Peer Community In Evolutionary Biology

Dating nodes in a phylogeny using inferred horizontal gene transfers

Tatiana Giraud *and* ***Toni Gabaldon*** *based on peer reviews by* ***Alexandros Stamatakis***, ***Mukul Bansal*** *and 2 anonymous reviewers*

Cédric Chauve, Akbar Rafiey, Adrian A. Davin, Celine Scornavacca, Philippe Veber, Bastien Boussau, Gergely J Szöllosi, Vincent Daubin, and Eric Tannier (2017) MaxTiC: Fast ranking of a phylogenetic tree by Maximum Time Consistency with lateral gene transfers. Missing preprint_server, ver. Missing article_version, peer-reviewed and recommended by Peer Community in Evolutionary Biology. 10.1101/127548

Submitted: 28 June 2017, Recommended: 07 November 2017

Cite this recommendation as:

Giraud, T. and Gabaldon, T. (2017) Dating nodes in a phylogeny using inferred horizontal gene transfers. *Peer Community in Evolutionary Biology*, 100037. 10.24072/pci.evolbiol.100037

Published: 07 November 2017

Dating nodes in a phylogeny is an important problem in evolution and is typically performed by using molecular clocks and fossil age estimates [1]. The manuscript by Chauve *et al.* [2] reports a novel method, which uses lateral gene transfers to help ordering nodes in a species tree. The idea is that a lateral gene transfer can only occur between two species living at the same time, which indirectly informs on node relative ages in a phylogeny: the donor species cannot be more recent than the recipient species. Horizontal gene transfers are increasingly recognized as frequent, even in eukaryotes, and especially in micro-organisms that have little fossil records [3-7]. Yet, such an important source of information has been very rarely used so far for inferring relative node ages in phylogenies. In this context, the method by Chauve *et al.* [2] represents an innovative and original approach to a difficult problem. An obvious limitation of the approach is that it relies on inferences of horizontal transfers, which detection is in itself a difficult problem. Incomplete taxon sampling, or the extinction of the true donor lineage may render patterns difficult to interpret in a temporary fashion. Yet, for clades with no fossils this may be the only piece of information we have at hand, and the growing amount of sequence data is likely to minimize issues derived from incomplete sampling. The developed method, MaxTiC (for Maximal Time Consistency) [2], represents a very nice application of theoretical developments on the well-known « Feedback Arc Set » computer science problem to the evolutionary question of ordering nodes in a phylogeny. MaxTiC uses as input a species tree and a set of time constraints based on lateral gene transfers inferred using other softwares, and minimizes conflicts between node ordering and these time constraints. The application of MaxTiC on simulated datasets indicated that node ordering was fairly accurate [2]. MaxTiC is implemented in a

freely available software, which represents original and relevant contribution to the field of evolutionary biology.

References:

[1] Donoghue P and Smith M, editors. 2003. Telling the evolutionary time. CRC press.

[2] Chauve C, Rafiey A, Davin AA, Scornavacca C, Veber P, Boussau B, Szöllősi GJ, Daubin V and Tannier E. 2017. MaxTiC: Fast ranking of a phylogenetic tree by Maximum Time Consistency with lateral gene transfers. bioRxiv 127548, ver. 6 of 6th November 2017. doi: [10.1101/127548](https://doi.org/10.1101/127548)

[3] Ropars J, Rodríguez de la Vega RC, Lopez-Villavicencio M, Gouzy J, Sallet E, Debuchy R, Dupont J, Branca A and Giraud T. 2015. Adaptive horizontal gene transfers between multiple cheese-associated fungi. Current Biology 19, 2562–2569. doi: [10.1016/j.cub.2015.08.025](https://doi.org/10.1016/j.cub.2015.08.025)

[4] Novo M, Bigey F, Beyne E, Galeote V, Gavory F, Mallet S, Cambon B, Legras JL, Wincker P, Casaregola S and Dequin S. 2009. Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast Saccharomyces cerevisiae EC1118. Proceeding of the National Academy of Science USA, 106, 16333–16338. doi: [10.1073/pnas.0904673106](https://doi.org/10.1073/pnas.0904673106)

[5] Naranjo-Ortíz MA, Brock M, Brunke S, Hube B, Marcet-Houben M, Gabaldón T. 2016. Widespread inter- and intra-domain horizontal gene transfer of d-amino acid metabolism enzymes in Eukaryotes. Frontiers in Microbiology 7, 2001. doi: [10.3389/fmicb.2016.02001](https://doi.org/10.3389/fmicb.2016.02001)

[6] Alexander WG, Wisecaver JH, Rokas A, Hittinger CT. 2016. Horizontally acquired genes in early-diverging pathogenic fungi enable the use of host nucleosides and nucleotides. Proceeding of the National Academy of Science USA. 113, 4116–4121. doi: [10.1073/pnas.1517242113](https://doi.org/10.1073/pnas.1517242113)

[7] Marcet-Houben M, Gabaldón T. 2010. Acquisition of prokaryotic genes by fungal genomes. Trends in Genetics. 26, 5–8. doi: [10.1016/j.tig.2009.11.007](https://doi.org/10.1016/j.tig.2009.11.007)

Reviews

Evaluation round #2

DOI or URL of the preprint: **10.1101/127548**
Version of the preprint: 3

Authors' reply, 02 November 2017

Download author's reply

Decision by Tatiana Giraud, posted 27 October 2017

minor revisions

I was pleased to see that this manuscript has been further carefully revised, and there remain only a few minor additional suggestions that should be addressed before the manuscript can be recommended by PCI

Reviewed by Mukul Bansal, 18 October 2017

The authors have addressed my major concerns and the updated manuscript is a clear improvement over the initial submission. The manuscript now provides an improved description of the heuristic algorithm and of the experimental analysis. The work is definitely interesting, and the proposed method has the potential to be quite useful for species tree dating in prokaryotes. I only have a single minor comment, which the authors can address as they see fit: The new text added to the manuscript has more grammatical errors than the original text from the initial submission. Carefully proofreading the newly added text would help.

Reviewed by anonymous reviewer 2, 07 October 2017

This paper seems now in mostly good shape, following the previous reviews and the revisions the authors have made in response to those earlier comments (I was not one of the previous reviewers). The idea of developing algorithms to rank nodes in trees using transfer events is a timely one, and the two algorithms described for minimizing conflicts (one heuristic and one exact) appear to be both new and sound. Accordingly I am happy to recommend publication, however, I have a few suggestions that will be easy for the authors to address.

1. The authors should cite and briefly discuss this paper A Method for Investigating Relative Timing Information on Phylogenetic Trees Daniel Ford, Frederick A. Matsen and Tanja Stadler Systematic biology, 58 (2): 167-183, 2009. While it doesn't directly deal with transfers, nevertheless the ideas in it are very relevant to this paper.

2. [optional] In the proof of Theorem 1, the authors could point out that the choice of a comb tree (line 3) is entirely arbitrary. Also, with slightly more work (and more "dummy" leaves o_i) one could also even ensure that each node is associated with just one arrow (unlike fig 3(b) where some nodes are associated with 2 and 3 arrows).

3. page 11 "Figure 6" - in my version I see the caption but no figure (>?!)

4. Proof of Theorem 2. I'd suggest starting it with \\em Proof\
   Maybe also flag at the start that the algorithm is based on dynamic programming techniques. Then replace "Indeed, call, ..., the sequence" -> "Let.... denote the sequence" line 4 of proof: "note $C*N$ the set" -> "let $C*N$ denote the set" Figure 4 - make the arrows on the end of the transfer arrows bigger next para: "Note $N*ij*\=... *Let then"* -> *"Let $N*\ij\= .., and let" page 9 - may put the usual square box for \end\proof\ just before the para "Applying the mixing..."

Evaluation round #1

DOI or URL of the preprint: **10.1101/127548**
Version of the preprint: 2

Authors' reply, 28 September 2017

Download author's reply

Decision by Tatiana Giraud, posted 09 August 2017

Revise

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 sofware. 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.

Reviewed by **Alexandros Stamatakis**, 14 July 2017

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.

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.

While I did not do this here, I usually also check the software that was developed with various tools (e.g., for C/C++ compiling with clang and all warnings enabled, checking with valgrind, checking for cyclomatic complexity etc etc.) to obtain a feeling for the respective code quality.

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.).

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.

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

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

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.

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

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.

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

Reviewed by anonymous reviewer 1, 08 August 2017

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.

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.

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.

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.

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?