Makarenkov, V., A. Boc and P. Legendre. 2014. A new algorithm for inferring hybridization events based on the detection of horizontal gene transfers.

Pp. 273–293 *in*: F. Aleskerov, B. Goldengorin and P. M. Pardalos [eds.] *Clusters, Orders, and Trees: Methods and Applications*. Springer Optimization and Its Applications, 92. Springer-Verlag, New York.

ISBN 978-1-4939-0742-7; doi: 10.1007/978-1-4939-0742-7

A new algorithm for inferring hybridization events based on the detection of horizontal gene transfers

Vladimir Makarenkov, Alix Boc, and Pierre Legendre

Abstract Hybridization and horizontal gene transfer are two major mechanisms of reticulate evolution. Both of them allows for a creation of new species by recombining genes or chromosomes of the existing organisms. An effective detection of hybridization events and estimation of their evolutionary significance have been recognized as one of the main hurdles of the modern computational biology. In this article, we underline common features characterizing horizontal gene transfer and hybridization phenomena and describe a new algorithm for the inference and validation of the diploid hybridization events, when the newly created hybrid has the same number of chromosomes as the parent species. A simulation study was carried out to examine the ability of the proposed algorithm to infer correct hybrids and their parents in various practical situations.

1 Introduction

Horizontal gene transfer and hybridization, which are often followed by genetic or chromosomic recombination, have been recognized as major forces contributing to the formation of new species. Both of these evolutionary mechanisms are important parts of the reticulate evolution phenomenon which requires a network topology for its correct graphical representation. Phylogenetic networks are a generalisation

Vladimir Makarenkov
Département d'Informatique, Université du Québec à Montréal, C.P.8888, succursale Centre Ville, Montréal, QC, Canada, H3C 3P8, e-mail: makarenkov.vladimir@uqam.ca

Alix Boc
Université de Montréal, C.P. 6128, succursale Centre-ville Montréal, QC, Canada, H3C 3J7, e-mail: alix.boc@umontreal.ca

Pierre Legendre
Université de Montréal, C.P. 6128, succursale Centre-ville Montréal, QC, Canada, H3C 3J7, e-mail: pierre.legendre@umontreal.ca

of phylogenetic (or additive) trees which have been systematically used in biological and bioinformatics studies since the publication of Darwin's *On the Origin of Species by Means of Natural Selection* [10] in order to represent the process of species evolution. Phylogenetic trees and networks are usually reconstructed according to similarities and differences between genetic or morphological characteristics of the observed species (i.e. taxa or objects). The tree reconstruction can rely either on distance-based methods [36] or on character-based methods [18]. When distance-based methods are considered, the tree building process is usually twofold: the distances are first estimated from character data and a tree is then inferred from the distance estimates. The character-based methods assume that genetic sequences evolve from a common ancestor by a process of mutation and selection without mixing (e.g. without horizontal gene transfer or hybridization events).

However, phylogenetic trees cannot be used to represent complex reticulate evolutionary mechanisms such as hybridization, horizontal gene transfer, recombination, or gene duplication followed by gene loss. Phylogenetic networks become the models of choice when reticulation events have influenced species evolution [19, 20]. One example of phylogenetic networks is a reticulogram, i.e. reticulated cladogram, which is an undirected connected graph capable of retracing reticulate evolutionary patterns existing among the given organisms [24]. Since their introduction in 2002, reticulograms have been used to portray a variety of phylogenetic and biogeographic mechanisms, including hybridization, microevolution of local populations within a species and historical biogeography dispersion events [24, 27].

Horizontal gene transfer (HGT), which is also called lateral gene transfer, is one of the main mechanisms contributing to the diversification of microbial genomes. HGT consists of a direct transfer of genetic material from one lineage to another. Bacteria and viruses have developed complex mechanisms of the acquisition of new genes by HGT to better adapt to changing environmental conditions [11, 41]. Two main HGT detection approaches exist in the literature. First of them proceeds by sequence analysis of the host genome in order to identify the genomic fragments with atypical GC content or codon usage patterns [23]. The second approach compares a morphology-based species tree, or a molecular tree inferred from a molecule which is supposed to be unaffected by horizontal gene transfer (e.g. 16S rRNA), with a phylogeny of the considered gene. When bacterial or viral data are examined, the observed topological differences between two trees can be often explained by HGT. The second approach comprises numerous heuristic algorithms, including the network-based models introduced by Hein [16], von Haeseler and Churchill [15], and Page and Charleston [32, 33]. Mirkin et al. [29] described a tree reconciliation method for integrating different gene trees into a unique species phylogeny. Maddison [26], and then Page and Charleston [33], were first to present the set of evolutionary constraints that should be satisfied when inferring HGT events. Several recently proposed methods deal with the approximation of the Subtree Prune and Regraft (SPR) distance which is used to estimate the minimum possible number of HGTs. Bordewich and Semple [8] showed that computing the SPR distance between two rooted binary trees is NP-hard. A model allowing for mapping several gene trees into a species tree was introduced by Hallett and Lagergren ([14], Lat-

Trans algorithm). Mirkin et al. [30] described an algorithm for the reconciliation of phyletic patterns with a species tree by simultaneously considering gene loss, gene emergence and gene transfer events. Mirkin et al. [30] showed that in each situation their algorithm, which can be seen as one of the main references in this field, provided a parsimonious evolutionary scenario for mapping gene loss and gain events into a species phylogenetic tree. Nakhleh et al. [31] and Than and Nakhleh [38] put forward the RIATA-HGT heuristic based on the divide-and-conquer approach. Boc et al. [5] introduced a new horizontal gene transfer inference algorithm, HGT-Detection, and showed that it is considerably faster than the exhaustive HGT detection strategy implemented in LatTrans, while being identical in terms of accuracy. HGT-Detection was also proved to be faster and generally more reliable than RIATA-HGT. The HGT-Detection algorithm will be considered as a backbone procedure for the hybrid detection technique that we introduce in this article.

Hybridization is another major process of reticulate evolution [2]. It is very common among plants, fish, amphibians and reptiles, and is rather rare among other groups of species, including birds, mammals and most arthropods [28]. The new species is created by the process of recombination of genomes of different parent species. When the new species have the same number of chromosomes as its parents, the process is called *diploid hybridization*. When the new species has the sum of the number of the parent's chromosomes, the process is called *polyploid hybridization*. In this study, we will assume that new species have been created by the process of diploid hybridization. Most of the hypotheses and conclusions about hybridization rely on morphological data, and in many situations, these hypotheses have not been rigorously tested by simulations [21]. The majority of the works addressing the issue of the hybrids detection aim at calculating the minimal number of hybridization events that are necessary to reconcile the given tree topologies [3, 8]. Some of them proceed by estimating the SPR distance between a pair of rooted trees [1, 39, 40]. The main drawback of these methods is that most of them can deal only with a small number of hybrids and none of them offers the possibility of a statistical validation of the obtained hybridization events.

In this article, we propose a new algorithm for inferring a minimum number of statistically validated hybridization events that are necessary to reconcile the set of gene trees belonging to different parents (i.e. male and female gene trees) under the hypothesis of diploid hybridization. The new method will use the common features characterizing horizontal gene transfer and hybridization processes by separating the task of detecting hybridization events into several sub-tasks, each of which could be tackled by solving an equivalent horizontal gene transfer detection problem. A statistical validation procedure allowing one to assess the bootstrap support of the proposed hybrids and their parents will be incorporated into the new algorithm. A simulation study along with an application example will be also presented in the article.

2 Definitions and basic concepts

This section recalls some basic definitions concerning phylogenetic trees and tree metrics following the terminology of Barthélemy and Guénoche [4]. The distance $\delta(x\ y)$ between two vertices x and y in a phylogenetic tree T is defined as the sum of the edge lengths of the unique path connecting x and y in T. Such a path is denoted (x,y). A leaf is a vertex of degree one.

Definition 1. Let X be a finite set of n taxa. A dissimilarity d on X is a non-negative function on $(X \mathrm{x} X)$ such that for any x, y from X:
(1) $d(x,y) = d(y,x)$, and
(2) $d(x,y) = d(y,x) \geq d(x,x) = 0$.

Definition 2. A dissimilarity d on X satisfies the four-point condition if for any x, y, z, and w from X:

$$d(x,y) + d(z,w) \leq Max\{d(x,z) + d(y,w); d(x,w) + d(y,z)\}. \quad (1)$$

Definition 3. For a finite set X, a phylogenetic tree (i.e. an additive tree or an X-tree) is an ordered pair $(T,\ \varphi)$ consisting of a tree T, with vertex set V, and a map φ: $X \rightarrow V$ with the property that, for all $x \in X$ with degree at most two, $x \in \varphi(X)$. A phylogenetic tree is called binary if φ is a bijection from X into the leaf set of T and every interior vertex has degree three. The main theorem linking the four-point condition and phylogenetic trees (i.e., phylogenies) is as follows:

Theorem 1. *(Zarestskii, Buneman, Patrinos, Hakimi and Dobson)*
Any dissimilarity satisfying the four-point condition can be represented by a phylogenetic tree such that for any x, y from X, $d(x,y)$ is equal to the length of the path linking the leaves x and y in T. This dissimilarity is called a tree metric. Furthermore, this tree is unique.

Figure 1 presents an example of a tree metric on the set X of five taxa and the corresponding phylogenetic tree. Note that raw biological data rarely give rise directly to a tree metric (i.e. to a phylogenetic tree) but rather to a dissimilarity not satisfying the four-point condition. Biologists have to infer tree metrics and the corresponding trees by fitting the given dissimilarity with a tree metric according to a specific criterion.

	x2	x3	x4	x5
x1	6	5	4	6
x2		5	6	4
x3			5	5
x4				6

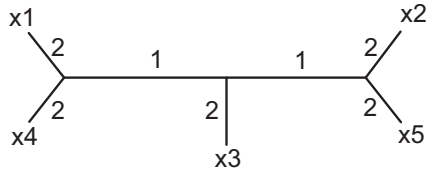

Fig. 1 A tree metric on the set X of 5 taxa and the associated phylogenetic tree with 5 leaves.

3 Horizontal gene transfer detection problem and related optimization criteria

The problem of finding the minimum number of horizontal gene transfers that are necessary to transform one phylogenetic tree into another (i.e. also known as *Subtree Transfer Problem*) has been shown to be NP-hard [17]. Here we recall the main features of the HGT-Detection algorithm [5] intended for inferring horizontal gene transfer events. This algorithm proceeds by a progressive reconciliation of the given species and gene phylogenetic trees, denoted T and T', respectively. Usually, the species tree T is inferred from the gene that is refractory to horizontal gene transfer and genetic recombination. This tree represents the direct, or tree-like, evolution. The gene tree T' represents the evolution of the given gene which is supposed to undergo horizontal transfers.

At each step of the algorithm, all pairs of edges of the species tree T are tested against the hypothesis that a horizontal gene transfer has occurred between them. Thus, the original species phylogenetic tree T is progressively transformed into the gene phylogenetic tree T' via a series of SPR moves (i.e. gene transfers). The topology of the gene tree T' is fixed throughout the transformation process. The goal of the method is to find the minimum possible sequence of trees T, T_1, T_2, ..., T' transforming T into T'. Obviously, a number of necessary biological rules should be taken into account. For example, the transfers within the same lineage should be prohibited [14, 26, 33]. The subtree constraint we consider here (see Appendix A) allows us to take into account all necessary evolutionary rules.

We will consider the four following optimization criteria which can be used to select optimal transfers at each step of the algorithm: least-squares, the Robinson and Foulds topological distance, the quartet distance and the bipartition dissimilarity. The first employed optimization criterion is the *least-squares function Q*. It is defined as follows:

$$Q = \sum_i \sum_j (d(i,j) - \delta(i,j))^2, \tag{2}$$

where $d(i,j)$ is the distance between the leaves i and j in the species tree T at the first step of the algorithm (or the transformed species trees at the following steps of the algorithm) and $\delta(i,j)$ is the distance between the leaves i and j in the gene tree T'.

The second criterion we use in the transfer detection part of our algorithm is the *Robinson and Foulds* (*RF*) *topological distance*. The RF metric [34] is an important and frequently used tool for comparing phylogenetic trees. This distance is equal to the minimum number of elementary operations, consisting of merging and splitting nodes, which are necessary to transform one tree into the other. This distance is twice the number of bipartitions present in one of the trees and absent in the other. When the RF distance is considered, we use it as the optimization criterion in the following way: all possible transformations of the species tree, consisting of SPR moves of its subtrees satisfying the biological constraints, are evaluated in such

a way that the RF distance between the transformed species tree T_1 and the gene tree T' is computed. The subtree transfer yielding the minimum of the RF distance between T_1 and T' is then selected.

The third considered criterion is the *quartet distance* (*QD*). QD is the number of quartets, or subtrees induced by four leaves, which differ between the compared trees. We can use this criterion in the same way that the RF metric.

The fourth optimization criterion is the *bipartition dissimilarity* (*BD*), first defined in Boc et al. [5]. Assume that T and T' are binary phylogenetic trees on the same set of leaves. A bipartition vector (i.e. split or bipartition) of the tree T is a binary vector induced by an internal edge of T. Let **BT** be the bipartition table of the internal edges of T and **BT'** be the bipartition table of the internal edges of T'. The bipartition dissimilarity *bd* between T and T' is defined as follows:

$$bd = (\sum_{a\in BT} \underset{b\in BT'}{Min}(Min(d(a,b);d(a,\bar{b}))) + \sum_{b\in BT'} \underset{a\in BT}{Min}(Min(d(b,a);d(b,\bar{a}))))/2, \quad (3)$$

where $d(a,b)$ is the Hamming distance between the bipartition vectors a and b ($\bar{a}$ and $\bar{b}$ are the complements of a and b, respectively). The bipartition dissimilarity can be seen as a refinement of the RF metric which takes into account only the identical bipartitions. For example, the bipartition dissimilarity between the trees T and T' with 6 leaves presented in Figure 2 is computed as follows: $bd(T,T) = ((0+1+1)+(0+1+2))/2 = 2.5$.

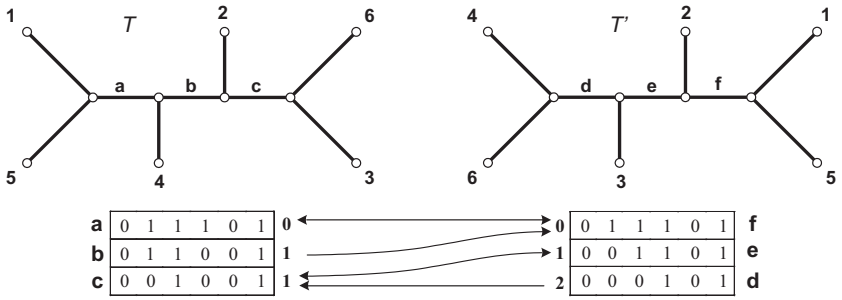

Fig. 2 Trees T and T' and their bipartition tables. Each row of the bipartition table corresponds to an internal edge of the tree. Arrows indicate the correspondence between the bipartition vectors in the two tables. Value in bold near each vector indicates the corresponding distance.

In our simulation study described below we presented the results obtained using the bipartition dissimilarity as the optimization criterion because it provided the best overall simulation performances compared to the RF and QD distances and least-squares.

4 Algorithm description

In this section we describe the main features of the new algorithm for detecting hybridization events. The statistical bootstrap validation will be performed for each hybrid species and only the hybrids with a significant bootstrap support will be included in the final solution. The new algorithm for identifying hybridization events proceeds by a reconciliation of the given pairs of gene trees, constructed for genes inherited from different parents. A modified version of the procedure for detecting horizontal gene transfers described in Boc et al. [5] will be integrated in our new algorithm. Let $\mathbf{G}_m$ be the set of genes that can be inherited from a male parent only and $\mathbf{G}_f$ be the set of the genes that can be inherited from a female parent only. In practice, nuclear and chloroplast genes often play the roles of $\mathbf{G}_m$ and $\mathbf{G}_f$, respectively. We assume that for each given gene there exists a set of orthologous gene sequences (i.e. sequences that originated from a single gene of the last common ancestor) that can be used to build a phylogenetic gene tree. Each gene is thus originally represented by a multiple sequence alignment of amino acids or nucleotides.

Step 1. For the multiple sequence alignments characterizing the male genes in $\mathbf{G}_m$ we infer a set of phylogenetic male gene trees $\mathbf{T}_m$ and for the multiple sequence alignments characterizing the female genes in $\mathbf{G}_f$ we infer a set of phylogenetic female gene trees $\mathbf{T}_f$; one gene tree by alignment is reconstructed. The trees can be inferred using methods such as Neighbor-Joining [35], PhyML [13], RaxML [37] or one of the phylogenetic inference algorithms from the PHYLIP package [12]. We then root all the trees in $\mathbf{T}_m$ and $\mathbf{T}_f$ according to biological evidence or using the outgroup or midpoint strategy and select the optimization criterion, which can be least-squares, the Robinson and Foulds topological distance [34], the quartet distance or the bipartition dissimilarity [5].

Step 2. For each pair of gene trees T and T', such that $T \in \mathbf{T}_m$ and $T' \in \mathbf{T}_f$, we use the HGT-Detection algorithm [5] to identify first horizontal gene transfers that are required to transform T into T'. The HGT-Detection program carries out a fast and accurate heuristic algorithm for computing a minimum-cost SPR transformation of the given (species) tree T into the given (gene) tree T'. Figure 3 shows how a species tree is transformed into a gene tree by applying a transfer (SPR move) between it's subtrees (i.e. edges adjacent to the species C and E). After this SPR move, T and T' have the identical topology. Second, we repeat the procedure by inversing the roles of T and T'. Now we look for horizontal gene transfers that are necessary to transform T' into T. The statistical bootstrap support of each obtained transfer is then assessed as defined in Boc et al. [5] and Boc and Makarenkov [6]. We identify as potential hybrids the species that receive transfers from different parents in T and T' (e.g. species H in Figure 4 which receives a transfer from the species C and B; here, C and B can be viewed as the parents of H).

Final step. All the obtained horizontal transfers are classified according to their statistical support to establish a ranked list of predicted hybrid species and their parents. In our algorithm, a confirmed hybrid species is a species that receives a transfer stemming from different parents in at least two gene trees (such that at least one of them is from $\mathbf{T}_m$ and at least one of them is from $\mathbf{T}_f$) with a fixed minimum

confidence score (i.e. average bootstrap support). When multiple trees from $\mathbf{T}_m$ and $\mathbf{T}_f$ are involved, this score is computed as the mean value of the average bootstrap scores found for the two groups of parents. If the genes trees are considered without uncertainties (i.e. no bootstrap validation performed), then all hybrid species found by the algorithm can be included in the final solution. The main steps of the new algorithm are presented below (see Algorithm 1). Its time complexity is the following:

$$O(m \times f \times r \times (C(Tree\_Inf) + n^4)), \tag{4}$$

where m and f are the cardinalities of the sets $\mathbf{G}_m$ and $\mathbf{G}_f$, respectively, r is the number of replicates in bootstrapping, $C(Tree\_Inf)$ is the time complexity of the tree inferring method used to infer trees from the gene sequences and n^4 is the time complexity of the HGT-Detection algorithm [5] applied to the given pair of species and gene trees with n leaves. Given that the time complexity of the PhyML [13] method which we used in our simulation study is $O(pnl)$, where p is the number of refinement steps being performed, n is the number of species and l is the sequence length, the exact time complexity of our implementation is the following:

$$O(m \times f \times r \times n \times (p \times l + n^3)). \tag{5}$$

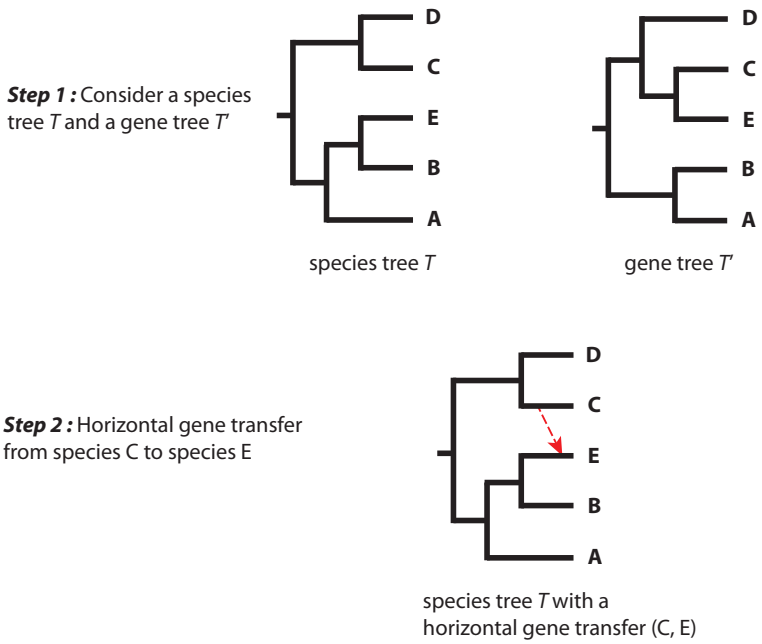

Fig. 3 Horizontal gene transfer (i.e. SPR move) from species C to species E is necessary to transform the topology of the species tree T into the topology of the gene tree T'.

Algorithm 1 Mains steps of the hybrids detection algorithm. See Appendix A for the definition of the subtree constraint which allows one to take into account all necessary biological rules and for Theorems 2 and 3 which allow one to select optimal transfers in different practical situations.

Infer all gene trees $\mathbf{T}_m$ for the set of the male genes $\mathbf{G}_m$ and all gene trees $\mathbf{T}_f$ for the set of the female genes $\mathbf{G}_f$;
Root all the trees in $\mathbf{T}_m$ and $\mathbf{T}_f$ according to biological evidence or using the outgroup or midpoint strategy;
Select the optimization criterion $OC = Q$ (least-squares), or RF (Robinson and Foulds distance), or QD (quartet distance), or BD (bipartition dissimilarity);

for each tree T from the set of the male gene trees $\mathbf{T}_m$ **do**
  for each tree T' from the set of the female gene trees $\mathbf{T}_f$ **do**
    if there exist identical subtrees with two or more leaves in T and T' **then**
      Decrease the size of the problem by collapsing them in both T and T';
    end if
    Compute the initial value of OC between T_0 and T';
    (*) $T_0 = T$; // or $T_0 = T$' - when repeated
    $k = 1$; // k is the step index

    while $OC \neq 0$ **do**
      Find the set of all eligible horizontal transfers (i.e., SPR moves) at step k (denoted as E_HT_k);
      The set E_HT_k contains only the transfers satisfying the subtree constraint;
      while transfers satisfying the conditions of Theorems 3 and 2 exist **do**
        if there exist transfers $\in E\_HT_k$ and satisfying the conditions of Theorem 3 **then**
          Carry out the SPR moves corresponding to these transfers;
        end if
        if there exist transfers $\in E\_HT_k$ and satisfying the conditions of Theorem 2 **then**
          Carry out the SPR moves corresponding to these transfers;
        end if
      end while
      Carry out all remaining SPR moves corresponding to transfers satisfying the subtree constraint;
      Compute the value of OC to identify the direction of each transfer;
      $k = k + 1$;
      Collapse the same subtrees in T_k and T'; // or in T_k and T - when repeated
      Compute the value of OC between T_k and T'; // or between T_k and T - when repeated
    end while

    Repeat the procedure above by inversing the roles of T and T', starting from (*);
    Identify species (potential hybrids) such that they receive transfers from different species in T and T';
  end for
end for

Classify all horizontal transfers and potential hybrids found;
Repeat the procedure above twice using the replicates of T and T' (obtained from the replicates of the multiple sequence alignments corresponding to T and T') to establish the list of predicted hybrid species and their parents with their bootstrap support.

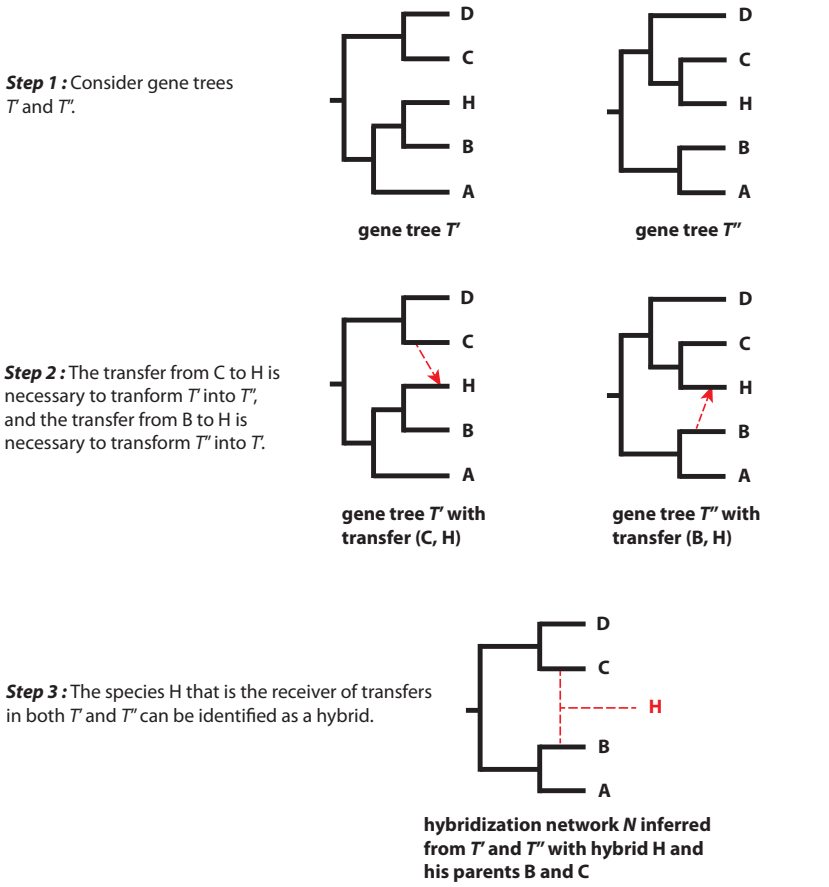

Fig. 4 The main idea of the hybrid inference method: the species H that receives transfers (Step 2) in both gene trees T' and T'' is identified as a hybrid. The hybridization network N is thus obtained (Step 3).

5 Simulation study

A Monte Carlo study was conducted to test the capacity of the new algorithm to identify correct hybrid species. Gene trees were assumed not to contain uncertainties and thus the simulations were carried out with tree-like data only (i.e. sequence data were not involved). We examined how the new algorithm performs depending on the number of observed species, the rate of hybridization and the number of hybrid species artificially added. The measured hybridization rate is the ratio of genes originating from the different parents (i.e. male and female species).

First, a binary gene tree T was generated using the random tree generation procedure described in [22]. An improved version of this procedure was included in our T-Rex package [7]. As we did not consider sequence data in these simulations, the edge lengths of the trees were not taken into account here. In each experiment, we considered 10 replicates of the gene tree T, assuming that some of them originated from the male and some of them from the female parent species.

Second, for a fixed hybridization rate h (h varied from 1 to 5 in our simulations) we randomly selected in the first h replicates of T the same species (or group of species) as Parent P_1 and in the remaining (10-h) replicates of T another species (or group of species) as Parent P_2. Obviously, when the groups were considered, all the species in P_1 were different from the species in P_2. A new edge with the hybrid species H was then added to each of the first h gene trees. It was connected to the edge separating P_1 from the rest the tree. Similarly, the edge with the same hybrid species H was added to each of the remaining (10-h) gene trees, and connected each time to the edge separating P_2 from the rest of the tree. This step was repeated sh times, where sh denotes the number of integrated hybrid species. In our simulations, sh varied from 1 to 10.

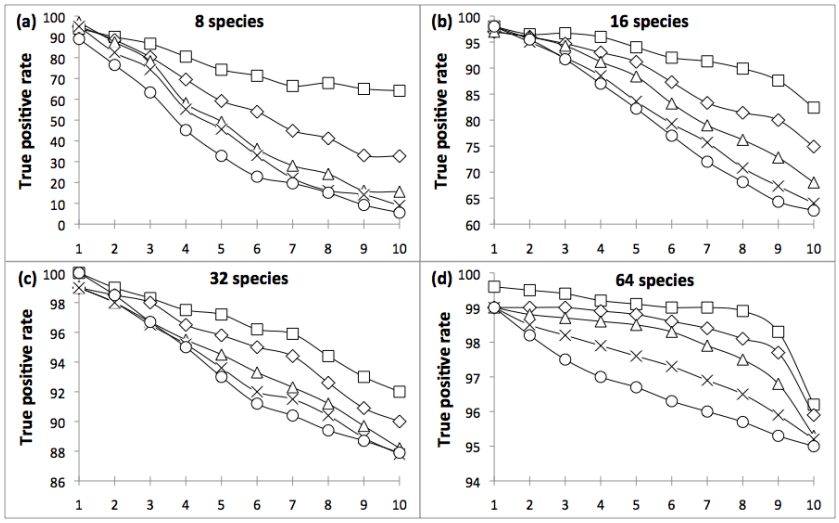

Fig. 5 Average true positive hybrid detection rate obtained for binary trees with 8 (a), 16 (b), 32 (c) and 64 (d) leaves. The five presented curves correspond to the hybridization rate h of 10% (□), 20% (◊), 30% (△), 40%(×) and 50% (○). The abscissa axis reports the number of hybrid species. Each presented value is an average computed over 1000 replicates.

Third, we carried out the introduced hybrid detection algorithm having as input 10 replicates of the gene tree T with the hybrids added as discussed above. As the gene trees were considered without uncertainties, all transfers detected in the process were considered as relevant and were taken into account in the final solution. The bipartition dissimilarity [5] was used as the optimization criterion in the HGT-

Detection procedure. The results illustrated in Figures 5 and 6 were obtained from simulations carried out with random binary phylogenetic trees with 8, 16, 32 and 64 leaves. For each tree size (8 to 64), each number of hybrid species (1 to 10) and each hybridization rate (10 to 50%), 1000 replicated data sets were generated.

The true detection rate (i.e. true positives) was measured as a percentage of the correctly recovered hybrid species that were generated. The performances of the new algorithm are more noticeable for large trees (see Figures 5 and 6, cases c-d) and a small number of hybrids. The quality of the obtained results decreases when the number of species decreases. For instance, to detect 10 hybrids in trees with 8 possible parental species seems to be a very tricky task, especially when the hybridization rate varies from 30 to 50% (i.e. h = 3, 4 and 5; see Figures 5 and 6, case a). Another general trend that could be noticed is that the number of true positives increases and the number of false positives decreases as the hybridization rate declines (i.e. the best results were always observed for h = 1 and 2).

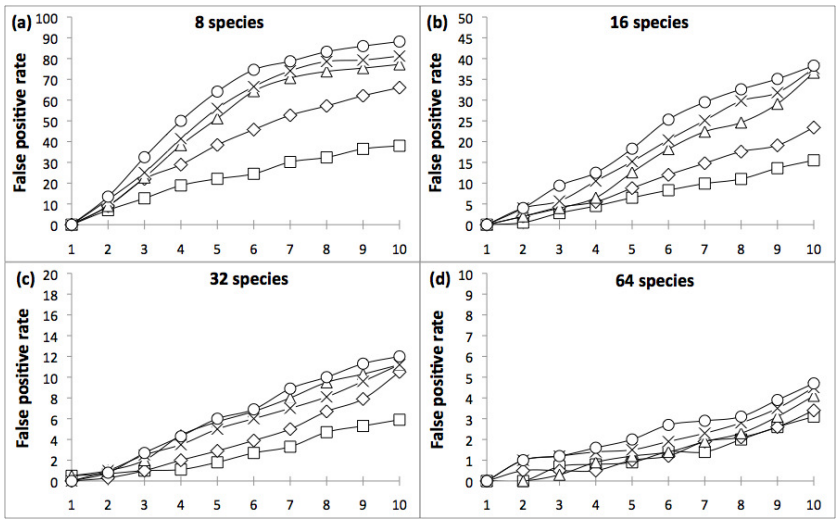

Fig. 6 Average false positive hybrid detection rate obtained for binary trees with 8 (a), 16 (b), 32 (c) and 64 (d) leaves. The five presented curves correspond to the hybridization rate h of 10% (□), 20% (◊), 30% (△), 40% (×) and 50% (○). The abscissa axis reports the number of hybrid species. Each presented value is an average computed over 1000 replicates.

6 Application example

Detecting hybrid species in the New Zealand's alpine Ranunculus dataset

We studied the evolution of 6 different genes belonging to 14 organisms of the alpine *Ranunculus* plants originally described in Lockhart et al. [25], and then analysed in Joly et al. [21]. The latter authors presented a novel parametric approach for statistically distinguishing hybridization from incomplete lineage sorting based on minimum genetic distances of nonrecombining genes. Joly and colleagues applied their method to detect hybrids among the New Zealand's alpine buttercups (*Ranunculus*). Fourteen individuals of *Ranunculus* belonging to six well-defined species were sequenced in five chloroplast regions (*trnC-trnD*, *trnL-trnF*, *psbA-trnH*, *trnD-trnT* and *rpL*16). Those sequences were concatenated in the analysis conducted by Joly et al. [21]. In this study, they will be analyzed separately using our new algorithm. Note that in most flowering plants, chloroplast genes are inherited by hybrids from the female parent only. In contrast, the sequences from another considered gene, the internal transcribed spacer (*nrITS*) region, were assumed to be inherited from the male parent only.

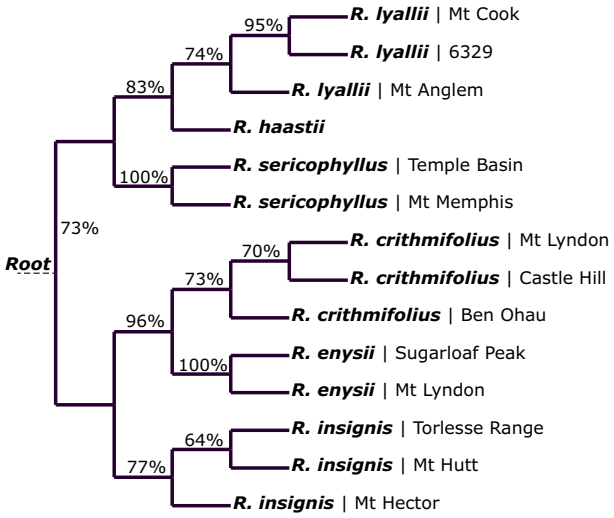

Fig. 7 Phylogenetic tree of the gene *nrITS* built for 14 organisms of alpine *Ranunculus* using the PhyML method [13]. The bootstrap scores of the internal edges of the tree are indicated.

We first reconstructed from the original sequences the topology of the *nrITS* gene tree (Figure 7) as well as those of the *psbA*, *rpL*16, *trnC*, *trnD* and *trnL* gene trees (Figure 8).

The hybrid species detection was performed by the new algorithm and 5 possible hybrid species were identified (see Table 1) along with their parents and the

Fig. 8 Phylogenetic trees of the genes *psbA*, *rpL*16, *trnC*, *trnD* and *trnL* built for 14 organisms of alpine *Ranunculus* using the PhyML method [13]. The bootstrap scores of the internal edges of the tree are indicated.

corresponding bootstrap scores. All transfers found, when gradually reconciling the *nrITS* gene tree with the *psbA*, *rpL*16, *trnC*, *trnD* and *trnL* gene trees, are illustrated in Figure 9. As a backbone tree topology here we used the species tree built with respect to the species chronogram of the alpine *Ranunculus* presented in ([21], Fig. 5). The most significant hybrid species we found was the *R. insignis* Mt Hutt. The species *R. crithmifolius* Ben Ohau and *R. crithmifolius* Mt Lyndon were identified as its parents with the bootstrap scores of 76% and 75%, respectively. Thus, the bootstrap support of this hybrid, computed as the average of its parents bootstrap scores, is equal to 75.5%.

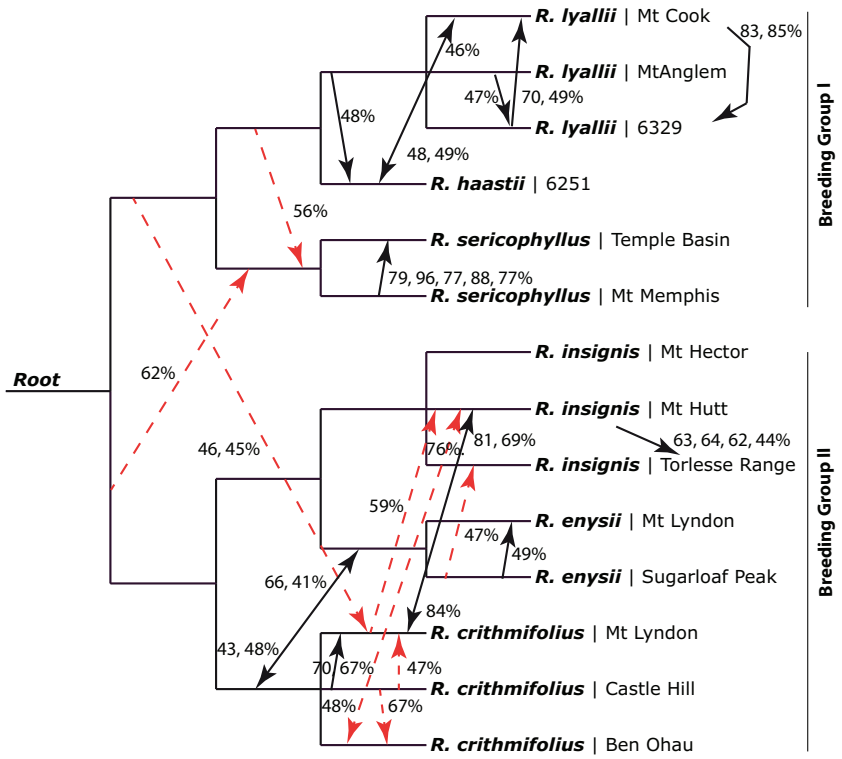

Fig. 9 Species tree for the 14 considered *Ranunculus* organisms with horizontal transfers mapped into it. Dashed arrows depict the transfers stemming from the gene *nrITS*. Full arrows depict the transfers stemming from the genes *psbA*, *rpL*16, *trnC*, *trnD* and *trnL*. A potential hybrid species should be a receiver of at least one dashed arrow and at least one full arrow originating from different sources.

Our algorithm also suggested multiple hypotheses for an eventual hybrid species *R. crithmifolius* Mt Lyndon. The first hypothesis assumes that its parents could be *R. crithmifolius* Castle Hill (47%) and *R. insignis* Mt Hutt (84%), combining for the average bootstrap support of 65.5%. The second hypothesis suggests that its parents could be the ancestor of *R. haastii*, *R. lyallii 6329*, *R. lyallii* Mt Anglem, *R. lyallii* Mt Cook, *R. sericophyllus* Mt Memphis and *R. sericophyllus* Temple Basin as the first parent, with the bootstrap of 45.5%, and *R. insignis* Mt Hutt as the second parent, with the bootstrap support of 84%, providing the average support of 64.5%. The third hypothesis concerning *R. crithmifolius* Mt Lyndon states that the parents of this organism could be in fact the ancestor of *R. haastii*, *R. lyallii 6329*, *R. lyallii* Mt Anglem, *R. lyallii* Mt Cook, *R. sericophyllus* Mt Memphis and *R. sericophyllus* Temple Basin, with the bootstrap score of 45.5%, and the species *R. crithmifolius* Castle Hill (47%), giving the average bootstrap support of 46%. As discussed in [21], hybridization is a likely hypothesis for the chloroplast lineage present in *R. crithmifolius* from Mt Lyndon and *R. insignis* from Mt Hutt. Our analysis supported both

these hypotheses while suggesting an additional hybrid possibility in this dataset, concerning *R. insignis* Torlesse Range (see Table 1). The latter species was also identified as a potential hybrid with the bootstrap support of 52.5%, whereas *R. insignis* Mt Hutt (58%) and *R. enysii* Sugarloaf Peak (47%) were categorized as its parents.

Table 1 Hypothetical hybrids of the considered alpine *Ranunculus* organisms based on the transfer scenarios presented in Figure 9. Each row reports the hybrid, two eventual parents and the corresponding bootstrap supports.

Hybride	*Parent 1*	*Parent 2*	*Average hybrid support*
R. insignis Mt Hutt	*R. crithmifolius* Ben Ohau (76%)	*R. crithmifolius* Mt Lyndon (75%)	75.5%
R. crithmifolius Mt Lyndon	*R. crithmifolius* Castle Hill (47%)	*R. insignis* Mt Hutt (84%)	65.5%
R. crithmifolius Mt Lyndon	Ancestor of (*R. haastii* , *R. lyallii 6329* , *R. lyallii* Mt Anglem, *R. lyallii* Mt Cook, *R. sericophyllus* Mt Memphis, *R. sericophyllus* Temple Basin) (45.5%)	*R. insignis* Mt Hutt (84%)	64.5%
R. insignis Torlesse Range	*R. insignis Mt Hutt (58%)*	*R. enysii* Sugarloaf Peak (47%)	52.5%
R. crithmifolius Mt Lyndon	Ancestor of (*R. haastii* , *R. lyallii 6329* , *R. lyallii* Mt Anglem, *R. lyallii* Mt Cook, *R. sericophyllus* Mt Memphis, *R. sericophyllus* Temple Basin) (45%)	*R. crithmifolius* Castle Hill (47%)	46%

7 Conclusion

We described a new algorithm for detecting and validating diploid hybridization events and thus for identifying the origins of hybrid species. To the best of our knowledge no algorithms including a statistical validation of the retraced hybrids and their parents by bootstrap analysis have been proposed in the literature. We showed that the problem of detecting horizontal gene transfers can be viewed as a sub-problem of a hybrid detection problem when multiple male and female genes are considered. The introduced algorithm subdivides the multi-gene reconciliation problem on several sub-problems searching for optimal scenarios of SPR moves that are required to reconcile gene trees associated with genes originating from different parents (male or female species). To find such optimal tree reconciliation scenarios, we use a specific version the HGT-Detection [5] algorithm, which is a fast and accurate heuristic for inferring horizontal gene transfer events. Our simulation study suggests that the best detection results are constantly obtained with large trees and

a small number of hybrids. Regarding the optimization criterion, the bipartition dissimilarity usually provided better results compared to the classical criteria, such as the Robinson and Foulds distance, the quartet distance and least-squares. As a future development, it would be interesting to see how the hybrid detection results would change if the trees with uncertainties (i.e. trees inferred from the sequence data) are be considered.

References

1. Albrecht, B., Scornavacca, C., Cenci, A., Huson, D.H.: Fast computation of minimum hybridization networks. Bioinformatics, **28**, 191–197 (2012)
2. Arnold ML.: Natural hybridization and evolution. Oxford: Oxford University Press (1997)
3. Baroni, M., Semple, C., Steel, M.: Hybrids in real time. Systematic Biology, **55(1)**, 46–56 (2006)
4. Barthélemy, J-P., Guénoche, A.: Trees and proximity representations. Wiley, New York (1991)
5. Boc, A., Philippe, H., Makarenkov,V.: Inferring and validating horizontal gene transfer events using bipartition dissimilarity. Syst. Biol., **59**, 195–211 (2010)
6. Boc, A., Makarenkov, V.: Towards an accurate identification of mosaic genes and partial horizontal gene transfers, Nucleic Acids Research, **39**, e144 (2011)
7. Boc, A., Diallo, Alpha B., Makarenkov, V.: T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks, Nucleic Acids Research, **40**, Web Server issue, W573–W579 (2012)
8. Bordewich, M., Semple, C.: On the computational complexity of the rooted subtree prune and regraft distance. Annals of Combinatorics, **8**, 409–423 (2004)
9. Charleston, M. A.: Jungle: a new solution to the host/parasite phylogeny reconciliation problem Math. Biosc., **149**, 191–223 (1998)
10. Darwin, C.: On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life, London: John Murray, pp. 502. (1859)
11. Doolittle,W.F. (1999) Phylogenetic classification and the universal tree. Science, **284**, 2124–2129 (1999)
12. Felsenstein, J.: PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics, **5**, 164–166 (1989)
13. Guindon, S., Gascuel, O.: A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52**, 696–704 (2003)
14. Hallett, M., Lagergren., J.: Efficient algorithms for lateral gene transfer problems. In: El-Mabrouk, N., Lengauer, T., Sankoff, D., (eds.), proceedings of the fifth annual international conference on research in computational biology, pp 149–156, ACM Press, New-York (2001)
15. von Haeseler, A., Churchill, G.A.: Network models for sequence evolution. J. of Mol. Evol., **37**, 77–85 (1993)
16. Hein, J.: A heuristic method to reconstructing the evolution of sequences subject to recombination using parsimony. Math. Biosciences, **98**, 185–200 (1990)
17. Hein, J., Jiang, T., Wang, L., Zhang, K.: On the Complexity of Comparing Evolutionary Trees. Discr. Appl. Math., **71**, 153–169 (1996)
18. Hennig, W.: Phylogenetic systematics (tr. D. Dwight Davis and Rainer Zangerl), University of Illinois Press. Urbana, Illinois (1966)
19. Huson, D.H., Bryant, D.: Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol., **23**, 254–267 (2006)
20. Huson, D.H., Rupp, R., Scornavacca, C.: Phylogenetic networks: Concepts, algorithms and applications. Cambridge University Press (2011)
21. Joly, S., McLenachan, P.A., Lockhart, P. J.: A statistical approach for distinguishing hybridization and incomplete lineage sorting. The American Naturalist., **174**, e54–e70 (2009)

22. Kuhner, M., Felsenstein, J.: A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol., **11**, 459–468 (1994)
23. Lawrence, J. G., Ochman, H.: Amelioration of bacterial genomes: rates of change and exchange. J. Mol. Evol., **44**, 383–397 (1997)
24. Legendre, P., Makarenkov, V.: Reconstruction of biogeographic and evolutionary networks using reticulograms. Syst. Biol., **51**, 199–216 (2002)
25. Lockhart, P. J., McLenachan, P. A., Havell, D., Gleny, D., Huson, D., Jensen, U.: Phylogeny, radiation, and transoceanic dispersal of New Zealand alpine buttercups: molecular evidence under split decomposition. Annals of the Missouri Botanical Garden **88**, 458–477 (2001)
26. Maddison, W. P.: Gene trees in species trees. Syst. Biol., **46**, 523–536 (1997)
27. Makarenkov, V., Legendre, P.: From a phylogenetic tree to a reticulated network. J. Comput. Biol., **11**, 195–212 (2004)
28. Makarenkov, V., Kevorkov, D., Legendre, P.: Phylogenetic Network Reconstruction Approaches, Applied Mycology and Biotechnology, International Elsevier Series, Bioinformatics, **6,** 61–97 (2006)
29. Mirkin, B. G., Muchnik, I., Smith, T.F.: A Biologically Consistent Model for Comparing Molecular Phylogenies. J. of Comp. Biol., **2**, 493–507 (1995)
30. Mirkin, B. G., Fenner, T. I., Galperin, M. Y., Koonin, E. V.: Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol. Biol., **3, 2**, (2003)
31. Nakhleh, L., Ruths, D., Wang., L.: RIATA-HGT: A fast and accurate heuristic for reconstructing horizontal gene transfer. Proceedings of the 11th International Computing and Combinatorics Conference, Kunming, Yunnan, China, 84–85 (2005)
32. Page, R. D. M.: Maps between trees and cladistic analysis of historical associations among genes, organism and areas. Syst. Biol., **43**, 58–77 (1994)
33. Page, R. D. M., Charleston, M. A.: Trees within trees: phylogeny and historical associations. Trends in Ecol. and Evol., **13**, 356–359 (1998)
34. Robinson, D. R., Foulds, L.R.: Comparison of phylogenetic trees. Math. Biosciences, **53**, 131–147 (1981)
35. Saitou, N., Nei., M.: The neighbor-joining method:a new method for reconstructing phylogenetic trees. Mol. Biol. Evol., **4**, 406–425 (1987)
36. Sneath, P.H.A., Sokal, R. R.: Numerical taxonomy The principles and practice of numerical classification. W. H. Freeman, San Francisco (1973)
37. Stamatakis, A.: RAxML-VI-HPC: Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. Bioinformatics, **22**, 2688–2690 (2006)
38. Than, C., D. Ruths, D., Nakhleh, L.: PhyloNet: A software package for analyzing and reconstructing reticulate evolutionary relationships, BMC Bioinf., **9**, 322 (2008)
39. Whidden, C., Zeh, N.: A unifying view on approximation and FPT of agreement forests. In Proceedings of WABI09, pages 390–402, Berlin, Heidelberg. Springer-Verlag (2009)
40. Whidden, C., Beiko, R. G., Zeh, N.: Fast FPT algorithms for computing rooted agreement forests: Theory and experiments. In P. Festa, editor, SEA, volume 6049 of Lecture Notes in Computer Science, pages 14–153. Springer (2010)
41. Zhaxybayeva, O., Lapierre,P., Gogarten, J.P.: Genome mosaicism and organismal lineages. Trends Genet., **20**, 254–260 (2004)

Appendix A

This Appendix includes the definition of the subtree constraint (Fig. A1) used in the hybrid detection algorithm (Algorithm 1). This constraint, originally formulated in [5], allows one to take into account all evolutionary rules that should be satisfied when inferring horizontal gene transfers. This Appendix also includes Theorems 2 and 3 allowing one to select optimal transfers during the execution of the hybrid detection algorithm (Algorithm 1) (see [5] for their proofs).

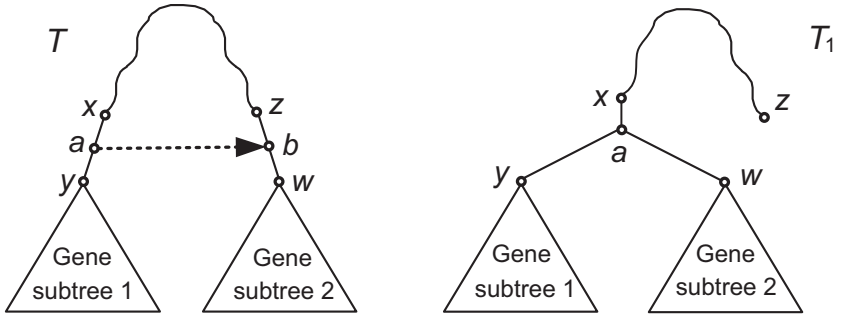

Fig. A1 Subtree constraint: the transfer between the branches (x,y) and (z,w) in the species tree T is allowed if and only if the cluster rooted by the branch (x,a), and regrouping both affected subtrees, is present in the gene tree. A single tree branch is depicted by a plane line and a path is depicted by a wavy line.

Theorem 2. *If the newly-formed subtree Sub_{yw} resulting from the HGT (horizontal gene transfer) is present in the gene tree T', and the bipartition vector associated with the branch (x,x1) in the transformed species tree T_1 (Fig. A2) is present in the bipartition table of T', then the HGT from (x,y) to (z,w), transforming T into T_1, is a part of a minimum-cost HGT scenario transforming T into T' and satisfying the subtree constraint.*

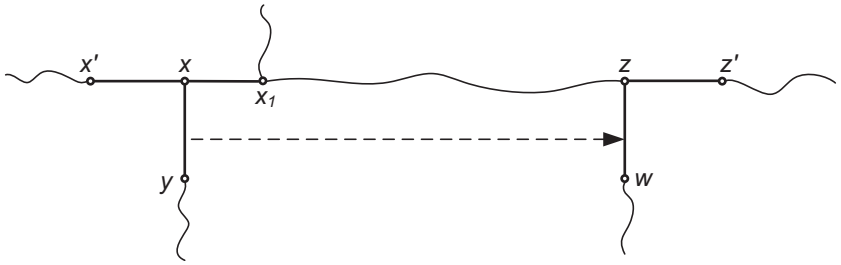

Fig. A2 HGT from the branch (x,y) to the branch (z,w) is a part of a minimum-cost HGT scenario transforming the species tree T into the gene tree T' if the bipartition corresponding to the branch (x,x_1) in the transformed species tree T_1 is present in the bipartition table of T' and the subtree Sub_{yw} is present in T'.

Theorem 3. *If the newly-formed subtree Sub_{yw} resulting from the HGT is present in the gene tree T', and all the bipartition vectors associated with the branches of the path (x',z') in the transformed species tree T_1 (Fig. A3) are present in the bipartition table of T', and the path (x',z') in T_1 consists of at least 3 branches, then the HGT from (x,y) to (z,w), transforming T into T_1, is a part of any minimum-cost HGT scenario transforming T into T' and satisfying the subtree constraint.*

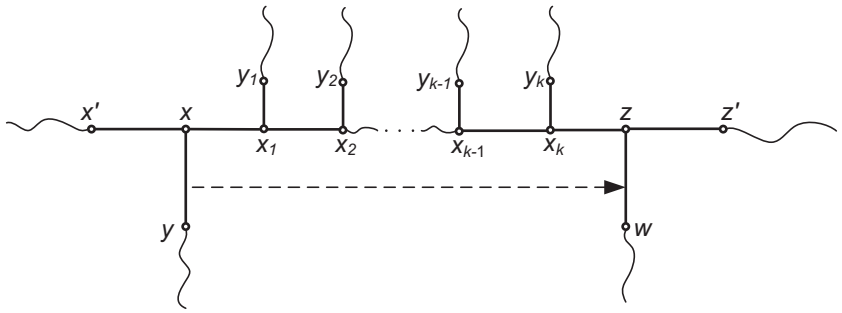

Fig. A3 HGT from the branch (x,y) to the branch (z,w) is a part of any minimum-cost HGT scenario transforming the species tree T into the gene tree T' if all the bipartitions corresponding to the branches of the path (x',z') in the transformed species tree T_1 are present in the bipartition table of T' and the subtree Sub_{yw} is present in the tree T'.