admit a single consensus if and only if they have more than half their edges in common.

Proof. Let T_1 and T_2 be two X-trees having n_1 and n_2 internal edges respectively. The generalized score of the two trees is $\mathcal{W}^2 = n_1 + n_2$. If they share k edges, $\mathcal{W}^1 = 2 \times k$ and so $\mathcal{W}^1 > \mathcal{W}^2$ if and only if $k > \frac{n_1 + n_2}{2}$. $\qquad\square$

One can formulate a similar result for m X-trees; if there are k majority edges, $\mathcal{W}^1 \geq m \times k \times \lfloor \frac{m}{2} \rfloor$ and $\mathcal{W}^1 > \mathcal{W}^m$ if and only if $k \times \lfloor \frac{m}{2} \rfloor > \frac{1}{m} \sum_{i=1}^{m} n_i$. One can conclude : for a profile of m resolved trees on n leaves, if $\lfloor \frac{m}{2} \rfloor > n - 3$, one majority edge is sufficient to assert that $\mathcal{W}^1 > \mathcal{W}^m$.

Methods

Maximizing $\mathcal{W}$ over $\mathcal{P}(\Pi)$ is not simple, since for each partition, the consensus tree of each class Π_i or at least its weight, $W_{\Pi_i}(C_i)$ must be computed. As the number of classes is not fixed, we have developed two hierarchical algorithms to build a series of nested partitions with $\{m, (m-1), \ldots, 1\}$ classes. For each one, its generalized score is computed and the partition within the hierarchy maximizing the score is retained. The atomic partition P_0 and the partition with a single class belonging to the series are competing. However, the best resulting score is not necessarily optimal over $\mathcal{P}(\Pi)$.

Average linkage hierarchical strategy

A classical approach in clustering consists in selecting a similarity function over the X-tree set and applying a hierarchical ascending method to build a series of nested partitions. We first recall two classical similarity measures between X-trees before using the average linkage algorithm (UPGMA) on one or the other similarity array. Other more sophisticated distances between trees [11] and other clustering algorithms, as in [9], can be used but the ones used here are fast enough to deal with hundreds of trees.

The Robinson-Foulds (RF) similarity

For any two trees T_i and T_j in Π, each tree being considered as the set of its internal edges, the number of common edges is first computed. The *Robinson-Foulds* similarity, derived from their distance [12], is twice this number divided by the number of edges in both trees.

$$S(T_i, T_j) = \frac{2 \times |\{a \in T_i \cap T_j\}|}{|T_i| + |T_j|}$$

The rate of common quartets (QS similarity)

The number of quartets in X having the same *topology* in two compared trees [13] are first counted. One similarity point will be assigned to quartet $\{x, y, z, t\}$ if, in both trees, either at least one internal edge separates the same pairs (for instance $\{x, y\}$ and $\{z, t\}$) or if they are both unresolved. Half a point is given when only one topology is resolved. If both are resolved and different, no similarity points are given.

Example 3. Coming back to the profile Π in Example 1, the *Robinson-Foulds* similarity is given in the left hand table of Figure 4, ignoring the denominators equal to 8 since all the trees are resolved. One can start joining T_1 and T_5 or T_3 and T_4 since their similarity values are equal. In both cases, the same hierarchy (represented in the dendrogram on the right side of Figure 4) is obtained.

The nested partitions are $(1|2|3|4|5)$, $(1,5|2|3|4)$, $(1,5|2|3,4)$, $(1,2,5|3,4)$ and $(1,2,3,4,5)$. The score values are $\mathcal{W}^5 = 20$, $\mathcal{W}^4 = 24$, $\mathcal{W}^3 = 28$, $\mathcal{W}^2 = 39$ et $\mathcal{W}^1 = 35$, respectively. So, it is the partition in two sub-profiles, detailed in Example 2, which maximizes the generalized score giving two consensus trees for Π.

Merging the two classes maximizing the generalized score function

Given k classes making a running partition of the hierarchical process, to maximize the generalized score of a nested partition with $k-1$ classes, one must join the two classes providing the best score value. Following this greedy principle, at each step, we evaluate the score value of all the fusions of any two classes and the two classes giving the maximum value are merged.

Example 4. Coming back to the profile Π in Example 1, the score values of the successive tested partitions are given in the arrays of Table 2 corresponding to the successive steps. The left array is the initial table containing the values of partitions joining just two trees. It forms class $\{T_1, T_5\}$ corresponding to consensus tree $T_{1,5}$. The central array leads to merging classes T_2 and $\{T_1, T_5\}$;

S	T_1	T_2	T_3	T_4
T_2	4			
T_3	2	0		
T_4	2	0	6	
T_5	6	2	4	2
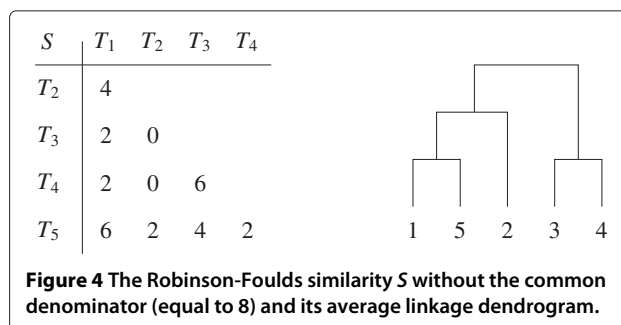
Figure 4 The Robinson-Foulds similarity S without the common denominator (equal to 8) and its average linkage dendrogram.

Table 2 The generalized score value of partitions of Π merging two classes

	T_1	T_2	T_3	T_4		$T_{1,5}$	T_2	T_3		$T_{1,2,5}$	T_3
T_2	20				T_2	35			T_3	28	
T_3	16	12			T_3	32	16		T_4	28	39
T_4	16	12	24		T_4	26	16	28			
T_5	24	16	20	13							

Finally the last union merges T_3 and T_4, giving the same hierarchy as in Figure 4. At the first step, one could also make class $\{T_3, T_4\}$ and the second step would be different since it leads to class $\{T_3, T_4, T_5\}$ giving another hierarchy and $\mathcal{W}^2 = 35$ which is less than the previous choice.

Results and discussion

On random trees

An heterogeneous profile Π can generate very few majority edges, and consequently a consensus tree with a small weight when it is not reduced to a star tree. In this case, the generalized score of the atomic partition must be larger than the single consensus. We verified this, testing many sets of random X-trees ; they had no majority internal edge, and $\mathcal{W}^1 = 0$. No proper class appeared and $\mathcal{W}^m$ is always maximum.

In a more precise test, we first selected three rooted X-tree topologies with 16 leaves: the balanced binary tree (any subdivision is balanced and there are 8 cherries), the caterpillar tree (any subdivision is one taxon against the remaining ones, providing just one cherry) and a random topology obtained by random hierarchical subdivisions of the 16 leaves. For each topology we derived 30 trees, simulating DNA sequence evolution with only substitutions (3/4 transition, 1/4 transversion), avoiding the alignment process. For each tree a random root sequence is fixed and substitutions are randomly selected according to the random branch lengths for each tree. The sequences are 1000 nucleotides long and the substitution rate is .25, which means that, on average, 1/4 characters of the root sequence are changed in the terminal ones. The Kimura distance [14] between the 16 terminal sequences is computed and an X-tree is established using the NJ algorithm. So, for the three topologies, 30 trees are established.

We first verified that each consensus tree follows its initial topology and that the generalized score strongly supported a single consensus within the families. We then selected 10 trees from each topological set multiple times to make a profile clearly composed of three classes. The generalized score always recognized the three classes corresponding to the three topologies, and $\mathcal{W}^3$ always gave the largest value.

Homogeneous trees

We first tested trees computed by Brown et al. [15]. Their abstract states: *Here we use large combined alignments of 23 orthologous proteins conserved across 45 species from all domains, to construct highly robust universal trees. Although individual protein trees are variable in their support of domain integrity, trees based on combined protein data sets strongly support separate monophyletic domains. Within the Bacteria, we placed spirochaetes as the earliest derived bacterial group. However, elimination from the combined protein alignment of nine protein data sets, which were likely candidates for horizontal gene transfer, resulted in trees showing thermophiles as the earliest evolved bacterial lineage.*

Since possibly divergent proteins have been eliminated, the single consensus must be strong and give the largest score. In fact, there are 22 majority internal edges over the 333 present in the 23 trees (with 45 species there are at most 989 bipartitions). The consensus tree has a weight of 430, revealing that each edge is supported by nearly 20 trees, so they are strongly majority, and the generalized score is $\mathcal{W}^1 = 9890$. The atomic consensus gives $\mathcal{W}^{23} = 964$. Decreasing the number of classes increases the scores, but they never reach the single consensus tree value. The best secondary value is obtained for 2 classes, isolating one singleton and giving $\mathcal{W}^2 = 8673$. It is, therefore, confirmed : there is a single consensus tree for this homogeneous tree set.

When the profile Π is homogeneous, the single consensus must give the largest score. It is what is expected from bootstrap trees corresponding to a single gene ; we first verified this on trees corresponding to *Escherichia Coli* strains.

9 genes on 30 *E. coli* strains

In a previous work with P. Darlu, we asked the same question of how to recognize divergent genes sequenced on the same X taxa set. We proposed a new method, *TreeOfTrees* [16], establishing a tree of which each leaf corresponds to a single gene (in fact bootstapped trees of this gene) and each internal edge receives a robustness coefficient allowing the separation of subsets of trees that are topologically different. It was the first attempt to statistically evaluate whether two trees are significantly closer to each other than to a third one. This method has been proved efficient on both simulated and real data.

The application was done on 9 genes (DNA sequences) in 30 strains of *Escherichia coli* [17]. Let X be the strain set and G the set of genes corresponding to 6 housekeeping genes (icd, pabB, polB, putP, trpA, trpB), plus 3 others, HPI, DR and UR (High Pathogenicity Island and its Downstream and Upstream regions), which are known to have been transferred. The corresponding sequences were

first aligned and 500 bootstrap trees were obtained with PHYML [18].

The *TreeOfTrees* method is based on the comparison of these bootstrap trees. At each iteration, corresponding to one bootstrap step, the algorithm compares $|G|$ X-trees, computing a distance between them and using a distance method (NJ) to define a G-tree, i.e. a tree whose leaves are the genes. At the end of the 500 iterations, 500 G-trees make a profile and a consensus tree is established indicating a bootstrap value for internal edges, as usual. When this value is high, one can deduce that the genes on both sides of this edge correspond to different gene tree sets, revealing different topologies. With these 9 genes, we have obtained the G-tree depicted in Figure 5 with bootstrap values displayed on the edges.

Based on these values in the consensus G-tree, we conclude that the X-trees (on the *E. coli* strains) built from the HPI, UR and DR sequences are significantly different from the others. The biological interpretation is discussed in Schubert et al. [17]. Before continuing with this data set, we would like to underline that the *TreeOfTrees* method does not make it possible to separate a single gene since the robustness coefficients are only defined for internal edges.

First, we have computed a consensus X-tree for each gene. The first thing to verify is that any 500 bootstrap tree set constitutes a homogeneous profile admitting a single consensus tree. This can be observed in Table 3 indicating, for each gene, the number of total bipartitions, the number of majority bipartitions, the weight of the consensus tree, the corresponding generalized score $\mathcal{W}^1$, the number of classes of the best partitioning of the profile in more than one class, (with the number of elements of the other classes) and its generalized score.

For all the genes, except icd, $\mathcal{W}^1$ is maximum. The competing partition has 1 or 2 extra classes which are very small ; gene putP is an exception since the second class has 80 elements, but $\mathcal{W}^2 \ll \mathcal{W}^1$. For icd, $\mathcal{W}^2 > \mathcal{W}^1$ but the two values are very close and the optimal partition has only one other class with 4 elements sharing 8 common bipartitions. So, the profiles generated by

Table 3 Results on bootstrap trees

	BiP	*Maj*	*W*	$\mathcal{W}^1$	*NbClas*	$\mathcal{W}^m$
UR	8	7	2623	1311500	2	1304768
trpB	28	15	6248	3124000	2	3114271
trpA	45	9	3824	1912000	3	1900390
putP	57	17	6608	3304000	2	2508400
polB	119	14	5331	2665500	2	2639187
icd	69	15	5681	2840500	2	2929008
HPI	76	13	4971	2485500	2	2467626
pabB	57	8	3667	1833500	2	1827846
DR	12	8	2685	1342500	2	1335146

For each set of bootstrap gene trees, the number of bipartitions (*BiP*), the number of majority bipartitions (*Maj*), the weight of the consensus tree (*W*), its generalized score ($\mathcal{W}^1$), the number of classes of the best multiple consensus (*NbClas*) and the generalized score of this partition ($\mathcal{W}^m$) are indicated.

bootstrapping are homogeneous and therefore recognized by the generalized score function.

These 9 consensus trees make our last profile. It generates 99 bipartitions, 3 of them being majority. The best generalized score obtained by the average linkage algorithm applied to the Robinson-Foulds similarity is shown on the first row of Table 4, and the quartet similarity on the second. The third row contains the generalized scores given by the second algorithm.

As can be seen, $\mathcal{W}^1$ is larger than $\mathcal{W}^9$, and the single consensus is better than the atomic one. But the single consensus tree score is greatly improved by the partition in 3 classes composed of {HPI, UR, DR}, {pabB, trpA, trpB, icd, PolB} and {putP}. It is compatible with the G-tree of Figure 5 in which {putP} cannot be separated. The best partition in 2 classes places {HPI, UR, DR, putP} apart from all the others ; its generalized score value (168) is greater than $\mathcal{W}^1$ but the optimal partition does not recognize class {UR, DR}. This closeness in Figure 5 may be due to the fact that UR and DR X-trees have a very low resolution, as do the whole set of bootstrap trees, since over 30 taxa, only 8 to 119 bipartitions can be observed.

Conclusion

We have described a simple and efficient method to decide if there is a single consensus between trees or not, and

Figure 5 The final tree on *E. coli* genes given by the *TreeOfTrees* method.
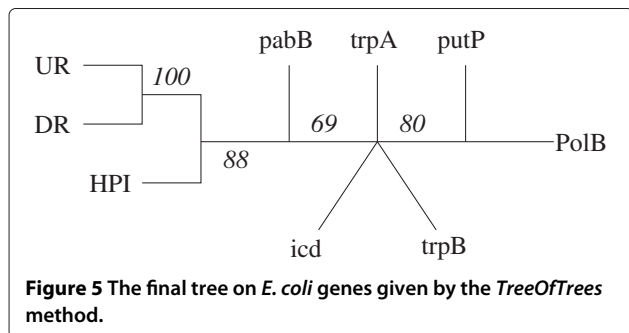
Table 4 Generalized scores of the *E. coli* genes for all possible numbers of classes

NbClas	1	2	3	4	5	6	7	8	9
AL(RF)	144	150	174	147	154	139	120	130	140
AL(QS)	144	150	135	159	169	136	146	129	140
GA	144	168	**182**	147	160	145	155	130	140

The two first rows correspond to the average linkage (*AL*) algorithm on both similarity indices and the third one corresponds to greedy algorithm (*GA*) merging the two classes maximizing the score function.

to establish a partitioning method that detects divergent genes. Applying a clustering hierarchical algorithm, the optimal partition is not certified. But it is sufficient to find a partition with a generalized score greater than $\mathcal{W}^1$ to assess the divergence of the profile and to search a decomposition in disjoint classes.

What remains, therefore, is to compare the consensus trees of classes in order to explain the divergence, suspected paralogy or possible transfers. More generally, the few, if any, discordant trees, can be removed to keep only genes that share the same evolutionary history and reflect the real tree of species. This method should also be extended to profiles made of trees connecting different taxa sets. The consensus tree notion must first be enlarged before combining trees connecting different subsets of X.

Endnote

[a]For instance, an X-tree computed from the sum of the unitary tree distances [19] which can be denoted as the average tree; the NJ-tree of this distance between trees of Figure 1 is identical to C_E.

References

1. de Vienne DM, Ollier S, Aguileta G: **Phylo-MCOA: a fast and efficient method to detect outlier genes an species in Phylogenomics using multiple co-inertia analysis.** *Mol Biol Evol* 2012, **29**(6):1587–1598.
2. Semple C, Steel M: *Phylogenetics*. Oxford: Oxford University Press; 2003.
3. Felsenstein J: *Inferring Phylogenies*. Sunderland: Sinauer Associates; 2002.
4. Felsenstein J: **Confidence-limits on phylogenies - an approach using the bootstrap.** *Evolution* 1985, **39**:783–791.
5. Bryant D: **A Classification of consensus methods for phylogenetics.** In *BioConsensus, DIMACS*. Edited by Janowitz M, Lapointe FJ, McMorris FR, Mirkin B, Roberts FS. Providence: AMS; 2003:163–184.
6. Garrey MR, Johnson DS: *Computers and Intractability: A Guide to Theory of NP-Completeness*. San Francisco: W.H Freeman; 1979.
7. Barthélemy JP, Mc Morris, F R: **The median procedure for n-trees.** *J Classif* 1986, **3**:329–334.
8. Maddison DR: **The discovery and importance of multiple islands of most-parcimonious trees.** *Syst Zool* 1991, **21**:1098–1109.
9. Stockham C, Wang LS, Warnow T: **Statistically based postprocessing of phylogenetic analysis by clustering.** *Bioinformatics* 2002, **18**:S285—S293.
10. Bonnard C, Berry V, Lartillot N: **Multipolar consensus for phylogenetic trees.** *Syst Biol* 2006, **55**(5):837–843.
11. Koperwas J, Walczak K: **Tree edit distance for leaf-labelled trees on free leafset and its comparison with frequent subsplit dissimilaritiy and popular distance measure.** *BMC Bioinformatics* 2011, **12**:204.
12. Robinson DF, Foulds LR: **Comparison of phylogenetic trees.** *Math Biosc* 1981, **53**:131–147.
13. Estabrook GF, McMorris FR, Meacham CA: **Comparison of undirected phylogenetic trees based on subtrees of 4 evolutionary units.** *Syst Zool* 1985, **34**:193–200.
14. Kimura M: **A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.** *J Mol Evol* 1980, **16**:111–120.
15. Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ: **Universal trees based on large combined protein sequence data sets.** *Nat Genet* 2001, **28**:281–285.
16. Darlu P, Guénoche A: **The TreeOfTrees method to evaluate the congruence between gene trees.** *J. of Classification* 2011, **28**(3):390–403.
17. Schubert S, Darlu P, Clermont O, et al.: **Role of intraspecies recombination in the spead of pathogenicity islands within the Escherichia coli species.** *PLoSpathogens* 2009, **5**(1):e1000257.
18. Guindon S, Gascuel O: **A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52**:696–704.
19. Lapointe JF, Cucumel G: **The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa.** *Syst Biol* 1997, **462**:306–312.