



ELSEVIER

Contents lists available at ScienceDirect

## Journal of Theoretical Biology

journal homepage: [www.elsevier.com/locate/yjtbi](http://www.elsevier.com/locate/yjtbi)

Letter to Editor

## The standard lateral gene transfer model is statistically consistent for pectinate four-taxon trees



## ARTICLE INFO

**Keywords:**  
Phylogenetic trees  
Lateral gene transfer  
Statistical consistency

## ABSTRACT

Evolutionary events such as incomplete lineage sorting and lateral gene transfers constitute major problems for inferring species trees from gene trees, as they can sometimes lead to gene trees which conflict with the underlying species tree. One particularly simple and efficient way to infer species trees from gene trees under such conditions is to combine three-taxon analyses for several genes using a majority vote approach. For incomplete lineage sorting this method is known to be statistically consistent; however, for lateral gene transfers it was recently shown that a zone of inconsistency exists for a specific four-taxon tree topology, and it was posed as an open question whether inconsistencies could exist for other four-taxon tree topologies? In this letter we analyze all remaining four-taxon topologies and show that no other inconsistencies exist.

© 2013 Elsevier Ltd. All rights reserved.

## 1. Introduction

A major problem in inferring species trees from gene trees is that different genes often suggest different evolutionary histories (Galtier and Daubin, 2008). This phenomenon is caused by incomplete lineage sorting and reticulate evolutionary events, i.e. hybridization and lateral gene transfer (LGT), and it naturally poses the question, whether the underlying species tree can be consistently reconstructed from a set of gene trees? In the case of hybridization, it is clear that no single tree can adequately describe the evolution of the species under study, and that a network is usually a more appropriate representation. For incomplete lineage sorting recent theoretical work based on the multi-species coalescent has shown that the most probable gene tree topology can differ from the species tree topology, when the number of species is greater than three (Degnan and Rosenberg, 2006). By contrast it has long been known that for triplets, the matching topology is the most probable topology (Tajima, 1983; Nei, 1987). Random models for LGT have been studied in a number of papers (Suchard, 2005; Galtier, 2007; Linz et al., 2007; Szöllösi et al., 2012, 2013; Roch and Snir, 2013; Steel et al., 2013), all assuming that random LGT events occur according to a Poisson process with the rate of transfers between two points in the tree either being constant or being dependent on the phylogenetic distance between the two points. Roch and Snir (2013) showed how a species tree can be reconstructed from a given number of gene trees, provided that the expected number of LGT events lies below a certain threshold; above this threshold, it becomes impossible to distinguish the underlying species tree from alternative trees. Complementary to this and to the results for incomplete lineage sorting, it was recently proved that, under the standard or extended models of lateral gene transfer, the matching gene tree topology is also the most probable topology for a tree with three species; but for the fork-shaped four-taxon tree topology there exist branch lengths for which the matching topology of a triplet has the lowest probability of the three possible topologies (Steel et al., 2013). In the original paper by Steel et al. (2013) it was posed as an open question whether this

could also be the case for other four-taxon tree topologies. In this letter we give an analysis of the remaining four-taxon tree topologies (the pectinate topologies), showing that in these cases, the matching topology for a set of three species is always the most probable topology, regardless of the location of the fourth species. This completes the four-taxon case and implies that four-taxon species trees can be consistently reconstructed using a triplet-based majority vote approach, provided that the branch lengths meet the conditions given by Steel et al. (2013). We end by presenting a curious corollary on this result.

## 2. Results

Linz et al. (2007) define what we will refer to as the *standard LGT model* with the following assumptions:

1. A binary, labeled, rooted and clocklike *species tree*  $T$  is given, as well as all the splitting times along this tree;
2. differences between a specific *gene tree* and  $T$  are only caused by LGT events;
3. the transfer rate is homogeneous per gene and unit time;
4. genes are transferred independently;
5. one copy of the transferred gene still remains in the donor genome; and
6. the transferred gene replaces any existing orthologous counterpart in the acceptor genome.

We will furthermore assume that no two transfer events occur at exactly the same time. With this model in mind, a *lateral gene transfer (LGT)* on  $T$  can be represented by a horizontal arc from a point  $p$  in  $T$  to a contemporaneous point  $p'$  in  $T$  where neither  $p$  nor  $p'$  are vertices of  $T$ . Such an arc describes the event that the gene which was present on the lineage at  $p'$  is replaced by the transferred gene from  $p$ . Thus given a species tree  $T$  and a sequence of transfer events on  $T$  for a specific gene, we obtain the associated gene tree by tracing the history of the gene from the

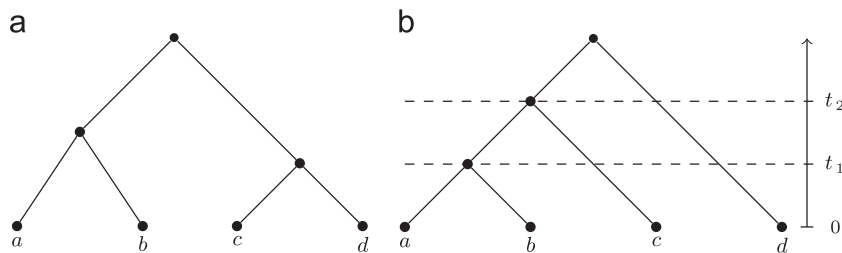


Fig. 1. The two four-taxon tree topologies. (a) The fork-shaped four-taxon tree topology and (b) The pectinate four-taxon tree topology ( $ab; c; d$ ).

present to the past, and at each encounter of an incoming horizontal arc into a lineage following this arc (against its direction). If  $A$  is a set of three species, we say that a sequence of LGT events induces a *match* on  $A$  if  $A$  induces the same topology in the associated gene tree as in the species tree.

For four-taxon trees there are two rooted binary tree topologies – the *fork-shaped* topology with two cherries as shown in Fig. 1 (a) and the *pectinate* tree topology shown in Fig. 1(b). The fork-shaped topology was studied thoroughly in Steel et al. (2013), and we will study the pectinate tree topology. The main result in this letter is the following theorem.

**Theorem 1.** *Let  $T$  be a pectinate four-taxon species tree, and let  $A$  be a set of three species in  $T$ . Then the probability, under the standard lateral gene transfer model, that a sequence of lateral gene transfer events in  $T$  induces a match on  $A$  is strictly greater than the probability that it induces either one of the two mismatch topologies (which have equal probability).*

For four species  $a, b, c, d$  we will write  $(ab; c; d)$  to denote the pectinate tree topology depicted in Fig. 1(b). This topology is symmetric to  $(ba; c; d)$ ,  $(d; c; ab)$  and  $(d; c; ba)$ , but no other symmetries hold. For any pectinate four-taxon tree we denote the time of the most recent common ancestor (MRCA) of the two most closely related species by  $t_1$ , and the time of the MRCA of the three most closely related species by  $t_2$  (with time increasing into the past such that  $t_1 < t_2$ ). Thus, for example if the tree has topology  $(ab; c; d)$ ,  $t_1$  is the time of the MRCA of  $a$  and  $b$ , and  $t_2$  is the time of the MRCA of  $a, b$  and  $c$  (see Fig. 1(b)). We now give a sketch of the proof for trees with topology  $(ab; c; *)$  (here  $*$  refers to the fourth species, the identity of which plays no role when we come to consider the topology of the triplet  $a, b, c$ ). The full proofs for trees of the topologies  $(ab; c; *)$ ,  $(ab; *; c)$ ,  $(a*; b; c)$  and  $(b*; a; c)$  are given in the supplementary material (Appendices A–C). All other topologies can be obtained from these four topologies by a permutation of the species, and the proofs of these cases therefore follow by symmetry.

**Proof.** Let  $T$  be a pectinate four-taxon species tree over the set of species  $X = \{a, b, c, *\}$  with topology  $(ab; c; *)$ , and let  $\sigma$  be a random sequence of LGT events on  $T$  generated by the standard LGT model with the rate of transfer events from a point  $p$  to any contemporaneous point  $p'$  being  $\lambda$ . We will show that the probability of  $\sigma$  inducing a match on  $A = \{a, b, c\}$  is greater than  $1/3$  and therefore greater than the probability that it induces either of the two mismatch topologies on  $A$  ( $ac|b$  and  $bc|a$ ), which have equal probability by the symmetry of  $a$  and  $b$  in the tree  $T$ .

The first thing to note is that any LGT event happening after time  $t_2$  cannot influence the topology of  $A$  in the gene tree, since by time  $t_2$  the topology of the three most closely related species becomes fixed. We can therefore ignore these transfers.

As in Steel et al. (2013), we can classify the remaining events as either  $A$ -joining,  $A$ -moving or neither of these two. An  $A$ -joining LGT event transfers a gene from a lineage that leads to a species in

$A$  to another lineage that leads to another species in  $A$ , while an  $A$ -moving LGT event transfers a gene from a lineage that does not lead to a species in  $A$  (i.e. in this case  $*$ ) to a lineage that leads to a species in  $A$ . For the precise definitions of these concepts, see Appendix B.2 in the supplementary material. Lateral gene transfer events of these two types can potentially change the topology of  $A$  in the associated gene tree, while the remaining transfers cannot. Thus, we can also ignore the events that are neither  $A$ -joining nor  $A$ -moving.

Now let  $\xi$  be the event that  $\sigma$  induces a match on  $A$ , and let  $J$  be the number of  $A$ -joining LGT events before  $t_1$ . Then by the law of total probability

$$\mathbb{P}(\xi) = \mathbb{P}(\xi|J > 0)\mathbb{P}(J > 0) + \mathbb{P}(\xi|J = 0)\mathbb{P}(J = 0). \quad (1)$$

To find  $\mathbb{P}(J > 0)$  and  $\mathbb{P}(J = 0)$  we observe that  $J$  has a Poisson distribution with mean  $2\lambda t_1$ , since at any moment in the interval  $[0; t_1]$  there are three lineages which lead to species in  $A$ , and for each of these the rate of transfers from that  $A$ -lineage to another  $A$ -lineage is  $\lambda \cdot 2/3$ . This means that the cumulative rate of  $A$ -joining transfers is  $3 \cdot \lambda \cdot 2/3 = 2\lambda$  at any given time in the interval  $[0; t_1]$ . Thus  $\mathbb{P}(J = 0) = e^{-2\lambda t_1}$  and  $\mathbb{P}(J > 0) = 1 - e^{-2\lambda t_1}$ .

Lemma 1.b(ii) in Steel et al. (2013) tells us that if there is at least one  $A$ -joining transfer in  $\sigma$ , then the first of these  $A$ -joining transfers determines the topology of  $A$  in the resulting gene tree (e.g. if the first such transfer joins  $x$  to  $y$  then the final topology of  $A = \{x, y, z\}$  will be  $xy|z$ ). Since two out of the six possible  $A$ -joining transfers lead to the matching  $ab|c$  topology, and each of the six are equally likely, the probability  $\mathbb{P}(\xi|J > 0)$  of getting a matching topology when  $J > 0$  is  $1/3$ .

It remains to find  $\mathbb{P}(\xi|J = 0)$ . When  $J = 0$  we consider the process of  $A$ -moving transfers between time  $t = 0$  and  $t = t_1$ . This is a Poisson process in which the rate at which any given species in  $A$  is moved is  $\frac{1}{3}\lambda$ , since each of the three  $A$ -lineages can be moved to only one ( $*$ ) out of three other lineages (otherwise it would be an  $A$ -joining transfer). Note that this process is independent of  $J$  as the source point of an  $A$ -joining transfer will always have an element of  $A$  as descendant, whereas an  $A$ -moving transfer would not. The walk in tree space corresponding to applying the  $A$ -moving transfers one at a time in increasing order by their time, corresponds to a simple random walk on the graph illustrated in Fig. 2 (a), in which the rate of moving from a state to each of its neighbors is  $\frac{1}{3}\lambda$ . As  $T$  has topology  $(ab; c; *)$ , the random walk starts in state 1 at time  $t = 0$ . Note that in Fig. 2(a)  $*$  is not the label of the fourth species but denotes an unlabeled lineage according to the construction in Appendix B.2 in the supplementary material. Now the probability of getting a match depends on which state the process is in at time  $t = t_1$ . Instead of analyzing the random walk on the graph in Fig. 2(a), we analyse the corresponding random walk  $Z_t$  on the graph in Fig. 2(b), where the rates denoted on the

Download English Version:

<https://daneshyari.com/en/article/6370719>

Download Persian Version:

<https://daneshyari.com/article/6370719>

[Daneshyari.com](https://daneshyari.com)