# Dating with transfers

Adrián Davín, Gergely J Szöllosi, Eric Tannier, Bastien Boussau, Vincent Daubin

# Dating with transfers

Adrián Davín * †1, Gergely Szöllősi[2], Éric Tannier[1,3], Bastien Boussau ‡ 1, Vincent Daubin § 1

[1] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I (UCBL), INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[2] Eötvös University – Hongrie
[3] INRIA Rhône-Alpes (INRIA Grenoble Rhône-Alpes) – INRIA – ZIRST 655 avenue de l'Europe, Montbonnot, F-38 334 Saint Ismier Cedex, France

Session phylogénie 2
mardi 28 17h10
Amphi Mérieux

To reconstruct the timing of the diversification of life on Earth, biologists combine fossil evidence with inferences drawn from the comparison of genome sequences. These inferences are based on the molecular clock or on a softer version of it, the relaxed molecular clock. This approach consists of estimating the divergence between sequences and then, assuming that mutations occur clock-wise, trying to determine the age of the ancestral sequence. This method can be refined by using diverse models that relax the hypothesis of constant pace of evolution and consider that sequences can evolve at different speeds. Some models assume that these rates of evolution are independent along the different branches of the tree relating the different sequences of DNA; some others consider that the rates are correlated among related branches, so the rate of a given branch is inherited to some extent from the parental one. Which is of these methods perform best is still heavily debated. In spite of the sophistication of the different models, calculating these rates is not a trivial problem and the best estimates of divergence time have usually very wide confidence intervals. To overcome this problem scientists can use fossils, that can be independently dated using methods such as stratigraphy or radiometry. Fossils are useful because they provide external information that can be used to constrain the positions of the nodes in a species tree, improving the accuracy of the estimates of the molecular clock. Combining relaxed molecular clock estimates and fossil is in active field of research in phylogenetics [1].

However, fossils are extremely scarce in the geological record. For about 80 % of the history of life, all organisms were unicellular, which means that finding fossils becomes an almost impossible task. Bones and hard shells are easy to be preserved but they just became frequent after the Cambrian explosion, when all the major animal clades appear at sudden. Before that Earth was dominated by bacteria and to a minor extent, small eukaryotes. These organisms are extremely small organisms with no hard parts that can fossilize easily. On top of that, for the few existent fossils we have there is very little certainty about the clades to which they belong, since morphological features cannot be used to place them in a phylogenetic tree. This means that if we are interested in studying what happened in the distant past, we have very little help coming from fossils and we must rely almost exclusively in the information conveyed by the DNA. As we previously stated, this is a hard problem since the estimates of the molecular clock can vary widely. We need accurate calculations if we want to know for example when did Eukaryotes diversify or when did cyanobacteria appear on Earth.

To overcome these problems, we propose a new method of dating, based on the DNA sequence complementary to the molecular clock. Lateral gene transfer (LGT) is a common and almost universal phenomenon in nature, where different species (sometimes even species belonging to different domains) exchange genes. This can be detected using differences between species trees and gene trees. We do this using ALE, a method to reconcile species trees and gene trees that allows

*. Intervenant
†. Corresponding author : adrian.arellano-davin@univ-lyon1.fr
‡. Corresponding author : bastien.boussau@univ-lyon1.fr
§. Corresponding author : vincent.daubin@univ-lyon1.fr

detecting gene duplication, transfers and loss with high accuracy [2]. ALE takes distribution of gene trees to consider the uncertainty in the tree topology and estimates event rates by taking into account these gene trees. By doing this, it leads to better estimates of lateral gene transfers than by using methods that rely only on the comparison between gene trees and species trees, which have been shown in simulations to consistently estimate an incorrect number of transfers.

These gene transfers events contain information that can be used to order the divergence of different clades, since an existing clade can only donate a gene to other contemporary clades. Put in other words, if we detect a transfer between A and B necessarily means that the ancestors of A are older than any descendant of B. This same type of analysis can be performed over many thousand families to detect a large number of transfers that are then converted to a large number of node order constraints. This complements molecular dating analyses for a time when we don't have any fossils to use and every possible source of information must be used [3].

We analyzed several data sets to investigate whether the dating information carried by transfers agree with the information carried by the relaxed molecular clock. For each data set, we built species trees using concatenates of universal genes alignments and bayesian inference. We then computed gene tree distributions and inferred transfer events using the software ALE. We find that in all cases, dates based on relaxed molecular clocks are more consistent with the relative constraints coming from the transfers we detect than random trees. Further, among relaxed molecular clock estimates, we find that a particular model of rate evolution, where branchwise rates are independently drawn from a Gamma distribution, agrees consistently better with the transfer-based constraints than other models of rate evolution.

Our results show that transfers carry a signal for dating species trees that is compatible with and therefore can advantageously complement existing methods based on relaxed molecular clock. Further, they suggest an approach for choosing between different models of the rate of molecular evolution

## References

[1] Lepage, T., Bryant, D., Philippe, H., & Lartillot, N. (2007). A general comparison of relaxed molecular clock models. *Molecular biology and evolution*, 24(12):2669-2680.

[2] Szöllősi, G. J., Rosikiewicz, W., Boussau, B., Tannier, É., & Daubin, V. (2013). Efficient exploration of the space of reconciled gene trees. *Systematic biology*, syt054.

[3] Szöllősi, G. J., Boussau, B., Abby, S. S., Tannier, É., & Daubin, V. (2012). Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proceedings of the National Academy of Sciences*, 109(43):17513-17518.

**Mots clefs :** Dating, molecular clock, phylogenetics, LGT, ALE