ABSTRACT
Motivation: Hybridization events in evolution may lead to incongruent gene trees. One approach to determining possible interspecific hybridization events is to compute a hybridization network that attempts to reconcile incongruent gene trees using a minimum number of hybridization events.

Results: We describe how to compute a representative set of minimum hybridization networks for two given bifurcating input trees, using a parallel algorithm and provide a user-friendly implementation. A simulation study suggests that our program performs significantly better than existing software on biologically relevant data. Finally, we demonstrate the application of such methods in the context of the evolution of the Aegilops/Triticum genera.

Availability and implementation: The algorithm is implemented in the program Dendroscope 3, which is freely available from www.dendroscope.org and runs on all three major operating systems.

Contact: scornava@informatik.uni-tuebingen.de;
huson@informatik.uni-tuebingen.de

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on July 13, 2011; revised on November 2, 2011; accepted on November 3, 2011

1 INTRODUCTION
Speciation by hybridization (Gross and Rieseberg, 2005; Mallet, 2007) is a widespread phenomenon not only in plants (Rieseberg et al., 2000; Soltis and Soltis, 2009), but also occurs in some other types of organisms (Giraud et al., 2008; Schwenk et al., 2008). When two individuals from distinct species hybridize and merge their sets of chromosomes, the hybrid organism is often sterile and does not produce any progeny. Differences among homologous chromosomes prevent correct meiotic pairing and viable gamete production. However, the eventual doubling of the chromosome number could restore a correct pairing (each chromosome pair with its double), and the fertile genotype could give rise to a new allopolyploid species. In this case, homologous chromosomes do not mix and the number of the genes is doubled. In other words, the genes from both parental species coexist in the polyploid genome and evolve independently.

In the case that the parental species are genetically similar enough, the pairing between homologous chromosomes is not completely prevented and balanced meiosis could take place. In this case, the hybrid genotype could produce viable hybrid gametes (containing portions of both parental chromosomes) and progeny. If the progeny remains reproductively isolated from the parental genotypes, it could give rise to a new species (homoploid hybrid speciation).

This implies that, in a hybrid species, different genes may have different evolutionary histories, in which case they will give rise to incongruent gene trees. Thus, using a single rooted phylogenetic tree to represent the evolutionary history of a set of taxa may be inaccurate in the presence of interspecific hybridization. A more precise description may be possible using a rooted phylogenetic network, in which internal tree nodes (nodes of indegree 1) represent putative speciation events, whereas reticulate nodes (nodes of indegree ≥ 2) represent possible hybridization events.

Suppose we are given a collection of species for which we suspect that hybridization events have played an important role in their evolution. One way to determine a set of possible hybridization events is to compute a hybridization network for a given set of gene trees that aims at explaining the incongruences between the different trees using a minimum number of putative hybridization events. A hybridization network for a set of trees T is simply a rooted phylogenetic network containing the trees in T. In computational terms, the problem can be formulated as follows: given a set T of two or more rooted phylogenetic trees, determine all hybridization networks for T that are minimum in the sense that they have a minimum reticulate number (see below). This problem is known to be a computationally hard problem even for the case of determining only one hybridization network for two bifurcating trees on the same set of taxa (Bordewich and Semple, 2007b).

In this article, we present an algorithm that takes as input two bifurcating, rooted phylogenetic trees $T_1$ and $T_2$ on the same taxon set $X$ and produces as output a representative set of minimum hybridization networks $N$ on $X$ that contain both trees. Such a representative set is defined to contain exactly one network derived from each of the possible maximum acyclic agreement forests (MAAF), as defined below. The number of reticulation nodes in any such minimum network is called the hybridization number, denoted by $h(T_1, T_2)$, for $T_1$ and $T_2$.

Our algorithm is based on previous work described in Baroni et al. (2006); Bordewich and Semple (2005, 2007a), which aims at computing the hybridization number, and on the work on the rooted SPR distance computation reported in Whidden and Zeh (2009); Whidden et al. (2010). We extend the published approach and provide a parallel implementation to compute a representative set of minimum hybridization networks containing one network per
MAAF. In this context, a number of theoretical issues arise and we will show how to address them in a forthcoming paper (Scornavacca et al., 2011).

In our experience, the number of resulting networks can be quite large and so we provide methods for showing how the two input trees are embedded in the networks and for determining the number of different networks that contain a specific reticulation. We also provide variants of the algorithm that can be used to compute the rooted SPR distance (defined below) or the hybridization number of two bifurcating, rooted phylogenetic trees.

We report on a simulation study that we have undertaken to compare our implementation with other competing methods. This study indicates that our approach is much faster than existing methods. Moreover, to illustrate how one may apply our method to a practical problem, we use it to investigate the evolution of the Aeglops/Triticum genera.

The algorithm presented in this article is implemented in our program Dendroscope 3 (Huson and Scornavacca, 2011), which is freely available from www.dendroscope.org and runs on all three major operation systems.

2 METHODS

Throughout this article, we follow the terminology and notation defined in Huson et al. (2011) and assume that the reader is familiar with graphs and related terminology. Let $T$ be a phylogenetic tree on $X$ and let $X' \subseteq X$ be a subset of taxa. We use $T(X')$ to denote the minimum connected subgraph of $T$ that contains all leaves that are labeled by elements of $X'$. The restriction of $T$ to $X'$ is defined as the phylogenetic tree $T(X')$ that is obtained from $T(X')$ by suppressing all nodes that have both in- and outdegree 1.

We define a rooted phylogenetic network on $X$ as a directed acyclic graph with a single node with indegree zero (the root), no nodes with both indegree and outdegree equal to 1, and nodes with outdegree zero (the leaves) bijectively labeled by the set $X$.

Given a rooted bifurcating phylogenetic tree $T$, a rooted subtree prune and regraft move (rSPR move for short) on $T$ is performed by first detaching a subtree of $T$ rooted at the target of $e_1 = (v_1, w_1)$ by deleting the edge $e_1$ and regrafting the subtree on a different branch $e_2$ of $T$, by first creating a new node $v_2$ in $e_2$ and then a new edge $(v_2, w_1)$. Finally, any node with both in- and outdegree 1 is suppressed. Note that, in the case of regrafting above the root, a new node $v'$ has to be created, as well as two new edges $(v', v_1)$ and $(v', w_1)$.

Let $T_1$ and $T_2$ be two rooted bifurcating phylogenetic trees on a taxon set $X$. The rSPR distance between $T_1$ and $T_2$ is defined as the minimum number of rSPR moves required to transform $T_1$ into $T_2$. The problem of computing the rSPR distance between two rooted bifurcating phylogenetic trees $T_1$ and $T_2$ on a taxon set $X$ is known to be NP-hard, but FPT (Bordewich and Semple, 2002a, 2002b; Huson et al., 2011). The hybridization number can be calculated by computing a maximum acyclic agreement forest. An agreement forest $F_\mathcal{X}(T_1, T_2)$ for $T_1$ and $T_2$ is called acyclic if its ancestor-descendant graph $AG(T_1, T_2)$ contains no directed cycle. This graph is defined as the directed graph whose vertex set is $F_\mathcal{X}(T_1, T_2)$ and for which an edge $(F_i, F_j)$ exists precisely when $i \neq j$, and either $(1)$ the root of $T_1(X')$ is an ancestor of the root of $T_2(X')$ in $T_1$, or

$(2)$ the root of $T_2(X')$ is an ancestor of the root of $T_2(X')$ in $T_2$, where $X, X' \subseteq X$ are the sets of taxa that appear in $F_i$ and $F_j$, respectively.

An acyclic agreement forest with a minimum number of components is called a MAAF. If a MAAF with $h$ components exists, then a hybridization network with reticulation number $h-1$ containing both $T_1$ and $T_2$ exists. For example, if a MAAF with only one component exists, then we have that $T_1$ and $T_2$ are congruent and 0 recombination events are needed.

2.1 The algorithm

In this section, we give a high level description of an algorithm that takes as input two rooted bifurcating phylogenetic trees $T_1$ and $T_2$ on the same taxon set $X$, and produces as output a representative set of minimum hybridization networks $\mathcal{N}$ for $T_1$ and $T_2$. An agreement forest with minimum cardinality is called a minimum agreement forest (MAF).

An agreement forest with minimum cardinality is called a minimum agreement forest (MAF). The concepts of rSPR distance and MAFs are closely related. Indeed, a pair of rooted bifurcating phylogenetic trees $(T_1, T_2)$ has an rSPR distance equal to $d$ if and only if there exists a MAF $F(T_1, T_2)$ with size $d+1$ (Bordewich and Semple, 2005; Hein et al., 1996). Hence, to determine the rSPR distance between $T_1$ and $T_2$, it suffices to compute a MAF for these two trees.

Recall that a hybridization network for two rooted bifurcating phylogenetic trees $T_1$ and $T_2$ on a taxon set $X$ is a rooted phylogenetic network that contains both trees. Given a rooted phylogenetic network $N = (V, E)$, the reticulate number of $N$ is defined as

$$r(N) = \sum_{v \in V: (v \neq \epsilon)} (\delta(v) - 1) + |\epsilon| - |V| + 1,$$

where $\delta(v)$ denotes the indegree of $v$. In the special case that $N$ is bicontractible, that is, all nodes have indegree at most two, then this is simply the number of reticulation nodes.

The hybridization number for $T_1$ and $T_2$ is the minimum reticulation number obtained over all hybridization networks $\mathcal{N}$ for $T_1$ and $T_2$. The problem of computing the hybridization number for two rooted bifurcating phylogenetic trees on the same taxon set is known to be NP-hard, but FPT (Bordewich and Semple, 2002a, 2002b; Huson et al., 2011). The hybridization number can be calculated by computing a maximum acyclic agreement forest. An agreement forest $F_\mathcal{X}(T_1, T_2)$ for $T_1$ and $T_2$ is called acyclic if its ancestor-descendant graph $AG(T_1, T_2)$ contains no directed cycle. This graph is defined as the directed graph whose vertex set is $F_\mathcal{X}(T_1, T_2)$ and for which an edge $(F_i, F_j)$ exists precisely when $i \neq j$, and either

$(1)$ the root of $T_1(X')$ is an ancestor of the root of $T_2(X')$ in $T_1$, or

$(2)$ the root of $T_2(X')$ is an ancestor of the root of $T_2(X')$ in $T_2$, where $X, X' \subseteq X$ are the sets of taxa that appear in $F_i$ and $F_j$, respectively.

An acyclic agreement forest with a minimum number of components is called a MAAF. If a MAAF with $h$ components exists, then a hybridization network with reticulation number $h-1$ containing both $T_1$ and $T_2$ exists. For example, if a MAAF with only one component exists, then we have that $T_1$ and $T_2$ are congruent and 0 recombination events are needed.

2.1 The algorithm

In this section, we give a high level description of an algorithm that takes as input two rooted bifurcating phylogenetic trees $T_1$ and $T_2$ on the same taxon set $X$, and produces as output a representative set of minimum hybridization networks $\mathcal{N}$ for $T_1$ and $T_2$. The problem of computing the rSPR distance or the hybridization number is reported, or the final hybridization number obtained over all hybridization networks $\mathcal{N}$ for $T_1$ and $T_2$.
Whidden and Zeh (2009), the authors also describe an FPT-algorithm for the implementation of the algorithm, each subproblem produced in this way for reducing a common cherry of \( R \) and \( c \), then the algorithm branches into a third path that corresponds to reducing the cherry \( \{a, c\} \) to a new leaf labeled \( \{a, c\} \) both in \( R \) and \( S \). If only one MAAF is required, then the algorithm is terminated as soon as the first MAAF is found.

### 3. Simulation Study

To study the performance of our approach and implementation, we undertook a simulation study. We generated 2000 synthetic datasets, each consisting of a pair of rooted bifurcating phylogenetic trees. The datasets are based on three parameters, namely the number of taxa \( n \), the number of rSPR moves \( k \) used to obtain the second tree from the first,\(^1\) and the tangled degree \( d \) (as defined below). For each of the following choices of the parameters, \( n \) in \( [20, 50, 100, 200] \), \( k \) in \( [5, 10, 20, 50] \) and \( d \) in \( [3, 5, 10, 15, 20] \), we constructed 10 pairs of trees, thus obtaining 2000 different datasets in total.

The tangled degree is an ad hoc concept that we introduce to control how tangled the resulting network will be and it influences the way we obtain the second tree from the first using rSPR moves. More precisely, let \( R \) be an rSPR move performed on a tree \( T \) by choosing two edges \( e_1 \) and \( e_2 \), pruning the subtree of \( T \) rooted at the source of \( e_2 = (v_1, v_2) \) and re-grafting it on \( e_1 = (v_2, v_3) \). We say that \( R \) respects a tangled degree of \( d \), if the path from the lowest common ancestor of the nodes \( v_1 \) and \( v_2 \) to the node \( v_3 \) contains at most \( d \) edges. The tangled degree is a useful concept because the smaller it is, the more likely it is that one or more cluster reductions can be performed on the resulting dataset. As we will see, the number of common clusters in the two input trees has a major effect on the performance of our algorithm. Indeed, the more cluster reductions...
Fig. 1. Comparison of the running time of Dendroscope 3 (on the left) and HybridNet (on the right). (a) and (b) Average running time as function of the number of leaves. (c) and (d) Average running time as function of the rSPR moves. (e) and (f) Average running time as function of the number of common clusters. In all plots, the percentages report the proportion of the executions that were completed within 20 mins.

we can perform, the smaller the problem instances are and the more effective our parallelization is.

In more detail, each pair $(T_1, T_2)$ of rooted bifurcating phylogenetic trees for a given set of parameters $n, k,$ and $d$ is created as follows: the first tree $T_1$ on $X = \{x_1, \ldots, x_n\}$ is generated by first creating a set of $n$ leaf nodes bijectively labeled by the set $X$. Then, two nodes $u$ and $v$, both with indegree 0, are randomly picked and a new node $w$, along with two new edges $(w, u)$ and $(w, v)$, is created. This is done until only one node with no ancestor, the root, is present.

The second tree $T_2$ is obtained from $T_1$ by applying $k$ rSPR moves, in each move respecting the given tangling degree $d$.

We compared our implementation with the best available software for computing the exact hybridization number between two rooted bifurcating trees on the same taxon set, which is HybridNET (Chen and Wang, 2010), available from http://www.cs.cityu.edu.hk/~lwang/software/Hn/. [The underlying algorithm is described in Chen and Wang (2011)]. We do not consider the program HybridInterleave (Collins et al., 2011), because it has been shown to be much slower than the HybridNET software (Chen and Wang, 2010). Both programs have been run on an AMD Phenom X4 955 Processor with 4 GB RAM.

We present the results of our simulation in Figures 1 and 2. Each program run that took >20 mins was aborted and then counted as if it took 20 mins. In Figure 1, aborted runs have been included in the averages and, in all plots, the percentages report the proportion of the executions that were completed within 20 min. In Figure 2, aborted runs have not been included in the averages because their hybridization number is unknown. In the latter figure, we report the number of the executions that were completed within 20 minutes.

Figures 1 and 2 show that the average running time (as function of the number of leaves, of the number of rSPRs or of the hybridization number) of our implementation is always lower than that of HybridNET. For both programs, the running time increases with the number of taxa (Fig. 1a and b), and also with the rSPR distance (approximated by the number of rSPR moves) (Fig. 1c...
Fig. 2. Average running time as function of the hybridization number for Dendroscope 3 (a) and HybridNet (b). In both plots, we report the number of the executions that were completed within 20 mins.

and d) and the hybridization number (Fig. 2). Figure 2 also shows that, in our simulations, HybridNet was unable to compute the hybridization number for any tree pair with hybridization number >21, while our program produces results for tree pairs having a hybridization number up to 40.

Note that the number of common clusters has a significant effect on the performance of our algorithm (Fig. 1e and f), and this explains much of the performance advantage over HybridNet. However, even in the case that no parallelization is performed, namely when the number of common clusters is equal to 1, our implementation is still faster than HybridNet (Fig. 3e and f). Note that the plot in Figure 1f is very erratic when the number of common clusters is >15. This is due to the fact that few datasets share such a high number of common clusters.

Using the same datasets, we also studied the performance of our implementation for the problems of computing the rSPR distance for two rooted bifurcating phylogenetic trees. We compared the running time of our implementation with that of the best available software for computing the exact rSPR distance between two rooted bifurcating trees on the same taxon set, i.e. rSPR (Whidden et al., 2010), available at http://kiwi.cs.dal.ca/Software/RSPR. In all conditions, the program rSPR performs better than our implementation (data not shown). Since the underlying algorithms are the same in this case, we suspect the difference in performance may be due to the fact that our program is implemented in Java, whereas the rSPR program is written in C++.

For the sake of completeness, in Supplementary Figure S1 we report the running time of Dendroscope 3 and HybridNET when computing the exact hybridization number and a set of hybridization networks, containing a network per MAAF, on a grass (Poaceae) dataset provided by the Grass Phylogeny Working Group (Grass Phylogeny Working Group, 2001). This dataset has been often used to evaluate programs computing the hybridization number or hybridization networks for two rooted bifurcating phylogenetic trees on the same taxon set.

Table 1. Accession numbers in GenBank of the sequences used for obtaining the trees in Figure 3.

<table>
<thead>
<tr>
<th>PinA</th>
<th>matK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triticum urartu TU55</td>
<td>EU307589</td>
</tr>
<tr>
<td>Triticum monococcum DP57</td>
<td>EU307591</td>
</tr>
<tr>
<td>Aegilops tauschii DP16</td>
<td>FJ898213</td>
</tr>
<tr>
<td>Aegilops comosa DP13</td>
<td>FJ898210</td>
</tr>
<tr>
<td>Aegilops uniaristata DP56</td>
<td>FJ898218</td>
</tr>
<tr>
<td>Aegilops bicornis DP18</td>
<td>FJ898215</td>
</tr>
<tr>
<td>Aegilops longissima DP17</td>
<td>FJ898214</td>
</tr>
<tr>
<td>Aegilops sharonensis DP53</td>
<td>FJ898216</td>
</tr>
<tr>
<td>Aegilops speltoides SP6</td>
<td>FJ898222</td>
</tr>
<tr>
<td>Hordeum vulgare (Morex)</td>
<td>AY643843</td>
</tr>
</tbody>
</table>

4 APPLICATION TO PHYLOGENY OF AEGILOP/TRITICUM GENERA

The Triticeae tribe (Family Poaceae) consists of diploid and polyploid grasses with the same basic haploid chromosome number (x = 7), and they are distributed worldwide. Phylogenetic analysis performed with different sequence datasets show significant inconsistencies (Kellogg et al., 1996; Mason-Gamer and Kellogg, 1996; Sasanuma et al., 2004). In particular, inconsistencies between chloroplast and genomic data are often detected (Sasanuma et al., 2004). We used two datasets of sequences from number of diploid species belonging to the close genera Triticum and Aegilops: matK and PinA, located on the chloroplast and the Triticeae chromosome 5, respectively. Sequence data were obtained from GenBank. The accession numbers can be found in Table 1.

The best-fit model of nucleotide substitution for each dataset was chosen using JModelTest (Posada, 2008), and PhyML 3.0 (Guindon et al., 2010) was used to obtain maximum likelihood phylogenetic trees (heuristic search with BIONJ starting tree, SPR and NNI swapping and 100 bootstrap replicates). Majority-rule extended
consensus trees were computed from 100 bootstrap replicates by the consense program of the PHYLIP package (Felsenstein, 2005), using the default parameters in the rooting setting. The two consensus trees are shown in Figure 3.

Two kinds of inconsistencies were found. One involves Aegilops speltoides and the ancestor of A.bicornis, A.longissima and A.sharonensis (hereafter referred to as the BLS species). A.speltoides has a basal position in both trees, but in the matK tree it is isolated, whereas in the PinA gene it is very close to the BLS. In the matK tree, the BLS species are close to species that radiated more recently. This deep phylogenetic inconsistency could be a consequence of a hybridization between A.speltoides and an ancestor of BLS species that introgressed into portions of A.speltoides genome. This hypothesis is supported by the observation of similar inconsistencies involving A.speltoides and A.longissima in a larger sample of nuclear genes (Escobar et al., 2011). Note that this hypothesis is present in all three hybridization networks obtained by Dendroscope 3 from the two consensus trees (Fig. 4). The other inconsistency involves the relative position of A.tauschii, of the ancestors of two groups of species, i.e. A.unicristata + A.comosa and of the BLS species. Three alternative hybridizations could be inferred (Fig. 4). However, this inconsistency is more recent and could also be explained by an incomplete allele sorting (that is, allelic variants coexisting in the common ancestor of all these species were randomly fixed in the derived species).

5 DISCUSSION
It has become standard practice to base evolutionary studies on multiple genes. When incongruences between different gene trees are small, then they are usually deemed insignificant and are dealt with by performing a consensus analysis. However, when the differences between the gene trees are more significant, and when mechanisms such as hybridization may have played an important role in the evolutionary history of a set of species, then an alternative approach may be to try to reconcile the different gene trees by combining them into a rooted phylogenetic network in which reticulation nodes represent possible hybridization events.

While some papers in the literature have focused on reconciling incongruent gene phylogenies in terms of a network (Koblmuller et al., 2007), a major problem has been the lack of software implementing an algorithm for computing and investigating such networks. In this article, we address the need for such software by providing an algorithm that runs fast on most practical problems and produces a representative set of minimum hybridization networks, containing one network for each MAAF. Our algorithm is implemented in the program Dendroscope 3, which allows the user to visualize and compare the resulting networks, both among each other, and also with the original input trees.

Our simulation study shows that our implementation is faster than existing implementations and the study of grasses, reported in Section 4, shows how one may use the software to obtain different possible hybridization scenarios.

ACKNOWLEDGEMENTS
The authors would like to thank Simone Linz for helpful discussions.

Conflict of Interest: none declared.

REFERENCES
Computation of minimum hybridization networks


