Summary of Changes on "Consistency of orthology and paralogy constraints in the presence of gene transfers"

Mark Jones, Manuel Lafond, Céline Scornavacca

Dear Dr. Holland and anonymous reviewers,

We thank all the reviewers for their insightful comments. We have revised the paper significantly and have considered every suggestion, which helped us improve the quality of the manuscript. We have clarified many definitions and added several references that were proposed by the reviewers.

Below is a detailed description of how all the comments were addressed. We also attach a copy of the revision in which every modification is highlighted in color. Note that in the bibliography, the reference numbers (e.g. [16]) of the papers have changed because we added references. The numbers in the reply below refer to the bibliography from the initial version, not the revision.

Once again, we thank the reviewers for their time and we are available to answer any inquiry,

Mark Jones, Manuel Lafond, and Céline Scornavacca

Comments from Dr. Holland

1. Page 1 Orthology and paralogy relations are often inferred by methods based on gene sequence similarity, which yield a graph depicting the relationships between gene pairs.

   Orthology and paralogy relations are often inferred by methods based on gene sequence similarity that yield a graph depicting the relationships between gene pairs.

   I always get confused about when to use which and that but here I think that is better (i.e. an essential clause)

   $-$ > Modified.

   Vertical descent with modification (speciation) constitutes only part of the events shaping a gene history;

I'd say that vertical descent with modification is different from speciation, i.e. it's evolution along an edge whereas speciation is a splitting event.

− > We prefer to keep the sentence as it is, since it is the accumulation of these mutations that leads to the splitting event.

2. page 3

   The authors ask, given a reconciled gene tree G that displays a given of relations, whether there is a species network N that be reconciled with G.

   The authors ask, given a reconciled gene tree G that displays a given of relations, whether there is a species network N that can be reconciled with G.

   − > Modified.

3. page 7 (last sentence) missing a reference It is worth mentioning the question studied in ??

   − > Added.

Comments from REVIEWER 1

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

   − > Agreed, we have added a reference to a survey paper.

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

   − > We replaced "reconciled" with "event-labeled" since the paper we refer to [19] uses this wording.

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

   − > Done.

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

   − > Done.

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

   − > "Lateral Gene Transfer", this has been added to the text.

6. line 8: What does it mean that a vertex is “contracted”, do you mean “suppress”?

   $->$ Changed.

7. 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?

   $->$ Good point, this is not intended since technically, the root has no coexisting species to transfer to. We now specify that the root of N' must have outdegree 2.

8. (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)).

   $->$ We have added an example of an acyclic forbidden structure in a time-consistent network.

   “Note that although time-consistency forbids directed cycles, not all directed acyclic graphs are time-consistent. For instance, one can easily construct an acyclic LGT network that contains two principal arcs (a, b) and (c, d), and secondary arcs (a, d) and (b, c); no time-consistent labeling is possible for a, b, c, d.”

9. Maybe out of scope, but is there a neat characterization of time-consistent DAGs?

   $->$ For LGT networks, there is a characterization in [16] and a linear time algorithm. We now mention this result. As for DAGs in general, we have not found an explicit result. The LGT network result could probably be extended, but it is indeed beyond the scope.

10. In addition, I cannot see in the proofs that this time-map for N is ever used except for Lemma 8.

    $->$ While the detailed construction of the time-map is only described in Lemma 8, we also require that the networks constructed in Lemma 5 and Algorithm 3 are time-consistent (which makes the hardness results stronger). A brief argument for time-consistency already appears after Algorithm 3; we also added “time-consistent” to the statement of Lemma 5 and added the following sentence to its proof: “Observe that N is time-consistent, since each time we insert a new secondary arc, its two endpoints are below every other previously inserted node.”

11. 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?

    $->$ Yes, because not all DAGs are time-consistent. We prefer to keep it as such since time-consistency is a stronger requirement that is closer to the biological reality.

12. (line -4) Def of gene tree: can you specify, what you mean with ”tree”? must it be binary, phylogenetic, rooted?

    $->$ Yes. Since we already say “All trees in this paper are assumed to be rooted and directed”, we simply added the word “binary” to the cited sentence.

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

    $->$ There are 3 similar definitions in [31]. We used Definition 3 from that paper (which does not deal with switching). This has been clarified in the text.

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

    $->$ Indeed, condition b.7 is intended to target reticulation nodes, specifically. We added a condition to b.5 to clarify this.

15. 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?

    $->$ We prefer to make it possible to map gene tree leaves to paths. By doing this, we rarely need to treat leaves as a special case, since they behave as internal nodes for the most part. That is, the path allows us to see the unobserved species between a gene tree leaf and its parent because of losses, as would be the case for internal nodes.

16. 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.

$->$ If we assume that N is time-consistent, the α should as well since genes must follow a path of N as we go down the gene tree. We prefer to omit going into full details, because that would require adding new definitions for the time consistency of an α map, among other things. But we did add a comment after Def 1:

"Moreover, it is not hard to see that for a given root-to-leaf path $g_1, \ldots, g_k$ of G, the concatenation of the $\alpha(g_i)$ paths correspond to a directed path in N (with some nodes that may occur multiple times in a row because of $\mathbb{D}$ nodes). Hence, if N is time-consistent, α ensures that genes evolve without going back in time."

17. 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?

    $->$ The μ essentially captures the same information, but remembers only last element of α (and may map to branches). We prefer not to make an in-depth comparison of α and μ here - as mentioned by other reviewers, the paper is already notation-heavy, and introducing the μ reconciliation map is not a necessity in terms of the main premise of the paper.

    We did add a remark and refer the reader to [17, Prop 1], which proposes a formal treatment of the relationship between α and μ.

18. 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.

    $->$ We have added such a figure.

19. Typo: $e(\alpha_1(b_1)) = e(\alpha_1(b_1)) = 0$.
    Typo: $e(\alpha(c_1))$ should be $e(\alpha_1(c_1))$

    $->$ Well spotted, we modified this.

20. 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?

$->$ Yes, here we mean that traditional predictors based on sequence similarity will output either orthologs or paralogs for xenologs. We rephrased to clarify.

21. 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?

    $->$ This is a good point and we have thought about this. In the end, we prefer to keep the e^* definition. One reason for this is that $e(...) \in \{\mathbb{S}, \mathbb{D}, \mathbb{T}\}$ is often used in the literature and should be familiar for those who worked on past models. The e^* then emphasizes how we propose to extend that model. Moreover, although e^* is not used directly in the proof, there are several places in section 4 and section 5 where we argue that $e(...) = \mathbb{T}$ must hold, with the corresponding $e^*(...) \in \{\mathbb{T}^{\mathbb{S}}, \mathbb{T}^{\mathbb{D}}\}$ being implicit from the context. Unifying e and e^* would require modifying the many occurrences of $\mathbb{T}$ to "$\mathbb{T}^{\mathbb{S}}$ or $\mathbb{T}^{\mathbb{D}}$", which could lead to mistakes and make the text a bit harder to read.

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

    $->$ Yes indeed, Def 1 specifies that the last α label must be in $\{\mathbb{S}, \mathbb{D}, \mathbb{T}\}$, and that $\mathbb{SL}, \mathbb{TL}$ are reserved for non-last α elements.

23. (Text below Def 2) ".. and that can be reconciled with N" replace by ".. and that can be reconciled with a given network N"

    $->$ Added.

24. 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.

    $->$ We have added a short argument for why e^* is determined by (G, α) and R.

25. (line -1) missing reference "??"

    $->$ Fixed.

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

$- >$ Agreed, we have added the reference at the beginning of the section.

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

    $- >$ The answer is no. As we observe in the last paragraph of section 2.3, N-consistency requires the existence of a DS-tree that displays the relations R. This DS-tree must actually be the cotree, and so R must be a cograph. In order to avoid having to introduce more definitions and proofs, we prefer to let the reader consult the references for that fact.

28. 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.

    $- >$ Indeed, the word 'relation' has a different meaning depending on the mathematical inclination of the reader. We replaced 'set of relations R' by 'relation graph R' to avoid this ambiguity.

29. 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?

    $- >$ The result is certainly worth stating, which we added at the beginning of the section. The most important difference is that transfer minimization is not considered in [17]. As in comment 17 above, the reader may consult [17, Prop 1] for the α vs μ question.

30. 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.

$- >$ It appears we had missed the AMB 2017 reference. We now cite it at the beginning of the section and mention the aforementioned results. Section 4 studies a different problem (minimizing transfers with a known network), so we do not aim to establish a precise correspondence with what we do.

31. 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)?

    $- >$ We are not sure to correctly understand this comment. If (D, ℓ) is N-reconcilable with some LGT network N, then by the definition of S-reconcilable, we have that (D, ℓ) is $T_0(N)$-reconcilable. In other words, N-reconcilable implies S-reconcilable for some S (by putting $S = T_0(N)$). Thus if (D, ℓ) is not S-reconcilable with any S, it isn't with any N. By [17, Prop 1], it doesn't appear that this would change in either type of map.

    Note that as suggested by another reviewer, S-reconcilable is now called S-base-reconcilable, which emphasizes that S should only be used as the base tree of the reconciled network.

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

    $- >$ Yes, it is. We now state this in the definition.

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

    $- >$ We do not think that this figure is highly needed and would add much to the paper. Also, drawing a gigantic D would take space.

34. (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 ‘ is already given. It seems however, that you change ‘ such that all internal nodes u satisfy $\ell(u) = T$. please clarify.

    $- >$ Agreed, we have applied the reviewer's suggestion.

35. 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?

$-$ > In the section “Proof of Theorem 4: NP-hardness of minimizing transfers with unknown transfer highways.” We now say this in the text.

Comments from REVIEWER 2

1. Instead of writing a paragraph with exemplary alpha mapping (in pg. 2, which seems to contain mistakes), I recommend providing a picture of G embedded into N with explanations. It would be beneficial in understanding the concept of gene tree-network reconciliations. The current approach might be too difficult for a reader without experience in such approaches.

   $-$ > We have added such a figure to accompany the paragraph.

2. Also, the labeling e^* should be explained directly in Definition 2.

   $-$ > We believe the reviewer intended to say that the labelling e should be explained directly in Definition 1. We now introduce e as part of Definition 1, and have moved the paragraph concerning notation around e to after Definition 1.

3. In Fig. 2, a comment should be on the presence of edge (c2,b2), since the edge seems not to represent an orthology relation from the exemplary reconciliation of G and N (which is confusing given the definition of orthology relation; however, it is formally correct, since the authors do not claim that R represents the relations from N).

   $-$ > R in Figure 1c is not $T_0(N)$-consistent but it is N-consistent using 1 transfers. We now say this.

4. Related to the above comment. Pg. 6, Sect. 2.2. Clarify that R is not the orthology graph for N (from Figure) or correct Fig. 1.

   $-$ > See answer above.

5. To be checked on page 6 (in the example):

   - in $e(alpha_1(b_1))$ repeated,

   - $e(alpha_1(b_2))$ missing,

   - $e(alpha_1(g_5)) = S$ (not T)

   - $e(alpha_2(g_5)) = T$ missing

   $-$ > Well spotted, thanks. Note that $e(alpha_1(g_5)) = SL$.

6. In Section 5.

   Definition 5 is conflicted with Definition 3. If S is a species tree, it is also a network with k=0 transfers. Also, "using k transfers", allows using 0 transfers. Thus, the notion of S-consistency is conflicted with Definition 3, when N is a species tree. The tricky part is that both notions are connected (also in the proof of Lemma 5). A careful reader can understand which definition must be applied, but it took me a while to untangle this issue.

   Suggestion: try to avoid using S-consistency and N-consistency, where N and S are defined as a network and a species tree, resp.; Maybe use "species tree-consistency"?

   $->$ Agreed, the consistencies could be more consistent. We have renamed S-consistency to S-base-consistency, making it clearer that S must be used as the base tree of an N-consistent network. We also use S-base-reconcilable.

7. - pg. 11. MWACT instead of ACT (2nd problem definition)

   $->$ Fixed

8. - The conditions on the weight functions are repeated several times (Lemma 3, Lemma 4, proofs, and other parts); I suggest introducing a new notion for the properties and removing the repeated lists.

   $->$ We prefer not to change this, as doing so runs the risk of introducing last-minute errors and would increase the already large amount of notation the reader has to remember.

9. - pg 6. the last line missing reference ??

   $->$ Added.

10. - pg 23. 7 line from the top, remove "edge"

    $->$ Removed.

11. - pg 28. 2nd line from the top, LAST subscript;

    $->$ Changed.

12. - pg 28. 3rd line "alpha ... are incomparable" - explain what does it mean in a network

    $->$ This is the same as in a tree, "i.e. none is an ancestor of the other.", c.f. page 11.

13. - inconsistent notation of edges: xy or (x,y) in several places

    $- >$ We use the xy notation for undirected graphs (i.e. the relation graph R) and (x, y) for directed graphs (i.e. the trees/networks). We clarify this when introducing R in the preliminary section.

14. The presented algorithms are clear and easy to understand Algorithm 2 can be improved by adopting more refined techniques from algorithmic papers on HGT reconciliation (see suggested papers at the end of the review), where the factor $O(|V|^2)$ can be replaced by O(1). Such an update requires the introduction of an additional formula, which for (g,s) returns the minimum cost under the assumption the g is mapped to s', where there is a path from s to s' in N, plus the cost of transfers on the best path from s to s' (note that t(s,s') will be not needed). I leave the decision to the authors on how to incorporate this observation into the results. Such an improvement is not crucial in the contribution (even if the improvement in the polynomial part of FPT algorithm is significant), so a comment would be sufficient.

    $- >$ Good point. We added a comment and a citation at the end of the section. We did not attempt to incorporate this acceleration, since we prefer to keep the algorithm as simple as possible.

15. Please provide space complexity analysis.

    $- >$ We have added it.

16. The suggested improvement in Alg. 2 is presented in several algorithmic papers on variants of reconciling a gene tree with a species tree with horizontal gene transfer e.g. by Bansal or Mykowiecka (DOI: 10.1109/TCBB.2017.2707083).

    $- >$ Thank you for the references, they are cited at the end of the DP algorithm section.

17. Modelling reconciliation with transfers: papers on H-trees by Gorecki et. al., which seems the most related to the reconciliation of (G,alpha) with the network N.

    $- >$ These are indeed related. We have added references to two H-tree papers before introducing the definition of reconciliation.

18. The question of consistency and existence of reconciliations relates to "reconciliation feasibility problems" which seem to be a simpler version of consistency problems: given sigma mapping and a species tree S, the question is whether there is a gene tree for sigma that reconciles with the S. Also, there

are more related questions e.g., on minimizing costs etc. See algorithmic papers by Eulenstein and others.

$->$ We are not sure which papers exactly the reviewer has in mind, but at the end of Section 2.3 we added references of the following papers:
`https://link.springer.com/chapter/10.1007/978-3-030-26176-4_17`
`https://link.springer.com/chapter/10.1007/978-3-319-94776-1_32`

19. Another feasibility-related paper: see M. Helmuth, 2017.

    $->$ Yes, it was also suggested by reviewer 1. We have added a reference.