Delft University of Technology

Polynomial-Time Algorithms for Phylogenetic Inference Problems

van Iersel, Leo; Janssen, Remie; Jones, Mark; Murakami, Yukihiro; Zeh, Norbert

DOI
10.1007/978-3-319-91938-6_4

Publication date
2018

Document Version
Final published version

Published in
Algorithms for Computational Biology - 5th International Conference, AlCoB 2018, Proceedings

Citation (APA)
van Iersel, L., Janssen, R., Jones, M., Murakami, Y., & Zeh, N. (2018). Polynomial-Time Algorithms for Phylogenetic Inference Problems. In J. Jansson, C. Martin-Vide, & M. A. Vega-Rodriguez (Eds.), *Algorithms for Computational Biology - 5th International Conference, AlCoB 2018, Proceedings* (pp. 37-49). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 10849 LNBI). Springer. https://doi.org/10.1007/978-3-319-91938-6_4

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Green Open Access added to TU Delft Institutional Repository

'You share, we take care!' – Taverne project

https://www.openaccess.nl/en/you-share-we-take-care

Polynomial-Time Algorithms for Phylogenetic Inference Problems

Leo van Iersel[1](✉), Remie Janssen[1], Mark Jones[1], Yukihiro Murakami[1], and Norbert Zeh[2]

[1] Delft Institute of Applied Mathematics, Delft University of Technology, Van Mourik Broekmanweg 6, 2628 XE Delft, The Netherlands
{L.J.J.vanIersel,R.Janssen-2,M.E.L.Jones,Y.Murakami}@tudelft.nl

[2] Faculty of Computer Science, Dalhousie University, 6050 University Ave, Halifax, NS B3H 1W5, Canada
nzeh@cs.dal.ca

Abstract. A common problem in phylogenetics is to try to infer a species phylogeny from gene trees. We consider different variants of this problem. The first variant, called Unrestricted Minimal Episodes Inference, aims at inferring a species tree based on a model of speciation and duplication where duplications are clustered in duplication episodes. The goal is to minimize the number of such episodes. The second variant, Parental Hybridization, aims at inferring a species *network* based on a model of speciation and reticulation. The goal is to minimize the number of reticulation events. It is a variant of the well-studied Hybridization Number problem with a more generous view on which gene trees are consistent with a given species network. We show that these seemingly different problems are in fact closely related and can, surprisingly, both be solved in polynomial time, using a structure we call "beaded trees". However, we also show that methods based on these problems have to be used with care because the optimal species phylogenies always have some restricted form. We discuss several possibilities to overcome this problem.

Keywords: Phylogenetic inference problems
Polynomial-time algorithms

1 Introduction

Phylogenetic trees are commonly used to represent the evolutionary history of a set of taxa. The leaves represent extant taxa; internal nodes represent speciation events that caused lineages to diverge. If we assume the only processes

Research funded in part by the Netherlands Organization for Scientific Research (NWO), including Vidi grant 639.072.602, the 4TU Applied Mathematics Institute, the Natural Sciences and Engineering Research Council of Canada and the Canada Research Chairs program.

© Springer International Publishing AG, part of Springer Nature 2018
J. Jansson et al. (Eds.): AlCoB 2018, LNBI 10849, pp. 37–49, 2018.
https://doi.org/10.1007/978-3-319-91938-6_4

are speciation and modification and that no incomplete lineage sorting occurs, then any gene will give a gene tree that is consistent with the species phylogeny. In such cases, there exist efficient algorithms to reconstruct a species tree from gene trees. There are, however, evolutionary processes beyond vertical inheritance of genetic material and speciation events that make it more challenging to reconstruct the real evolutionary history. Examples of such processes are hybridization, horizontal gene transfer, and duplication. Each of these processes can result in discordance between gene trees.

This leads to a number of problems in which the task is to minimize the number of such complicating events. In *reconciliation problems*, we are given the gene trees together with the species phylogeny, and the task is to find optimal embeddings of the gene trees into the species phylogeny. Such methods are for example used to estimate dates of duplications, to discover relations between duplicate genes [7], and to reconstruct the infection history of parasites [19]. In *inference problems*, only the gene trees are given and we aim to find a species phylogeny that minimizes the discordance with the gene trees. Such problems are relevant when the species phylogeny is not yet known with certainty.

Duplication Minimization Problems. Gene duplications happen as a consequence of errors in the DNA replication process. This leads to a species having multiple copies of the same gene. There exist many types of gene duplication, which depend on the positions of errors within the replication process [20]. The scale of gene duplications is determined by the number of genes that get duplicated. An extreme example of a large-scale duplication is *Whole Genome Duplication (WGD)*, in which every gene in the genome is duplicated. This process, also known as polyploidization, occurs as a result of an error in separation of chromosomes during gamete production. It is most common in plants but has also occurred in animals [22], and there are two WGD events even in the evolutionary history leading to humans [8]. Large-scale duplications provide species with diversification potential, giving them the ability to quickly adapt to a changing environment [10].

In their seminal paper [11], Goodman et al. pioneered the parsimony approach to reconciling gene trees with species trees. This has motivated researchers to explore reconciliation through different models, whilst optimizing some measure of the number of duplication events.

These models can be categorized according to how duplication events are clustered to form duplication episodes and which restrictions are put on the possible locations of duplications [21]. We focus on the *minimal episodes (ME)* model where duplications can be clustered if they occur on the same branch of the species phylogeny and have no ancestor-descendant relationship in a gene tree (see Fig. 1). We believe this model to be most relevant since it can cluster duplications that can be part of a single (large-scale) duplication event. We consider the *unrestricted* variant of this model, which does not put any restrictions on the locations of gene duplications (called the FHS-model in [21]).

Reconciliation problems have been studied intensively, especially models without clustering. Several reconciliation problems with clustering have been

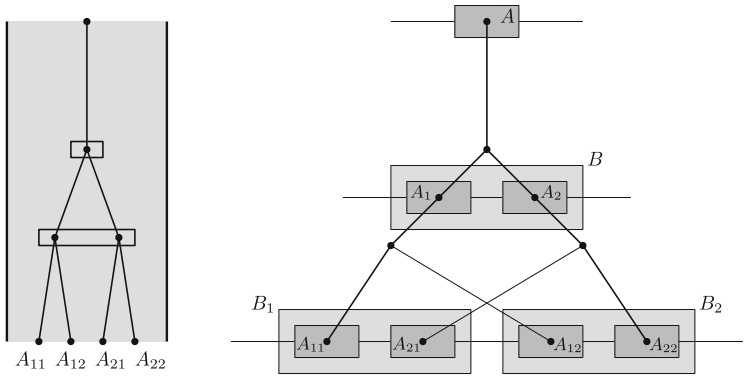

Fig. 1. Left: A gene tree embedded into a branch of a species tree with duplications clustered as in the Minimal Episodes model. Duplication clusters are shown as rectangles. Right: A representation of the DNA of the species at different points in the species tree (at corresponding heights). In the first duplication, the gene A (dark rectangle) is duplicated, forming A_1 and A_2. In the second duplication, the block B (light rectangle) comprising A_1 and A_2 is duplicated. This results in four homologous copies of gene A using only two duplication episodes. The gene tree is also drawn through the depictions of the DNA.

proven to be computationally intractable [9,17], whereas for others there are polynomial-time [3,6] or even linear-time [16,18,21] algorithms. For unrestricted ME reconciliation, there only exists an exponential-time algorithm [21], while the computational complexity of this problem is still unknown.

It has also been attempted to use reconciliation as a basis for inferring species phylogenies. For the unrestricted ME model, Burleigh et al. [5] used a brute-force approach on all possible species phylogenies. They observed that the unrestricted ME model fails to rank the true species tree among the top third of all topologies. It was suggested that a possible reason for this anomaly is that duplication episodes near the root are overly powerful under this model. A similar observation was made in a more recent reconciliation study [21]. However, neither article gives a mathematical explanation for this phenomenon. It should also be noted that, since the number of possible species phylogenies grows extremely quickly with the number of species, brute-force approaches are only feasible for very small data sets.

Inference problems are generally assumed to be computationally intractable. However, NP-hardness has been proven only for some restricted inference problem without clustering [17]. For an inference problem with restricted clustering (called gene duplication (GD) clustering in [21]), NP-hardness was suggested in [9] but not proven. Because of the suspected intractability of these problems, some heuristic inference approaches have been attempted using efficient algorithms for reconciliation (see, e.g., [12]).

Reticulation Minimization Problems. Another possible cause of discordance between gene trees is *reticulate evolution*, such as hybridization or horizontal gene transfer. In such cases, the evolutionary history is represented by a *phylogenetic network* rather than a tree.

Reticulate evolution can occur in nature when genetic material from one species is transmitted to some other species. In asexual species, such transfers are called *horizontal gene transfers (HGT)*. In bacteria, for example, this happens in nature by transformation (take-up from the environment) or conjugation (transmission from another bacterium). In sexual species, a cause for such transmissions can be *hybridization*, where individuals from different but related taxa mate. There is also evidence that horizontal gene transfers occur between multicellular sexual species. HGT can even happen between more distant species.

Gene trees that appear to be inconsistent may in fact simply take different paths through the network. This leads to a family of inference problems in which the aim is to find a phylogenetic network that is consistent with the gene trees and has the minimum number of *reticulation events* (nodes in the network with two ancestral branches). A phylogenetic network is often taken to be consistent with a gene tree if that tree is *displayed* by the network, which, roughly speaking, means that the gene tree can be drawn inside the network in such a way that each network branch contains at most one lineage of the gene tree. A more generous definition is to count a network as consistent with a gene tree if the tree is *weakly displayed* by the network [13,23]. Roughly speaking, this means that different lineages of the gene tree may "travel down" the same branch of the network, as long as any branching node in the tree coincides with a branching node in the network. In this case, the tree is also called a *parental tree* of the network. This models situations where a species has individuals carrying multiple homologous copies of a gene.

The Hybridization Number problem, in which we seek a network with the minimum number of reticulations displaying all input trees, has been well-studied. It has been shown that Hybridization Number is NP-hard already when the input consists of only two gene trees [4]. Furthermore, there are theoretical FPT algorithms for any fixed number of gene trees [15], but there are no practical algorithms that can handle instances with more than two input trees unless the number of taxa is extremely small.

In contrast, the Parental Hybridization problem, in which we seek a network with the minimum number of reticulations that weakly displays each input tree, was introduced only recently [23] and its computational complexity was open prior to this article. Our motivation for studying this problem is threefold:

(i) Since Hybridization Number is NP-hard, it is interesting whether relaxing the notion of a tree displayed by a network leads to an easier problem.
(ii) Since reticulation can lead to multiple homologous copies of a gene in a species, requiring that each gene tree is displayed by the network may lead us to overestimate the number of reticulations.

(iii) The problem of finding an optimal network that weakly displays a set of phylogenies arises as a crucial subproblem in a recent heuristic approach for constructing phylogenetic networks in the presence of hybridization and incomplete lineage sorting [23].

Structural Assumptions. We restrict to binary trees and networks. Unlike many papers in this area, we allow a network to contain *parallel arcs*, that is, pairs of arcs that join the same pair of nodes. Parallel arcs are normally omitted because, for most problems, it can either be shown that there exists an optimal solution without parallel arcs or it can be assumed that a realistic solution contains no parallel arcs. For example, any set of gene trees has an optimal hybridization network without parallel arcs. For the problems studied in this paper, however, an optimal solution may require parallel arcs. Considering this problem with the added restriction that parallel arcs are forbidden may be an interesting mathematical challenge; however, we do not believe it is biologically meaningful.

Explicit reasons to allow parallel arcs in networks are abundant. We give three: First, if one restricts a large network to a subset of the taxa, the natural restriction could have parallel arcs. Second, phylogenetic Markov models for character evolution behave differently if parallel arcs are suppressed. Third, polyploidization events often result from a sort of interspecific or intraspecific hybridization [2]; an intraspecific hybridization is most naturally represented by parallel arcs in the network.

Throughout this paper, we allow input trees to be multi-labeled, that is, each species may appear as a label of multiple leaves in a tree. This is natural for the problems we study, as gene duplication and reticulation can both lead to multiple homologous genes appearing in the genome of a single species.

Our Contributions. We show that both Unrestricted Minimal Episodes Inference and Parental Hybridization reduce to the problem Beaded Tree, which we introduce in this paper. Using this reduction, we show that both problems can be solved in polynomial time by adapting Aho et al.'s classic algorithm for testing gene tree consistency [1]. Thereby, we provide the first polynomial-time algorithm for an inference problem with a duplication cluster model. Furthermore, we provide the first polynomial-time algorithm for constructing a phylogenetic *network* from gene trees.

We also show that optimal solutions to Beaded Tree have a restricted structure and this has corresponding implications for the optimal solutions to Unrestricted Minimal Episodes Inference and Parental Hybridization that our algorithms produce. Moreover, we show that, in fact, *all* optimal solutions to Unrestricted Minimal Episodes Inference have a restricted structure. Therefore, this model should be used with care. We end with a discussion of different ways to overcome these issues.

See [14] for the full version of this paper.

2 Preliminaries

We begin by defining *multi-labeled trees*, which form the input for all problems considered in this paper.

Definition 1. *Let X be a set of species. A* multi-labeled tree (MUL-tree) *on X is a directed acyclic graph with one node of in-degree 0 and out-degree 1 (the* root*) and with all other nodes having either in-degree 1 and out-degree 2 (*tree nodes*) or in-degree 1 and out-degree 0 (*leaf nodes *or* leaves*). Each leaf is labeled with an element of X. If each element of X labels at most one leaf, we call the MUL-tree a* tree*.*

Next, we define a *duplication tree*, which represents the evolutionary history of a set of species, including points at which duplication events occurred.

Definition 2. *Let X be a set of species. A* duplication tree *on X is a directed acyclic graph D with one node of in-degree 0 and out-degree 1 (the* root*), $|X|$ nodes of in-degree 1 and out-degree 0 (*leaf nodes *or* leaves*), and all other nodes having either in-degree 1 and out-degree 2 (*tree nodes*) or in-degree 1 and out-degree 1 (*duplication nodes*). The leaves are bijectively labeled with the elements of X. The* duplication number *of D is the number of duplication nodes it contains.*

Informally, a MUL-tree T is *consistent* with a duplication tree D if T can be drawn inside D so that branches of T duplicate only at duplication nodes of D, in the sense that both out-edges of a node of T may follow the same out-edge of the duplication node (see Fig. 2). We formalize this as follows:

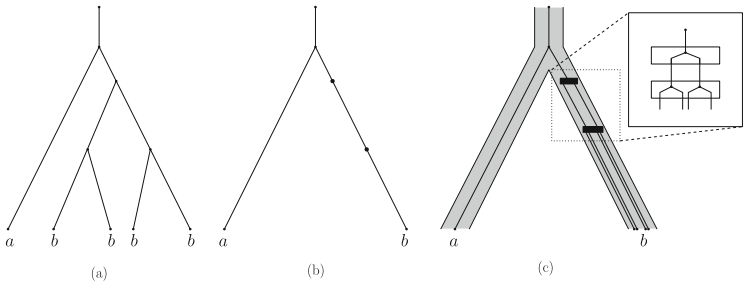

Fig. 2. (a) A MUL-tree T on $X = \{a, b\}$. (b) A duplication tree D that is consistent with T. (c) An illustration showing how T can be drawn inside D, and a zoomed-in portion to illustrate what happens at the duplication nodes. This shows how two or more incoming branches may duplicate simultaneously at a duplication node (according to the Minimal Episodes model).

Definition 3. *Given a MUL-tree T on X and a duplication tree D on X, a* duplication mapping *from T to D is a function $M : V(T) \to V(D)$ such that*

- *For each leaf $l \in L(T)$, $M(l)$ is a leaf of D labeled with the same species as l,*
- *For each edge $uv \in E(T)$, $M(u)$ is a strict ancestor of $M(v)$, and*
- *For each internal node u of T with children v, v', either $M(u)$ is the least common ancestor of $M(v)$ and $M(v')$, or $M(u)$ is a duplication node.*

This is illustrated in Fig. 3. We say that D is consistent with *T if there is a duplication mapping from T to D.*

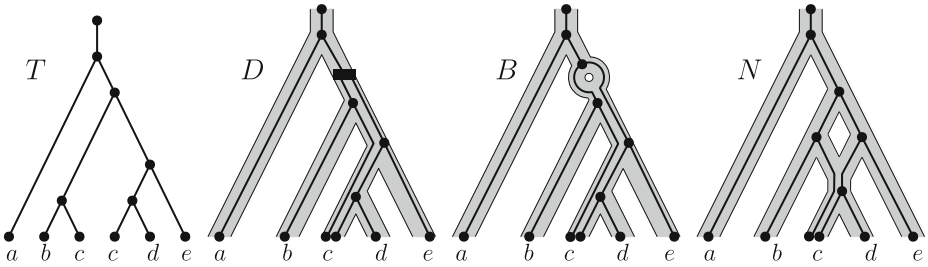

Fig. 3. A MUL-tree T, a duplication mapping from T to a duplication tree D, and weak embeddings of T into a beaded tree B and into a phylogenetic network N.

Let S be the species tree derived from D by suppressing duplication nodes. Then a duplication mapping from T to D represents a reconciliation of T with S under the Minimal Episodes model. Each duplication node in D represents a cluster of duplications, which is called a *duplication episode*. Internal nodes in T are treated as duplications if they are mapped to duplication nodes of D, and as speciations otherwise. Duplications are clustered together if they are mapped to the same duplication node of D. The properties of a duplication tree and duplication mapping ensure that duplications that are clustered occur on the same branch of the species phylogeny and have no ancestor-descendant relationship in a gene tree, as required by the Minimal Episodes model. We are now ready to define the following problem:

Unrestricted Minimal Episodes Inference
Input: A set $\mathcal{T} = \{T_1, \ldots, T_t\}$ of MUL-trees with label sets $X_1, \ldots, X_t \subseteq X$.
Output: A duplication tree D on X with minimum duplication number such that D is consistent with each tree in $\mathcal{T}$.

Next, we introduce the concept of *phylogenetic networks*, which are central to the problem Parental Hybridization:

Definition 4. *Let X be a set of species. A* (rooted binary) phylogenetic network *N on X is a directed acyclic multigraph with one node of in-degree 0 and out-degree 1 (the* root*), $|X|$ nodes of in-degree 1 and out-degree 0 (*leaves*), and all other nodes having either in-degree 1 and out-degree 2 or in-degree 2 and out-degree 1 (*reticulation nodes*). The leaves are bijectively labeled with the elements of X. The* reticulation number *of N is the number of reticulation nodes it contains.*

Definition 5. *Given a set X of species, let N be a phylogenetic network, and T a MUL-tree on X. A* weak embedding *of T into N is a function h that maps every node of T to a node of N, and every edge in T to a directed path in N such that*

- *for each leaf $l \in L(T)$, $h(l)$ is a leaf of N labeled with the same species;*
- *for each edge $xy \in E(T)$, the path $h(xy)$ is a path from $h(x)$ to $h(y)$ in N;*
- *for each internal node x in T with children y, y', the paths $h(xy)$ and $h(xy')$ start with different out-edges of $h(x)$.*

This is illustrated in Fig. 3. We say that N weakly displays *T if there is a weak embedding of T into N.*

We note that N weakly displays T if and only if T is a *parental tree inside* N as defined in [23], hence the name Parental Hybridization. The notion of a tree *weakly displayed* by a network was first introduced in [13], where it was shown that T is weakly displayed by N if and only if there exists a *locally separated reconciliation* from T to N, which is equivalent to our definition of a weak embedding. We now formally define the Parental Hybridization problem:

Parental Hybridization
Input: A set $\mathcal{T} = \{T_1, \ldots, T_t\}$ of MUL-trees with label sets $X_1, \ldots, X_t \subseteq X$.
Output: A phylogenetic network N on X with minimum reticulation number such that N weakly displays all trees in $\mathcal{T}$.

Next, we define a certain type of phylogenetic network that, together with the corresponding computational problem defined below, turns out to be the key to both Unrestricted Minimal Episodes Inference and Parental Hybridization.

Definition 6. *A* bead *in a phylogenetic network N is a pair of nodes (u, v) such that there are two parallel edges from u to v. A* beaded tree *is a phylogenetic network B in which every reticulation node is in a bead (see Fig. 3).*

Beaded Tree
Input: A set $\mathcal{T} = \{T_1, \ldots, T_t\}$ of MUL-trees with label sets $X_1, \ldots, X_t \subseteq X$.
Output: A beaded tree B on X with minimum reticulation number that weakly displays all trees in $\mathcal{T}$.

3 Reduction to Beaded Trees

The two problems Unrestricted Minimal Episodes Inference and Parental Hybridization are in fact both reducible to Beaded Tree. This allows us to focus on the Beaded Tree problem in the rest of the paper.

Lemma 7. *Let X be a set of species and $\mathcal{T} = \{T_1, \ldots, T_t\}$ a set of MUL-trees on subsets of X. For any integer k, there exists a solution to* Unrestricted Minimal Episodes Inference *on $\mathcal{T}$ with k duplications if and only if there exists a solution to* Beaded Tree *on $\mathcal{T}$ with k beads.*

Lemma 8. *For any instance $\mathcal{T}$ of* Parental Hybridization*, there exists an optimal solution B that is a beaded tree.*

We can also show that any instance of Beaded Tree has an optimal solution with a certain interesting structure.

Theorem 9. *Given an instance $\mathcal{T}$ of* Beaded Tree*, there exists an optimal solution B such that all reticulations are on a single path.*

Moreover, *any* optimal solution to an instance of Beaded Tree must satisfy certain structural properties.

Theorem 10. *Given any optimal solution B to an instance $\mathcal{T}$ of* Beaded Tree*, there exists a path from the root to a leaf of B, such that for any node u not on this path, there is at most one reticulation strictly descended from u.*

4 Beaded Tree Algorithm

Let Supertree denote an algorithm that takes as input a set of MUL-trees $\mathcal{T}$, and returns either a tree T weakly displaying $\mathcal{T}$, or the value None if no such tree exists. A simple modification of the algorithm of [1] can be used for this.

Given a phylogenetic network N on X and a subset $S \subseteq X$, let $N \setminus S$ denote the network derived from N by deleting every leaf in S, and then exhaustively deleting unlabelled nodes of out-degree 0 and suppressing nodes of in-degree 1 out-degree 1. Let $N|_S$ denote the network $N \setminus (X \setminus S)$.

Given a set $\mathcal{T}$ of MUL-trees, let $F_1(\mathcal{T})$ denote the set of trees derived by, roughly speaking, splitting each tree of $\mathcal{T}$ into two by deleting the root.

Definition 11. *Let $\{T_1, \ldots, T_t\}$ be a set of MUL-trees and X the union of their label sets. The* split partition *$\{S_1, \ldots, S_s\}$ of $\{T_1, \ldots, T_t\}$ is the partition of X into minimal sets such that, if x and y appear within the same MUL-tree in $F_1(\mathcal{T})$ and $x \in S_j$, then $y \in S_j$.*

The beaded tree algorithm is described in Algorithm 1 and an example is given in Fig. 4.

Theorem 12. *Algorithm 1 finds an optimal solution to the* Beaded Tree *problem with input $\mathcal{T}$ in time $O((|X|^3 + |X|^2 k)n)$, with n the total number of vertices of the trees in $\mathcal{T}$ and k the reticulation number of an optimal solution.*

Data: $\mathcal{T} = \{T_1, \ldots, T_t\}$
Result: Beaded tree B that weakly displays $\mathcal{T}$ with minimum number of reticulations
if $|X| = 1$ *and* $\max_{i \in [t]} |L(T_i)| = 1$ **then**
  return a tree with 1 leaf on X;
end
else
  Calculate the split partition $S_1, \ldots, S_s$ of $\mathcal{T}$;
  for $i \in [s]$ **do**
    Let $T = \textsc{Supertree}(\mathcal{T}|_{S_i})$;
    if T *is not* NONE **then**
      Let $B' = \textsc{Beaded-Tree}(\mathcal{T} \setminus S_i)$;
      Construct B by joining B' and T with a new root;
      return B
    end
  end
  Let $B' = \textsc{Beaded-Tree}(F_1(\mathcal{T}))$;
  Construct B by adding a bead whose child is the root of B';
  return B
end

Algorithm 1. Algorithm BEADED-TREE($\mathcal{T}$).

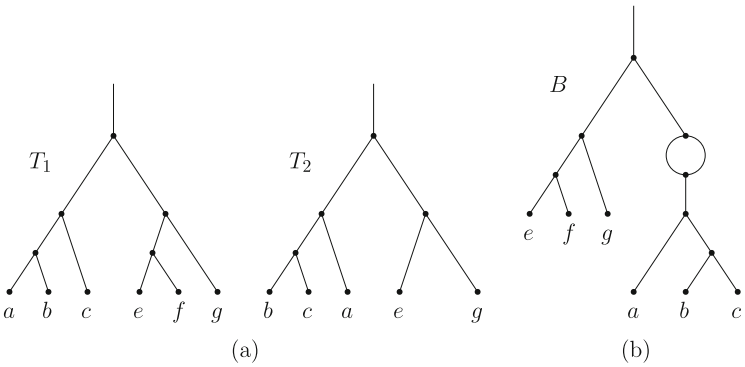

Fig. 4. (a) An instance $\mathcal{T} = \{T_1, T_2\}$ of BEADED TREE. (b) The beaded tree B constructed by running algorithm BEADED-TREE on $\mathcal{T}$. Initially, the split partition is $\{a, b, c\}, \{e, f, g\}$. As SUPERTREE returns a tree on $\{e, f, g\}$, the top tree node of B has that tree as one of its children. To construct the other side of B, we run BEADED-TREE on $\mathcal{T}|_{\{a,b,c\}}$, and SUPERTREE does not return a tree on this set. Therefore this side of B begins with a bead.

5 Concluding Remarks

Although we have shown that the UNRESTRICTED MINIMAL EPISODES INFERENCE and PARENTAL HYBRIDIZATION problems are polynomial-time solvable,

we have also shown that the phylogenies produced by solving these problems have severely restricted structures.

The optimal phylogenetic network that our algorithm produces for the Parental Hybridization problem is always a phylogenetic tree with "beads", where a bead consists of a speciation directly followed by a reticulation. Such solutions are not necessarily the most realistic or likely ones since they contain a lot of "extra lineages", i.e. multiple lineages of an input tree travelling through the same branch of the phylogenetic network. Minimizing the total number of extra lineages, the *XL-score*, irrespective of the reticulation number, is also not ideal, since there always exists a solution with zero extra lineages and possibly a very high reticulation number. Therefore, the most relevant open problem that needs to be solved is to find a phylogenetic network that minimizes a weighted sum of the XL-score and the reticulation number of the network. Another alternative problem formulation that seems reasonable is to minimize the total number of parental trees that the constructed phylogenetic network has in addition to the input trees.

Another option would be to completely exclude beads in the solutions. However, although this is an interesting theoretical open problem, we do not see a reason why the resulting optimal solutions would by any more realistic, or why it would be reasonable to assume that a speciation cannot be followed by a reticulation.

Regarding Unrestricted Minimal Episodes Inference, the situation is in some sense even worse. We have shown that *all* optimal solutions have a very specific structure: there is one main path from the root to a taxon containing potentially many duplication episodes, while each path branching off this main path contains at most one duplication episode. Although such scenarios are not to be excluded (for example see the eukaryotic species phylogeny from [12]), it is unrealistic to expect all phylogenies to look like this. Therefore, we have proposed an alternative problem in [14], which miminizes the "duplication depth": the maximum number of duplication episodes that lie on any directed path. This problem can also be solved in polynomial time and we expect it to produce more realistic solutions. Note moreover that, although the problem definition does not exclude unnecessary duplication episodes as long as they do not increase the duplication depth, our algorithm will not create such redundant duplication episodes. Nevertheless, to properly assess the two algorithms, it is necessary to implement both algorithms and extensively test them on simulated and real biological datasets.

Interestingly, the problem Unrestricted Minimal Episodes Reconciliation, where the species tree is given, is *not* known to be polynomial-time solvable. There is only an exponential-time algorithm [21]. Could it be possible to adapt our algorithm for Unrestricted Minimal Episodes Inference to solve also the reconciliation variant?

Finally, it would be interesting to study more general models, which simultaneously take different processes into account, such as duplication episodes, hybridization, gene loss and transfers. Although such problems have been stud-

ied in a reconciliation setting where the species tree is (assumed to be) known, there has been less work on variants where the species tree or network needs to be inferred. Although such problems seem daunting, we have shown here that not knowing the species tree can actually make computational problems easier.

References

1. Aho, A.V., Sagiv, Y., Szymanski, T.G., Ullman, J.D.: Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. SIAM J. Comput. **10**, 405–421 (1981)
2. Albertin, W., Marullo, P.: Polyploidy in fungi: evolution after whole-genome duplication. Proc. Roy. Soci. Lond. B Biol. Sci. **279**(1738), 2497–2509 (2012)
3. Bansal, M.S., Eulenstein, O.: The multiple gene duplication problem revisited. Bioinformatics **24**(13), i132–i138 (2008)
4. Bordewich, M., Semple, C.: Computing the minimum number of hybridization events for a consistent evolutionary history. Discrete Appl. Math. **155**(8), 914–928 (2007)
5. Burleigh, J.G., Bansal, M.S., Eulenstein, O., Vision, T.J.: Inferring species trees from gene duplication episodes. In: Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology, pp. 198–203. ACM (2010)
6. Burleigh, J.G., Bansal, M.S., Wehe, A., Eulenstein, O.: Locating multiple gene duplications through reconciled trees. In: Vingron, M., Wong, L. (eds.) RECOMB 2008. LNCS, vol. 4955, pp. 273–284. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78839-3_24
7. Chan, Y.B., Ranwez, V., Scornavacca, C.: Reconciliation-based detection of co-evolving gene families. BMC Bioinform. **14**(1), 332 (2013)
8. Dehal, P., Boore, J.L.: Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. **3**(10), e314 (2005)
9. Fellows, M., Hallett, M., Stege, U.: On the multiple gene duplication problem. In: Chwa, K.-Y., Ibarra, O.H. (eds.) ISAAC 1998. LNCS, vol. 1533, pp. 348–357. Springer, Heidelberg (1998). https://doi.org/10.1007/3-540-49381-6_37
10. Glasauer, S.M., Neuhauss, S.C.: Whole-genome duplication in teleost fishes and its evolutionary consequences. Mol. Genet. Genomics **289**(6), 1045–1060 (2014)
11. Goodman, M., Czelusniak, J., Moore, G.W., Romero-Herrera, A.E., Matsuda, G.: Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst. Biol. **28**(2), 132–163 (1979)
12. Guigo, R., Muchnik, I., Smith, T.F.: Reconstruction of ancient molecular phylogeny. Mol. Phylogenet. Evol. **6**(2), 189–213 (1996)
13. Huber, K.T., Moulton, V., Steel, M., Wu, T.: Folding and unfolding phylogenetic trees and networks. J. Math. Biol. **73**(6–7), 1761–1780 (2016)
14. van Iersel, L., Janssen, R., Jones, M., Murakami, Y., Zeh, N.: Polynomial-time algorithms for phylogenetic inference problems (2018). arXiv:1802.00317 [q-bio.PE]
15. van Iersel, L., Kelk, S., Scornavacca, C.: Kernelizations for the hybridization number problem on multiple nonbinary trees. J. Comput. Syst. Sci. **82**(6), 1075–1089 (2016)
16. Luo, C.W., Chen, M.C., Chen, Y.C., Yang, R.W., Liu, H.F., Chao, K.M.: Linear-time algorithms for the multiple gene duplication problems. IEEE/ACM Trans. Comput. Biol. Bioinf. **8**(1), 260–265 (2011)

17. Ma, B., Li, M., Zhang, L.: From gene trees to species trees. SIAM J. Comput. **30**(3), 729–752 (2000)
18. Mettanant, V., Fakcharoenphol, J.: A linear-time algorithm for the multiple gene duplication problem. In: National Computer Science and Engineering Conference (Thailand) (2008)
19. Page, R.D.: Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. Syst. Biol. **43**(1), 58–77 (1994)
20. Panchy, N., Lehti-Shiu, M., Shiu, S.H.: Evolution of gene duplication in plants. Plant Physiol. **171**(4), 2294–2316 (2016)
21. Paszek, J., Gorecki, P.: Efficient algorithms for genomic duplication models. IEEE/ACM Trans. Comput. Biol. Bioinform. (2017)
22. Zhang, J.: Evolution by gene duplication: an update. Trends Ecol. Evol. **18**(6), 292–298 (2003)
23. Zhu, J., Yu, Y., Nakhleh, L.: In the light of deep coalescence: revisiting trees within networks. BMC Bioinform. **17**(14), 415 (2016)