Journal of Theoretical Biology 335 (2013) 295–298

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: 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.

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; Szöllősi 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

0022-5193/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2013.07.002

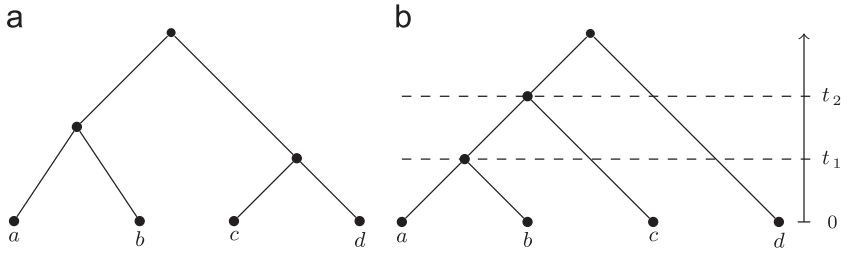

Fig. 1. The two four-taxon tree topologies. (a) The fork-shaped four-taxon tree topology and (b) The pectinate four-taxon tree topology ($ab; c; d$).

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 $\underline{\sigma}$ 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 λ. We will show that the probability of $\underline{\sigma}$ inducing a match on $A = \{a, b, c\}$ is greater that 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 ξ be the event that $\underline{\sigma}$ 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 $\underline{\sigma}$, 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

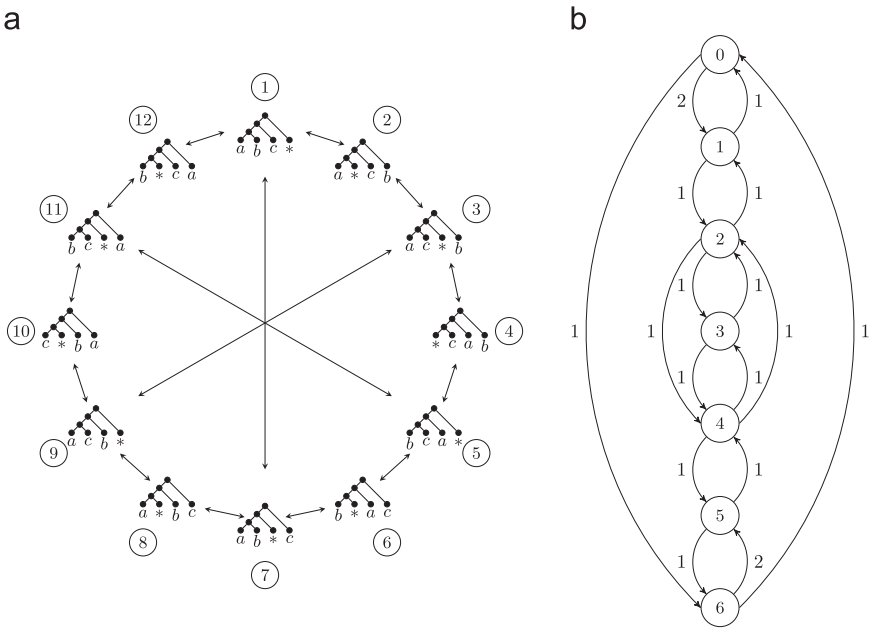

Fig. 2. The 12-state Markov chain describing the walk in tree space corresponding to applying the A-moving LGT events one at a time, and the corresponding 7-state Markov chain obtained by grouping states 2 and 12, 3 and 11, 4 and 10, 5 and 9, and 6 and 8.

edges of the graph are factors of $\frac{1}{3}\lambda$. This graph is constructed from the 12-state graph in Fig. 2(a) by grouping states 2 and 12, 3 and 11, 4 and 10, 5 and 9, and 6 and 8, and the random walk starts in state 0. Let $p_t(i)$ be the probability $\mathbb{P}(Z_t=i)$ of Z_t being in state i at time t. Then $\mathbb{P}(\xi|J=0)=\sum_{i=0}^{6}p_{t_1}(i)\mathbb{P}(\xi|J=0,Z_{t_1}=i)$, and thus

$$\mathbb{P}(\xi)=\frac{1}{3}\left(1-e^{-2\lambda t_1}\right)+e^{-2\lambda t_1}\left(\sum_{i=0}^{6}p_{t_1}(i)\mathbb{P}(\xi|J=0,Z_{t_1}=i)\right) \tag{2}$$

We find $p_{t_1}(i)$ for $i=0,1,\ldots,6$ by analyzing the continuous-time Markov chain associated with Z_t (see Appendix A in supplementary material), and we compute $\mathbb{P}(\xi|J=0,Z_{t_1}=i)$ for $i=0,1,\ldots,6$ by a case analysis of the individual states in Fig. 2(a) (see Appendix C in supplementary material). By doing this we arrive at

$$\mathbb{P}(\xi)=\tfrac{1}{3}\left(1+e^{-7\mu}\left(\tfrac{3}{4}e^{-B-4\mu}+\left(1-\tfrac{1}{2}e^{-B}\right)e^{-2\mu}+\left(1-\tfrac{1}{4}e^{-B}\right)\right)\right), \tag{3}$$

where $\mu=\frac{1}{3}\lambda t_1$ and $B=3\lambda(t_2-t_1)$. From this it is easy to see that $\mathbb{P}(\xi)>\frac{1}{3}$ for all positive values of μ and B.

The probabilities of getting either of the two mismatching topologies $ac|b$ or $bc|a$ can be computed in a very similar way. Specifically, as for the matching topology $\mathbb{P}(\zeta)=\frac{1}{3}\left(1-e^{-2\lambda t_1}\right)+e^{-2\lambda t_1}\left(\sum_{i=0}^{6}p_{t_1}(i)\mathbb{P}(\zeta|J=0,Z_{t_1}=i)\right)$, where ζ denotes one of the events $\zeta_{ac|b}$ or $\zeta_{bc|a}$ of getting topology $ac|b$ or $bc|a$, respectively. And by analyzing the states of the random walk in Fig. 2(a) we see that $\mathbb{P}(\zeta_{ac|b}|J=0,Z_{t_1}=i))=\mathbb{P}(\zeta_{bc|a}|J=0,Z_{t_1}=i))$ for $i=0,1,\ldots,6$. Thus $\mathbb{P}(\zeta_{ac|b})=\mathbb{P}(\zeta_{bc|a})$ and therefore

$$\mathbb{P}(\xi)>\tfrac{1}{3}>\mathbb{P}(\zeta_{ac|b})=\mathbb{P}(\zeta_{bc|a}), \tag{4}$$

which completes the outline of the proof. □

A plot of the probability of a random sequence of LGT events inducing a match on a triplet in a species tree of topology $(ab;c;*)$ (see (3)) as a function of $\mu=\frac{1}{3}\lambda$ and $B=3\lambda(t_2-t_1)$ is shown in Fig. 3.

It is interesting and reaffirming to study the limits of the probability of gene trees with the matching topology when t_1, t_2 or t_2-t_1 approaches 0. While a full study of these limits is given in the supplementary material, we will just pose the following corollary here.

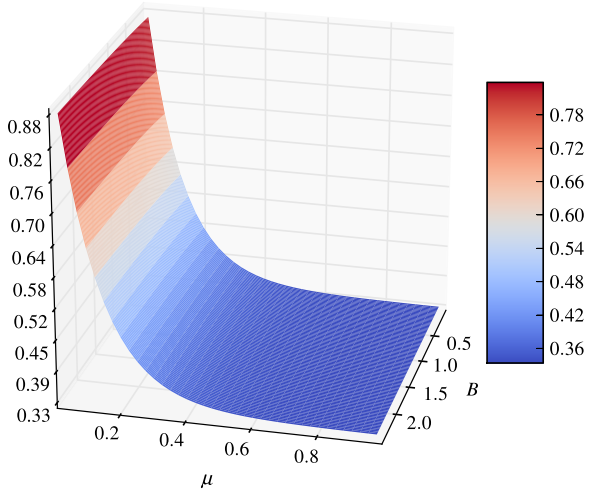

Fig. 3. The probability that a sequence of LGT events induces a match on three species a, b and c for four-taxon trees with the $(ab;c;*)$ topology as a function of $B=3\lambda(t_2-t_1)$ and $\mu=\frac{1}{3}\lambda t_1$.

Corollary 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 always approaches a value strictly greater than $\frac{1}{3}$ as t_2-t_1 tends to zero.*

This corollary may at first seem surprising, as it is in contrast to more familiar stochastic processes in phylogenetics – such as incomplete lineage sorting and site substitution models – where shrinking an interior branch length to zero results in a convergence to $\frac{1}{3},\frac{1}{3},\frac{1}{3}$ in support for the three resolutions of the resulting trifurcation. However, it is caused by the fact that, no matter how infinitesimally small t_2-t_1 is, the matching topology will be induced with probability 1 if no LGT events happen before t_1. Moreover, this apparent support for the true species tree as t_2-t_1 converges to zero

is somewhat artifactual in practice. This is because we assume here that the gene tree for A is correctly inferred, however as t_2-t_1 tends to zero, the posterior support for this tree due to either incomplete lineage sorting or to tree estimation from sequence data (or both) will converge to 1/3.

A further feature of an LGT analysis that differs from incomplete lineage sorting is that lineages beyond those connecting the species in question (e.g. involving unsampled or extinct species) can still play an important role in determining the statistical signal for different trees. This feature was highlighted recently by Szöllősi et al. (2013) who estimated that more than a quarter of LGT events appeared to involve genetic material from extinct species in a study of 36 cyanobacteria (Szöllősi et al., 2013). Results from Steel et al. (2013) show that even if just *one* additional unsampled species is present, this changes the statistical support for a tree on three species, and can lead to inconsistency in inferring a tree on that triplet, if the underlying species tree is fork shaped. But in this letter we have shown that this cannot happen if the underlying species tree is a pectinate tree. An interesting future task would be to determine precisely when triplet inconsistency can occur (in both the pectinate and fork-shaped cases), when more than one additional species is unsampled.

Acknowledgements

We thank the two anonymous reviewers of this paper for several helpful recommendations.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at http://dx.doi.org.10.1016/j.jtbi.2013.07.002.

References

Degnan, J.H., Rosenberg, N.A., 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2 (5), e68.
Galtier, N., 2007. A model of horizontal gene transfer and the bacterial phylogeny problem. Syst. Biol. 56 (4), 633–642.
Galtier, N., Daubin, V., 2008. Dealing with incongruence in phylogenomic analyses. Philos. Trans. R. Soc. B: Biol. Sci. 363 (1512), 4023–4029.
Linz, S., Radtke, A., von Haeseler, A., 2007. A likelihood framework to measure horizontal gene transfer. Mol. Biol. Evol. 24 (6), 1312–1319.
Nei, M., 1987. Molecular Evolutionary Genetics. Columbia University Press.
Roch, S., Snir, S., 2013. Recovering the treelike trend of evolution despite extensive lateral genetic transfer: a probabilistic analysis. J. Comput. Biol. 20 (2), 93–112.
Steel, M., Linz, S., Huson, D.H., Sanderson, M.J., 2013. Identifying a species tree subject to random lateral gene transfer. J. Theor. Biol. 322, 81–93.
Suchard, M.A., 2005. Stochastic models for horizontal gene transfer taking a random walk through tree space. Genetics 170 (1), 419–431.
Szöllősi, G.J., Boussau, B., Abby, S.S., Tannier, E., Daubin, V., 2012. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. Proc. Natl. Acad. Sci. 109 (43), 17513–17518.
Szöllösi, G.J., Tannier, E., Lartillot, N., Daubin, V., 2013. Syst. Biol. 32 (3), 386–397.
Tajima, F., 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105 (2), 437–460.

Andreas Sand[a,b,*], Mike Steel[c]
[a]*Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark*
[b]*Department of Computer Science, Aarhus University, Aarhus, Denmark*
[c]*Allan Wilson Centre for Molecular Ecology and Evolution, University of Canterbury, Christchurch, New Zealand*

Received 14 May 2013

Available online 13 July 2013

* Corresponding author at: Bioinformatics Research Centre, Aarhus University, Denmark. Tel.: +45 45 871 55557
E-mail address: asand@birc.au.dk