AleRax: A tool for gene and species tree co-estimation and reconciliation under a probabilistic model of gene duplication, transfer, and loss

Benoit Morel,1,2 Tom A. Williams,3 Alexandros Stamatakis4,1,2 and Gergely J. Szöllősi5,6,7*

1Computational Molecular Evolution group, Heidelberg Institute for Theoretical Studies, 69118, Heidelberg, Germany, 2Institute for Theoretical Informatics, Karlsruhe Institute of Technology, 76131, Karlsruhe, Germany, 3School of Biological Sciences, University of Bristol, BS8 1TQ, Bristol, UK, 4Biodiversity Computing Group, Institute of Computer Science, GR-70013 Heraklion, Greece, 5ELTE-MTA “Lendület”, Evolutionary Genomics Research Group, Pázmány P. stny. 1A, H-1117, Budapest, Hungary, 6Institute of Evolution, HUN-REN Centre for Ecological Research, Konkoly-Thege M. út 29-33, H-1121, Budapest, Hungary and 7Model-Based Evolutionary Genomics Unit, Okinawa Institute of Science and Technology Graduate University, 904-0495, Okinawa, Japan

* To whom correspondence should be addressed.

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

**Motivation:** Genomes are a rich source of information on the pattern and process of evolution across biological scales. How best to make use of that information is an active area of research in phylogenetics. Ideally, phylogenetic methods should not only model substitutions along gene trees, which explain differences between homologous gene sequences, but also the processes that generate the gene trees themselves along a shared species tree. To conduct accurate inferences, one needs to account for uncertainty at both levels, that is, in gene trees estimated from inherently short sequences and in their diverse evolutionary histories along a shared species tree.

**Results:** We present AleRax, a software that can infer reconciled gene trees together with a shared species tree using a simple, yet powerful, probabilistic model of gene duplication, transfer, and loss. A key feature of AleRax is its ability to account for uncertainty in the gene tree and its reconciliation by using an efficient approximation to calculate the joint phylogenetic - reconciliation likelihood and sample reconciled gene trees accordingly. Simulations and analyses of empirical data show that AleRax is one order of magnitude faster than competing gene tree inference tools while attaining the same accuracy. It is consistently more robust than species tree inference methods such as SpeciesRax and ASTRAL-Pro 2 under gene tree uncertainty. Finally, AleRax can process multiple gene families in parallel thereby allowing users to compare competing phylogenetic hypotheses and estimate model parameters, such as DTL probabilities for genome-scale datasets with hundreds of taxa.

**Availability and Implementation:** GNU GPL at https://github.com/BenoitMorel/AleRax and data are made available at https://cme.h-its.org/exelixis/material/alerax_data.tar.gz.

**Contact:** Gergely.Szollosi@oist.jp

**Supplementary information:** Supplementary material is available at Bioinformatics online.

1. Introduction

Simultaneously inferring gene trees and the species tree is challenging. Genomes contain abundant information about evolutionary history, but single gene sequences are often too short to reliably resolve gene trees (Haag et al., 2022). Moreover, gene trees are not species trees, but each is the unique result of series of evolutionary events. Starting from an individual site in a genome up to the species level, a hierarchy of evolutionary processes generate genomic sequences, with each level of the
AleRax is obtained by summing over all possible a priori G(2) given estimated and P be the species tree, and S be an alignment of homologous gene trees. DupTree (Wehe et al., 2008) searches for the species tree with the most parsimonious reconciliation cost, measured as the number of duplication events. STAG (Emms and Kelly, 2018) infers a species tree by applying a distance method to each gene family that covers all species, and subsequently builds a consensus tree from all these distance-based trees. FastMuRFs (Molloy and Warnow, 2020) extends the definition of the Robinson-Foulds (RF) distance to multi-copy gene trees and strives to minimize this distance between the species tree and all input gene trees. ASTRAL-Pro 2 (Zhang and Mirarab, 2022) infers a species tree from multi-copy gene trees by maximizing a duplication-aware quartet score. We recently developed SpeciesRax (Morel et al., 2022), a tool that models gene duplication, transfer, and loss (DTL) events and that estimates a maximum likelihood (ML) rooted species tree under the so-called UndatedDTL model. However, all those methods only consider a single estimated gene tree per gene family as input, and are thus sensitive to gene tree uncertainty. Alternative methods such as Phyldog (Boussau et al., 2013) jointly estimate the species tree and the gene trees that evolved along it via speciation, duplication, and loss. However, Phyldog does not scale to large datasets, does not model gene transfer, and is less accurate than SpeciesRax and Astral-Pro on simulations, even in the absence of gene transfers (Morel et al., 2022).

Another challenge is to infer a reconciled gene tree: that is, a gene tree together with a reconciliation, or series of DTL and speciation events that trace its evolution along the species tree. To this end, we previously developed GeneRax (Morel et al., 2020), a probabilistic method that searches for the reconciled gene trees that maximize the joint likelihood, defined as the product of the phylogenetic likelihood (the probability of the multiple sequence alignment (MSA) given the gene tree) and the reconciliation likelihood (the probability of the gene tree given the species tree under UndatedDTL). However, GeneRax does not provide a confidence measure for the inferred gene trees. Amalgamated Likelihood Estimation (ALE) (Szöllősi et al., 2013) is an alternative two-step approach that sums the joint likelihood over all reconciled gene trees. The probabilities of the gene trees are estimated based on the conditional clade probabilities of their constituent clades (Höhna and Drummond, 2012; Larget, 2013). Confidence in a particular DTL event can then be estimated by counting how frequently it appears in the different reconciled gene trees. However, ALE is not able to simultaneously process multiple gene families, or to infer the species tree.

Here we present AleRax, a novel probabilistic method for phylogenetic tree inference that can perform both species tree inference and reconciled gene tree inference from a sample of gene trees. We first present the method and the key features of AleRax. Then, we present the results of our benchmarks: we found that AleRax is on par with ALE in terms of gene tree reconstruction accuracy, but an order of magnitude faster and robust to numerical underflow that sometimes causes ALE to fail. Based on simulations, AleRax is 25% more accurate than SpeciesRax and twice as accurate as ASTRAL-Pro 2 for species tree inference, but at least one order of magnitude slower than those two methods.

2. Method

2.1. Likelihood definition and computation

Let S be the species tree, and A be an alignment of homologous genes, that is, a gene family alignment. The probability of observing A given S is obtained by summing over all possible rooted gene trees G for alignment A:

\[ P(A|S) = \sum_{G} P(A|G)P(G|S) \]  \hspace{1cm} (1)

(Felsenstein, 1988; Szöllősi et al., 2013).

AleRax approximates \( P(A|S) \) by using the ALE (Szöllősi et al., 2013) algorithm, which takes as input a sample of gene trees for each gene family, ideally computed via a Bayesian inference tool such as Phylobayes (Lartillot et al., 2013) or MrBayes (Ronquist et al., 2012) prior to running AleRax. The gene trees in the samples can be rooted or unrooted. Instead of naïvely computing the sum \( \sum_{G} P(A|G)P(G|S) \), AleRax uses the sample of gene trees provided to estimate a nonzero probability for all gene tree topologies that can be "amalgamated" (David and Alm, 2011) from clades present in the sample using conditional clade probabilities (CCPs) (Höhna and Drummond, 2012; Larget, 2013). This corresponds to the maximum entropy distribution given marginal split frequencies in the sample of gene trees provided as input (Szöllősi et al., 2013). Using CCPs allows AleRax to exploit a joint dynamic programming recursion that efficiently and accurately approximates Equation 1 (Szöllősi et al., 2013). An in-depth description of this algorithm is provided in the Supplementary Material. Importantly, unlike ALE, AleRax uses arbitrary-precision floating point values when necessary, allowing it to process datasets that cause ALE to fail due to numerical underflow. AleRax can also process multiple gene families in parallel given a sample of genes for each family allowing it to infer the species tree as well as more realistic DTL models, e.g., branch-wise DTL parameters on the species tree, by maximizing the likelihood of the species tree, which can be written as the product over all per-family alignment probabilities:

\[ L(S|A) = \prod_i P(A_i|S). \]  \hspace{1cm} (2)

2.2. Species tree inference

AleRax searches for the rooted ML species tree. It implements the same tree search strategy as SpeciesRax (Morel et al., 2022), which applies hill-climbing subtree prune and regraft (SPR) moves until it cannot find a move that yields a tree with a better likelihood. However, unlike SpeciesRax, which relies on a single gene tree per gene family, AleRax uses a sample of per family gene trees to integrate over gene tree uncertainty by approximating \( P(A_i|S) \), AleRax recomputes the model parameters (described in the next subsection) after each round of SPR moves. The starting species
Tree can be generated at random, estimated with MiniNJ (Morel et al., 2022) (a duplication-aware distance method), or specified by the user.

2.3. Model parameter estimation

In general, the reconciliation model used by AleRax, the so-called UndatedDTL model (Morel et al., 2020) can have three free parameters per gene family $k$ and per species tree branch $e$: $\delta^k_e$, $\tau^k_e$, and $\lambda^k_e$, representing the duplication, transfer, and loss probabilities, respectively. However, estimating such a large number of free parameters (proportional to the product of the number of species in $S$ multiplied by the number of gene families) can lead to over-parameterization. Hence, AleRax implements three distinct approaches to parameter estimation that can be chosen by the user. In the global parameters mode, all families and all species share the same set of parameters \( \{\delta, \tau, \lambda\} \), resulting in only three free parameters. In the per-family parameters mode, AleRax optimizes a different set of parameters \( \{\delta_k, \tau_k, \lambda_k\} \) for each gene family $k$. This corresponds to $3K$ free parameters, where $K$ is the number of gene families. In the per-species parameter mode, the user can provide a list of species groupings that share the same set of parameters. In this mode, the model has $3n$ free parameters, where $n$ is the number of groupings defined. In each mode, AleRax optimizes the model parameter values with respect to the likelihood via gradient descent. Our experiments suggest that AleRax performs similarly in terms of gene tree and species tree accuracy under the different modes.

2.4. Sampling reconciled gene trees

AleRax samples, per gene family, $r$ reconciled gene trees proportional to their joint likelihood using stochastic backtracking following the dynamic programming recursion (Szüllösi et al., 2013) used to approximate Equation 1. We describe the algorithm in the Supplementary Material. The number of gene trees $r$ to sample is specified by the user. For each gene family, AleRax outputs the $r$ reconciled gene trees, the rooted majority-rule consensus gene tree, and the number of gene events per species, averaged over all output samples for this specific gene family. The reconciliations are stored in RecPhyloXML format (Duchemin et al., 2018) and can be converted into SVG figures using Thirdkind (Penel et al., 2022). AleRax also outputs the species pairs involved in horizontal gene transfers, sorted by the number of
AleRax is on average 25% more accurate than SpeciesRax and twice as accurate as ASTRAL-Pro 2. For instance, Fig. 1 compares the accuracy of the different methods when varying the average gene sequence lengths. Our explanation is that using fixed gene tree estimates does not capture all the information contained in the gene sequences. Treating gene trees as latent variables allows AleRax to avoid this information loss, and thus to be more robust to phylogenetic uncertainty. When using gene tree distributions with 1000 gene trees per gene family, AleRax is from one to two orders of magnitude slower than SpeciesRax (Fig. 1d).

To investigate the performance of rooted species tree inference with AleRax on real data, we returned to an earlier phylogenetic analysis of the archaeal domain of life (Williams et al., 2017; Davín et al., 2018). The Archaea constitute one of the two prokaryotic domains of life, but their early evolution and the position of the root remains poorly understood. Previous studies, including a 2017 analysis by some of us, have placed the root between the genome-reduced DPANN Archaea and the other lineages (Williams et al., 2017; Dombrowski et al., 2020; Aouad et al., 2022). In the 2017 study, we used ALE to infer the optimal root position based on a fixed unrooted topology estimating using concatenation of a set of 45 vertically-evolving marker genes. This hybrid strategy was employed because ALE cannot search the space of rooted species trees. We performed AleRax, SpeciesRax and ASTRAL-Pro-2 analyses on the set of 5,379 gene families, using the same MCMC gene tree samples to represent gene tree uncertainty as in the original study in the case of AleRax. On a 40-cores machine, Astral-Pro 2 finished after 15s, SpeciesRax after 167s, and AleRax after 9,215s. AleRax inferred a rooted species tree closely similar to the original study and to other recent analyses (Dombrowski et al., 2020; Aouad et al., 2022), with a root between DPANN Archaea and the rest, and recovering the monophyly of major archaeal groups including DPANN, TACK Archaea, and most of the Euryarchaeota (see Fig. 2), although two incomplete genomes (SAG I15 and Thaumarchaeota E09) were likely misplaced. The SpeciesRax tree was heterodox, recovering a mixture of TACK Archaea, Euryarchaeota and DPANN on both sides of the root (Fig. 2). The unrooted species tree inferred by ASTRAL-Pro-2 was very similar to the topology of the AleRax tree (Fig. 2). If current views about archaeal phylogeny are broadly correct, this analysis suggests that AleRax can be substantially more accurate than SpeciesRax when inferring the root and topology of species trees from real data, likely because of its ability to model gene tree uncertainty. The analysis also demonstrates that reconciliation methods can capture information about the root and topology of species trees, and that in the case of the Archaea this signal is congruent with that from traditional concatenation-based phylogenetics. In this case study, AleRax and ASTRAL-Pro-2 both recovered species trees that are consistent with current views. While ASTRAL-Pro-2 was much faster, the AleRax analysis provided an estimate of the root of the species tree, reconciliation scenarios for each gene family, and gene content reconstructions for each internal node of the tree. These comparative genomic outputs are an important part of the use case for AleRax and similar reconciliation tools, and can be used (for example) to propose hypotheses about ancestral gene repertoires and organismal characteristics, including metabolisms (Williams et al., 2017).
4. Conclusion and future work

AleRax is an efficient and user-friendly tool for species tree inference and gene tree-species tree reconciliation that can be applied to datasets from across the tree of life. It is particularly useful in biological applications where not only the species tree, but also the histories individual of gene families are of interest. We showed that AleRax is on par with ALE in terms of reconciled gene tree accuracy, while being one order of magnitude faster and more robust to numerical errors. Furthermore, it infers more accurate species trees than SpeciesRax and ASTRAL-Pro 2, because it can accommodate gene tree uncertainty. In the future, we plan to improve our model to accommodate DTL rate heterogeneity over species, incomplete lineage sorting, and horizontal gene transfer time constraints.

Acknowledgement

BM and AS are financially supported by the Klaus Tschira Foundation, by DFG grant STA 860/6-2, and by the European Union (EU) under Grant Agreement No 101087081 (Compet-Biodiv-GR). GSz received funding from the European Research Council under the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 714774 and the grant GINOP-2.3.2-15-2016-00057. This work was funded by the Gordon and Betty Moore Foundation through grant GBMF9741 to TAW and GSz.

References


