Relative time constraints improve molecular dating

Gergely J. Szöllősi*, Sebastian Höhna, Tom A. Williams, Dominik Schrempf, Vincent Daubin, Bastien Boussau*

*: corresponding authors: ssolo@elte.hu; bastien.boussau@univ-lyon1.fr

Affiliations:

Gergely J. Szöllősi:
MTA-ELTE "Lendület"' Evolutionary Genomics Research Group, Pázmány P. stny. 1A, H-1117 Budapest, Hungary;
Department of Biological Physics, Eötvös University,Pázmány P. stny. 1A, H-1117 Budapest, Hungary;
Evolutionary Systems Research Group, Centre for Ecological Research, Klebelsberg Kunó u. 3, H-8237 Tihany, Hungary.
Email: ssolo@elte.hu

Sebastian Höhna:
GeoBio-Center LMU, Ludwig-Maximilians-Universität München, Richard-Wagner Straße 10, 80333 Munich, Germany;
Department of Earth and Environmental Sciences, Paleontology & Geobiology, Ludwig-Maximilians-Universität München, Richard-Wagner Straße 10, 80333 Munich, Germany. *Email:* hoehna@lmu.de

Tom A. Williams: School of Biological Sciences, University of Bristol, 24 Tyndall Ave, Bristol, BS8 1TH, United Kingdom.
Email: tom.a.williams@bristol.ac.uk

Dominik Schrempf: Dept. Biological Physics, Eötvös University, Pázmány P. stny. 1A., H-1117 Budapest, Hungary.
Email: schrempfd35@univie.ac.at

Vincent Daubin: Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR 5558, F-69622 Villeurbanne, France.
Email: vincent.daubin@univ-lyon1.fr

Bastien Boussau: Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR 5558, F-69622 Villeurbanne, France.
Email: bastien.boussau@univ-lyon1.fr

Abstract

Dating the tree of life is central to understanding the evolution of life on Earth. Molecular clocks calibrated with fossils represent the state of the art for inferring the ages of major groups. Yet, other information on the timing of species diversification can be used to date the tree of life. This is the case for instance for horizontal gene transfer events and ancient coevolutionary relationships such as (endo)symbioses, which can imply temporal relationships between two nodes in a phylogeny (Davín et al. 2018). This can be particularly helpful when the geological record is sparse, e.g. for microorganisms, which represent the vast majority of extant and extinct biodiversity.

Here, we demonstrate that relative age constraints, when combined with fossil calibrations, can significantly improve both the accuracy and resolution of molecular

clock estimates. We provide an implementation of relative age constraints in RevBayes (Höhna et al. 2016) that can be combined in a modular manner with the wide range of molecular dating methods available in the software.

To validate our method in a realistic data setting we apply it to two data sets of 40 Cyanobacteria and 62 Archaea respectively, and provide cross-validations of fossil calibrations and relative age constraints.

Introduction

Dated species trees (chronograms or timetrees, in which branch lengths are measured in units of real time) are used in all areas of evolutionary biology. Their construction typically involves collecting molecular sequence data, which are then analyzed using probabilistic models. Commonly, a *relaxed molecular clock* approach is adopted. Such methods typically combine at three components: a model of sequence evolution, a model of rate variation across the phylogeny, and priors on node ages. Inference is typically performed using Bayesian MCMC algorithms.

Inferring the age of speciations based on molecular data is challenging because it amounts to factoring divergence between sequences, estimated in units of substitutions per site, into time (ages of splits) on one hand, and rates of evolution on the other hand. Additional information on ages and rates must be provided through additional data, or priors. Information on node ages is provided through *calibrated* nodes, *i.e.* nodes that can be associated to a date in the past, usually with some uncertainty. The information on how to calibrate a node can come from fossils, but any information about a date in the past that can be associated to nodes can be used. External data is rarely available to inform rate inference, so the prior on the rate is usually loose. Information on the rate of evolution then derives from the information contained in the analyzed sequence data, and in the node age calibrations. Details of the model of rate evolution therefore matter a lot. When rates can be considered to be constant throughout the phylogeny, *i.e.* when the strict molecular clock hypothesis (Zuckerkandl and Pauling 1962) can be applied, a single rate needs to be estimated. For data sets that do not fit the strict molecular clock hypothesis, different rates need to be used to model sequence evolution in different parts of the tree. Several such relaxed clock models have been proposed (Thorne, Kishino, and Painter 1998; Drummond et al. 2006; Heath, Holder, and Huelsenbeck 2012; Lepage et al. 2007; Lartillot, Phillips, and Ronquist 2016) to account for rate variation across the phylogeny. Some assume that branch-wise rates are drawn independently of each other from some prior distribution (Drummond et al. 2006; Lepage et al. 2007; Heath, Holder, and Huelsenbeck 2012). Others assume that neighboring branches are expected to have more similar rates than distant branches (Thorne, Kishino, and Painter 1998), and a model that can accommodate both situations has recently been proposed (Lartillot, Phillips, and Ronquist 2016). The sophistication of relaxed clock models comes at a price: inference is computationally more demanding than under the strict molecular clock. This is because some of their parameters are highly correlated. However, they typically provide better model fit, but can result in wider credibility intervals (Pybus 2006).

Since the information on the rate of evolution extracted by relaxed clock models is weak, dating a phylogeny relies heavily on the calibrations that are used to anchor the nodes in time (Pybus 2006; dos Reis, Donoghue, and Yang 2015). Unfortunately, fossils are rare and unevenly distributed in the tree of life. Microbes, in particular, leave few fossils that can be

unambiguously assigned to known species or clades. Therefore, entire clades cannot be reliably dated because they lack such information. For example, a recent dating analysis of the tree of life (Betts et al. 2018) used 10 fossil calibrations, 7 of which could be assigned to eukaryotes, 3 to bacteria, and none to archaea. Clearly, incorporating new sources of information into dating analyses would be very useful, especially for microbial clades.

Recently it has been shown that gene transfers could help date species trees, because they contain information on the chronological order of speciation nodes (Szöllosi et al. 2012; Davín et al. 2018). Transfers provide *node order constraints*, *i.e.* they specify that a given node in the phylogeny is necessarily older than another node, even though the older node is not an ancestor of the descendant node (Fig. 1a). (Davín et al. 2018) showed that the dating information provided by these constraints was consistent with information provided with (calibrated) relaxed molecular clocks, which suggests that node calibrations could be combined with node order constraints to date species trees more accurately. The benefit of including transfer-based constraints may be particularly noticeable in microbial clades, where transfers can be frequent (Doolittle 1999; Abby et al. 2012; Szöllosi et al. 2012; Davín et al. 2018). Transfer-based constraints may thus compensate for the lack of fossil calibrations in microbial clades. However, constraints may also be derived from other events, such as the transfer of a parasite or symbiont between hosts, endosymbioses or other obligatory relationships (Fig. 1b).

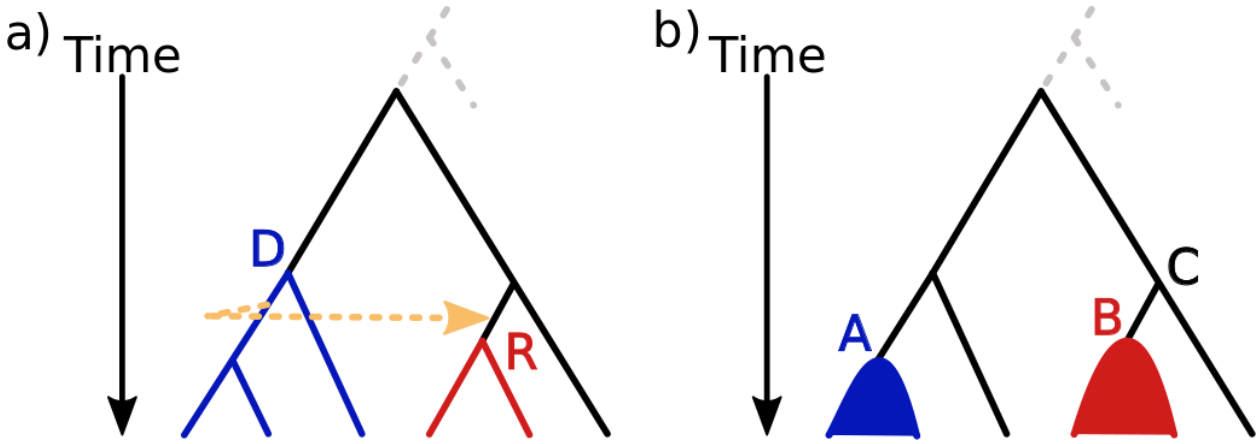


Fig. 1: Gene transfers and dependences between species provide relative dating information. a) A gene transfer from a descendent of D to an ancestor of R implies that D is older than R. b) Species from clade A depend upon the existence of species from clade B for their life. We cannot know for sure that we have sampled the oldest node in clade B, so all we can deduce is that the ancestor of clade A is younger than node C.

Here, we present a method to combine node order constraints with the dating machinery made of node age calibrations and (relaxed) molecular clocks, examine its performance in simulations, and evaluate its benefits on 3 empirical data sets.

Materials and Methods

Bayesian MCMC dating with calibrations and constraints

Informal description

Relaxed clock dating methods are often implemented in a Bayesian MCMC framework. Briefly, prior distributions are specified for (1) a diversification process (e.g., a birth-death prior), (2) the parameters of a model of sequence evolution (e.g., the HKY model, (Hasegawa, Kishino, and Yano 1985)), (3) calibration ages, and (4) the parameters of a model of rate heterogeneity along the tree. Such models may consider that neighboring branches have correlated rates of evolution (e.g., the autocorrelated lognormal model, (Thorne, Kishino, and Painter 1998)), or that each branch is associated to a rate drawn from a shared distribution (e.g. the uncorrelated gamma model (Thorne, Kishino, and Painter 1998; Drummond et al. 2006)). Calibrations are associated with prior distributions that account for the uncertainty associated with their age (dos Reis, Donoghue, and Yang 2015), and sometimes for the uncertainty associated with their position in the species tree (Heath, Huelsenbeck, and Stadler 2014). Our method introduces relative node age constraints as a new type of information that can be incorporated into this framework.

We chose to treat relative node order constraints as data without uncertainty, in the same way that topological constraints have been implemented in e.g. mrBayes (Ronquist and Huelsenbeck 2003). This way of treating node order constraints departs from how calibrations of node ages are treated in that calibrations are associated with distributions that convey the uncertainty associated with their age or position. This decision provides us with a simple way to incorporate constraints in the model: during the MCMC, any tree that does not satisfy a constraint is given a prior probability of 0, and is thus rejected during the Metropolis-Hastings step. Therefore, only trees that satisfy all relative node age constraints have a non-zero posterior probability.

Formal description

Let A be a sequence alignment, *Ca* be a set of fossil calibrations, *Co* be a set of node order constraints, *T* be a species tree, and θ be a set of all other parameters (e.g., sequence evolution, diversification and substitution rates, calibration times, etc.). Hence, the posterior probability of our model is

$$P(T, \theta | A,\ Ca,\ Co) = \frac{P(A, Ca | T, \theta) \times P(Co|T) \times P(T) \times P(\theta)}{P(A, Ca, Co)}.$$

The tree prior P(T) is typically based on a birth-death process (Rannala and Yang 1996) or the coalescent model (Kingman 1982). Node order constraints are accounted for by

$$P(T|Co) = \delta(T,\ Co),$$

where $\delta(T,\ Co)$ is the indicator function that is unity if T satisfies the node order constraints Co, and zero otherwise.

Implementation

We implemented this model in RevBayes so that it can be combined with other (relaxed) molecular clock models and models of sequence evolution and species diversification which are available in the software. Using the model in a Rev script implies calling two additional functions: one to read the constraints from a file, and another one to specify a tree prior that accounts for these constraints. Scripts are available at https://github.com/Boussau/DatingWithConsAndCal. We also provide a tutorial to guide RevBayes users: https://boussau.github.io/tutorials/relative_time_constraints/ .

Two-step inference of timetrees

Dating a phylogeny involves factoring branch lengths, specified in expected numbers of substitutions per site, into rate and time parameters. During a full Bayesian MCMC analysis of a sequence alignment, rate parameters and node ages are sampled together during the iterations, in one step. This inference problem is difficult, even when the topology of the tree is fixed, and MCMC chains typically have to be run for many iterations to obtain a good approximation of the posterior distribution. To reduce the computational cost, we decided to use a two-step approach.

In the first step, the posterior distribution of branch lengths is obtained from an MCMC analysis of unrooted trees. In the second step, the posterior distribution of branch lengths from the first MCMC chain is used as input to a second MCMC chain to infer distributions of rate parameters and of divergence times. Besides its speed, another advantage of this two-step approach is that complex state-of-the-art substitution models such as the CAT model, which is currently available only in PhyloBayes (Lartillot et al. 2013), can be used in the first step.

Following the above scheme, during the first step we sampled the posterior distribution of branch lengths using standard MCMC methods under a fixed topology, and calculated the posterior means and variances for each branch length. In the second step, we then approximated the phylogenetic likelihood using a composite likelihood composed of the product of per-branch Gaussian distributions with the estimated posterior means and variances of the branch lengths.

Simulations

General framework

We generated an artificial tree. We gathered calibration points by recording true node ages in this artificial tree. We also gathered relative constraints by recording true relative orders between the nodes. Then we simulated a DNA sequence alignment on the tree. Based on this sequence alignment, we used the two-step approach described above to infer timetrees. We then compared the reconstructed node ages to the true node ages from the artificial tree.

Simulating an artificial timetree

To obtain a tree with realistic speciation times, we decided to simulate a tree that has the same speciation times as in the timetree of life from (Betts et al. 2018). To do so, we gathered the speciation times from that timetree and produced an artificial tree by first randomly joining tips to produce coalescent events, and second assigning the speciation times from the empirical timetree to these coalescent events. We call the resulting tree a “shuffled tree” (Fig. 2). This shuffled tree has total depth from root to tips 45.12 units of time, as the timetree of life from (Betts et al. 2018).

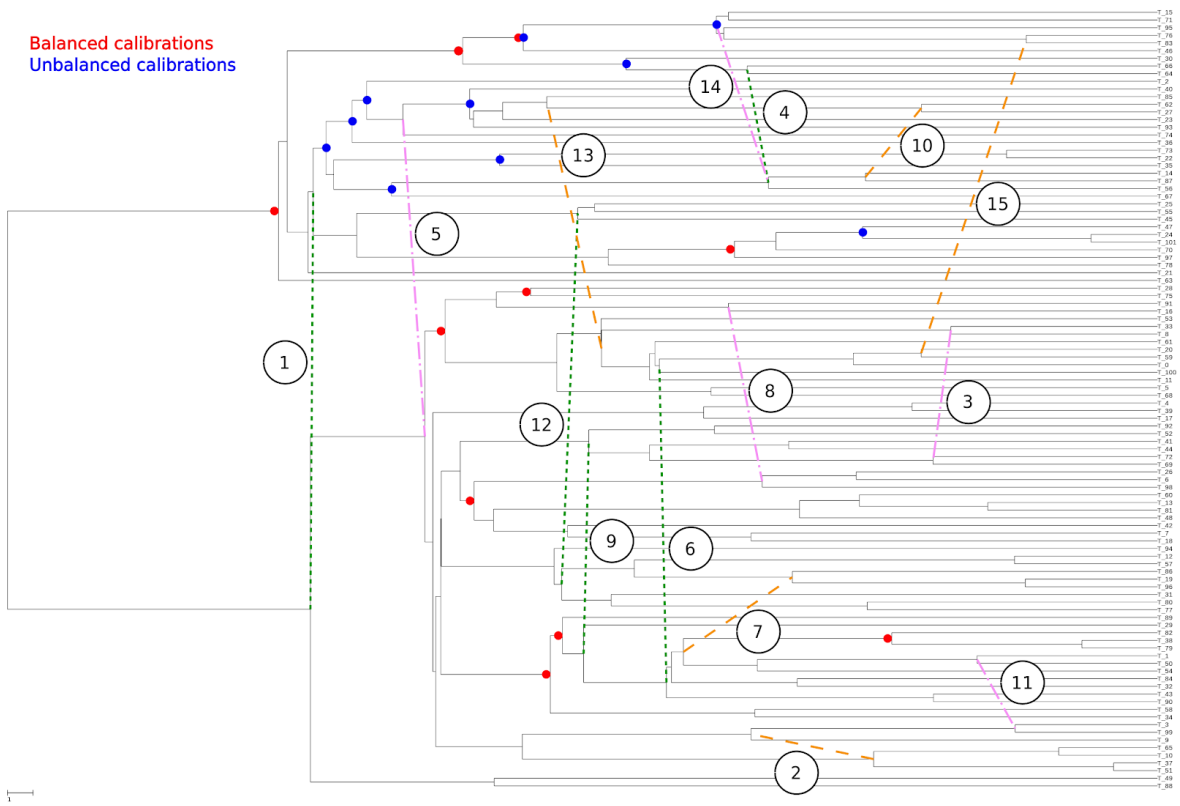


Figure 2: Shuffled tree, calibrated nodes and node order constraints. Calibrated nodes are shown with red dots when they are part of the set of 10 balanced calibrations, and with blue dots when they are part of the set of 10 unbalanced calibrations. Handpicked constraints have been numbered from 1 to 15, according to the order in which they were used (e.g. constraint 1 was used when only one constraint was included, constraints 1 to 5 when 5 constraints were included, and so on). Constraints have been colored according to their characteristics: green constraints are the 5 constraints between nodes with most similar ages (proximal), orange constraints are the 5 constraints between nodes with least similar ages (distal), and purple constraints are in between.

Building calibration times and node order constraints

We chose to use 10 internal node calibrations plus one calibration at the root node, as in (Betts et al. 2018). We used two configurations: one *balanced* configuration where calibrations

are placed on both sides of the root, and one *unbalanced* configuration where calibrations are found only on one side of the root (Fig. 1).

We used different sets of handpicked constraints, containing between 1 and 15 constraints by gathering true relative node orders from the shuffled tree. In choosing our sets of constraints we avoided redundant constraints, *i.e.* constraints that were already implied by previously included constraints. We expected that the informativeness of a constraint may depend on the age difference between the two nodes involved in the constraint. We investigated two types of constraints: proximal and distal. Proximal constraints specify the relative order of two nodes that are close in time, while distal constraints specify the order of two nodes with very different ages. We investigated the informativeness of proximal vs distal constraints by picking 5 constraints in each case (green and orange constraints in Fig. 1, respectively).

We built calibration times from the artificial trees by gathering the true speciation time, and associating it a prior distribution to convey uncertainty. The prior distribution we chose is uniform between [true age - (true age/5) ; true age + (true age/5)] and decays according to the tails of a normal distribution with standard deviation 2.5 beyond these boundaries (with 2.5% of the prior weight in each tail). 10 calibration points were chosen both in the balanced and unbalanced cases (Fig. 1). In addition, the tree root was calibrated with a uniform distribution between [root age - (root age/5) ; root age + (root age/5)].

Simulations of deviations from the clock

The shuffled tree was rescaled to yield branch lengths that can be interpreted as numbers of expected substitutions (its length from root to tip was 0.451). Then it was traversed from root to tips, and rate changes were randomly applied to the branches. 2 types of rate changes were allowed: small rate changes, that occur frequently, and large rate changes, that occur rarely. The magnitudes of small and large rate changes were drawn from lognormal distributions with parameters (mean=0.0, variance=0.1) and (mean=0.0, variance=0.2), respectively, and their rates were 33 and 1, respectively. After this process, branches smaller than 0.01 were set to 0.01. A Python code using the ete3 library (Betts et al. 2018; Huerta-Cepas, Serra, and Bork 2016) is available at https://github.com/Boussau/DatingWithConsAndCal, along with the command lines used and plots of the resulting tree with altered branch lengths.

Alignment simulation

The tree rescaled with deviations from the clock was used to simulate alignments 1000 bases long according to a HKY model (Hasegawa, Kishino, and Yano 1985), with ACGT frequencies {0.18, 0.27, 0.33, 0.22} and with a transition/transversion ratio of 3. A Gamma distribution discretized into 20 categories with an alpha parameter equal to 0.3 was used to simulate rate heterogeneity across sites. The Rev script is available at https://github.com/Boussau/DatingWithConsAndCal.

Inference based on simulated data

Inference of timetrees based on the simulated alignments was performed in two steps as explained above. Both steps were performed in RevBayes (Höhna et al. 2016), with scripts available at https://github.com/Boussau/DatingWithConsAndCal.

We inferred branch length distributions under a Jukes-Cantor model (Jukes and Cantor 1969) without rate heterogeneity across sites to make our test more realistic in that the reconstruction

model is simpler than the process generating the data. The tree topology was fixed to the true unrooted topology.

The obtained posterior distributions of branch lengths were then summarized by their mean and variance per branch. These means and variances were given as input to a script that provides timetrees according to a birth-death prior on the tree topology and node ages, an uncorrelated Gamma prior on the rate of sequence evolution through time (Drummond et al. 2006), and using the calibrations and constraints gathered in previous steps (see above).

Results

Simulations

Two-step inference provides an efficient and flexible method to estimate time trees

We compared posterior distributions of node ages obtained using the classical full Bayesian MCMC approach to those obtained using our two-step approximation. As shown in Supplementary Figs. S1-4, the two posterior distributions of node ages are practically indistinguishable. Further, the impact of the approximation is negligible in comparison to the choice of the model of rate evolution.

Constraints improve dating accuracy

We used two statistics to evaluate the accuracy of node age estimates. First, we computed the normalized root mean square deviation (RMSD) between the true node ages used in the simulation and the node ages estimated in the Maximum A Posteriori tree (Fig. 2a), and normalized it by the true node ages. This provides measures of the error as a percentage of the true node ages. Second, we computed the coverage probability, *i.e.* how frequently the 95% High Posterior Density (HPD) intervals on node ages contained the true node ages (Fig. 2b).

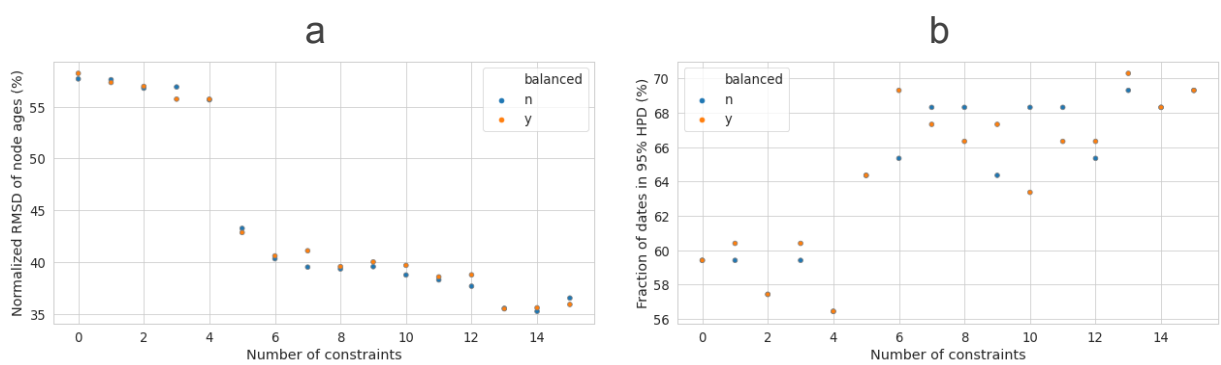


Figure 3: Increasing the number of constraints improves node age estimation. a) Average normalized RMSD over all internal node ages is shown in orange for 10 balanced calibrations and blue for 10 unbalanced calibrations. This is a measure of the error as a percentage of the true node ages. b) The percentage of nodes with true age in 95% High Posterior Density (HPD) interval is shown (colors as in a).

As the number of constraints increases, Fig. 3a shows that the error in node ages decreases, and Fig. 3b shows that the 95% HPD intervals include the true node ages more often. When 0 or only 1 constraint is used, the true node age is contained in only ~55% of the 95% HPD intervals, suggesting that the mismatch between the model used for simulation and the model used for inference has a noticeable impact. Poor mixing could also explain these results, but it is unlikely to occur in our experiment for two reasons. First, the Expected Sample Sizes for the node ages are typically above 300. Second, if the same moves are used in the MCMC, but the simulation model is changed to fit the inference model, about 95% of the true node ages end up in 95% HPD intervals, as expected for well-calibrated Bayesian methods and well-mixing MCMC chains (see Supp. Fig. S5 and associated section).

Results improve markedly with 5 or more constraints, with a strong effect when moving from 4 to 5 constraints, and then a slower improvement. There is no obvious feature of constraint 5 that would make it substantially more helpful than other constraints for dating.

The results obtained with the balanced set of calibrations are similar to the results obtained with the unbalanced set of calibrations, in particular in terms of RMSD in node ages.

Constraints reduce credibility intervals

The additional information provided by constraints results in smaller credibility intervals, as shown in Fig. 4. The improvement in coverage probability observed in Fig. 3b therefore occurs despite smaller credibility intervals.

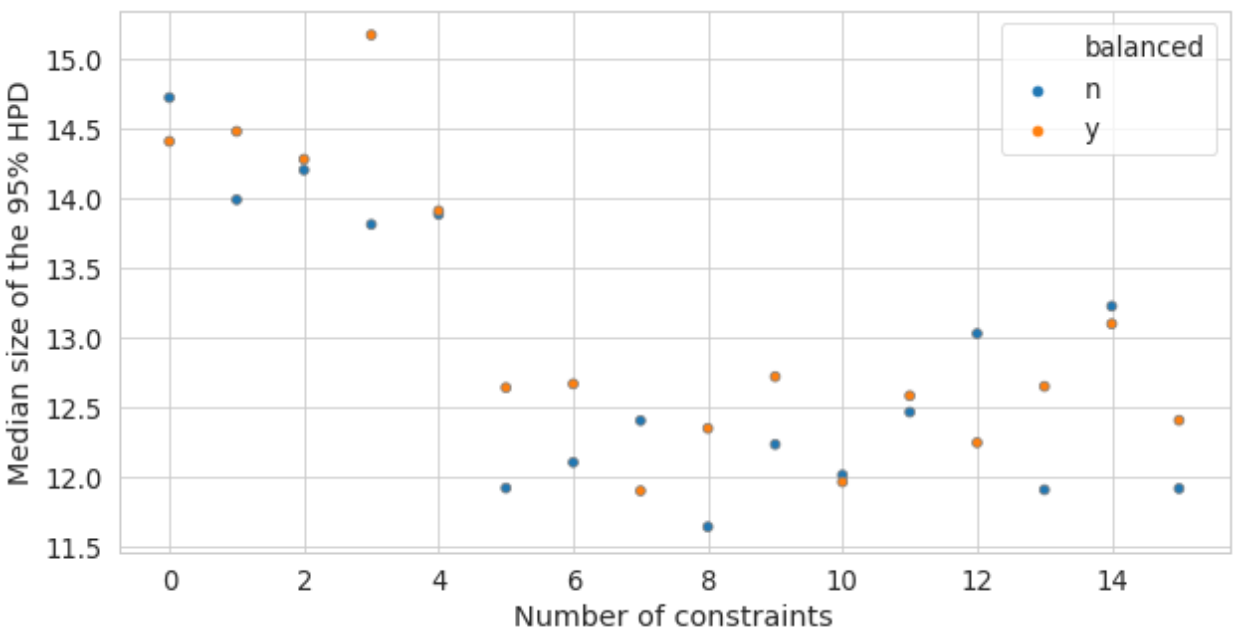
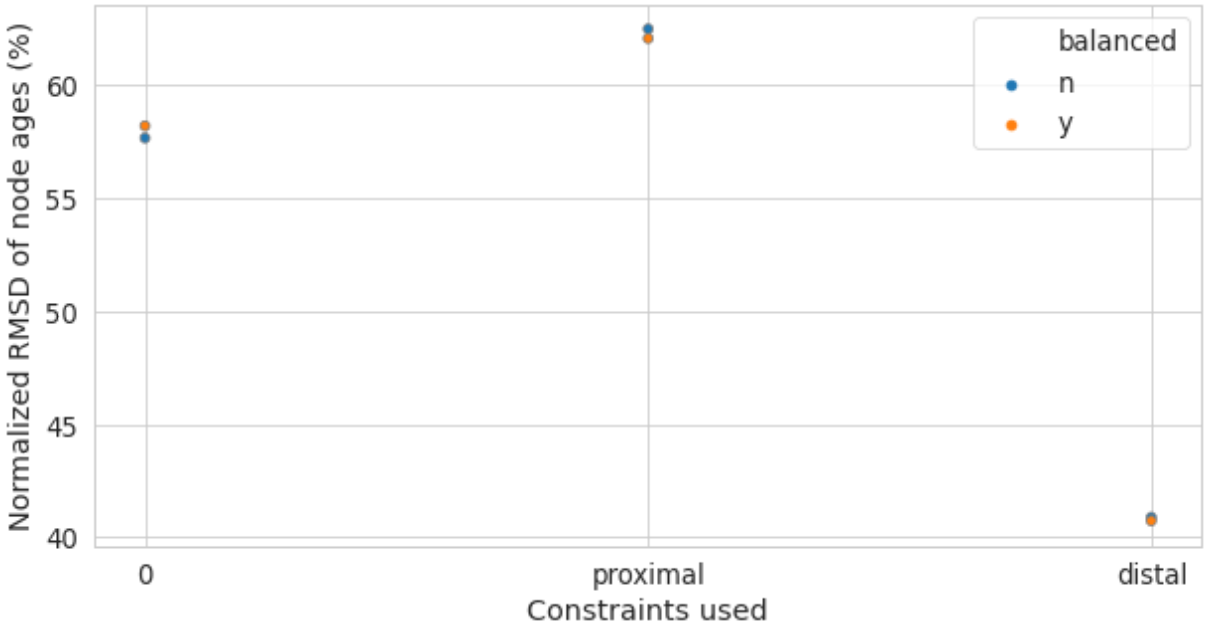


Figure 4: The 95% HPD intervals on node ages become smaller as the number of constraints increases. The sizes are given in units of time; for reference, the total depth for the true tree is 45.12 units of time. Colors as in Fig. 2.

Investigating the informativeness of constraints

To investigate the effect of the age difference between nodes, we compared two sets of constraints. The first set contains constraints between nodes of similar ages (we call them proximal), the second set contains constraints between nodes of more different ages (we call

them distal). Fig. 5 shows that using distal constraints provides stronger benefit over not using constraints than using proximal constraints. In fact, in this experiment, it appears that using proximal constraints has brought no benefit. It would seem that constraints between nodes whose ages are similar are in fact not useful: it makes little difference to the ages of the nodes in the entire tree to know that A is only slightly older than B. Constraints between nodes whose ages are very different make a difference: they forbid some dating configurations, which may have consequences over the whole tree.



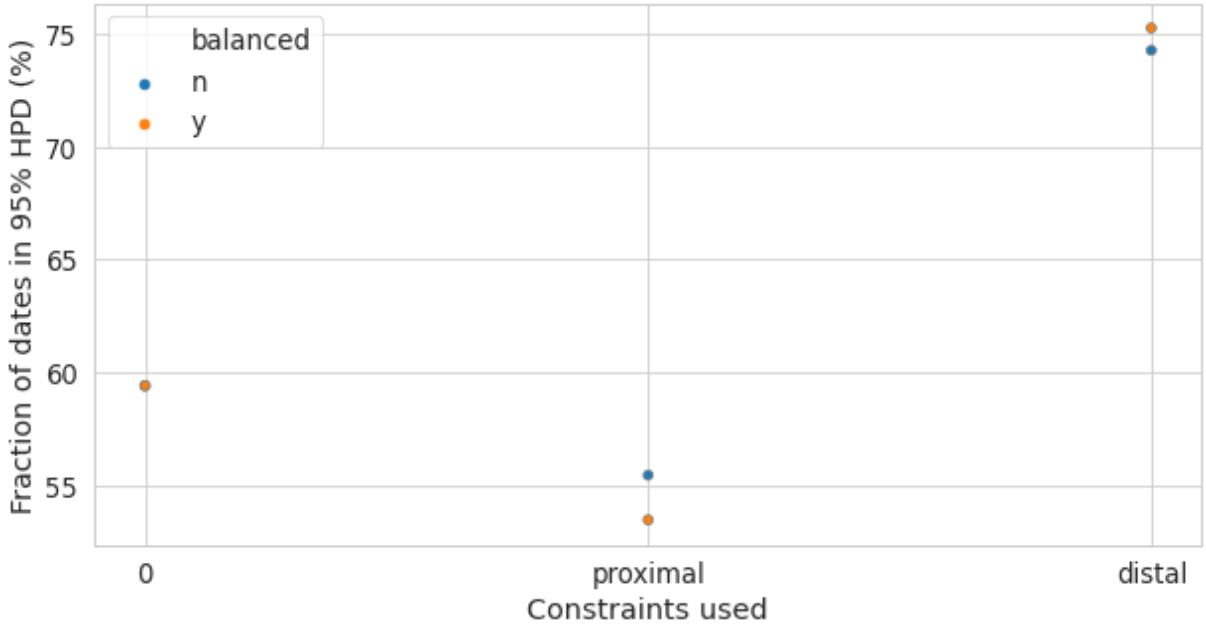


Figure 5: Distal constraints help more than proximal constraints. Colors as in Fig. 2.

Analyses of empirical data

(Davín et al. 2018) showed that gene transfers contain dating information that is consistent with relaxed molecular clock models. We used a phylogeny of Cyanobacterial genomes presented in Davin et al., and a phylogeny of Archaeal genomes from (Williams et al. 2017) to investigate the individual and cumulative impacts of fossil calibrations and relative constraints on the inference of time trees.

Relative constraints agree with fossil calibration on the age of akinete-forming multicellular Cyanobacteria

(Davín et al. 2018) analyzed a set of 40 cyanobacteria spanning most of their species diversity. Cyanobacteria likely originated more than 2 billion years ago, but a review of the literature suggests that there is only a single reliable fossil calibration that we can place on the species tree: a minimum bound for akinete-forming multicellular Cyanobacteria from (Tomitani et al. 2006). These authors reported a series of fossils that they assign to filamentous Cyanobacteria producing both specialized cells for nitrogen fixation (heterocysts) and resting cells able to endure environmental stress (akinetes).

We investigated whether relative node order constraints could recover the effect of the available fossil calibration by comparing several dating protocols: fossil calibration with no relative age constraints (Fig. 6a), no fossil calibration and no relative age constraints (Fig. 6b), relative age constraints with no fossil calibration, (Fig. 6c), and both calibrations and constraints (Fig. 6d). Fossil calibration corresponded to a minimum age for fossil akinetes at 1.956 GYa (dashed red line Arrow on Fig. 6a and d). Reflecting our uncertainty regarding the age of the root, we tried two alternatives for the maximum root age (*i.e.* age of crown cyanobacteria), 2.45 Gy and 2.7 Gy, corresponding to the “Great Oxygenation Event” and the “whiff of Oxygen” (Holland 2006) respectively. In Fig 6. we show results obtained with the 2.45 Gy root calibration, while Fig. 7a presents the age of key nodes for both choices of root maximum age.

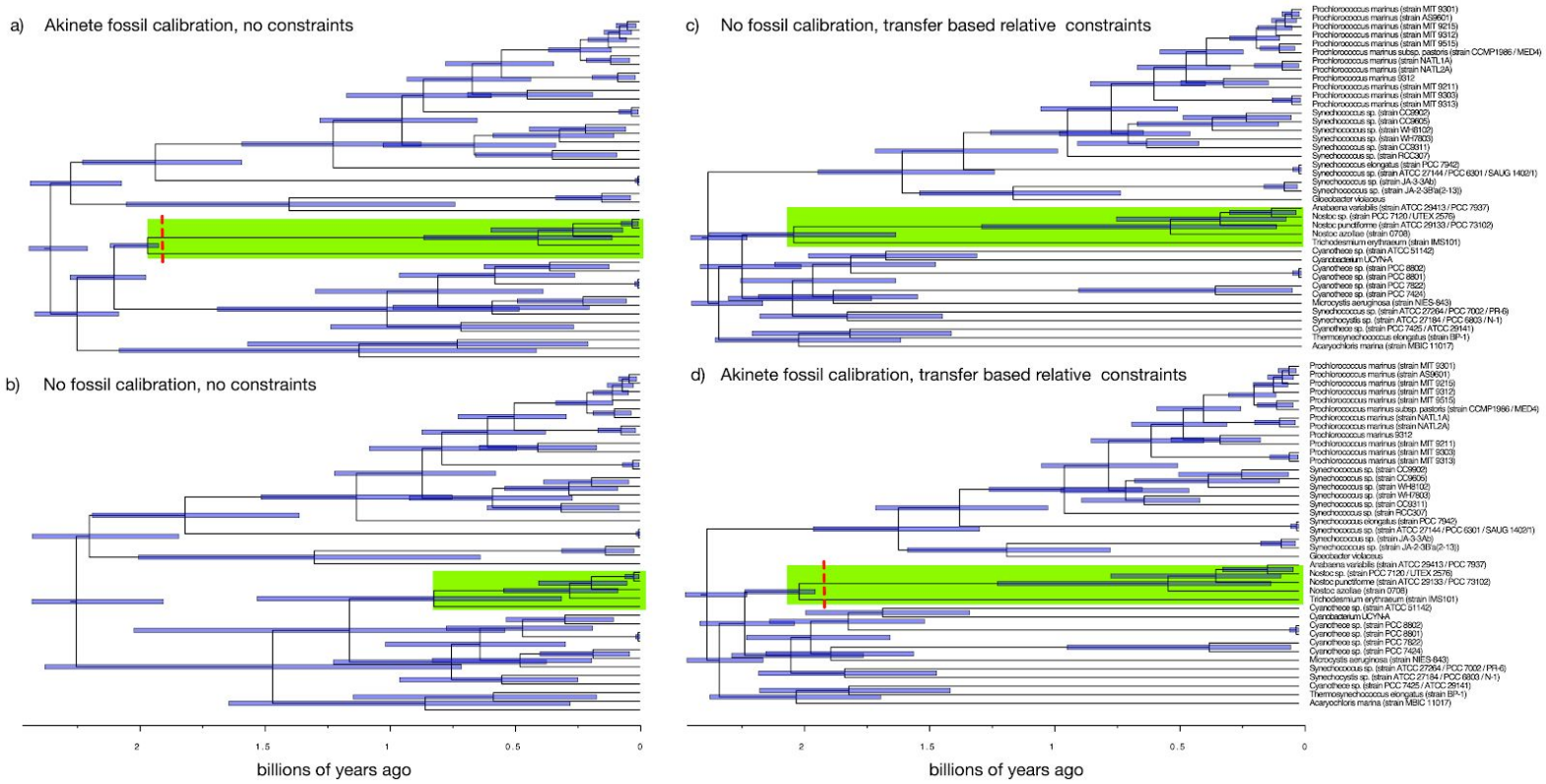


Figure 6. **Relative age constraints agree with the akinete fossil calibration that akinete-forming multicellular Cyanobacteria are likely older than suggested by sequence data alone**. We compared four dating protocols for the 40 cyanobacteria from Davin et al. [cite]: a) fossil calibration (dashed red line) with no relative age constraints, b) no fossil calibration and no relative age constraints, c) relative age constraints, with no fossil calibration and d) simultaneous fossil calibration and constraints (Fig. 5d). All four chronograms were inferred with a root maximum age of 2.45 Gya with an uncorrelated gamma rate prior, and a birth-death prior on divergence times. Clade highlighted in green corresponds to akinete-forming multicellular cyanobacteria.

Comparison between Figs. 6a and 6b shows that including the minimum calibration increases the age of the clade containing akinete-forming multicellular Cyanobacteria (green clade) by about 1 Gy. Interestingly, the inclusion of constraints compensates for the absence of a minimum calibration (Fig. 6c) and places the age of clade of akinete forming multicellular Cyanobacteria close to its age when a fossil-based minimum age calibration is used (Fig.6a) calibrations are used. The information provided by constraints thus agrees with the fossil age for multicellular Cyanobacteria. As a result, the combination of both calibrations and constraints produces a chronogram with smaller credibility intervals (Fig. 6d).

To further characterize the effect of constraints on the age of akinete forming multicellular Cyanobacteria, we plotted the distributions of its age based on different sources of dating information. In Figure 7a we show the age of akinete forming multicellular Cyanobacteria (green clade in Fig 6) estimated based on i) only the rate and divergence time priors, ii) priors and sequence divergence only, iii) priors and relative age constraints only and iv) priors and both sequence divergence and relative age constraints. Comparison of the age distributions shows

that relative age constraints convey information that complements sequence divergence and is coherent with the fossil record on the age of akinete-forming Cyanobacteria.

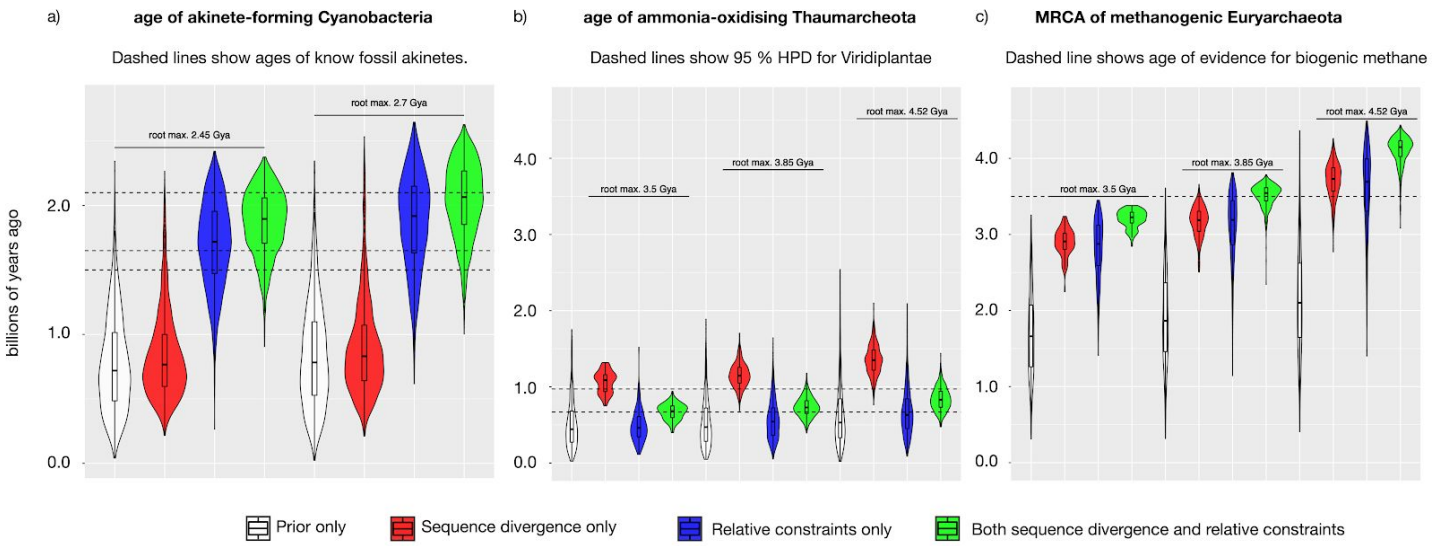


Figure 7: Distributions of key node ages according to different sources of dating information. We show the age of a) akinete-forming Cyanobacteria, b) Thaumarchaeota and c) the most recent common ancestor of methanogenic Archaea. Distributions in white are based on solely the maximum root age and the rate and divergence time priors, distributions in red are informed by sequence divergence, distributions in blue include relative age constraints, but not sequence divergence, while distributions in green rely on both. Dashed lines indicate, respectively, a) age of fossils of putative akinete forming multicellular cyanobacteria, b) age of Viridiplantae and c) age of evidence for biogenic methane.

Relative constraints refine the time tree of Archaea

We next investigated divergence times of the Archaea, one of the primary domains of life (Woese, Kandler, and Wheelis 1990). We used the data from (Williams et al. 2017) containing 62 species. Most analyses place the root of the entire tree of life between Archaea and Bacteria (Woese, Kandler, and Wheelis 1990; Iwabe et al. 1989; Gogarten et al. 1989; Gouy, Baurain, and Philippe 2015), suggesting that the Archaea are likely an ancient group. However, there are no unambiguous fossil Archaea and so the history of the group in geological time is poorly constrained. Methanogenesis is a hallmark metabolism of some members of the Euryarchaeota, and so the discovery of biogenic methane in 3.46Gya rocks (Ueno et al. 2006) might indicate that Euryarchaeota already existed at that time. However, the genes required for methanogenesis have also been identified in genomes of other archaeal groups including Korarchaeota (McKay et al. 2019) and Verstraetearchaeota (Vanwonterghem et al. 2016), and it is difficult to exclude the possibility that methanogenesis maps to the root of the Archaea (Berghuis et al. 2019). Thus, ancient methane might have been produced by Euryarcheota, another extant archaeal group, a stem archaeon or even by Cyanobacteria (Bižić et al. 2020).

In the absence of strong geochemical constraints, can relative constraints help to refine the time tree of Archaea? We investigated two nodes on the archaeal tree from (Williams et al. 2017): the common ancestor of ammonia-oxidising (AOA) Thaumarchaeota and the common ancestor of methanogenic Euryarchaeota (that is, the common ancestor of all Euryarchaeota except for the Thermococcus/Pyrococcus clade). While we lack absolute constraints for these lineages, dating hypotheses have been proposed on the basis of individually identified and curated gene transfers to, or from, other lineages for which fossil information does exist. These include the transfer of a DnaJ-Fer fusion gene from Viridiplantae (land plants and green algae) into the common ancestor of AOA Thaumarchaeota (Petitjean et al. 2012), and a transfer of three SMC complex genes from within one clade of Euryarchaeota (Methanotecta, including the class 2 methanogens) to the root of Cyanobacteria (Wolfe and Fournier 2018). Note that, in the following analyses, we use relative constraints derived from inferred within-Archaea gene transfers; therefore, these constraints are independent of the transfers used to propose the hypotheses we test.

As the age of the root is uncertain, we explored the impact on our inferences of three different choices: a relatively young estimate of 3.5Gya from the analysis of (Wolfe and Fournier 2018; Betts et al. 2018); the end of the late heavy bombardment at 3.85Gya (Boussau and Gouy 2012); and the age of the solar system at 4.52Gya (Barboni et al. 2017).

We found that, despite the uncertainty in the age of the root, the estimated age of AOA Thaumarchaeota informed by relative age constraints is consistent with the hypothesis that AOA are younger than stem Viridiplantae (Petitjean et al. 2012), with a recent estimate for the age of Viridiplantae between 972.4-669.9 Mya (Petitjean et al. 2012; Morris et al. 2018); (Figure 7b). As in the case of Cyanobacteria, information from relative constraints had a substantial impact on the analysis; sequence data alone (in combination with the root age prior) suggest a somewhat older age of AOA Thaumarchaeota, consistent with recent molecular clock analyses (Ren et al. 2019).

In the case of methanogenic Euryarchaeota, inference both with and without relative constraints was strongly influenced by the choice of root prior (Figure 7c), and so the results do not clearly distinguish between hypotheses about the age of archaeal methanogenesis or the potential source of ancient biogenic methane. With those caveats in mind, the information from relative constraints supported moderately older age distributions than inference from sequence data alone across all root priors. The results are consistent with an early origin of methanogenic Euryarchaeota within the archaeal domain (Wolfe and Fournier 2018) and, for the moderate (3.85Gya) and older (4.52Gya) priors, indicate that these archaea are a potential source of biogenic methane at 3.46Gya (Ueno et al. 2006).

Discussion

Constraints are a new and reliable source of information for dating phylogenies

(Ueno et al. 2006; Davín et al. 2018) showed that gene transfers contained reliable information about node ages. They also used this information in an *ad hoc* two-step process to

provide approximate age estimates for a few nodes in 3 clades. Here we built upon these results to develop a fully Bayesian method that accounts for both relative node order constraints and absolute time calibrations within the MCMC algorithm by extending the standard relaxed clock approach. We also introduced a fast and accurate two-step method for incorporating branch length distributions inferred under complex substitution models into relaxed molecular clock analyses.

To test our method, we performed sequence simulations and analyzed three empirical data sets. We simulated sequences according to a model that differs from the inference model so as to emulate the typical situation with empirical data, where the process that generated the data differs from our inference models. As expected under these conditions, node age coverage probabilities, *i.e.* the percentage of true node ages that fall within inferred 95% credibility intervals, are much lower than 95%. We used a realistic phylogeny for simulating sequences by drawing node ages from a previously published dated tree of life (Betts et al. 2018) but by rearranging the tree topology. We then investigated the effect of sampling node age and node relative order constraints on dating accuracy. A single tree topology and a single simulated alignment were used overall, which might adversely affect the generality of our results. However, this tree topology is large (102 tips) and realistic, and the results on empirical data suggest that our method is useful across the tree of life. Further, using a single alignment allowed us to estimate branch length distributions only once and then use our fast two-step inference to reduce our computational footprint.

The simulations show that relative node order constraints improve the accuracy of node ages and coverage probabilities. We further found that constraints between nodes of similar ages were less useful than constraints between nodes of differing ages. This is encouraging since it should be easier to find transfers between nodes whose ages differ widely (distal transfers) than between nodes with similar ages, because large age differences give more time for transfers to occur and to leave a detectable footprint in extant genomes.

Results obtained on empirical data sets show that relative node order constraints extracted from dozens of gene transfers contain information that can compensate for the lack of fossil calibrations. This shows promise for dating phylogenies for which fossils are scant, i.e. the great majority of the tree of life.

One limitation of the method presented here is that relative constraints are treated as though they are known with certainty. Only trees that satisfy all of the input constraints will have non-zero probability, and so incorrect input constraints will result in incorrect age estimates. We therefore suggest that only the most reliable constraints should be used when dating a species tree using transfers. One practical approach, which we have used in our empirical analyses of genomic data, is to use only those constraints that are highly supported (Davín et al. 2018). A clear direction for future work will be to treat relative constraints probabilistically, perhaps as a function of the number and quality of inferred gene transfers that support them, or with a probability p that constraints are matched, which would be estimated in the course of the MCMC.

Dating phylogenies is a challenging statistical problem where data is limiting, since only fossils and rates of molecular evolution provide information. Here we have developed a new method to exploit information contained in gene transfers, which are particularly numerous in clades where fossil information is lacking. Gene transfers define relative node order constraints. We have shown on simulation that using node order constraints improves node age estimates and reduces credibility intervals. We have also used our method on three empirical data sets to show that node order constraints can compensate for the absence of a fossil calibration: ages obtained without a fossil calibration but with constraints match those obtained with the fossil

calibration, and incorporating both sources of time information further refines the inferred divergence times. Looking forward we envision that our method will be useful to date parts of the tree of life where node ages have so far remained very uncertain.

Supplementary material

Supplementary Material is available at BioRxiv:
https://www.biorxiv.org/content/10.1101/2020.10.17.343889v1.supplementary-material

Data availability

Scripts and data used to run the simulation analyses are available at
https://github.com/Boussau/DatingWithConsAndCal
Data for the empirical data analysis has been deposited at:
https://doi.org/10.5061/dryad.s4mw6m958

And a tutorial is available at: https://boussau.github.io/tutorials/relative_time_constraints/

Author contributions

GJS, VD and BB initiated the project. BB, GJS and SH implemented the model in RevBayes. GJS ran the empirical analyses, and analyzed them with TAW. BB ran the simulations. DS, GJS and BB wrote the tutorial. BB, GJS, TAW and VD wrote the manuscript. All authors read and commented on the manuscript.

Acknowledgements

DS and GJSz received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program under Grant Agreement 714774, GJS received funding Grant GINOP-2.3.2.-15-2016- 00057. TAW is supported by a Royal Society University Research Fellowship and NERC grant NE/P00251X/1. BB and TAW acknowledge

support from a “Projet de Recherche Collaborative” co-funded by the CNRS and the Royal Society. We thank Eric Tannier for fruitful discussions.

REFERENCES

Abby, Sophie S., Eric Tannier, Manolo Gouy, and Vincent Daubin. 2012. “Lateral Gene Transfer as a Support for the Tree of Life.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (13): 4962–67.

Barboni, Melanie, Patrick Boehnke, Brenhin Keller, Issaku E. Kohl, Blair Schoene, Edward D. Young, and Kevin D. McKeegan. 2017. “Early Formation of the Moon 4.51 Billion Years Ago.” *Science Advances* 3 (1): e1602365.

Berghuis, Bojk A., Feiqiao Brian Yu, Frederik Schulz, Paul C. Blainey, Tanja Woyke, and Stephen R. Quake. 2019. “Hydrogenotrophic Methanogenesis in Archaeal Phylum Verstraetearchaeota Reveals the Shared Ancestry of All Methanogens.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (11): 5037–44.

Betts, Holly C., Mark N. Puttick, James W. Clark, Tom A. Williams, Philip C. J. Donoghue, and Davide Pisani. 2018. “Integrated Genomic and Fossil Evidence Illuminates Life’s Early Evolution and Eukaryote Origin.” *Nature Ecology & Evolution* 2 (10): 1556–62.

Bižić, M., T. Klintzsch, D. Ionescu, M. Y. Hindiyeh, M. Günthel, A. M. Muro-Pastor, W. Eckert, T. Urich, F. Keppler, and H-P Grossart. 2020. “Aquatic and Terrestrial Cyanobacteria Produce Methane.” *Science Advances* 6 (3): eaax5343.

Boussau, Bastien, and Manolo Gouy. 2012. “What Genomes Have to Say about the Evolution of the Earth.” *Gondwana Research*. https://doi.org/10.1016/j.gr.2011.08.002.

Davín, Adrián A., Eric Tannier, Tom A. Williams, Bastien Boussau, Vincent Daubin, and Gergely J. Szöllősi. 2018. “Gene Transfers Can Date the Tree of Life.” *Nature Ecology & Evolution* 2 (5): 904–9.

Doolittle, W. F. 1999. “Phylogenetic Classification and the Universal Tree.” *Science*. https://doi.org/10.1126/science.284.5423.2124.

Drummond, Alexei J., Simon Y. W. Ho, Matthew J. Phillips, and Andrew Rambaut. 2006. “Relaxed Phylogenetics and Dating with Confidence.” *PLoS Biology* 4 (5): e88.

Gogarten, J. P., H. Kibak, P. Dittrich, L. Taiz, E. J. Bowman, B. J. Bowman, M. F. Manolson, et al. 1989. “Evolution of the Vacuolar H+-ATPase: Implications for the Origin of Eukaryotes.” *Proceedings of the National Academy of Sciences of the United States of America* 86 (17): 6661–65.

Gouy, Richard, Denis Baurain, and Hervé Philippe. 2015. “Rooting the Tree of Life: The Phylogenetic Jury Is Still out.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 370 (1678): 20140329.

Hasegawa, M., H. Kishino, and T. Yano. 1985. “Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA.” *Journal of Molecular Evolution* 22 (2): 160–74.

Heath, Tracy A., Mark T. Holder, and John P. Huelsenbeck. 2012. “A Dirichlet Process Prior for Estimating Lineage-Specific Substitution Rates.” *Molecular Biology and Evolution* 29 (3): 939–55.

Heath, Tracy A., John P. Huelsenbeck, and Tanja Stadler. 2014. “The Fossilized Birth-Death Process for Coherent Calibration of Divergence-Time Estimates.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (29): E2957–66.

Höhna, Sebastian, Michael J. Landis, Tracy A. Heath, Bastien Boussau, Nicolas Lartillot, Brian

R. Moore, John P. Huelsenbeck, and Fredrik Ronquist. 2016. “RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language.” *Systematic Biology* 65 (4): 726–36.

Holland, Heinrich D. 2006. “The Oxygenation of the Atmosphere and Oceans.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 361 (1470): 903–15.

Huerta-Cepas, Jaime, François Serra, and Peer Bork. 2016. “ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data.” *Molecular Biology and Evolution* 33 (6): 1635–38.

Iwabe, N., K. Kuma, M. Hasegawa, S. Osawa, and T. Miyata. 1989. “Evolutionary Relationship of Archaebacteria, Eubacteria, and Eukaryotes Inferred from Phylogenetic Trees of Duplicated Genes.” *Proceedings of the National Academy of Sciences of the United States of America* 86 (23): 9355–59.

Jukes, Thomas H., and Charles R. Cantor. 1969. “Evolution of Protein Molecules.” *Mammalian Protein Metabolism*. https://doi.org/10.1016/b978-1-4832-3211-9.50009-7.

Kingman, J. F. C. 1982. “The Coalescent.” *Stochastic Processes and Their Applications*. https://doi.org/10.1016/0304-4149(82)90011-4.

Lartillot, Nicolas, Matthew J. Phillips, and Fredrik Ronquist. 2016. “A Mixed Relaxed Clock Model.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 371 (1699). https://doi.org/10.1098/rstb.2015.0132.

Lartillot, Nicolas, Nicolas Rodrigue, Daniel Stubbs, and Jacques Richer. 2013. “PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment.” *Systematic Biology* 62 (4): 611–15.

Lepage, Thomas, David Bryant, Hervé Philippe, and Nicolas Lartillot. 2007. “A General Comparison of Relaxed Molecular Clock Models.” *Molecular Biology and Evolution* 24 (12): 2669–80.

McKay, Luke J., Mensur Dlakić, Matthew W. Fields, Tom O. Delmont, A. Murat Eren, Zackary J. Jay, Korinne B. Klingelsmith, Douglas B. Rusch, and William P. Inskeep. 2019. “Co-Occurring Genomic Capacity for Anaerobic Methane and Dissimilatory Sulfur Metabolisms Discovered in the Korarchaeota.” *Nature Microbiology* 4 (4): 614–22.

Morris, Jennifer L., Mark N. Puttick, James W. Clark, Dianne Edwards, Paul Kenrick, Silvia Pressel, Charles H. Wellman, Ziheng Yang, Harald Schneider, and Philip C. J. Donoghue. 2018. “The Timescale of Early Land Plant Evolution.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (10): E2274–83.

Petitjean, Céline, David Moreira, Purificación López-García, and Céline Brochier-Armanet. 2012. “Horizontal Gene Transfer of a Chloroplast DnaJ-Fer Protein to Thaumarchaeota and the Evolutionary History of the DnaK Chaperone System in Archaea.” *BMC Evolutionary Biology* 12 (November): 226.

Pybus, Oliver G. 2006. “Model Selection and the Molecular Clock.” *PLoS Biology* 4 (5): e151.

Rannala, B., and Z. Yang. 1996. “Probability Distribution of Molecular Evolutionary Trees: A New Method of Phylogenetic Inference.” *Journal of Molecular Evolution* 43 (3): 304–11.

Reis, Mario dos, Philip C. J. Donoghue, and Ziheng Yang. 2015. “Bayesian Molecular Clock Dating of Species Divergences in the Genomics Era.” *Nature Reviews. Genetics* 17 (2): 71–80.

Ren, Minglei, Xiaoyuan Feng, Yongjie Huang, Hui Wang, Zhong Hu, Scott Clingenpeel, Brandon K. Swan, et al. 2019. “Phylogenomics Suggests Oxygen Availability as a Driving Force in Thaumarchaeota Evolution.” *The ISME Journal* 13 (9): 2150–61.

Ronquist, Fredrik, and John P. Huelsenbeck. 2003. “MrBayes 3: Bayesian Phylogenetic Inference under Mixed Models.” *Bioinformatics* 19 (12): 1572–74.

Szöllosi, Gergely J., Bastien Boussau, Sophie S. Abby, Eric Tannier, and Vincent Daubin. 2012. “Phylogenetic Modeling of Lateral Gene Transfer Reconstructs the Pattern and Relative Timing of Speciations.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (43): 17513–18.

Thorne, J. L., H. Kishino, and I. S. Painter. 1998. “Estimating the Rate of Evolution of the Rate of Molecular Evolution.” *Molecular Biology and Evolution*. https://doi.org/10.1093/oxfordjournals.molbev.a025892.

Tomitani, Akiko, Andrew H. Knoll, Colleen M. Cavanaugh, and Terufumi Ohno. 2006. “The Evolutionary Diversification of Cyanobacteria: Molecular-Phylogenetic and Paleontological Perspectives.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (14): 5442–47.

Ueno, Yuichiro, Keita Yamada, Naohiro Yoshida, Shigenori Maruyama, and Yukio Isozaki. 2006. “Evidence from Fluid Inclusions for Microbial Methanogenesis in the Early Archaean Era.” *Nature* 440 (7083): 516–19.

Vanwonterghem, Inka, Paul N. Evans, Donovan H. Parks, Paul D. Jensen, Ben J. Woodcroft, Philip Hugenholtz, and Gene W. Tyson. 2016. “Methylotrophic Methanogenesis Discovered in the Archaeal Phylum Verstraetearchaeota.” *Nature Microbiology* 1 (12): 1–9.

Williams, Tom A., Gergely J. Szöllősi, Anja Spang, Peter G. Foster, Sarah E. Heaps, Bastien Boussau, Thijs J. G. Ettema, and T. Martin Embley. 2017. “Integrative Modeling of Gene and Genome Evolution Roots the Archaeal Tree of Life.” *Proceedings of the National Academy of Sciences of the United States of America* 114 (23): E4602–11.

Woese, C. R., O. Kandler, and M. L. Wheelis. 1990. “Towards a Natural System of Organisms: Proposal for the Domains Archaea, Bacteria, and Eucarya.” *Proceedings of the National Academy of Sciences of the United States of America* 87 (12): 4576–79.

Wolfe, Joanna M., and Gregory P. Fournier. 2018. “Horizontal Gene Transfer Constrains the Timing of Methanogen Evolution.” *Nature Ecology & Evolution* 2 (5): 897–903.

Zuckerkandl, Emile, and Linus Pauling. 1962. *Molecular Disease, Evolution, and Genic Heterogeneity*.