Dear PCI Evolutionary Biology editor,

All authors thank you for your encouraging and helpful work concerning our manuscript *MaxTiC: Fast ranking of a phylogenetic tree by Maximum Time Consistency with lateral gene transfers*.

We have uploaded a revision of our manuscript. We considerably modified the manuscript, to improve the presentation, usability and reproducibility of our method, guided by the remarks we received. In particular, we extensively revised the abstract and introduction, added two figures for the methodology description, added pseudo-code for the main heuristic. We now provide a documentation of the software and data for testing it.

One important note is that our group has submitted another paper under revision, which contains an application of MaxTiC to three biological datasets from the three domains of life, and the comparison of the results to other dating techniques based on the molecular clock. This is probably what reviewer 1 asks for. We chose to cut the study into two, because the description of the tool and the methodological aspects can be read independently of the application. This other paper is available on Biorxiv https://doi.org/10.1101/193813.

Here is a detailed account of our modifications, along answers to the reviewers' comments.

Decision & reviews

The manuscript has been evaluated by two referees, who agree that this method using lateral gene transfers to help finding the temporal ordering (or ranking) of nodes in a given species tree is sound and should be of interest for scientists in evolutionary biology. The referees nevertheless raise concerns about the possible target journal, about lack of sufficient details and of clarity and suggest some improvements. I have to agree that, as it stands, the manuscript may not be readable for most biologists who could use this interesting method, which could prevent a wide use of the software. I would therefore recommend writing the abstract and introduction for a broader audience and explain there the method more intuitively. The conclusion does a better job in this regards than the abstract, but could still be improved. To sum up, there is potential for an interesting and relevant contribution to the field of evolutionary biology. However, the paper needs careful revision along the lines above. If you are able to accommodate these points, I would encourage resubmission to PCI Evol Biol.

Reply: We have rewritten the abstract, introduction and conclusion to better target an audience of biologists. Indeed, biologists and bioinformaticians interested in using the methods described in the paper are our intended audience, even if, as noticed by the reviewers, computer scientists and bioinformaticians can also be interested by the use of the feedback arc set problem, the mixing heuristic and the problems opened by this study.

We removed computer science technical vocabulary from the abstract and introduction. However, we maintained a certain level of formalism to keep the explanation of the principle clear enough.

We also revised the method section to make it more accessible and detailed. For example, we removed the definition section and placed the definitions where needed in the text. We also added two figures to follow the execution of the main algorithm on an example, and a pseudo-code so that the heuristic can be fully implemented without ambiguities.

Reviewed by Alexandros Stamatakis, 2017-07-14 10:47

The authors present a very nice application of theoretical computer science results (the feedback arc problem) to a real biological problem. They develop a heuristic for minimizing the number of conflicts between a ranked order of nodes in the species tree and corresponding time constraints as obtained by programs for detecting lateral gene transfer.

The paper is overall nicely written and in general I would recommend acceptance as a reviewer. However, it is unclear for which journal this would be appropriate. The algorithms and theory are not described in sufficient detail (see some comments below) to merit publication in a more theoretical CS-style journal (like Journal of Theoretical Biology or BMC Algorithms for Molecular Biology) . In addition, there is too much algorithms and not enough biology for a journal like Syst Bio or MBE. So, I believe, the options here are to either make it more biological by moving most of the algorithms stuff to an on-line supplement and analyzing some recently published high-profile biological datasets or describe the algorithms in more detail and opt for a more theoretical journal.

Reply: This work is indeed a hybrid, as the reviewer notes. The method presented was conceived to address a biological problem, but such a usage is not fully worked through here. We must clarify that we have also submitted a biological case study based on this method. This study applies the detection of transfers as described here to three sets of genomes from the three domains of life, and compares the results with those of several other methods based on the molecular clock. This latter work, however, is too large and autonomous to be described in the present manuscript as an application of the method. Just as the current work is too large and autonomous to be the method section of this mentioned biological work. So we have two currently submitted papers, one more methodological here and one more biological, that we now refer to as (Davin et al, 2017), and that we have deposited on Biorxiv as https://doi.org/10.1101/193813.
Our option here is to present the paper as a methodological work and a tool which can be used for a biological work. We revised the manuscript in order to fill this objective, with the goal of being understandable by a wide community in biology and bioinfomatics in the abstract, introduction and conclusion, and to detail the method in the appropriate section.

If we were to send it to a journal, it would probably be a bioinformatics or computational biology journal, or the resource section of a biology journal. However if we get the PCI Evolutionary Biology recommendation we don't intend to submit it to any journal, because we feel the recommendation would be a sufficient proof for whom it concerns that it has passed a selective peer review process.

Detailed comments:

The link to the github repo with the python scripts is insufficient for reproducing the results. The authors should describe in detail how APE etc. needs to be installed, how the python scripts were executed, where the simulated datasets can be downloaded etc. etc., i.e. a full transcript that allows for easily reproducing the results must be put together.

Reply: The criticism is entirely justified and we apologize for this oversight when submitting the first version of the manuscript. We have now documented the software so that it can be easily used. It does not need any library and it is self contained, provided that the user has python 2 installed and downloads two files.

We also provide a documentation and the data, along with the scripts, to run the analyses described in the manuscript. The result of a simulation and a small biological dataset are fully reproducible from it. We also provide the program we use to transform transfers given by ALE into constraints.

We hope to achieve much better usability and reproducibility standards with this version.

---

page 3: The authors should provide a more extended rationale regarding the simulation settings with SimPhy (why 1000 gene trees, why pop size between 2 and 10^6, why a transfer rate from 10^-9 t0 10^-6, etc.).

Reply: We now provide in the manuscript full explanations on our choices. The ranges of parameters were guided by the willing to show some tendencies in the behavior of our tool and pipeline, and to reproduce some measures we had on biological datasets.

---

page 5: the proof and algorithm description needs at least 2-3 additional figures that would make everything much easier to follow, e.g., Theorem I needs a figure, the mixing principle needs a figure, the dynamic programming algorithm needs a figure.

Reply: We added two figures to illustrate the theorem, the mixing principle and the dynamic programming algorithm.

---

page 5: the log n approximation should be mentioned earlier in the sentence where you mention that there is no constant factor approximation.

Reply: Done. We also corrected a typo in the approximation ratio.

page 6: For the sake of completeness: provide (i) time and space complexities (ii) pseudocode of the algorithm

Reply: Both done.

page 6: The description of the local search is a bit fuzzy and incomplete, e.g., I don't understand when it terminates and how exactly it works, apart from the fact that it apparently does some sort of randomized search.

Reply: We dedicated an additional paragraph to the description of the local search, and added a reference for our choices in this regard.

page 7: would it be possible to design a program that solves the problem exhaustively on small instances and use it on some empirical dataset, e.g., the small yeast genome dataset from Antonis Rokas?

Reply: We implemented this suggestion on a small dataset constructed as a subset of the cyanobacteria dataset we already use in the paper, available from a 2015 publication (Szollosi et al, *Proc Roy Soc* 2015). We chose this dataset instead of another publicly available one for coherence with the rest of the paper, and because searching for dating information in another public dataset would add an unnecessary complication.

We added a paragraph presenting the results of the heuristic and of the exhaustive search on this small dataset.

As already stated above, I believe that this manuscript could become more interesting to the user community if you showed that the method produces "interesting" results on some recently published phylogenomic studies.

Reply: As mentioned earlier we have another paper in revision which reports such a study and uses the method presented here (Davin et al, 2017, Biorxiv https://doi.org/10.1101/193813). We chose to split the study into two because they were conducted as different research programs, and both contain autonomous messages.

page 10: Why did you fix the transfer rate to 10^-6 for assessing uncertainties in the species tree?

Reply: We have precised in the text that this value has been taken as a reference because it is one of the rates which gave the measured number of transfers per family that compared well with what we measured on the two biological datasets.

Reviewed by anonymous reviewer, 2017-08-08 17:46

This paper introduces a technique for using lateral gene transfers (LGTs) to estimate the temporal ordering (or ranking) of nodes in a given species tree. The technique is based on the idea that any correctly inferred LGT must be compatible with the true ranking of the species tree (i.e. donor species could not have lived more recently than the recipient species). The paper proposes a heuristic algorithm that takes as input an unranked species tree and a weighted list of LGTs, inferred using existing methods, and computes a ranking of the species nodes that is compatible with a maximum weight subset of the LGTs. An experimental study using simulated data suggests that the objective of seeking a ranking of the species nodes that is compatible with a maximum weight subset of the LGTs is generally reasonable, even though the true ranking often does not maximize the weight of compatible LGTs. The experiments also show that the heuristic algorithm generally produces fairly accurate rankings.

Some aspects of the algorithm description and experimental setup can be improved as follows.

a. The paper vaguely suggests, but does not prove, that the proposed heuristic algorithm is a log n-approximation algorithm for the maximum compatibility problem. Since the species tree can be unbalanced, it is not clear if this is the case. This should be clarified in the text.

Reply: We made it clear that it is not the case, unless the species tree is somewhat balanced (and we corrected a typo: we refer to log^2 n approximation ratios). As now clarified in the text, we feel that this loss in theoretical guarantee is counterbalanced by the description of an exact algorithm that mixes two ranked trees, a solution which can be interesting *per se*.

---

b. The description of the "mixing" problem in the abstract and in section 3 is confusing. It should be clarified that the mixing step only solves the constrained problem where the given orders for the two subproblems are preserved. The current description suggests that an optimal ranking is computed, which is not the case.

Reply: We clarified this by providing two pseudo-codes: One describes the heuristic on the general problem and calls the second one, which is an exact solution to the constrained sub-problem. The headlines of the pseudocodes should be explicit enough that the reader can understand that we solve the constrained problem exactly, and the general problem heuristically.

c. The experimental study is interesting and informative but uses an overly simplified model of evolution. The paper also claims that the data was generated "under conditions comparable to published biological datasets", but this is not correct. In simulating the gene trees, no gene duplications or gene losses are allowed. This makes the simulation study a bit unrealistic. There should at least be a reasonable lost rate used (approximately equal to the LGT rate), even if gene duplications are not allowed.

Reply: We agree that the simulated datasets cannot be claimed realistic. Even complexifying the scenarios by adding duplications and losses would not make them realistic for many reasons, including the fact that all available simulators that we can use assume the clustering of sequences into given families, although this is in itself a hard problem with no good solution in practice. We changed the sentence claiming that conditions were comparable in simulations and reality, which was indeed inappropriate. The only thing that we can claim is that we compare some outputs of the simulated dataset with comparable measures made on biological datasets, and that we gain some understanding of the behavior of our tool from this comparison.

To what point we should complexify simulations in order to produce difficult datasets is indeed a crucial question. Note that in the current simulations, transfers are "replacing transfers", so they involve a loss in the recipient lineage. In that view, loss rate is approximately equal to LGT rate in all ALE inferences even though we explicitly set it to zero in Simphy.

The choice to set duplications and loss rates to zero is made to restrict the exploration of the parameter space, and construct a situation similar to a dataset which has been restricted to universal or near universal unicopy genes, as it is often the case in whole genome studies. Instead of modeling all possible events in a genome, we focus on the families where discrepancies with species history can be attributed with great confidence to transfers. We propose this as a trade-off between the expense in computation needed to perform simulations and the knowledge we gain from it.

d. To properly understand normalized Kendall similarity, it would help to include the average normalized Kendall similarity for a random ranking of the nodes in the species tree. It looks like Figure 8 might include this information, but the description is confusing. This information should be included in the main text and the description of Figure 8 should also be clarified.

Reply: We clarified the caption of figure 10 (figure 8 in the first submission) and the related text. The problem with including it in the other figures is that the distribution changes with the species tree, and the species tree is different in every simulation. It is doable for Fig 10 where the reference species tree is unique for all the analysis. Even if we make it vary a bit it is not re-simulated from scratch for each point, as it is the case for all other case studies. So the random distribution does not vary much between the different points in that precise case.

What we did in order to provide the user with such an information is that we added to the program the possibility to generate random rankings and compare the input tree with the random ones, with a p-value.

---

e. The authors investigate the relationship between the number of input LGTs and the accuracy of the ranking. However, from the perspective of an end user, it would still be difficult to determine if the input set of LGTs is sufficient to confidently rank the entire species tree. Is it possible to extend the heuristic algorithm to only output the portions of the ranking that are well-supported by the input LGTs?

Reply: There are several ways of providing such an information. First, we output a total order while the real solution is indeed a partial order because some relations are not supported by the input and thus are randomly chosen. This can be discriminated in the present program, which outputs the constraints that are compatible with the total order, which define the partial order supported by the data. Second, one can ask a statistical quantification of the support of some patterns of the total order. We have achieved this for the companion paper by jackknifing the constraints.