Answer to the Reviewers' Major Points for : A Linear Time Solution to the Labeled Robinson-Foulds Distance Problem

Samuel Briand[1], Christophe Dessimoz[2,3,4,5,6], Nadia El-Mabrouk[1], and Yannis Nevers[2,3,6]

November 11, 2020

1. DIRO, Université de Montréal
2. Department of Computational Biology, University of Lausanne
3. Center for Integrative Genomics, University of Lausanne
4. Centre for Life's Origins and Evolution, Genetics Evolution and Environment, University College London
5. Department of Computer Science, University College London
6. SIB Swiss Institute of Bioinformatics

Editor (Céline Scornavacca)

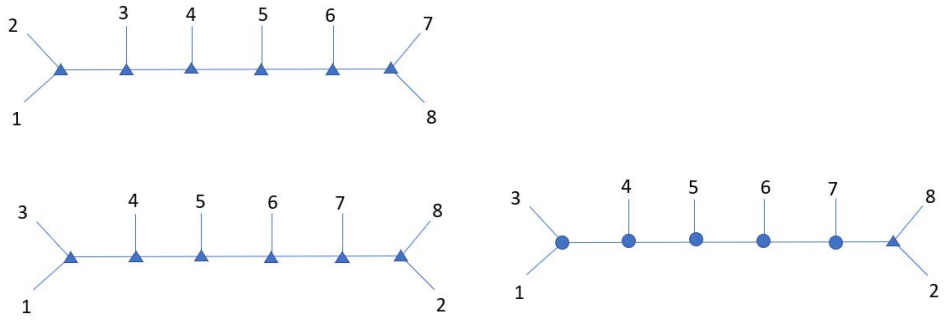
The paper has been reviewed by four (!) reviewers, who did a tremendous job, reading the paper in depth and providing constructive comments. They all agree that the paper is generally well-written and that deserves publication after fixing some issues that makes it more complex than needed (e.g. the rooted vs unrooted issue, some missing details in the proofs, a better explanation of the algorithm and of the purpose of Section 5.1, ...). When preparing the revision, the authors should answer to the major points of each reviewer in a separate text and provide a file where the modifications are highlighted (e.g. using difflatex). Also, they should compile the paper using an "plain" latex template (no Bioinformatics logo, please) and put the paper on a preprint server (e.g. arxiv).

- Thank you for your and the reviewers' detailed feedback. We have addressed all the points in this revised version of the manuscript (deposited on bioRxiv at `https://doi.org/10.1101/2020.09.14.293522`). Our point-by-point replies are provided below. As requested, we also provide a version of the manuscript with changes highlighted.

Reviewer 1 (Barbara Holland)

Major Points

The authors are quite open about the fact that the edit operations and distance defined is not necessarily very biologically reasonable. Indeed, it does seem a bit odd that e.g. the top tree below has the same distance to both of the lower trees.



- Thank you for the detailed summary of our paper. Yes, you are right, with the new LRF distance, the top tree has the same distance to both of the lower trees, which was not the case with the previous ELRF distance. This can be seen as an advantage of the previous ELRF distance, although the edit operations defined for ELRF, as those defined for RF, are not necessarily biologically meaningful either. The new LRF definition is a more direct generalization of RF leading to a linear-time algorithm, which is an important property to have, that we have found worth this additional simplification to the ELRF distance.

While the distance has been developed with application to gene-tree species tree reconciliation in mind, the applications are obviously much broader. It would be interesting to repeat the simulation study in this paper with a focus on ancestral state reconstruction to address the question of whether denser sampling improves accuracy there too.

- We agree this would be an interesting question. Given the focus of this paper on trees, we did not include this idea in the outlook, but we too would be interested in the outcome of such a study.

Minor Points

Pg 2, column 1, I am not sure what you mean by "node labels in a given tree are pairwise different", just that no labels are repeated?

- Yes. We added a precision in the introduction.

Pg 2, column 1, To address the distance drawbacks $\longrightarrow$ To address the distance's drawbacks

- Corrected.

Pg 2, col 1 . Is the node flip operation called a 'flip' because there can only be two kinds of labels? If so, might be worth mentioning this.

- Done. We explain prior to the flip operation that the target of the previous extension are trees with binary node labeling.

Pg 3, column 1 IF the node insertion operation was changed to restrict lambda to be lambda(y) would the operations then be equivalent to those for ELRF?

- No, as the contraction operation in ELRF requires $\lambda(x) = \lambda(y)$. We now make this clearer in the preamble to Lemma 3.

Pg 3, col 2, lemma 1: Is it also worth showing that the space is connected? I had a brief worry about this when I saw that operations were not allowed at the root, but some scribbling convinced me that this doesn't create an issue. I might just be worth adding a sentence or two.

- Done. We no longer apply the metric directly to rooted trees (explained in Sec 3 par 1), and therefore no longer need to add details to lemma 1.

Pg 4, col 1 It may be a start tree $\longrightarrow$ It may be a star tree

- Corrected.

Pg 5, col 1 Allow us considering each $\longrightarrow$ Allow us to consider each

- Corrected.

Pg 5, col 2 Not clear what you mean by "leftmost".

- The proof was rewritten and no longer uses "leftmost"

Were $e = x, y \longrightarrow$ Where $e = x, y$

- Corrected.

Pg 6, col The next stem $\longrightarrow$ the next step

- Corrected.

Pg 6, col 2 Does labelled gene tree inference benefits from $\longrightarrow$ Does labelled gene tree inference benefit from.

- Corrected.

Reviewer 2 (Gabriel Cardona)

Major Points

Generic: In my opinion, labeled trees usually refer to trees whose nodes are uniquely labeled by a given set. Moreover, in the context of generalizing the RF distance, bipartitions admit an straightforward generalization to (uniquely) labeled trees, even when not all internal nodes are labeled. Hence, when reading the title of the paper, my first thought (and I guess that also that of many other potential readers) was this particular generalization. Therefore, I'd suggest using a different name.

- We have added a precision in the abstract to make it clear what we mean by a labeled tree (and in particular that we are talking about the internal nodes). That being said, we still believe "Labeled Robinson Foulds" remains a reasonable shorthand for the problem. Note furthemore that we did not find any other, inconsistent use of "Labeled Robinson Foulds" in the literature.

P1,top: The authors should remove the Bioinformatics/Oxford logo in order to be published in PCI.

- We now use a plain template.

P3,C1,par 3: "can then be deduced from the RF distance of the 'unrooted version' ": This depends on how you define the "unrooted version" of a tree. With your notations it should mean to forget the root and eliminate it if it has degree two. Two different rooted trees (hence at distance > 0) may have the same unrooted version (hence at distance 0). Therefore, the "rooted version" of the distance cannot be deduced from the "unrooted version".

- You are completely right—our explanation was wrong. We have removed this paragraph and corrected what we mean, in the first paragraph of section 3. Here what we mean: "The Robinson-Foulds (RF) distance is defined in the literature for rooted and unrooted trees. Moreover, as mentioned in [5], the problem of computing the RF distance for two rooted trees can be reduced to computing the RF distance for the two corresponding unrooted trees obtained by grafting an edge linking the root to a dummy leaf. Therefore, in this paper we restrict ourselves to unrooted trees."

P2,C1,par 4: "In this paper, we focus on unrooted trees, thus avoiding the special case of the root. Therefore, from now on, all trees are considered unrooted.": One of the things that sometimes makes this manuscript a little hard to read is when they try to write all definitions and results in Section 2 suitable for both rooted and unrooted trees. If in the end they only consider unrooted trees, I'd suggest making this Section 2 more specific to unrooted trees. Note, however, that islands, defined below, are rooted if I understood it correctly (or they allow for nodes of degree two).

- Yes, as indicated in the previous comment, we now state from the beginning of Section 3 (previously Section 2, Notations) that the paper only deals with unrooted trees. We also simplified this section by removing all the notations that are specific to rooted trees in the lemmas and proofs. Finally, to the last point, no, the islands are not rooted, so we don't need these notations.

P4,C1,definition 3: I find this definition more intricate than needed. I'd say that islands are exactly the connected components obtained by removing the internal good edges such that they contain at least a leaf of the original tree. Also, with the given definition it is not clear if a node all whose incident edges are good internal edges constitutes by its own an island (but the definition allows for it).

- It is perhaps a bit counterintuitive, but the way we defined islands is such that each good edge belongs to exactly two neighbouring islands. Thus, connected components of the tree obtained by removing the internal good edges are *not* islands. We stress this for instance in Figure 2, where the dotted lines belong to the two adjacent islands. Note that we could have defined the islands as you suggest, but we found it easier to handle them the way we did. As for your question, yes a node for which all incident edges are good edges is an island. This is stated in the paragraph following Definition 3.

P4,C2,top: Please give a precise reference for the Lemma. Also, I see this formulation too intricate: why not simply say that the partition on the set of leaves induced by the islands is the same in both trees?

- We added the precise reference (lemma 3 from [5]). As stated above, we cannot talk about partitions as each internal good edge of T belongs to two islands.

P5,C1,par 5: "we clearly require at least $\epsilon(I)+\epsilon(I')$ node removals": I think it is true, but more details should be given: it must use that all edges are bad, for instance.

- We added details accordingly.

P5,C2,par 2: "This sequence of operations then leads to the tree T'_j, which is the same as T_j except possibly the two labels of x and y": This should be proved.

- Thank you for pointing out a shortcut in the proof that was not correct. We completely modified the formulation and the proof of Lemma 6.

P5,C2,par 5: "and thus P can be reordered in the form...": It should be justified that operations can be reordered.

- Actually, as mentioned by another reviewer, we don't need to reorder the operations of P. We reformulated the proof without this reordering.

P5,C1,par 2,3,4: The three paragraphs should be rewritten and expanded. Notice that the purpose of the manuscript is giving a linear-time algorithm. Hence, all these steps of the algorithm have to be fully explained, including the proofs of the running times of each step.

- All the steps are now explained to justify the linear-time claim.

Also, there is in my opinion a problem on how the iteration is done (in terms of good edges): First, some edges may not be adjacent to islands, and if they are, it has not been mentioned that these islands are unique (as the pseudocode assumes). I'd suggest iterating over islands instead of good edges; it would also avoid the problem of having to check if an island has been visited or not.

- We have clarified certain definitions and the pseudocode. We still choose to iterate over good edges because iterating over the islands of one tree makes it cumbersome to identify the matching island of the other tree. The pairing is given by the good edges, which are defined in the two trees.

P5,C1,algorithm: There is a strange mixture of lines with extremely detailed pseudocode and other ones too vague. For instance, although it may be clear from the context what getBipartitions, getIslands, islandPair do, it should at least be explicitly stated. Maybe it makes no sense to write it as a latex algorithm. The same information can be given with an itemize (nested, if needed), so that more details can be given on what is exactly computed, combining the information in the algorithm with the description given in the rest of the section.

- We have expanded and clarified the pseudocode and the description of our algorithms. In particular, we now present and describe in detail the getIsland() recursion. The only part which is not described is the identification of common bipartitions, which is needed to compute the Robinson Foulds distance and is thus well established (we provide a reference).

Minor Points

P2,C2,par 2: "admits a single,..." sounds strange: it is a tree with a distinguished node called its root "Now an internal node x is binary": specify that internal and different from the root.

- Notations on rooted trees have been removed.

P2,C2,par 3: "y is a descendant of x if y is on the path...": It is easier to say that the path from r to y passes through x

- Notations on rooted trees have been removed.

P2,C2,par 7: "As recalled in Briand et al...": Please give the original reference (where it was first proved).

- This has been removed.

P2,C2, par -3: "become the children...": Seems as if the children of y were replaced by those of x; in fact, the children of x (except for y) are added to the children of y. "Del(T, x, y)": The "Del" looks bad (typographically). It should be an operatorname or DeclareMathOperator. Same for the other operations. "removing the edge x, z...": Not needed; when you remove x, all these edges are removed by the definition of node removal.

- We have replaced "children" by "neighbors".

P3,C1,par 2: "In the case of rooted trees, the RF distance is defined as the symmetric difference between the clades of the two trees.": Repeated (appears above)

- This has been removed.

P3,C1,par 3: "The only thing that can make bipartitions and clades differ in number is rooting into a bad edge.": I don't understand what this sentence means.

- This has been removed.

P3,C2,proof of Lemma 1: It is an edit distance. Only the reversibility of the operations has to be remarked; the rest is a classical result.

- We have shortened the proof.

P3,C2,proof of Lemma 2: Maybe I miss some subtle detail, but I see this result as straightforward: since there is a single label, the operation Sub cannot be applied.

- We agree it is straightforward and we have removed the proof.

P3,C2,par -4: "Edge contraction Cont(T, x, y) similar to...": Similar or equal?

- It has been changed to equal.

P3,C2,par -3: "Node flip Flip(x, λ)": The other operations have T as their first argument.

- This was indeed an oversight. We have added T.

P4,C1,lemma 3: I'd suggest giving an example where the inequality is strict.

- We added an example (new Figure 2).

P4,C1,definition 3: "..., and all terminal edges of I are good edges of T.": I think this condition is not needed, since all terminal edges are good edges.

- Explained above.

P4,C1,par -1: " while each good edge belongs to exactly two islands of T". I'd say that it belongs to no island at all, but maybe this "belongs" is defined to make it happen. In any case, it should be clarified. See also my objection on the definition of island above and notice that the caption of Fig. 2 also uses this notion of "belongs".

- Explained above.

P5,C2,lemma 6: It should be stated in terms of node deletions, or else define what it means here the deletion of an edge.

- All edge "delete/deletion/deleting" changed to node "remove/removal/removing" since we defined the latter.

P5,C2,par 2: "...a path transforming T into T ." $\longrightarrow$ "... a path transforming T into T and assume that it involves the deletion of a good edge." In the description of cases (1)...(3): Why not do it the other way?: start with the concrete cases (2),(3) and then say "otherwise o'k=ok" Clarify what "does not affect node y" and "rename z as x" means.

- This proof has been completely rewritten. See above.

P5,C2,par 5: "As islands can only share good edges, ...": It should be stated what it means.

- Any bad edge belongs to a single island, while each internal good edge belongs to two islands. It has been clarified.

P6,C1,par 3: "stem" $\longrightarrow$ "steps"

- Corrected.

P6,C1,par 4: "is implemented by adding" $\longrightarrow$ "is updated by adding"

- Corrected.

P6, section 5.1: I don't see the relevance of this experiment.

- As we now explain at the beginning of that section, the purpose of this experiment is to get a first sense of *LRF*'s ability to measure the actual number of edits between two trees, which may be of interest to potential users.

P6,C2,par 1: "Finally, we showed that the new distance is computable for an arbitrary number of label types associated with internal nodes of the tree." I don't understand this sentence. I'd suggest modifying the previous sentence "...being a metric and reducing to..." $\longrightarrow$ "...being a metric, even for an arbitrary number of labels, and reducing to..."

- Done: the last sentence of the paragraph was removed and "being a metric and reducing to..." was changed to "...being a metric, even for an arbitrary number of label types, and reducing to...".

P6,C2,par 3: I don't see why "Our experimental results illustrate the utility of computing tree distances taking labels into account, as the conventional RF distance is blind to label changes." It is obvious that RF distance does not take labels into account, and it is independent of any experimental result.

- You are right. We no longer make this claim.

Reviewer 3 (Jean-Baka Domelevo Entfellner)

Though the authors duly cite the body of literature relative to the TED distance, including the 1989 paper by Zhang and Shasha and the 1992 paper by Zhang, Statman and Shasha, they don't explain why a linear solution is demonstrable for the calculation of the LRF edit distance on phylogenetic trees, while the 1992 reference showed that the TED on unordered labeled trees (trees in which the neighbours of a node constitute an unordered set) is NP-complete. This is because the literature on TED considers a non-constant cost function on edit operations, while for the RF and LRF distances, every operation has constant unitary cost. This should be made explicit in the paper.

- Thank you for bring up this point. We have added a sentence in the introduction to explain that the NP-complete result is for non-constant cost function on edit operations.

Similarly, as not all the readers will be familiar with the formulation of the RF distance as a total number of alpha (edge or node deletion) and alpha$^($ $-$ 1) (edge or node insertion) operations on a transformation path, it would be good to say a few precise words about it in section 2.1, for instance when the authors say "The Robinson-Foulds or Edit distance [...] is the length of a shortest path of node edit operations transforming [...]": the authors should state clearly which were the edit operations originally devised by Robinson and Foulds in their 1981 paper.

- Yes we added a precision to make the link : "is the size of a shortest path of node edit operations (i.e. edge extensions and edge contractions) transforming... "

There is a bit of a confusing or uncomfortable back-and-forth between rooted and unordered trees at the beginning of the paper, mainly for historical reasons (the TED edit distance having been used in communities, like the one of classification, where trees are naturally rooted trees). This lasts until the end of section 2, when the authors say "Therefore, from now on, all trees are considered unrooted." And yet, in the following sections, the authors keep using the words "child" and "children", while they should have used the word "neighbour(s)". Sticking to the vocables of rooted trees while talking of unrooted trees may confuse the reader.

- Notations on rooted trees have been removed and Child(ren) has been changed to Neighbour(s).

Throughout the paper, the authors seem to have turned a blind eye to the fact that unrooted trees may still contain nodes of degree 2, while the whole algorithmic machinery developed in the paper CANNOT accommodate such trees. Accepting trees in which at least one internal node has degree 2 gives birth to situations in which, according to the definition of the node edit operations by the authors in the present paper, there exists a transformation path from T to T' but NOT from T' to T, which obviously annihilates the proof that makes LRF a metric. Although internal nodes of degree 2 are usually not seen as valid phylogenetic trees, for the sake of mathematical accuracy, the authors should mention somewhere that they consider only trees whose internal nodes all have degree 3 or greater.

- We now mention in the paper that we assume that all internal nodes are at least of degree 3.

The authors should be careful, when they define subtrees in the fourth paragraph of section 2, to pay attention that their current definition, as written in the version I reviewed, does not preserve connectivity.

- Connectivity is preserved by the fact that a subtree is defined as a tree.

In several occurrences, the authors talk about the symmetric difference between sets A and B while they actually mean the size of that difference (i.e. the number of elements in the union of A-B and B-A).

- "The symmetric difference" replaced by "The size of the symmetric difference".

In definition 3 (section 3.1), please pay attention to the fact that, contrary to what the authors wrote, islands do not form a partition of the tree: any two islands share a good edge and its two connected nodes.

- Yes right. This has been corrected.

Lemmas 1, 2, 3, 5 and 6 all come with their proofs in this paper. In contrast, the proof for lemma 4 does not feature here. This is puzzling, and the reader needs to go to Briand and al (2020) to find the proof (?). In case it wasn't included here for the sake of brevity, please mention this fact, together with a few words describing a rough summary of the proof.

- The proof was added.

In the second paragraph of the proof for lemma 5, when the authors say: "On the other hand, since an edit operation can remove or insert at most one edge, [...] we clearly require [...]", they should rather say that the grounding for that part of the proof comes from the fact that all internal nodes in an island are bad ones, and therefore need to be removed.

- We added: On the other hand, as all the edges of I are bad edges, they should be all removed, before reinserting those of I'.

Please write "label-disjoint" everywhere, altering the few occurrences where "label disjoint" is written without hyphenation.

- Corrected.

The proof of lemma 6 is quite difficult to follow and it leaves the reader under the impression that weaknesses exist in there that are not addressed by the authors. The proof relies on the construction of an alternative sequence of edit operations transforming T into T', and it makes assumptions whose validity it is uneasy to check. For instance, where is the guarantee that at that stage, in the tree T_{k-1}, those z and x nodes will be neighbours? That is not straightforward, and in my opinion, the proof needs a bit of reworking or rewriting to clarify this point.

- Thank you. Your and reviewer's questions prompted us to completely rethink the formulation and proof of Lemma 6.

In the proof for lemma 6, B1 and B2 are defined as subsets of leaves (taxa), but are used as subtrees. Although every reader will understand your point there, please clarify this for the sake of mathematical correctness.

- As mentioned in the previous point, we completely changed the formulation and proof of Lemma 6.

In the logical description of the algorithm (section 4), on line 8 of the pseudocode, $(x1, y1)$ and $(x2, y2)$ are ill-defined; we understand each of those four elements denotes an island, but we don't understand why pairs of islands would be examined in pairs. The text should explain this (more) clearly.

- We have modified the pseudocode and no longer use the function called on line 8 (islandPair).

In Figure 5, it would be informative to display in the top graph (relative to the RF distance) the line with equation $y = 0.7 * x$, since an average 30% of the random edit operations are node label substitutions, to which the RF distance is totally blind. In that sense, the line with equation $y = x$ is not the "expected"/"fair" regression line here.

- Thank you for this astute suggestion, which we have implemented.

Reviewer 4 (Anonymous)

Major Points

It is not clear why RF is introduced for rooted trees in Section 2.1. and it is not always clear where the authors refer to rooted and where to unrooted trees. As the main focus of the paper is on unrooteed trees, I recommend only talking about such trees and not about rooted trees, unless the results of the paper can equally be applied to rooted trees, in which case this should be mentioned.

- The beginning of section 2 is clearer and allows us to work solely on unrooted trees.

In the proof of Thm 1 it is not necessary to show that the order of Pi can be changed on the path P. It is sufficient to show that all Pi are shortest paths and there is a shortest path preserving all good internal edges.

- Yes true. We removed the reordering in the proof of the theorem.

To me it is not clear what the aim of Section 5.1 is. Robinson Foulds, ELRF, and LRF are compared by taking trees, randomly performing edit operations that define LRF (or ELRF) and then computing RF, ELRF, and LRF distance between the computed trees. The results are as one would expect – which is resulting from the fact that for ELRF and LRF the corresponding edit operations have been used while for RF not only the operations corresponding to RF have been used. It hence is not obvious to me what the purpose of this empirical comparison of RF, ELRF, and LRF is.

- Indeed, this was not clear. We now state that the purpose of this empirical analysis is to get a first sense of LRF's ability to measure the actual number of edits between two trees, which may be of interest to potential users, compared to RF and $ELRF$.

Minor Points

Def 1, Node deletion: not only the edge x,z, but also the edge x,y should be deleted in Del(T,x,y).

- By removing x,z for each z in Neighbour (x) we remove x,y.

Def 2, Node label substitution: It is not clear if the label lambda of x can be replaced by the same label lambda by such a move.

- It can as λ can be any label of the defined finite set Λ.

Paragraph after Def 2: 'The two following lemma state [...]' $\longrightarrow$ should be 'lemmas'.

- Corrected.

Proof Lemma 2, Paragraph 2: 'Conversely, Let P be a path labeled node edit [...]' $\longrightarrow$ should be 'Conversely, let P be a path of labeled node edit [...]'

- Corrected.

Paragraph after Def 3: 'start tree' $\longrightarrow$ should be 'star tree'

- Corrected.

Same paragraph: good internal edges belong to exactly two islands - from my understanding only the endpoints belong to islands, the edges itself are not in any of the islands.

- Actually each internal good edge of a tree is a terminal edge of two islands.

Fig 2 (Now Figure 3): It is very hard to see what the islands of these trees are and the caption does not help there.

- A color code has been added to the islands of the trees, and the caption slightly edited.

Fig 4: There are a number of leaf labels missing and the number of inversions between the two topmost trees on the right is wrong (should be 2 instead of 3).

- Corrected.

Proof Lemma 6, 3rd line: first Bx and By are used, then B1 and B2.

- We completely modified the formulation and the proof of Lemma 6.

Proof Thm 1: 3rd line: 'islands share good edges' – it is not clear what that means.

- We hope we've now made it clear.