Review

Title: Consistency of orthology and paralogy constraints in the presence of gene transfers
Authors: Mark Jones, Manuel Lafond, Celine Scornavacca

This paper deals with a quite important and interesting problem in phylogenetics, namely the construction of event-labeled gene trees and species networks to explain empirical estimates of orthology (genes that diverged after a speciation event ($\mathbb{S}$)) or paralogy data (genes that diverged after a duplication event ($\mathbb{D}$)). Such data is associated with a relation graph R whose vertex set is the set of considered genes and that contains an edge xy precisely if x and y are estimated orthologs (resp., paralogs). Relations graphs that can be "explained" by a gene tree having two labels $\mathbb{S}$ and $\mathbb{D}$ that can also be reconciled with a species tree have been characterized in the last years. Moreover, methods to correct such estimates to a closest relation graphs that can be explained in such a manner have been established.

In general, however, one may expect that errors in such relations graphs R may arise due to the existence of lateral gene transfer and the fact that the underlying species history is not tree-like. In this case, one may ask, whether there is a gene tree with an additional label "transfer" ($\mathbb{T}$) that can be reconciled with a species network in a time-consistent manner. This question is addressed by the authors in the current paper.

It is shown that the problem of determining whether R is "N-consistent" is NP-hard. N-consistent means, for a given relation R and given network N, that there is a gene tree G together with reconciliation map α such that the type of events inferred by G, α and N determine the structure of R. The latter type of problem extends to allowing at most k-transfers, a problem that is shown to be W[1]-hard. Furthermore, a dynamic programming approach is provided to determine whether an event-labeled gene tree (G, ℓ) can be reconciled with a given network using a minimum number of transfer events. This algorithms runs in $O(f(k)p(|V(D)|, |V(N)|))$ time with f being a function on k (the maximum degree in the gene tree) and p a polynomial on the number of vertices in D and N and, thus, the problem is FPT. Finally, a characterization of relations graphs that are S-consistent is given and a polynomial time algorithm to decide if R is S-consistent is given for the case that S is a species tree. The latter problem becomes NP-hard, when only k transfers are allowed.

The results are very interesting and the paper is well-written. The proofs are correct as far as I checked. However, there are some issues (listed below) that should be addressed before publishing the paper.

Comments:

page 2, line 19 "using sequence similarity [29,7,among others]" seems a bit sloppy - maybe provide further references or a survey here.

page 2, line 2-4: [line 2] remove "reconciled" from "given a reconciled gene tree"

[line 3] add "set" to "displays a given [set] of relations"

[line 4] add "can" to "that [can] be reconciled"

page 4 [line 2] what does "LGT" abbreviate?

[line 8] What does it mean that a vertex is "contracted", do you mean "suppress"?

[line 5-10] Is it possible that one arc incident to the root of N is contained in E_S? In this case, the root of N' has only one outgoing arc since N is binary. However, to obtain $T_0(N)$ only vertices with in- and out-degree 1 are "contracted", which means that $T_0(N)$ may have a root with a single arc. Is this intended and possible, or does this yield problems in upcoming proofs?

[def "time-consistent"]. It took me some time to understand, when a DAG is not time-consistent. Maybe provide a small example for this case (e.g. 3-vertex DAG with arcs $(a,b), (b,c), (a,c)$). Maybe out of scope, but is there a neat characterization of time-consistent DAGs?

In addition, I cannot see in the proofs that this time-map for N is ever used except for Lemma 8. In Lemma 8, you write "add secondary arcs to S in a time-consistent manner". It seems, that you show - by using the time-map as a vehicle - that you create a DAG. So is time-consistency needed here at all?

[line -4] Def of gene tree: can you specify, what you mean with "tree"? must it be binary, phylogenetic, rooted?

page 5, Def 1 Def 1 seems to be different from the definition in [31] where switched-on/off edges are used - clarify.

The constraint (b.7) seems to be redundant, since then (b.5) is already satisfied - clarify

The definition of α allows to map leaves of the gene tree to paths in the network, as also used in your example and then they get label $\mathbb{SL}$ or $\mathbb{TL}$. Why not mapping every leaf directly to the species in which it resides and thus, forbid to map gene-leaves to paths in N?

Moreover, a reconciliation map between gene tree and species network should be time-consistent to ensure that genes do not travel through time when mapped onto the species network. I guess that the map α is always time-consistent, but this needs some verification.

What is the difference / similarities between the map μ as e.g. used in Ref [26,30] and your map α when N is restricted to be a tree?

page 6, 3rd paragraph This example does not help without a figure, that is, an explicit drawing of the gene tree embedded into the network (the reader must do this either way to understand your example). Please, provide such a drawing.

Typo: $e(\alpha_1(b_1)) = e(\alpha_1(b_1)) = \emptyset$.

Typo: $e(\alpha(c_1))$ should be $e(\alpha_1(c_1))$

page 7, line 1 "xenologs could be "interpreted" as either orthologs [] or paralogs."

This sentence is confusing, since you wrote before Sec 2.2 that it is defined based on the labeling of the lca - in which case there is no room for interpretation.

Do you mean, when inferred from sequence data? Why could orthologs or paralogs not be interpreted as transfers?

page 7, Def 2 The definition of $e^*(u,i)$ seems only be used in the proof of Lemma 2 as a vehicle and there it is defined a 2nd time. Is there a way to streamline Def 2 by just using $e(u,i)$ instead?

Moreover, it is not obvious that Def 2 covers all cases, or to be more precise: what happens if e.g. $e^*(\text{LCA}_G(x,y), \text{LAST}) = \mathbb{SL}$? is this forbidden by definition?

[Text below Def 2] ".. and that can be reconciled with N" replace by ".. and that can be reconciled with a given network N"

page 7 Why does the statement hold: "Note that, if (G,α) and R are known, there is only one relabeling e^* that ensures that (G,α) displays R"? Please, give a reference or verify.

[line -1] missing reference "??"

page 8 All the theory in Sec 2.3 goes back to the seminal paper

Böcker S, Dress AWM (1998) Recovering symbolically dated, rooted trees from symbolic ultrametrics. Adv Math 138: 105–125

and should be cited here. Maybe a few words to the structural properties (e.g. cograph) would be good to have for the reader.

Are there examples of non-cographs R that are N-consistent? If so, can you provide one?

page 13 1st sentence in Sec 4. "we show that given a set of relations R" - I think R is a relation but not a set of relations.

page 13, Sec, 4, 1st paragraph In Ref [17] it was shown that a DS-tree can always be reconciled with some network (Thm 6) and it is characterized when a DS-tree can be reconciled with a given network N (Thm 7) – at least in terms of the reconciliation map used in [17] – again how does your α differs from the map μ used there and how do these result fit into your results?

In this context it might also be worth to say that, given an event-labeled gene tree (G,ℓ) where also transfer edges in G have been specified, it is possible to determine in polynomial time if (D,ℓ) is S-reconcilable with some species tree S (even if S is not known a priory), see the work of *Hellmuth M. Biologically feasible gene trees, reconciliation maps and informative triples. Algorithms Mol*

Biol. 2017;12(1):23. together with the work [30] and [26]. In other words, the problem of finding a species tree S and a time-consistent reconciliation map between a given gene tree (G,ℓ) gets easy, if G and its event-labels incl. transfer edges are known. In this case, a time-consistent network can readily be found just by adding arcs in S on which a transfer happens (=two comparable genes in G for which their images are mapped in an incomparable way in S).

To this end, however, it would be nice to see the differences / similarities between the map μ as e.g. used in Ref [17,26,30] and your map α when N is restricted to be a tree.

Can you provide an example of an event-labeled gene tree (D,ℓ) that is not S-reconcilable with any species tree S (where S-reconcilable is in terms of the map defined in [26,30]) but N-reconcilable with some species network (latter reconcilable w.r.t. α)?

page 16, Def 5 is the species network considered in this definition still an LGT-network? please clarify.

page 16 [1st paragraph below Def 5.] Can you give an example-figure for such a "peculiar case"?

[1st paragraph below Lemma 6.] "We make every internal node of D a transfer node." This sentence is misleading, since (D,ℓ) and thus the labeling ℓ is already given. It seems however, that you change ℓ such that all internal nodes u satisfy $\ell(u) = \mathbb{T}$. please clarify.

page 18 "*We invite the interested reader to consult the Appendix for the details.*"

Can you explain where the details can be found in the appendix?