Phylogenomics with Paralogs

Marc Hellmuth [1], Nicolas Wieseke[2], Markus Lechner[3], Hans-Peter Lenhof[1], Martin Middendorf[2], and Peter F Stadler[4-9]

[1]Center for Bioinformatics, Saarland University, Building E 2.1, D-66041 Saarbrücken, Germany
[2]Parallel Computing and Complex Systems Group, Department of Computer Science, Leipzig University, Augustusplatz 10, D-04109 Leipzig, Germany
[3]Institut für Pharmazeutische Chemie, Philipps-Universität Marburg, Marbacher Weg 6, D-35032 Marburg, Germany
[4]Bioinformatics Group, Department of Computer Science; and Interdisciplinary Center of Bioinformatics, Leipzig University, Härtelstraße 16-18, D-04107 Leipzig, Germany
[5]Max-Planck-Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany
[6]Fraunhofer Institute IZI, Perlickstraße 1, Leipzig, Germany
[7]Inst. f. Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria
[8]RTH, University of Copenhagen, Grønnegårdsvej 3, Frederiksberg, Denmark
[9]Santa Fe Institute, 1399 Hyde Park Rd., NM 87501 Santa Fe, USA

Abstract

Phylogenomics heavily relies on well-curated sequence data sets that consist, for each gene, exclusively of 1:1-orthologous. Paralogs are treated as a dangerous nuisance that has to be detected and removed. We show here that this severe restriction of the data sets is not necessary. Building upon recent advances in mathematical phylogenetics we demonstrate that gene duplications convey meaningful phylogenetic information and allow the inference of plausible phylogenetic trees provided orthologs and paralogs can be distinguished with a degree of certainty. Starting from tree-free estimates of orthology, cograph editing can sufficiently reduce the noise in order to find correct event-annotated gene trees. The information of gene trees can then directly be translated into constraints on the species trees. While the resolution is very poor for individual gene families, we show that genome-wide data sets are sufficient to generate fully resolved phylogenetic trees, even in the presence of horizontal gene transfer.

1 Introduction

Molecular phylogenetics is primarily concerned with the reconstruction of evolutionary relationships between species based on sequence information. To this end alignments of protein or DNA sequences are employed whose evolutionary history is believed to be congruent to that of the respective species. This property can be ensured most easily in the absence of gene duplications. Phylogenetic studies thus judiciously select families of genes that rarely exhibit duplications (such as rRNAs, most ribosomal proteins, and many of the housekeeping enzymes). In phylogenomics, elaborate automatic pipelines such as `HaMStR` [19], are used to filter genome-wide data sets to at least deplete sequences with detectable paralogs (homologs in the same species).

In the presence of gene duplications, however, it becomes necessary to distinguish between the evolutionary history of genes (*gene trees*) and the evolutionary history of the species (*species trees*) in which these genes reside. Leaves of a gene tree represent genes. Their inner nodes represent two kinds of evolutionary events, namely the duplication of genes within a genome – giving rise to paralogs – and speciations, in which the ancestral gene complement is transmitted to two daughter lineages. Two genes are (co-)orthologous if their last common ancestor in the gene tree represents a speciation event, while they are paralogous if their last common ancestor is a duplication event, see [20] and [21] for a more recent discussion on orthology and paralogy relationships. Speciation events, in turn, define the inner vertices of a species tree. However, they depend on both, the gene and the species phylogeny, as well as the reconciliation between the two. The latter identifies speciation vertices in the gene tree with a particular speciation event in the species tree and places the gene duplication events on the edges of the species tree. Intriguingly, it is nevertheless possible in practice to distinguish orthologs and paralogs with acceptable accuracy without constructing either gene or species trees [2]. Many tools of

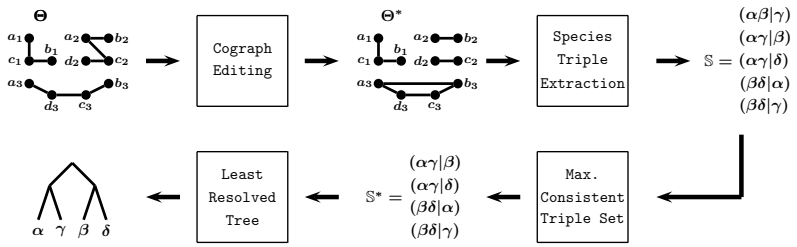


Figure 1: Outline of the computational framework. Starting from an estimated orthology relation Θ, its graph representation G_Θ is edited to obtain the closest cograph G_{Θ^*}, which in turn is equivalent to a (not necessarily fully resolved) gene tree T and an event labeling t. From (T,t) we extract the set $\mathbb{S}$ of all relevant species triples. As the triple set $\mathbb{S}$ need not to be consistent, we compute the maximal consistent subset $\mathbb{S}^*$ of $\mathbb{S}$. Finally, we construct a least resolved species tree from $\mathbb{S}^*$.

this type have become available over the last decade, see [34, 15] for a recent review. The output of such methods is an estimate Θ of the true orthology relation Θ^*, which can be interpreted as a graph G_Θ whose vertices are genes and whose edges connect estimated (co-)orthologs.

Recent advances in mathematical phylogenetics have led to the conclusion that the estimated orthology relation Θ contains information on the structure of the species tree. Intriguingly, the accessible phylogenetic information is entirely encoded in the duplication events, i.e., the paralogs (the complement of orthologs). Building upon the theory of symbolic ultrametrics [5] we showed that a symmetric relation R on a set of genes is an orthology relation if and only if R yields a cograph [26]. Cographs can be generated from the single-vertex graph K_1 by complementation and disjoint union. Moreover, they are associated with a unique tree, known as the cotree, which represents the cographs topology [13]. The corresponding cotree, which can be computed efficiently from the cograph, is a homeomorphic image of the gene tree (in which adjacent events of the same type are collapsed to a common vertex). A key observation is that certain triples of genes from three different species must appear in the same relative arrangement in the species tree [27]. The estimated orthology relation Θ of every gene family that contains gene duplications thus provide some information on the gene tree. Estimates Θ of the true orthology relation Θ^* for many gene families, i.e., data that are commonly computed in phylogenomic studies for the purpose of filtering the input data, therefore might provide sufficient information to reconstruct the species phylogeny on its own.

This idea cannot be turned immediately into a practicable method for data analysis because of the inaccuracies in the estimates of the true orthology relation Θ^*. Work on the cograph-editing problem, which asks for the cograph most similar to an arbitrary input graph [37, 38], however points out an avenue to correcting the noise in the estimate Θ. Although this enables us to compute a collapsed event-labeled gene tree for each gene family, these trees will not necessarily be congruent due to incorrectly edited cographs or because of horizontal gene transfer. A conceptually elegant solution is provided by the theory of supertrees in the form of the largest set of consistent triples [32, 24]. The final step is to compute the least resolved estimate of a species tree consistent with this triple set so that the end result does not pretend to have a higher resolution than actually supported by the data. Fig. 1 illustrates the interconnection between these problems as utilized in this work.

All three combinatorial optimization problems (cograph editing [38], maximal consistent triple set [8, 44, 30] and least resolved supertree [31]) are NP-hard. We show here that they are nevertheless amenable to formulations as Integer Linear Programs (ILP) that can be solved for real-life data sets comprising genome-scale protein sets for dozens of species. The workflow in Fig. 1 reconstructs the history of species at acceptable levels of accuracy from estimates of paralogs and orthologs, resp., even in the presence of horizontal gene transfer.

1.1 Preliminaries

Phylogenetic Trees: We consider a set $\mathfrak{G}$ of at least three genes from a non-empty set $\mathfrak{S}$ of species. We denote genes by lowercase Roman and species by lowercase Greek letters. We assume that for each gene its species of origin is known. This is encoded by the surjective map $\sigma : \mathfrak{G} \to \mathfrak{S}$ with $a \mapsto \sigma(a)$. A *phylogenetic tree (on L)* is a rooted tree $T = (V,E)$ with leaf set $L \subseteq V$ such that no inner vertex $v \in V^0 := V \setminus L$ has outdegree one and whose root $\rho_T \in V$ has indegree zero. A phylogenetic tree T is called *binary* if each inner vertex has outdegree two. A phylogenetic tree on $\mathfrak{G}$, resp., on $\mathfrak{S}$, is called *gene tree*, resp., *species tree*. A (inner) vertex y is an ancestor of $x \in V$, in symbols $x \prec_T y$ if $y \neq x$ lies on the unique path connecting x with ρ_T. The *most recent common ancestor* $\mathrm{lca}_T(L')$ of a subset $L' \subseteq L$ is the unique vertex in T that is the least upper bound of L' under the partial order $\preceq_T$. We write $L(v) := \{y \in L | y \preceq_T v\}$ for the set of leaves in the subtree $T(v)$ of T rooted in v. Thus, $L(\rho_T) = L$ and $T(\rho_T) = T$.

Rooted Triples: Rooted triples [18] are a key concept in the theory of supertrees [40, 3]. A rooted triple $r = (\mathsf{xy}|\mathsf{z})$

with leaf set $L_r = \{x,y,z\}$ is *displayed* by a phylogenetic tree T on L if (i) $L_r \subseteq L$ and (ii) the path from x to y does not intersect the path from z to the root ρ_T. Thus $\mathrm{lca}_T(x,y) \prec_T \mathrm{lca}_T(x,y,z)$. A set R of triples is *(strict) dense* on a given leaf set L if for each set of three distinct leaves there is (exactly) one triple $r \in R$. We denote by $\mathfrak{R}(T)$ the set of all triples that are displayed by the phylogenetic tree T. A set R of triples is *consistent* if there is a phylogenetic tree T on $L_R := \cup_{r\in R} L_r$ such that $R \subseteq \mathfrak{R}(T)$, i.e., T displays (all triples of) R. If no such tree exists, R is said to be *inconsistent*.
Given a triple set R, the polynomial-time algorithm `BUILD` [1] either constructs a phylogenetic tree T displaying R or recognizes that R is inconsistent. The problem of finding a phylogenetic tree with the smallest possible number of vertices that is consistent with every rooted triple in R, i.e., a *least resolved* tree, is an NP-hard problem [31]. If R is inconsistent, the problem of determining a maximum consistent subset of an inconsistent set of triples is NP-hard and also APX-hard, see [10, 42]. Polynomial-time approximation algorithms for this problem and further theoretical results are reviewed by [11].

1.2 Theory

1.2.1 Triple Closure Operations and Inference Rules.

If R is consistent it is often possible to infer additional consistent triples. Denote by $\langle R\rangle$ the set of all phylogenetic trees on L_R that display R. The *closure* of a consistent set of triples R is $\mathrm{cl}(R) = \cap_{T\in\langle R\rangle}\mathfrak{R}(T)$, see [9, 23, 8, 29, 4]. We say R is *closed* if $R = \mathrm{cl}(R)$ and write $R \vdash (\mathsf{xy}|\mathsf{z})$ iff $(\mathsf{xy}|\mathsf{z}) \in \mathrm{cl}(R)$. The closure of a given consistent set R can be computed in in $O(|R|^5)$ time [9]. Extending earlier work of Dekker [17], Bryant and Steel [9] derived conditions under which $R \vdash (\mathsf{xy}|\mathsf{z}) \implies R' \vdash (\mathsf{xy}|\mathsf{z})$ for some $R' \subseteq R$. Of particular importance are the following so-called *2-order* inference rules:

$$\{(\mathsf{ab}|\mathsf{c}),(\mathsf{ad}|\mathsf{c})\} \vdash (\mathsf{bd}|\mathsf{c}) \tag{i}$$
$$\{(\mathsf{ab}|\mathsf{c}),(\mathsf{ad}|\mathsf{b})\} \vdash (\mathsf{bd}|\mathsf{c}),(\mathsf{ad}|\mathsf{c}) \tag{ii}$$
$$\{(\mathsf{ab}|\mathsf{c}),(\mathsf{cd}|\mathsf{b})\} \vdash (\mathsf{ab}|\mathsf{d}),(\mathsf{cd}|\mathsf{a}). \tag{iii}$$

Inference rules based on pairs of triples $r_1, r_2 \in R$ can imply new triples only if $|L_{r_1} \cap L_{r_2}| = 2$. Hence, in a strict dense triple set only the three rules above may lead to new triples. The following two results (see [24] and Supplemental Material) play a key role for the ILP formulation of triple consistency:

Theorem 1. *A strict dense triple set R on L with $|L| \geq 3$ is consistent if and only if $\mathrm{cl}(R') \subseteq R$ holds for all $R' \subseteq R$ with $|R'| = 2$.*

Theorem 2. *If the tree T inferred from the triple set R by means of `BUILD` is binary, then the closure $\mathrm{cl}(R)$ is strict dense. Moreover, T is unique and hence, a least resolved tree for R.*

1.2.2 Orthology Relations and Cographs.

An empirical orthology relation $\Theta \subset \mathfrak{G} \times \mathfrak{G}$ is a symmetric, irreflexive relation that contains all pairs (x,y) of orthologous genes. Here, we assume that $x,y \in \mathfrak{G}$ are *paralogs* if and only if $x \neq y$ and $(x,y) \notin \Theta$. This amounts to ignoring horizontal gene transfer. Orthology detection tools often report some weight or confidence value $w(x,y)$ for x and y to be orthologs from which Θ is estimated using a suitable cutoff. Importantly, Θ is symmetric, but not transitive, i.e., it does in general not represent a partition of $\mathfrak{G}$.

Given Θ we aim to find a gene tree T with an "event labeling" $t : V^0 \to \{\bullet, \square\}$ at the inner vertices so that, for any two distinct genes $x,y \in L$, $t(\mathrm{lca}_T(x,y)) = \bullet$ if $\mathrm{lca}_T(x,y)$ corresponds to a speciation and hence $(x,y) \in \Theta$ and $t(\mathrm{lca}_T(x,y)) = \square$ if $\mathrm{lca}_T(x,y)$ is a duplication vertex and hence $(x,y) \notin \Theta$. If such a tree T with event-labeling t exists for Θ, we call the pair (T,t) a *symbolic representation* of Θ. We write $(T,t;\sigma)$ if in addition the species assignment map σ is given. A detailed and more general introduction to the theory of symbolic representations is given in the Supplemental Material.

Empirical estimates of the orthology relation Θ will in general contain errors in the form of false-positive orthology assignments, as well as false negatives e.g. due to insufficient sequence similarity. Horizontal gene transfer adds to this noise. Hence an empirical relation Θ will in general not have a symbolic representation. In fact, Θ has a *symbolic representation* (T,t) if and only if G_Θ is a cograph [26], from which (T,t) can be derived in linear time, see also Theorem 5 in the Supplemental Material. Cographs have simple characterization as P_4-free graphs, that is, no four vertices induce a simple path. We refer to [7] for a survey of cographs and many other equivalent characterizations. Cographs can be recognized in linear time [14, 25]. However, the *cograph editing problem*, which aims to convert a given graph $G(V,E)$ into a cograph $G^* = (V,E^*)$ with the minimal number $|E \triangle E^*|$ of inserted or deleted edges, is an NP-complete problem [37, 38]. The symbol $\triangle$ denotes the symmetric difference of two sets. As shown in the Supplemental Material, it is therefore NP-complete to decide for a given Θ and a positive integer K whether there is an orthology relation Θ^* that has a (discriminating) symbolic representation such that $|\Theta \triangle \Theta^*| \leq K$.

In our setting the problem is considerably simplified by the structure of the input data. The gene set of every living organism consists of hundreds or even thousands of non-homologous gene families. Thus the initial estimate of G_Θ

already consists of a large number of connected components. As shown in Lemma 8 in the Supplemental Material, it suffices to solve the cograph editing for each connected component separately.

1.2.3 Triples and Reconciliation Maps.

A phylogenetic tree $S = (W, F)$ on $\mathfrak{S}$ is a species tree for a gene tree $T = (V, E)$ on $\mathfrak{G}$ if there is a reconciliation map $\mu : V \to W \cup F$ that maps genes $a \in \mathfrak{G}$ to species $\sigma(a) = \alpha \in \mathfrak{S}$ such that the ancestor relation $\preceq_S$ is implied by the ancestor relation $\preceq_T$. A more formal definition is given in the Supplemental Material. Inner vertices of T that map to inner vertices of S are speciations, while vertices of T that map to edges of S are duplications. Hernandez et al. [27] investigated the conditions for the existence of a reconciliation map μ from T to S. Given $(T, t; \sigma)$, consider the triple set $\mathbb{G}$ consisting of all triples $r = (\mathsf{ab}|\mathsf{c}) \in \mathfrak{R}(T)$ so that (i) all genes $a, b, c \in L_r$ belong to different species, and (ii) the event at the most recent common ancestor of L_r is a speciation event, $t(\mathrm{lca}_T(a, b, c)) = \bullet$. From $\mathbb{G}$ and σ, one can construct the following set of species triples:

$$\mathbb{S} = \{(\alpha\beta|\gamma) \,|\, \exists(\mathsf{ab}|\mathsf{c}) \in \mathbb{G} \text{ with } \sigma(a) = \alpha, \sigma(b) = \beta, \sigma(c) = \gamma\} \tag{1}$$

The main result of [27] establishes that there is a species tree on $\sigma(\mathfrak{G})$ for (T, t, σ) if and only if the triple set $\mathbb{S}$ is consistent. In this case, a reconciliation map can be found in polynomial time. No reconciliation map exists if $\mathbb{S}$ is inconsistent.

In order to compute an estimate for the species tree in practice, we therefore have to compute a maximum consistent subset of triples $\mathbb{S}^* \subset \mathbb{S}$ and to compute a least resolved tree S from $\mathbb{S}^*$. As discussed above, both of these problems are NP-hard.

1.3 ILP Formulation

Since we have to solve three intertwined NP-complete optimization problems we cannot realistically hope for an efficient exact algorithm. We therefore resort to ILP as the method of choice for solving the problem of computing a least resolved species tree S from an empirical estimate of the orthology relation G_Θ. We will use binary variables throughout. Table 1.3 summarizes the definition of the ILP variables and provides a key to the notation used in this section. In the following we summarize the ILP formulation. A detailed description and proofs for the correctness and completeness of the inequality constraints can be found in the Supplemental Material.

Sets & Constants	**Definition**
$\mathfrak{G}$	Set of genes
$\mathfrak{S}$	Set of species
Θ_{ab}	Genes $a, b \in \mathfrak{G}$ are estimated orthologs: $\Theta_{ab} = 1$ iff $(a, b) \in \Theta$.
Binary Variables	**Definition**
E_{xy}	Edge set of the cograph $G_{\Theta^*} = (\mathfrak{G}, E_{\Theta^*})$ of the closest relation Θ^* to Θ: $E_{xy} = 1$ iff $\{x, y\} \in E_{\Theta^*}$ (thus, iff $(x, y) \in \Theta^*$).
$T_{(\alpha\beta\|\gamma)}$	Rooted (species) triples in obtained set $\mathbb{S}$: $T_{(\alpha\beta\|\gamma)} = 1$ iff $(\alpha\beta\|\gamma) \in \mathbb{S}$.
$T'_{(\alpha\beta\|\gamma)}$, $T^*_{(\alpha\beta\|\gamma)}$	Rooted (species) triples in auxiliary strict dense set $\mathbb{S}'$, resp., maximal consistent species triple set $\mathbb{S}^*$: $T^\bullet_{(\alpha\beta\|\gamma)} = 1$ iff $(\alpha\beta\|\gamma) \in \mathbb{S}^\bullet$, $\bullet \in \{\prime, *\}$.
$M_{\alpha p}$	Set of clusters: $M_{\alpha p} = 1$ iff $\alpha \in \mathfrak{S}$ is contained in cluster $p \in \{1, \ldots, \lvert\mathfrak{S}\rvert - 2\}$.
$N_{\alpha\beta,p}$	Cluster p contains both species α and β: $N_{\alpha\beta,p} = 1$ iff $M_{\alpha p} = 1$ and $M_{\beta p} = 1$
$C_{p,q,\Gamma\Lambda}$	Compatibility: $C_{p,q,\Gamma\Lambda} = 1$ iff cluster p and q have gamete $\Gamma\Lambda \in \{01, 10, 11\}$.
Y_p	Non-trivial clusters: Y_p=1 iff cluster $p \neq \emptyset$.

Table 1: The notation used in our ILP formulation.

1.3.1 From Estimated Orthologs to Cographs.

Our first task is to compute a cograph G_{Θ^*} that is as similar as possible to G_Θ (Eq. (ILP 1) and (ILP 3)) with the additional constraint that no pair of genes within the same species is connected by an edge, since no pair of orthologs can be found in the same species (Eq. (ILP 2)). Binary variables E_{xy} express (non)edges in G_{Θ^*} and binary constants Θ_{ab} (non)pairs of the input relation Θ. This ILP formulation requires $O(|\mathfrak{G}|^2)$ binary variables and $O(|\mathfrak{G}|^4)$ constraints. In practice, the effort is not dominated by the number of edges, since the connected components of G_Θ can be treated independently.

$$\min \sum_{(x,y)\in\mathfrak{G}\times\mathfrak{G}} (1-\Theta_{xy})E_{xy} + \sum_{(x,y)\in\mathfrak{G}\times\mathfrak{G}} \Theta_{xy}(1-E_{xy}) \tag{ILP 1}$$

$$E_{xy} = 0 \text{ for all } \{x,y\} \text{ with } \sigma(x) = \sigma(y). \tag{ILP 2}$$

$$E_{wx} + E_{xy} + E_{yz} - E_{xz} - E_{wy} - E_{wz} \leq 2 \tag{ILP 3}$$
$$\forall \text{ ordered tuples } (w,x,y,z) \text{ of distinct } w,x,y,z \in \mathfrak{G}$$

1.3.2 Extraction of All Species Triples.

The construction of the species tree S is based upon the set $\mathbb{S}$ of species triples that can be derived from the set of gene triples $\mathbb{G}$, as explained in the previous section. Although the problem of determining such triples is not NP-hard, we give in the Supplemental Material an ILP formulation due to the sake of completeness. However, as any other approach can be used to determine the species triples we omit here the ILP formulation, but state that it requires $O(|\mathfrak{S}|^3)$ variables and $O(|\mathfrak{G}|^3 + |\mathfrak{S}|^4)$ constraints.

1.3.3 Maximal Consistent Triple Set.

An ILP approach to find maximal consistent triple sets was proposed in [12]. It explicitly builds up a binary tree as a way of checking consistency. Their approach, however, requires $O(|\mathfrak{S}|^4)$ ILP variables, which limits the applicability in practice. By Theorem 1, a strict dense triple set R is consistent if, for all two-element subsets $R' \subseteq R$, the closure $\mathrm{cl}(R')$ is contained in R. This observation allows us to avoid the explicit tree construction and makes is much easier to find a maximal consistent subset $\mathbb{S}^* \subseteq \mathbb{S}$. Of course, neither $\mathbb{S}^*$ nor $\mathbb{S}$ need to be strict dense. However, since $\mathbb{S}^*$ is consistent, Lemma 7 (Supplemental Material) guarantees that there is a strict dense triple set $\mathbb{S}'$ containing $\mathbb{S}^*$. Thus we have $\mathbb{S}^* = \mathbb{S}' \cap \mathbb{S}$, where $\mathbb{S}'$ must be chosen to maximize $|\mathbb{S}' \cap \mathbb{S}|$. We define binary variables $T'_{(\alpha\beta|\gamma)}$, $T^*_{(\alpha\beta|\gamma)}$, resp., binary constants $T_{(\alpha\beta|\gamma)}$ to indicate whether $(\alpha\beta|\gamma)$ is contained in $\mathbb{S}'$, $\mathbb{S}^*$, resp., $\mathbb{S}$. The ILP formulation that uses $O(|\mathfrak{S}|^3)$ variables and $O(|\mathfrak{S}|^4)$ constraints is as follows.

$$\max \sum_{(\alpha\beta|\gamma)\in\mathbb{S}} T'_{(\alpha\beta|\gamma)} \tag{ILP 4}$$

$$T'_{(\alpha\beta|\gamma)} + T'_{(\alpha\gamma|\beta)} + T'_{(\beta\gamma|\alpha)} = 1. \tag{ILP 5}$$

$$2T'_{(\alpha\beta|\gamma)} + 2T'_{(\alpha\delta|\beta)} - T'_{(\beta\delta|\gamma)} - T'_{(\alpha\delta|\gamma)} \leq 2 \tag{ILP 6}$$

$$0 \leq T'_{(\alpha\beta|\gamma)} + T_{(\alpha\beta|\gamma)} - 2T^*_{(\alpha\beta|\gamma)} \leq 1 \tag{ILP 7}$$

This ILP formulation can easily be adapted to solve a *"weighted" maximum consistent subset* problem: Denote by $w(\alpha\beta|\gamma)$ the number of connected components in G_{Θ^*} that contain three vertices $a,b,c \in \mathfrak{G}$ with $(\mathsf{ab}|\mathsf{c}) \in \mathbb{G}$ and $\sigma(a) = \alpha, \sigma(b) = \beta, \sigma(c) = \gamma$. These weights can simply be inserted into the objective function Eq. (ILP 4)

$$\max \sum_{(\alpha\beta|\gamma)\in\mathbb{S}} T'_{(\alpha\beta|\gamma)} * w(\alpha\beta|\gamma). \tag{ILP 8}$$

to increase the relative importance of species triples in $\mathbb{S}$ if they are observed in multiple gene families.

1.3.4 Least Resolved Species Tree.

We finally have to find a least resolved species tree from the set $\mathbb{S}^*$ computed in the previous step. Thus the variables $T^*_{(\alpha\beta|\gamma)}$ become the input constants. For the explicit construction of the tree we use some of the ideas of [12].

To build an arbitrary tree for the consistent triple set $\mathbb{S}^*$, one can use one of the fast implementations of BUILD [40]. If this tree is binary, then Theorem 2 implies that the closure $\mathrm{cl}(\mathbb{S}^*)$ is strict dense and that this tree is a unique and least

resolved tree for $\mathbb{S}^*$. Hence, as a preprocessing step `BUILD` is used in advance, to test whether the tree for $\mathbb{S}^*$ is already binary. If not, we proceed with the following ILP approach that uses $O(|\mathfrak{S}|^3)$ variables and constraints.

$$\min \sum_p Y_p \tag{ILP 9}$$

$$0 \leq Y_p|\mathfrak{S}| - \sum_{\alpha\in\mathfrak{S}} M_{\alpha p} \leq |\mathfrak{S}| - 1. \tag{ILP 10}$$

$$0 \leq M_{\alpha p} + M_{\beta p} - 2N_{\alpha\beta,p} \leq 1. \tag{ILP 11}$$

$$1 - |\mathfrak{S}|(1 - T^*_{(\alpha\beta|\gamma)}) \leq \sum_p N_{\alpha\beta,p} - \frac{1}{2}N_{\alpha\gamma,p} - \frac{1}{2}N_{\beta\gamma,p}. \tag{ILP 12}$$

$$\begin{aligned} C_{p,q,01} &\geq -M_{\alpha p} + M_{\alpha q} \\ C_{p,q,10} &\geq M_{\alpha p} - M_{\alpha q} \\ C_{p,q,11} &\geq M_{\alpha p} + M_{\alpha q} - 1 \end{aligned} \tag{ILP 13}$$

$$C_{p,q,01} + C_{p,q,10} + C_{p,q,11} \leq 2 \;\forall p,q \tag{ILP 14}$$

Since a phylogenetic tree S is equivalently specified by its *hierarchy* $\mathscr{C} = \{L(v) \mid v \in V(S)\}$ whose elements are called *clusters* (see Supplemental Material or [40]), we construct the clusters induced by all triples of $\mathbb{S}^*$ and check whether they form a hierarchy on $\mathfrak{S}$. Following [12], we define the binary $|\mathfrak{S}| \times (|\mathfrak{S}| - 2)$ matrix M, whose entries $M_{\alpha p} = 1$ indicates that species α is contained in cluster p, see Supplemental Material. The entries $M_{\alpha p}$ serve as ILP variables. In contrast to the work of [12], we allow *trivial* columns in M in which all entries are 0. Minimizing the number of *non-trivial* columns then yields a least resolved tree.

For any two distinct species α,β and all clusters p we introduce binary variables $N_{\alpha\beta,p}$ that indicate whether two species α,β are both contained in the same cluster p or not (Eq. (`ILP 11`)). To determine whether a triple $(\alpha\beta|\gamma)$ is contained in $\mathbb{S}^* \subseteq \mathbb{S}$ and displayed by a tree, we need the constraint Eq. (`ILP 12`). Following, the ideas of Chang et al. we use the "three-gamete condition" Eq. (`ILP 13`) and (`ILP 14`) ensures that M defines a "partial" hierarchy (any two clusters satisfy $p \cap q \in \{p,q,\emptyset\}$) of compatible clusters. A detailed discussion how these conditions establish that M encodes a "partial" hierarchy M can be found in the Supplemental Material.

Our aim is to find a least resolved tree that displays all triples of $\mathbb{S}^*$. We use the $|\mathfrak{S}| - 2$ binary variables $Y_p = 1$ to indicate whether there are non-zero entries in column p (Eq. (`ILP 10`)). Finally, Eq. (`ILP 9`) captures that the number of non-trivial columns in M, and thus the number of inner vertices in the respective tree, is minimized.

1.4 Implementation and Data Sets

Details on implementation and test data sets can be found in the Supplemental Material. Simulated data were computed with and without horizontal gene transfer using both the method described in [28] and the `Artificial Life Framework` (`ALF`) [16]. As real-life data sets we used the complete protein complements of 11 *Aquificales* and 19 *Enterobacteriales*, resp. The initial orthology relation are estimated with `Proteinortho` [35]. The ILP formulation of Fig. 1 is implemented in the Software `ParaPhylo` using `IBM ILOG CPLEX`™ Optimizer 12.6. `ParaPhylo` is freely available from `http://pacosy.informatik.uni-leipzig.de/paraphylo`.

2 Results and Discussion

The key result of the theory layed out in more rigorous way in *Materials and Methods* is that estimates of the orthology relations within gene families with paralogs contains useful phylogenetic information. This insight arises as the combination of several abstract mathematical results: (1) *In the absence of horizontal gene transfer, the orthology relation of each gene family is a cograph.* Since cographs have a very special restrictive structure this is a very strong constraint that can be used to reduce the noise and inaccuracies of empirical estimates of orthology from pairwise sequence comparison. To this end, the empirically estimated orthology assignments are edited to the nearest cograph in such a way that a minimal number of edges (i.e., orthology assignments) is introduced or removed. (2) It is well known that *each cograph is equivalently represented by its cotree*. In our context this cotree is an incompletely resolved gene-tree endowed with the additional information that for each interior node it is unambiguously known whether the branch points are a speciation or a duplication events. Even though adjacent speciations or adjacent duplications cannot be resolved, the tree faithfully encodes the relative order of any pair of duplication and speciation. In the presence of horizontal gene transfer G_Θ may deviate from the structural requirements of a cograph. Still, the situation can be described in terms of edge-colored graphs

whose subgraphs are cographs [5, 26], so that the cograph structure remains an acceptable approximation. (3) *Every triple in this cotree that has leaves from three species and is rooted in a speciation event also appears in the underlying species tree.* This result allows us to collect from the cotrees for each gene family partial information on the underlying species tree. Interestingly, only gene families that harbor duplications, and thus have a non-trivial cotree, are informative. If no paralogs exist, then the orthology relation is a clique (i.e., every family member is orthologous to every other family member) and the corresponding cotree is completely unresolved, and hence contains no triple.

Taken together in a genome-wide approach, the informative triples rooted in speciation events comprise the information implicit in the orthology relation. To obtain fully resolved species trees, a sufficient number of gene duplications must have occurred, distributed across the many gene families. To reconstruct the species phylogeny it therefore suffices to solve the supertree problem for this triple set. More precisely, the best estimate of the species phylogeny is the least resolved tree that contains all informative triples. Despite the variance reduction due to cograph editing, noise in the data, as well the as occasional introduction of contradictory triples as a consequence of horizontal gene transfer is unavoidable. Thus the collected triple set will in general not be consistent with a single tree. Our best estimate is therefore the least resolved tree covering the largest subset of compatible triples.

We use here an exact ILP formulation, outlined above and described in full detail in the Supplemental Material, to compute species trees from empirically estimated orthology assignments. The corresponding workflow is summarized in Fig. 1. As a proof of concept, we use simulated data to demonstrate that it is indeed feasible in practice to obtain correct gene trees directly from empirical estimates of orthology.

For more than 300 gene families the average TT distance [6] was always smaller than 0.09, independently from the number of species, see Fig. 2(a). The latter result implies that the reconstructed species trees are almost identical. Other tree distances are discussed in the Supplemental Material. Moreover, 80%, 56%, 24%, and 11% of the species trees could be reconstructed perfectly (fully resolved) for 5, 10, 15, and 20 species, respectively, using 500 gene families. This comes with no surprise, given the low amount of paralogs in the simulations (7.5% to 11.2%), and the high amount of extremely short branches in the generated species trees – on 11.3% to 17.9% of the branches, less then one duplication is expected to occur.

In order to evaluate the robustness of the species trees in response to noise in the input data we used simulated gene families with different noise models and levels: (i) insertion and deletion of edges in the orthology graph (homologous noise), (ii) insertion of edges (orthologous noise), (iii) deletion of edges (paralogous noise), and (iv) modification of gene/species assignments (xenologous noise). We observe a substantial dependence of the accuracy of the reconstructed species trees on the noise model. The results are most resilient against overprediction of orthology (noise model ii), while missing edges in Θ have a larger impact, see Fig. 2(c) for TT distance, and Supplemental Material for the other distances. This behavior can be explained by the observation that many false orthologs (overpredicting orthology) lead to an orthology graph whose components are more clique-like and hence, yield few informative triples. Incorrect species triples thus are reduced, while missing species triples often can be supplemented through other gene families. On the other hand, if there are many false paralogs (underpredicting orthology) more false species triples are introduced, resulting in inaccurate trees. Xenologous noise (model iv), simulated by changing gene/species associations with probability p while retaining the original gene tree, amounts to an extreme model for horizontal transfer. Our model, in particular in the weighted version, is quite robust for small amounts of HGT of 5% to 10%. Although some incorrect triples are introduced in the wake of horizontal transfer, they are usually dominated by correct alternatives observed from multiple gene families, and thus excluded during computation of the maximal consistent triple set. Only large scale concerted horizontal transfer, which may occur in long-term endosymbiotic associations [33], thus pose a serious problem.

Simulations with `ALF`[16] show that our method is resilient against errors resulting from mis-predicting xenology as orthology, see Figure 2(b) right, even at horizontal gene transfer rates of 39.5%. Assuming perfect paralogy knowledge, i.e., assuming that all xenologs are mis-predicted as orthologs, the correct trees are reconstructed essentially independently from the amount of HGT for 69.75% of the data sets, and the triple distance to the correct tree remain minute in the remaining cases. This is consistent with noise model (ii), i.e., a bias towards overpredicting orthology. Tree reconstruction based directly on the estimated orthology relation computed with `Proteinortho` are of course more inaccurate, Figure 2(b) left. Even extreme rates of HGT, however, have no discernible effect on the quality of the inferred species trees. Our approach is therefore limited only by quality of initial orthology prediction tools.

The fraction s of all triples obtained from the orthology relations that are retained in the final tree estimates serves as a quality measure similar in flavor e.g. to the retention index of cladistics. Bootstrapping support values for individual nodes are readily computed by resampling either at the level of gene families or at the level of triples (see Supplemental Material).

With the *Aquificales* data set `Proteinortho` predicts 2856 gene families, from which 850 contain duplications. The reconstructed species tree (see Fig. 3, support $s = 0.61$) is almost identical to the tree presented in [36]. All species are clustered correctly according to their taxonomic families. A slight difference refers to the two *Sulfurihydrogenibium* species not being directly clustered. These two species are very closely related. With only a few duplicates exclusively found in one of the species, the data was not sufficient for the approach to resolve this subtree correctly. Additionally, *Hydrogenivirga sp.* is misplaced next to *Persephonella marina*. This does not come as a surprise: Lechner *et al.* [36]

already suspected that the data from this species was contaminated with material from *Hydrogenothermaceae*.

The second data set comprises the genomes of 19 *Enterobacteriales* with 8218 gene families of which 15 consists of more than 50 genes and 1342 containing duplications. Our orthology-based tree shows the expected groupings of *Escherichia* and *Shigella* species and identifies the monophyletic groups comprising *Salmonella*, *Klebsiella*, and *Yersinia* species. The topology of the deeper nodes agrees only in part with the reference tree from PATRIC database [43], see Supplemental Material for additional information. The resulting tree has a support of 0.53, reflecting that a few of the deeper nodes are poorly supported.

Data sets of around 20 species with a few thousand gene families, each having up to 50 genes, can be processed in reasonable time, see Table S1. However, depending on the amount of noise in the data, the runtime for cograph editing can increase dramatically even for families with less than 50 genes.

3 Conclusion

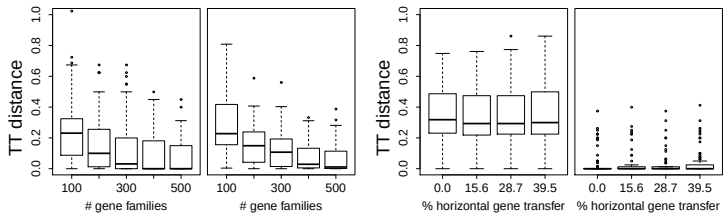
We have shown here both theoretically and in a practical implementation that it is possible to access the phylogenetic information implicitly contained in gene duplications and thus to reconstruct a species phylogeny from information of paralogy only. This source of information is strictly complementary to the sources of information employed in phylogenomics studies, which are always based on alignments of orthologous sequences. In fact, 1:1 orthologs – the preferred data in sequence-based phylogenetics – correspond to cographs that are complete and hence have a star as their cotree and therefore do not contribute *at all* to the phylogenetic reconstruction in our approach. Access to the phylogenetic information implicit in (co-)orthology data requires the solution of three NP-complete combinatorial optimization problems. This is generally the case in phylogenetics, however: both the multiple sequence alignment problem and the extraction of maximum parsimony, maximum likelihood, or optimal Bayesian trees is NP-complete as well. Here we solve the computational tasks exactly for moderate-size problems by means of an ILP formulation. Using phylogenomic data for *Aquificales* and *Enterobacteriales* we demonstrated that non-trivial phylogenies can indeed be re-constructed from tree-free orthology estimates alone. Just as sequence-based approaches in molecular phylogeny crucially depend on the quality of multiple sequence alignments, our approach is sensitive to the initial estimate Θ of the orthology relation. Horizontal gene transfer, furthermore, is currently not included in the model but rather treated as noise that disturbs the phylogenetic signal. Simulated data indicate that the method is rather robust and can tolerate surprisingly large levels of noise in the form of both mis-predicted orthology and horizontal gene transfer, provided a sufficient number of independent gene families is available as input data. Importantly, horizontal gene-transfer can introduce a bias only when many gene families are simultaneously affected by horizontal transfer. Lack of duplications, on the other hand, limits our resolution at very short time scales, a regime in which sequence-based approaches work very accurately.

We have used here an exact implementation as ILP to demonstrate the potential of the approach without confounding it with computational approximations. The current implementation thus does not easily scale to very large data sets. Paralleling the developments in sequence-based phylogenetics, where the NP-complete problems of finding a good input alignment and of constructing tree(s) maximizing the parsimony score, likelihood, or Bayesian posterior probability also cannot be solved exactly for large data sets, it will be necessary in practice to settle for heuristic solutions. In sequence-based phylogenetics, these have improved over decades to the point where they are no longer a limiting factor in phylogenetic reconstruction. Several polynomial time heuristics and approximation algorithms have been devised *already* for the triple consistency problem [22, 39, 10, 41]. The cograph editing problem and the least resolved tree problem, in contrast, have received comparably little attention so far, but constitute the most obvious avenues for boosting computational efficiency. Empirical observations such as the resilience of our approach against overprediction of orthologs in the input will certainly be helpful in designing efficient heuristics.

In the long run, we envision that the species tree S, and the symbolic representation of the event-annotated gene tree (T,t) may serve as constraints for a refinement of the initial estimate of Θ, solely making use only of (nearly) unambiguously identified branchings and event assignments. A series of iterative improvements of estimates for Θ, (T,t), and S, and, more importantly, methods that allow to accurately detect paralogs, may not only lead to more accurate trees and orthology assignments, but could also turn out to be computationally more efficient.

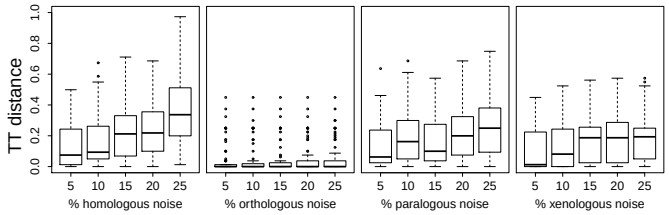
Acknowledgments

We thank Jiong Guo, Leo van Iersel, Daniel Stöckel, and Jakob L. Andersen for helpful comments on the cograph editing problem and the ILP formulation. This work was funded by the German Research Foundation (DFG) (Proj. No. MI439/14-1).



(a) Accuracy as function of number of independent gene families.

(b) Accuracy as function of amount of HGT.



(c) Accuracy as function of noise type and level.

Figure 2: Accuracy of reconstructed species trees in simulated data sets. *(a)* Dependence on the number of gene families: 10 (left), and 20 (right) species and 100 to 500 gene families are generated using `ALF` with duplication/loss rate 0.005 and horizontal gene transfer rate 0.0. *(b)* Dependence on the intensity of horizontal gene transfer: Orthology estimated with `Proteinortho` (left), and assuming perfect paralogy knowledge (right); 10 species and 1000 gene families are generated using `ALF` with duplication/loss rate 0.005 and horizontal gene transfer rate ranging from 0.0 to 0.0075. *(c)* Dependence on the type and intensity ($p = 5 - 25\%$) of noise in the raw orthology data Θ: 10 species and 1000 gene families are generated using `ALF` with duplication/loss rate 0.005 and horizontal gene transfer rate 0.0. Tree distances are measured by the triple metric (TT); all box plots summarize 100 independent data sets.

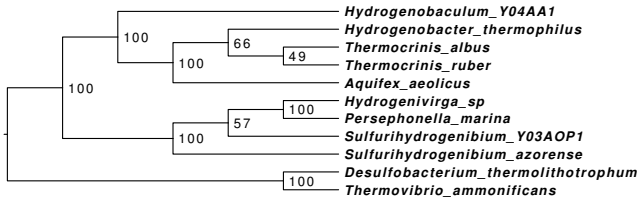


Figure 3: Phylogenetic tree of eleven *Aquificales* species inferred from paralogy. Internal node labels indicate triple-based bootstrap support.

References

[1] A. V. Aho, Y. Sagiv, T. G. Szymanski, and J. D. Ullman. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM J. Comput.*, 10:405–421, 1981.

[2] A. M. Altenhoff and C. Dessimoz. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol.*, 5:e1000262, 2009.

[3] O.R.P Bininda-Emonds. *Phylogenetic Supertrees*. Kluwer Academic Press, Dordrecht, The Netherlands, 2004.

[4] S. Böcker, D. Bryant, A.W.M. Dress, and M.A. Steel. Algorithmic aspects of tree amalgamation. *Journal of Algorithms*, 37(2):522 – 537, 2000.

[5] S. Böcker and A.W.M. Dress. Recovering symbolically dated, rooted trees from symbolic ultrametrics. *Adv. Math.*, 138:105–125, 1998.

[6] D. Bogdanowicz, K. Giaro, and B. Wróbel. Treecmp: Comparison of trees in polynomial time. *Evolutionary Bioinformatics Online*, 8:475, 2012.

[7] A. Brandstädt, V.B. Le, and J.P. Spinrad. *Graph Classes: A Survey*. SIAM Monographs on Discrete Mathematics and Applications. Soc. Ind. Appl. Math., Philadephia, 1999.

[8] D. Bryant. *Building trees, hunting for trees, and comparing trees: theory and methods in phylogenetic analysis.* PhD thesis, University of Canterbury, 1997.

[9] D. Bryant and M. Steel. Extension operations on sets of leaf-labelled trees. *Adv. Appl. Math.*, 16(4):425–453, 1995.

[10] J. Byrka, P. Gawrychowski, K. T. Huber, and S. Kelk. Worst-case optimal approximation algorithms for maximizing triplet consistency within phylogenetic networks. *J. Discr. Alg.*, 8:65–75, 2010.

[11] J. Byrka, S. Guillemot, and J. Jansson. New results on optimizing rooted triplets consistency. *Discr. Appl. Math.*, 158:1136–1147, 2010.

[12] W-C Chang, G.J. Burleigh, D.F. Fernández-Baca, and O. Eulenstein. An ilp solution for the gene duplication problem. *BMC bioinformatics*, 12(Suppl 1):S14, 2011.

[13] D. G. Corneil, H. Lerchs, and L. Steward Burlingham. Complement reducible graphs. *Discr. Appl. Math.*, 3:163–174, 1981.

[14] D. G. Corneil, Y. Perl, and L. K. Stewart. A linear recognition algorithm for cographs. *SIAM J. Computing*, 14:926–934, 1985.

[15] D. A. Dalquen, A.M. Altenhoff, G.H. Gonnet, and C. Dessimoz. The impact of gene duplication, insertion, deletion, lateral gene transfer and sequencing error on orthology inference: A simulation study. *PLoS ONE*, 8(2):e56925, 02 2013.

[16] D. A. Dalquen, M. Anisimova, G. H. Gonnet, and C. Dessimoz. ALF–a simulation framework for genome evolution. *Mol. Biol. Evol.*, 29(4):1115–1123, Apr 2012.

[17] M. C. H. Dekker. Reconstruction methods for derivation trees. Master's thesis, Vrije Universiteit, Amsterdam, Netherlands, 1986.

[18] A.W.M. Dress, K.T. Huber, J. Koolen, V. Moulton, and A. Spillner. *Basic phylogenetic combinatorics*. Cambridge University Press, 2012.

[19] I. Ebersberger, S. Strauss, and A. von Haeseler. HaMStR: Profile hidden markov model based search for orthologs in ESTs. *BMC Evol. Biol.*, 9:157, 2009.

[20] W.M. Fitch. Homology: a personal view on some of the problems. *Trends Genet.*, 16:227–231, 2000.

[21] T. Gabaldón and EV. Koonin. Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.*, 14(5):360–366, 2013.

[22] L. Gasieniec, J. Jansson, A. Lingas, and A. Ostlin. On the complexity of constructing evolutionary trees. *J. Comb. Optim.*, 3:183–197, 1999.

[23] S. Grünewald, M. Steel, and M.S. Swenson. Closure operations in phylogenetics. *Mathematical Biosciences*, 208(2):521 – 537, 2007.

[24] S. Guillemot and M. Mnich. Kernel and fast algorithm for dense triplet inconsistency. *Theoretical Computer Science*, 494(0):134 – 143, 2013. Theory and Applications of Models of Computation (TAMC 2010).

[25] M. Habib and C. Paul. A simple linear time algorithm for cograph recognition. *Discrete Applied Mathematics*, 145(2):183–197, 2005.

[26] M. Hellmuth, M. Hernandez-Rosales, K.T. Huber, V. Moulton, P.F. Stadler, and N. Wieseke. Orthology relations, symbolic ultrametrics, and cographs. *Journal of Mathematical Biology*, 66(1-2):399–420, 2013.

[27] M. Hernandez-Rosales, M. Hellmuth, N. Wieseke, K.T. Huber, V. Moulton, and P.F. Stadler. From event-labeled gene trees to species trees. *BMC Bioinformatics*, 13(Suppl 19):S6, 2012.

[28] M. Hernandez-Rosales, M. Hellmuth, N. Wieseke, and P. F. Stadler. Simulation of gene family histories. *BMC Bioinformatics*, 15(Suppl 3):A8, 2014.

[29] K.T. Huber, V. Moulton, C. Semple, and M. Steel. Recovering a phylogenetic tree using pairwise closure operations. *Applied mathematics letters*, 18(3):361–366, 2005.

[30] J. Jansson. On the complexity of inferring rooted evolutionary trees. *Electronic Notes Discr. Math.*, 7:50–53, 2001.

[31] J. Jansson, R.S. Lemence, and A. Lingas. The complexity of inferring a minimally resolved phylogenetic supertree. *SIAM J. Comput.*, 41:272–291, 2012.

[32] J Jansson, J. H.-K. Ng, K. Sadakane, and W.-K. Sung. Rooted maximum agreement supertrees. *Algorithmica*, 43:293–307, 2005.

[33] P. J. Keeling and J. D. Palmer. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, 9:605–618, 2008.

[34] D.M. Kristensen, Y.I. Wolf, A.R. Mushegian, and E.V. Koonin. Computational methods for gene orthology inference. *Briefings in Bioinformatics*, 12(5):379–391, 2011.

[35] M. Lechner, S. Findeiß, L. Steiner, M. Marz, P.F. Stadler, and S.J. Prohaska. Proteinortho: Detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics*, 12:124, 2011.

[36] M. Lechner, A. Nickel, S. Wehner, K. Riege, N. Wieseke, B. Beckmann, R. Hartmann, and M. Marz. Genomewide comparison and novel ncrnas of aquificales. *BMC Genomics*, 15(1):522, 2014.

[37] Y. Liu, J. Wang, J. Guo, and J. Chen. Cograph editing: Complexity and parametrized algorithms. In B. Fu and D. Z. Du, editors, *COCOON 2011*, volume 6842 of *Lect. Notes Comp. Sci.*, pages 110–121, Berlin, Heidelberg, 2011. Springer-Verlag.

[38] Y. Liu, J. Wang, J. Guo, and J. Chen. Complexity and parameterized algorithms for cograph editing. *Theoretical Computer Science*, 461(0):45 – 54, 2012.

[39] K. Maemura, J. Jansson, H. Ono, K. Sadakane, and M. Yamashita. Approximation algorithms for constructing evolutionary trees from rooted triplets. In *Proceedings of 10th Korea-Japan Joint Workshop on Algorithms and Computation*, 2007. Gwangju, Korea.

[40] C. Semple and M. Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, UK, 2003.

[41] S.J. Tazehkand, S.N. Hashemi, and H. Poormohammadi. New heuristics for rooted triplet consistency. *Algorithms*, 6:396–406, 2013.

[42] L. van Iersel, S. Kelk, and M. Mnich. Uniqueness, intractability and exact algorithms: reflections on level-*k* phylogenetic networks. *J. Bioinf. Comp. Biol.*, 7:597–623, 2009.

[43] Wattam et al. Patric, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Research*, 2013.

[44] B.Y. Wu. Constructing the maximum consensus tree from rooted triples. *J. Comb. Optimization*, 8:29–39, 2004.

Phylogenomics with Paralogs: SUPPLEMENTAL MATERIAL

S 1 Theory

In this section we give an expanded and more technical account of the mathematical theory underlying the relationships between orthology relations, triple sets, and the reconciliation of gene and triple sets. In particular, we include here the proofs of the key novel results outline in the main text. The notation in the main text is a subset of the one used here. Theorems, remarks, and ILP formulations have the same numbers as in the main text. As a consequence, the numberings in this supplement may not always be in ascending order.

S 1.1 Notation

For an arbitrary set X we denote with $\binom{X}{n}$ the set of n-elementary subsets of X. In the remainder of this paper, L will always denote a finite set of size at least three. Furthermore, we will denote with $\mathfrak{G}$ a set of genes and with $\mathfrak{S}$ a set of species and assume that $|\mathfrak{G}| \geq 3$ and $|\mathfrak{S}| \geq 1$. Genes contained in $\mathfrak{G}$ are denoted by lowercase Roman letters $a, b, c, \ldots$ and species in $\mathfrak{S}$ by lower case Greek letters $\alpha, \beta, \gamma \ldots$. Furthermore, let $\sigma : \mathfrak{G} \to \mathfrak{S}$ with $x \mapsto \sigma(x)$ be a mapping that assigns to each gene $x \in \mathfrak{G}$ its corresponding species $\sigma(x) = \chi \in \mathfrak{S}$. With $\sigma(\mathfrak{G})$ we denote the image of σ. W.l.o.g. we can assume that the map σ is surjective, and thus, $\sigma(\mathfrak{G}) = \mathfrak{S}$. We assume that the reader is familiar with graphs and its terminology, and refer to [23] as a standard reference.

S 1.2 Phylogenetic Trees

A tree $T = (V, E)$ is a connected cycle-free graph with vertex set $V(T) = V$ and edge set $E(T) = E$. A vertex of T of degree one is called a *leaf* of T and all other vertices of T are called *inner* vertices. An edge of T is an *inner* edge if both of its end vertices are inner vertices. The sets of inner vertices of T is denoted by V^0. A tree T is called *binary* if each inner vertex has outdegree two. A *rooted tree* $T = (V, E)$ is a tree that contains a distinguished vertex $\rho_T \in V$ called the *root*.

A *phylogenetic tree* T *(on* L*)* is a rooted tree $T = (V, E)$ with leaf set $L \subseteq V$ such that no inner vertex has in- and outdegree one and whose root $\rho_T \in V$ has indegree zero. A phylogenetic tree on $\mathfrak{G}$, resp., on $\mathfrak{S}$, is called *gene tree*, resp., *species tree*.

Let $T = (V, E)$ be a phylogenetic tree on L with root ρ_T. The ancestor relation $\preceq_T$ on V is the partial order defined, for all $x, y \in V$, by $x \preceq_T y$ whenever y lies on the (unique) path from x to the root. Furthermore, we write $x \prec_T y$ if $x \preceq_T y$ and $x \neq y$. For a non-empty subset of leaves $L' \subseteq L$, we define $\mathrm{lca}_T(L')$, or the *most recent common ancestor of* L', to be the unique vertex in T that is the least upper bound of L' under the partial order $\preceq_T$. In case $L' = \{x, y\}$, we put $\mathrm{lca}_T(x, y) := \mathrm{lca}_T(\{x, y\})$ and if $L' = \{x, y, z\}$, we put $\mathrm{lca}_T(x, y, z) := \mathrm{lca}_T(\{x, y, z\})$. If there is no danger of ambiguity, we will write $\mathrm{lca}(L')$ rather then $\mathrm{lca}_T(L')$.

For $v \in V$, we denote with $L(v) := \{y \in L | y \preceq_T v\}$ the set of leaves in the subtree $T(v)$ of T rooted in v. Thus, $L(\rho_T) = L$ and $T(\rho_T) = T$.

It is well-known that there is a one-to-one correspondence between (isomorphism classes of) phylogenetic trees on L and so-called hierarchies on L. For a finite set L, a *hierarchy on* L is a subset $\mathcal{C}$ of the power set $\mathbb{P}(L)$ such that

(i) $L \in \mathcal{C}$

(ii) $\{x\} \in \mathcal{C}$ for all $x \in L$ and

(iii) $p \cap q \in \{p, q, \emptyset\}$ for all $p, q \in \mathcal{C}$.

The elements of $\mathcal{C}$ are called clusters.

Theorem 3 ([49]). *Let $\mathcal{C}$ be a collection of non-empty subsets of L. Then, there is a phylogenetic tree T on L with $\mathcal{C} = \{L(v) \mid v \in V(T)\}$ if and only if $\mathcal{C}$ is a hierarchy on L.*

The following result appears to be well known. We include a simple proof since we were unable to find a reference for it.

Lemma 1. *The number of clusters $|\mathcal{C}|$ in a hierarchy $\mathcal{C}$ on L determined by a phylogenetic tree $T = (V, E)$ on L is bounded by $2|L| - 1$.*

Proof. Clearly, the number of clusters $|\mathcal{C}|$ is determined by the number of vertices $|V|$, since each leaf $v \in L$, determines the singleton cluster $\{v\} \in \mathcal{C}$ and each inner node v has at least two children and thus, gives rise to a new cluster $L(v) \in \mathcal{C}$. Hence, $|\mathcal{C}| = |V|$.

First, consider a binary phylogenetic tree $T = (V, E)$ on $|L|$ leaves. Then there are $|V| - |L|$ inner vertices, all of out-degree two. Hence, $|E| = 2(|V| - |L|) = |V| - 1$ and thus $|V| = 2|L| - 1$. Hence, T determines $|\mathcal{C}| = 2|L| - 1$ clusters and has in particular $|L| - 1$ inner vertices.

Now, its easy to verify by induction on the number of leaves $|L|$ that an arbitrary phylogenetic tree $T' = (V', E')$ has $n_0 \leq |L| - 1$ inner vertices and thus, $|\mathcal{C}'| = |V'| = n_0 + |L| \leq 2|L| - 1$ clusters. □

S 1.3 Rooted Triples

S 1.3.1 Consistent Triple Sets

Rooted triples, sometimes also called rooted triplets [25], constitute an important concept in the context of supertree reconstruction [49, 5] and will also play a major role here. A rooted triple $r = (\mathsf{xy}|\mathsf{z})$ is *displayed* by a phylogenetic tree T on L if $x, y, z \in L$ pairwise distinct, and the path from x to y does not intersect the path from z to the root ρ_T and thus, having $\mathrm{lca}_T(x, y) \prec_T \mathrm{lca}_T(x, y, z)$. We denote with L_r the set of the three leaves $\{x, y, z\}$ contained in the triple $r = (\mathsf{xy}|\mathsf{z})$, and with $L_R := \cup_{r \in R} L_r$ the union of the leaf set of each $r \in R$. For a given leaf set L, a triple set R is said to be *(strict) dense* if for each $x, y, z \in L$ there is (exactly) one triple $r \in R$ with $L_r = \{x, y, z\}$. For a phylogenetic tree T, we denote by $\mathfrak{R}(T)$ the set of all triples that are displayed by T. A set R of triples is *consistent* if there is a phylogenetic tree T on L_R such that $R \subseteq \mathfrak{R}(T)$, i.e., T displays all triples $r \in R$.

Not all sets of triples are consistent, of course. Given a triple set R there is a polynomial-time algorithm, referred to in [49] as `BUILD`, that either constructs a phylogenetic tree T displaying R or recognizes that R is not consistent or *inconsistent* [1]. Various practical implementations have been described starting with [1], improved variants are discussed in [48, 38]. The problem of determining a maximum consistent subset of an inconsistent set of triples, however, is NP-hard and also APX-hard, see [14, 51] and the references therein. We refer to [15] for an overview on the available practical approaches and further theoretical results.

For a given consistent triple set R, a rooted phylogenetic tree that has as few inner vertices as possible and which is consistent with every rooted triplet in R is called a *least resolved* tree (for R). Finding a tree with a minimal number of inner nodes for a given consistent set of rooted triples is also an NP-hard problem, see [40].

S 1.3.2 Graph Representation of Triples

There is a quite useful representation of a set of triples R as a graph also known as *Aho graph*, see [1, 37, 12]. For given a triple set R and an arbitrary subset $\mathcal{L} \subseteq L_R$, the graph $[R, \mathcal{L}]$ has vertex set $\mathcal{L}$ and two vertices $x, y \in \mathcal{L}$ are linked by an edge, if there is a triple $(\mathsf{xy}|\mathsf{z}) \in R$ with $z \in \mathcal{L}$. Based on connectedness properties of the graph $[R, \mathcal{L}]$ for particular subsets $\mathcal{L} \subseteq L_R$, the algorithm `BUILD` recognizes if R is consistent or not. In particular, this algorithm makes use of the following well-known theorem.

Theorem 4 ([1, 12]). *A set of rooted triples R is consistent if and only if for each subset $\mathcal{L} \subseteq L_R$, $|\mathcal{L}| > 1$ the graph $[R, \mathcal{L}]$ is disconnected.*

Lemma 2 ([37]). *Let R be a dense set of rooted triples on L. Then for each $\mathcal{L} \subseteq L$, the number of connected components of the Aho graph $[R, \mathcal{L}]$ is at most two.*

Lemma 2 implies that the tree computed with `BUILD` based on the Aho graph for a consistent dense set of rooted triples must be binary. We will use the Aho graph and its key properties as a frequent tool in upcoming proofs.

For later reference, we recall

Lemma 3 ([12]). *If R' is a subset of the triple set R and L is a leaf set, then $[R', L]$ is a subgraph of $[R, L]$.*

S 1.3.3 Closure Operations and Inference Rules

The requirement that a set R of triples is consistent, and thus, that there is a tree displaying all triples, allows to infer new triples from the set of all trees displaying all triples of R and to define a *closure operation* for R, which has been extensively studied in the last decades, see [12, 28, 11, 36, 6]. Let $\langle R \rangle$ be the set of all phylogenetic trees on L_R that display all the triples of R. The closure of a consistent set of rooted triples R is defined as

$$\mathrm{cl}(R) = \bigcap_{T \in \langle R \rangle} \mathfrak{R}(T).$$

This operation satisfies the usual three properties of a closure operator, namely: $R \subseteq \mathrm{cl}(R)$; $\mathrm{cl}(\mathrm{cl}(R)) = \mathrm{cl}(R)$ and if $R' \subseteq R$, then $\mathrm{cl}(R') \subseteq \mathrm{cl}(R)$. We say R is *closed* if $R = \mathrm{cl}(R)$. Clearly, for any tree T it holds that $\mathfrak{R}(T)$ is closed. The brute force computation of the closure of a given consistent set R runs in $O(|R|^5)$ time [12]: For any three leaves $x, y, z \in L_R$ test whether exactly one of the sets $R \cup \{(\mathsf{xy}|\mathsf{z})\}$, $R \cup \{(\mathsf{xz}|\mathsf{y})\}$, $R \cup \{(\mathsf{zy}|\mathsf{x})\}$ is consistent, and if so, add the respective triple to the closure $\mathrm{cl}(R)$ of R.

For a consistent set R of rooted triples we write $R \vdash (\mathsf{xy}|\mathsf{z})$ if any phylogenetic tree that displays all triples of R also displays $(\mathsf{xy}|\mathsf{z})$. In other words, $R \vdash (\mathsf{xy}|\mathsf{z})$ iff $(\mathsf{xy}|\mathsf{z}) \in \mathrm{cl}(R)$. In a work of Bryant and Steel [12], in which the authors extend and generalize the work of Dekker [22], it was shown under which conditions it is possible to infer triples by using only subsets $R' \subseteq R$, i.e., under which conditions $R \vdash (\mathsf{xy}|\mathsf{z}) \implies R' \vdash (\mathsf{xy}|\mathsf{z})$ for some $R' \subseteq R$. In particular, we will make frequent use of the following inference rules:

$$\{(\mathsf{ab}|\mathsf{c}), (\mathsf{ad}|\mathsf{c})\} \vdash (\mathsf{bd}|\mathsf{c}) \tag{i}$$

$$\{(\mathsf{ab}|\mathsf{c}), (\mathsf{ad}|\mathsf{b})\} \vdash (\mathsf{bd}|\mathsf{c}), (\mathsf{ad}|\mathsf{c}) \tag{ii}$$

$$\{(\mathsf{ab}|\mathsf{c}), (\mathsf{cd}|\mathsf{b})\} \vdash (\mathsf{ab}|\mathsf{d}), (\mathsf{cd}|\mathsf{a}). \tag{iii}$$

Remark 3. *It is an easy task to verify, that such inference rules based on two triples $r_1, r_2 \in R$ can lead only to new triples, whenever $|L_{r_1} \cap L_{r_2}| = 2$. Hence, the latter three stated rules are the only ones that lead to new triples for a given pair of triples in a strict dense triple set.*

For later reference and the ILP formulation, we give the following lemma.

Lemma 4. *Let R be a strict dense set of rooted triples. For all $L' = \{a, b, c, d\} \subseteq L_R$ we have the following statements:*

All triples inferred by rule (ii) *applied on triples $r \in R$ with $L_r \subset L'$ are contained in R if and only if all triples inferred by rule* (iii) *applied on triples $r \in R$ with $L_r \subset L'$ are contained in R.*

Moreover, if all triples inferred by rule (ii) *applied on triples $r \in R$ with $L_r \subset L'$ are contained in R then all triples inferred by rule* (i) *applied on triples $r \in R$ with $L_r \subset L'$ are contained in R.*

Proof. The first statement was established in [30, Lemma 2].

For the second statement assume that for all pairwise distinct $L' = \{a, b, c, d\} \subseteq L_R$ it holds that all triples inferred by rule (ii), or equivalently, by rule (iii) applied on triples $r \in R$ with $L_r \subset L'$ are contained in R. Assume for contradiction that there are triples $(\mathsf{ab}|\mathsf{c}), (\mathsf{ad}|\mathsf{c}) \in R$, but $(\mathsf{bd}|\mathsf{c}) \notin R$. Since R is strict dense, we have either $(\mathsf{bc}|\mathsf{d}) \in R$ or $(\mathsf{cd}|\mathsf{b}) \in R$. In the first case and since $(\mathsf{ab}|\mathsf{c}) \in R$, rule (ii) implies that $(\mathsf{ac}|\mathsf{d}) \in R$, a contradiction. In the second case and since $(\mathsf{ab}|\mathsf{c}) \in R$, rule (iii) implies that $(\mathsf{cd}|\mathsf{a}) \in R$, a contradiction. □

We are now in the position to prove the following important and helpful lemmas and theorem. The final theorem basically states that consistent strict dense triple sets can be characterized by the closure of any two element subset of R. Note, an analogous result was established by [30]. However, we give here an additional direct and transparent proof.

Lemma 5. *Let R be a strict dense set of triples on L such that for all $R' \subseteq R$ with $|R'| = 2$ it holds $\mathrm{cl}(R') \subseteq R$. Let $x \in L$ and $L' = L \setminus \{x\}$. Moreover, let $R_{|L'} \subset R$ denote the subset of all triples $r \in R$ with $L_r \subseteq L'$. Then $R_{|L'}$ is strict dense and for all $R' \subseteq R_{|L'}$ with $|R'| = 2$ it holds $\mathrm{cl}(R') \subseteq R_{|L'}$.*

Proof. Clearly, since R is strict dense and since $R_{|L'}$ contains all triples except the ones containing x it still holds that for all $a, b, c \in L'$ there is exactly one triple $r \in R_{|L'}$ with $a, b, c \in L_r$. Hence, $R_{|L'}$ is strict dense.

Assume for contradiction, that there are triples $r_1, r_2 \in R_{|L'} \subset R$ with $\mathrm{cl}(r_1, r_2) \not\subseteq R_{|L'}$. By construction of $R_{|L'}$, no triples $r_1, r_2 \in R_{|L'}$ can infer a new triple r_3 with $x \in L_{r_3}$. This immediately implies that $\mathrm{cl}(r_1, r_2) \not\subseteq R$, a contradiction. □

Lemma 6. *Let R be a strict dense set of triples on L with $|L| = 4$. If for all $R' \subseteq R$ with $|R'| = 2$ holds $\mathrm{cl}(R') \subseteq R$ then R is consistent.*

Proof. By contraposition, assume that R is not consistent. Thus, the Aho graph $[R, \mathcal{L}]$ is connected for some $\mathcal{L} \subseteq L$. Since R is strict dense, for any $\mathcal{L} \subseteq L$ with $|\mathcal{L}| = 2$ or $|\mathcal{L}| = 3$ the Aho graph $[R, \mathcal{L}]$ is always disconnected. Hence, $[R, \mathcal{L}]$ for $\mathcal{L} = L$ must be connected. The graph $[R, L]$ has four vertices, say a, b, c and d. The fact that R is strict dense and $|L| = 4$ implies that $|R| = 4$ and in particular, that $[R, L]$ has three or four edges. Hence, the graph $[R, L]$ is isomorphic to one of the following graphs G_0, G_1 or G_2.

The graph G_0 is isomorphic to a path $x_1 - x_2 - x_3 - x_4$ on four vertices; G_1 is isomorphic to a chordless square; and G_2 is isomorphic to a path $x_1 - x_2 - x_3 - x_4$ on four vertices where the edge $\{x_1, x_3\}$ or $\{x_2, x_4\}$ is added. W.l.o.g. assume that for the first case $[R, L] \simeq G_0$ has edges $\{a, b\}$, $\{b, c\}$, $\{c, d\}$; for the second case $[R, L] \simeq G_1$ has edges $\{a, b\}$, $\{a, c\}$, $\{c, d\}$ and $\{b, d\}$ and for the third case assume that $[R, L] \simeq G_2$ has edges $\{a, b\}$, $\{a, c\}$, $\{c, d\}$ and $\{a, d\}$.

Let $[R, L] \simeq G_0$. Then there are triples of the form $(\mathsf{ab}|*)$, $(\mathsf{bc}|*)$, $(\mathsf{cd}|*)$, where one kind of triple must occur twice, since otherwise, $[R, L]$ would have four edges. Assume that this is $(\mathsf{ab}|*)$. Hence, the triples $(\mathsf{ab}|\mathsf{c}), (\mathsf{ab}|\mathsf{d}) \in R$ since $|R| = 4$. Since R is strict dense, $(\mathsf{bc}|*) = (\mathsf{bc}|\mathsf{d}) \in R$, which implies that $(\mathsf{cd}|*) = (\mathsf{cd}|\mathsf{a}) \in R$. Now, $R' = \{(\mathsf{ab}|\mathsf{c}), (\mathsf{bc}|\mathsf{d})\} \vdash (\mathsf{ac}|\mathsf{d})$. However, since R is strict dense and $(\mathsf{cd}|\mathsf{a}) \in R$ we can conclude that $(\mathsf{ac}|\mathsf{d}) \notin R$, and therefore $\mathrm{cl}(R') \not\subseteq R$. The case with triples $(\mathsf{cd}|*)$ occurring twice is treated analogously. If triples $(\mathsf{bc}|*)$ occur twice, we can argue the same way to obtain obtain $(\mathsf{bc}|\mathsf{a}), (\mathsf{bc}|\mathsf{d}) \in R$, $(\mathsf{ab}|*) = (\mathsf{ab}|\mathsf{d})$, and $(\mathsf{cd}|*) = (\mathsf{cd}|\mathsf{a})$. However, $R' = \{(\mathsf{bc}|\mathsf{a}), (\mathsf{cd}|\mathsf{a})\} \vdash (\mathsf{bd}|\mathsf{a}) \notin R$, and thus $\mathrm{cl}(R') \not\subseteq R$.

Let $[R, L] \simeq G_1$. Then there must be triples of the form $(\mathsf{ab}|*)$, $(\mathsf{ac}|*)$, $(\mathsf{cd}|*)$, $(\mathsf{bd}|*)$. Clearly, $(\mathsf{ab}|*) \in \{(\mathsf{ab}|\mathsf{c}), (\mathsf{ab}|\mathsf{d})\}$. Note that not both $(\mathsf{ab}|\mathsf{c})$ and $(\mathsf{ab}|\mathsf{d})$ can be contained in R, since then $[R, L] \simeq G_0$. If $(\mathsf{ab}|*) = (\mathsf{ab}|\mathsf{c})$ and since R is strict dense, $(\mathsf{ac}|*) = (\mathsf{ac}|\mathsf{d})$. Again, since R is strict dense, $(\mathsf{cd}|*) = (\mathsf{cd}|\mathsf{b})$ and this implies that $(\mathsf{bd}|*) = (\mathsf{bd}|\mathsf{a})$. However, $R' = \{(\mathsf{ab}|\mathsf{c}), (\mathsf{ac}|\mathsf{d})\} \vdash (\mathsf{ab}|\mathsf{d}) \notin R$, since R is strict dense and $(\mathsf{bd}|\mathsf{a}) \in R$. Thus, $\mathrm{cl}(R') \not\subseteq R$. If $(\mathsf{ab}|*) = (\mathsf{ab}|\mathsf{d})$ and since R is strict dense, we can argue analogously, and obtain, $(\mathsf{bd}|*) = (\mathsf{bd}|\mathsf{c})$, $(\mathsf{cd}|*) = (\mathsf{cd}|\mathsf{a})$ and $(\mathsf{ac}|*) = (\mathsf{ac}|\mathsf{b})$. However, $R' = \{(\mathsf{ab}|\mathsf{d}), (\mathsf{bd}|\mathsf{c})\} \vdash (\mathsf{ad}|\mathsf{c}) \notin R$, and thus $\mathrm{cl}(R') \not\subseteq R$.

Let $[R, L] \simeq G_2$. Then there must be triples of the form $(\mathsf{ab}|*)$, $(\mathsf{ac}|*)$, $(\mathsf{cd}|*)$, $(\mathsf{ad}|*)$. Again, $(\mathsf{ab}|*) \in \{(\mathsf{ab}|\mathsf{c}), (\mathsf{ab}|\mathsf{d})\}$. By similar arguments as in the latter two cases, if $(\mathsf{ab}|*) = (\mathsf{ab}|\mathsf{c})$ then we obtain, $(\mathsf{ac}|*) = (\mathsf{ac}|\mathsf{d})$, $(\mathsf{ad}|*) = (\mathsf{ad}|\mathsf{b})$ and $(\mathsf{cd}|*) = (\mathsf{cd}|\mathsf{b})$. Since $R' = \{(\mathsf{ab}|\mathsf{c}), (\mathsf{ac}|\mathsf{d})\} \vdash (\mathsf{bc}|\mathsf{d}) \notin R$, we can conclude that $\mathrm{cl}(R') \not\subseteq R$. If $(\mathsf{ab}|*) = (\mathsf{ab}|\mathsf{d})$ we obtain analogously, $(\mathsf{ad}|*) = (\mathsf{ad}|\mathsf{c})$, $(\mathsf{cd}|*) = (\mathsf{cd}|\mathsf{b})$ and $(\mathsf{ac}|*) = (\mathsf{ac}|\mathsf{b})$. However, $R' = \{(\mathsf{ab}|\mathsf{d}), (\mathsf{ad}|\mathsf{c})\} \vdash (\mathsf{bd}|\mathsf{c}) \notin R$, and thus $\mathrm{cl}(R') \not\subseteq R$. □

Theorem 1. *Let R be a strict dense triple set on L with $|L| \geq 3$. The set R is consistent if and only if $\mathrm{cl}(R') \subseteq R$ holds for all $R' \subseteq R$ with $|R'| = 2$.*

Proof. $\Rightarrow$: If R is strict dense and consistent, then for any triple triple $(\mathsf{ab}|\mathsf{c}) \notin R$ holds $R \cup (\mathsf{ab}|\mathsf{c})$ is inconsistent as either $(\mathsf{ac}|\mathsf{b})$ or $(\mathsf{bc}|\mathsf{a})$ is already contained in R. Hence, for each $a, b, c \in L$ exactly one $R \cup \{(\mathsf{ab}|\mathsf{c})\}$, $R \cup \{(\mathsf{ac}|\mathsf{b})\}$, $R \cup \{(\mathsf{bc}|\mathsf{a})\}$ is consistent, and this triple is already contained in R. Hence, R is closed. Therefore, for any subset $R' \subseteq R$ holds $\mathrm{cl}(R') \subseteq \mathrm{cl}(R) = R$. In particular, this holds for all $R' \subseteq R$ with $|R'| = 2$.

$\Leftarrow$: *(Induction on $|L|$.)*
If $|L| = 3$ and since R is strict dense, it holds $|R| = 1$ and thus, R is always consistent. If $|L| = 4$, then Lemma 6 implies that if for any two-element subset $R' \subseteq R$ holds that $\mathrm{cl}(R') \subseteq R$, then R is consistent. Assume therefore, the assumption is true for all strict dense triple sets R on L with $|L| = n$.

Let R be a strict dense triple set on L with $|L| = n + 1$ such that for each $R' \subseteq R$ with $|R'| = 2$ it holds $\mathrm{cl}(R') \subseteq R$. Moreover, let $L' = L \setminus \{x\}$ for some $x \in L$ and $R_{|L'} \subset R$ denote the subset of all triples $r \in R$ with $L_r \subset L'$. Lemma 5 implies that $R_{|L'}$ is strict dense and for each $R' \subseteq R_{|L'}$ with $|R'| = 2$ we have $\mathrm{cl}(R') \subseteq R_{|L'}$. Hence, the induction hypothesis can be applied for any such $R_{|L'}$ implying that $R_{|L'}$ is consistent. Moreover, since $R_{|L'}$ is strict dense and consistent, for any triple $(\mathsf{xy}|\mathsf{z}) \notin R_{|L'}$ holds that $R_{|L'} \cup (\mathsf{xy}|\mathsf{z})$ is inconsistent. But this implies that $R_{|L'}$ is closed, i.e., $\mathrm{cl}(R_{|L'}) = R_{|L'}$. Lemma 2 implies that the Aho graph $[R_{|L'}, \mathcal{L}]$ has exactly two connected components C_1 and C_2 for each $\mathcal{L} \subseteq L'$ with $|\mathcal{L}| > 1$. In the following we denote with $\mathcal{L}_i = V(C_i)$, $i = 1, 2$ the set of vertices of the connected component C_i in $[R_{|L'}, \mathcal{L}]$. Clearly, $\mathcal{L} = \mathcal{L}_1 \dot\cup \mathcal{L}_2$. It is easy to see that $[R, \mathcal{L}] \simeq [R_{|L'}, \mathcal{L}]$ for any $\mathcal{L} \subseteq L'$, since none of the graphs contain vertex x. Hence, $[R, \mathcal{L}]$ is always disconnected for any $\mathcal{L} \subseteq L'$. Therefore, it remains to show that, for all $\mathcal{L} \cup \{x\}$ with $\mathcal{L} \subseteq L'$ holds: if for any $R' \subseteq R$ with $|R'| = 2$ holds $\mathrm{cl}(R') \subseteq R$, then $[R, \mathcal{L} \cup \{x\}]$ is disconnected and hence, R is consistent.

To proof this statement we consider the different possibilities for $\mathcal{L}$ separately. We will frequently use that $[R_{|L'}, \mathcal{L}]$ is a subgraph of $[R, \mathcal{L}]$ for every $\mathcal{L} \subseteq L$ (Lemma 3).

Case 1. If $|\mathcal{L}| = 1$, then $\mathcal{L} \cup \{x\}$ implies that $[R, \mathcal{L} \cup \{x\}]$ has exactly two vertices and clearly, no edge. Thus, $[R, \mathcal{L} \cup \{x\}]$ is disconnected.

Case 2. Let $|\mathcal{L}| = 2$ with $\mathcal{L}_1 = \{a\}$ and $\mathcal{L}_2 = \{b\}$. Since R is strict dense, exactly one of the triples $(\mathsf{ab}|\mathsf{x})$, $(\mathsf{ax}|\mathsf{b})$, or $(\mathsf{xb}|\mathsf{a})$ is contained in R. Hence, $[R, \mathcal{L} \cup \{x\}]$ has exactly three vertices where two of them are linked by an edge. Thus, $[R, \mathcal{L} \cup \{x\}]$ is disconnected.

Case 3. Let $|\mathcal{L}| \geq 3$ with $\mathcal{L}_1 = \{a_1, \ldots, a_n\}$ and $\mathcal{L}_2 = \{b_1, \ldots, b_m\}$. Since $R_{|L'}$ is consistent and strict dense and by construction of $\mathcal{L}_1$ and $\mathcal{L}_2$ it holds $\forall a_i, a_j \in \mathcal{L}_1, b_k \in \mathcal{L}_2, i \neq j : (\mathsf{a_i a_j}|\mathsf{b_k}) \in R_{|L'} \subseteq R$ and $\forall a_i \in \mathcal{L}_1, b_k, b_l \in \mathcal{L}_2, k \neq l : (\mathsf{b_k b_l}|\mathsf{a_i}) \in R_{|L'} \subseteq R$. Therefore, since R is strict dense, there cannot be any triple of the form $(\mathsf{a_i b_k}|\mathsf{a_j})$ or $(\mathsf{a_i b_k}|\mathsf{b_l})$ with $a_i, a_j \in \mathcal{L}_1, b_k, b_l \in \mathcal{L}_2$ that is contained R. It remains to show that R is consistent. The following three subcases can occur.

3.a) The connected components C_1 and C_2 of $[R_{|L'}, \mathcal{L}]$ are connected in $[R, \mathcal{L} \cup \{x\}]$. Hence, there must be a triple $(\mathsf{ab}|\mathsf{x}) \in R$ with $a \in \mathcal{L}_1$ and $b \in \mathcal{L}_2$. Hence, in order to prove that R is consistent, we need to show that there is no triple $(\mathsf{cx}|\mathsf{d})$ contained R for all $c, d \in \mathcal{L}$, which would imply that $[R, \mathcal{L} \cup \{x\}]$ stays disconnected.

3.b) The connected component C_1 of $[R_{|L'}, \mathcal{L}]$ is connected to x in $[R, \mathcal{L} \cup \{x\}]$. Hence, there must be a triple $(\mathsf{ax}|\mathsf{c}) \in R$ with $a \in \mathcal{L}_1$, $c \in \mathcal{L}$. Hence, in order to prove that R is consistent, we need to show that there are no triples $(\mathsf{b_k x}|\mathsf{a_i})$ and $(\mathsf{b_k x}|\mathsf{b_l})$ for all $a_i \in \mathcal{L}_1$, $b_k, b_l \in \mathcal{L}_2$, which would imply that $[R, \mathcal{L} \cup \{x\}]$ stays disconnected.

3.c) As in Case 3.*b*), the connected component C_2 of $[R_{|L'}, \mathcal{L}]$ might be connected to x in $[R, \mathcal{L} \cup \{x\}]$ and we need to show that there are no triples $(\mathsf{a_i x}|\mathsf{b_k})$ and $(\mathsf{a_i x}|\mathsf{a_j})(\mathsf{a_i x}|\mathsf{a_j})$ for all $a_i, a_j \in \mathcal{L}_1$, $b_k \in \mathcal{L}_2$ in order to prove that R is consistent.

Case 3.a) Let $(\mathsf{ab}|\mathsf{x}) \in R$, $a \in \mathcal{L}_1$, $b \in \mathcal{L}_2$. First we show that for all $a_i \in \mathcal{L}_1$ holds $(\mathsf{a_i b}|\mathsf{x}) \in R$. Clearly, if $\mathcal{L}_1 = \{a\}$ the statement is trivially true. If $|\mathcal{L}_1| > 1$ then $\{(\mathsf{ab}|\mathsf{x}), (\mathsf{a_i a}|\mathsf{b})\} \vdash (\mathsf{a_i b}|\mathsf{x})$ for all $a_i \in \mathcal{L}_1$. Since the closure of all two element subsets of R is contained in R and $(\mathsf{ab}|\mathsf{x}), (\mathsf{a_i a}|\mathsf{b}) \in R$ we can conclude that $(\mathsf{a_i b}|\mathsf{x}) \in R$. Analogously one shows that for all $b_k \in \mathcal{L}_2$ holds $(\mathsf{ab_k}|\mathsf{x}) \in R$.
Since $\{(\mathsf{a_i a}|\mathsf{b_k}), (\mathsf{ab_k}|\mathsf{x})\} \vdash (\mathsf{a_i b_k}|\mathsf{x})$ and $(\mathsf{a_i a}|\mathsf{b_k}), (\mathsf{ab_k}|\mathsf{x}) \in R$ we can conclude that $(\mathsf{a_i b_k}|\mathsf{x}) \in R$ for all $a_i \in \mathcal{L}_1$, $b_k \in \mathcal{L}_2$. Furthermore, $\{(\mathsf{a_i a_j}|\mathsf{b}), (\mathsf{a_i b}|\mathsf{x})\} \vdash (\mathsf{a_i a_j}|\mathsf{x})$ for all $a_i, a_j \in \mathcal{L}_1$ and again, $(\mathsf{a_i a_j}|\mathsf{x}) \in R$ for all $a_i, a_j \in \mathcal{L}_1$. Analogously, one shows that $(\mathsf{b_k b_l}|\mathsf{x}) \in R$ for all $b_k, b_l \in \mathcal{L}_2$.
Thus, we have shown, that for all $c, d \in \mathcal{L}$ holds that $(\mathsf{cd}|\mathsf{x}) \in R$. Since R is strict dense, there is no triple $(\mathsf{cx}|\mathsf{d})$ contained in R for any $c, d \in \mathcal{L}$. Hence, $[R, \mathcal{L} \cup \{x\}]$ is disconnected.

Case 3.b) Let $(\mathsf{ax}|\mathsf{c}) \in R$ with $a \in \mathcal{L}_1$, $c \in \mathcal{L}$. Assume first that $c \in \mathcal{L}_1$. Then there is triple $(\mathsf{ac}|\mathsf{b}) \in R$. Moreover, $\{(\mathsf{ax}|\mathsf{c}), (\mathsf{ac}|\mathsf{b})\} \vdash (\mathsf{ax}|\mathsf{b})$ and thus, $(\mathsf{ax}|\mathsf{b}) \in R$. This implies that there is always some $c' = b \in \mathcal{L}_2$ with $(\mathsf{ax}|\mathsf{c'}) \in R$. In other words, w.l.o.g. we can assume that for $(\mathsf{ax}|\mathsf{c}) \in R$, $a \in \mathcal{L}_1$ holds $c \in \mathcal{L}_2$.
Since $\{(\mathsf{ax}|\mathsf{b}), (\mathsf{aa_i}|\mathsf{b})\} \vdash (\mathsf{a_i x}|\mathsf{b})$ and $(\mathsf{ax}|\mathsf{b}), (\mathsf{aa_i}|\mathsf{b}) \in R$ we can conclude that $(\mathsf{a_i x}|\mathsf{b}) \in R$ for all $a_i \in \mathcal{L}_1$. Moreover, $\{(\mathsf{a_i x}|\mathsf{b}), (\mathsf{bb_k}|\mathsf{a_i})\} \vdash (\mathsf{a_i x}|\mathsf{b_k})$ and by similar arguments, $(\mathsf{a_i x}|\mathsf{b_k}) \in R$ for all $a_i \in \mathcal{L}_1, b_k \in \mathcal{L}_2$. Finally, $\{(\mathsf{a_i x}|\mathsf{b_k}), (\mathsf{b_l b_k}|\mathsf{a_i})\} \vdash (\mathsf{b_k b_l}|\mathsf{x})$, and therefore, $(\mathsf{b_k b_l}|\mathsf{x}) \in R$ for all $b_k, b_l \in \mathcal{L}_2$. To summarize, for all $a_i \in \mathcal{L}_1, b_k, b_l \in \mathcal{L}_2$ we have $(\mathsf{a_i x}|\mathsf{b_k}) \in R$ and $(\mathsf{b_k b_l}|\mathsf{x}) \in R$. Since R is strict dense there cannot be triples $(\mathsf{b_k x}|\mathsf{a_i})$ and $(\mathsf{b_k x}|\mathsf{b_l})$ for any $a_i \in \mathcal{L}_1$, $b_k, b_l \in \mathcal{L}_2$, and hence, $[R, \mathcal{L} \cup \{x\}]$ is disconnected.

Case 3.c) By similar arguments as in Case 3.*b*) and interchanging the role of $\mathcal{L}_1$ and $\mathcal{L}_2$, one shows that $[R, \mathcal{L} \cup \{x\}]$ is disconnected.

In summary, we have shown that $[R, \mathcal{L} \cup \{x\}]$ is disconnected in all cases. Therefore R is consistent. □

Theorem 2. *Let R be a consistent triple set on L. If the tree obtained with* `BUILD` *is binary, then the closure* $\mathrm{cl}(R)$ *is strict dense. Moreover, this tree T is unique and therefore, a least resolved tree for R.*

Proof. Note, the algorithm `BUILD` relies on the Aho graph $[R, \mathcal{L}]$ for particular subsets $\mathcal{L} \subseteq L$. This means, that if the tree obtained with `BUILD` is binary, then for each of the particular subsets $\mathcal{L} \subseteq L$ the Aho graph $[R, \mathcal{L}]$ must have exactly two components. Moreover, R is consistent, since `BUILD` constructs a tree.

Now consider arbitrary three distinct leaves $x, y, z \in L$. Since T is binary, there is a subset $\mathcal{L} \subseteq L$ with $x, y, z \in \mathcal{L}$ in some stage of `BUILD` such that two of the three leaves, say x and y are in a different connected component than the leaf z. This implies that $R \cup (\mathsf{xy}|\mathsf{z})$ is consistent, since even if $\{x, y\} \notin E([R, \mathcal{L}])$, the vertices x and y remain in the same connected component different from the one containing z when adding the edge $\{x, y\}$ to $[R, \mathcal{L}]$. Moreover, by the latter argument, both $R \cup (\mathsf{xz}|\mathsf{y})$ and $R \cup (\mathsf{yz}|\mathsf{x})$ are not consistent. Thus, for any three distinct leaves $x, y, z \in L$ exactly one of the sets $R \cup \{(\mathsf{xy}|\mathsf{z})\}$, $R \cup \{(\mathsf{xz}|\mathsf{y})\}$, $R \cup \{(\mathsf{zy}|\mathsf{x})\}$ is consistent, and thus, contained in the closure $\mathrm{cl}(R)$. Hence, $\mathrm{cl}(R)$ is strict dense.

Since a tree T that displays R also displays $\mathrm{cl}(R)$ and because $\mathrm{cl}(R)$ is strict dense and consistent, we can conclude that $\mathrm{cl}(R) = \mathfrak{R}(T)$ whenever T displays R. Hence, T must be unique and therefore, the least resolved tree for R. □

Lemma 7. *Let R be a consistent set of triples on L. Then there is a strict dense consistent triple set R' on L that contains R.*

Proof. Let $\mathrm{Aho}(R)$ be the tree constructed by `BUILD` from a consistent triple set R. It is in general not a binary tree. Let T' be a binary tree obtained from $\mathrm{Aho}(R)$ by substituting a binary tree with k leaves for every internal vertex with $k > 2$ children. Any triple $(ab|c) \in \mathfrak{R}(\mathrm{Aho}(R))$ is also displayed by T' since unique disjoint paths $a - b$ and $c - \rho$ in $\mathrm{Aho}(R)$ translate directly to unique paths in T', which obviously are again disjoint. Furthermore, a binary tree T' with leaf set L displays exactly one triple for each $\{a, b, c\} \in \binom{L}{3}$; hence R' is strict dense. □

Remark 4. *Let T be a binary tree. Then $\mathfrak{R}(T)$ is strict dense and hence, $\mathfrak{R}(T) \cup \{r\}$ is inconsistent for any triple $r \notin \mathfrak{R}(T)$. Since $\mathfrak{R}(T) \subseteq \mathfrak{R}(\mathrm{Aho}(\mathfrak{R}(T))$ by definition of the action of `BUILD` and there is no consistent triple set that strictly contains $\mathfrak{R}(T)$, we have $\mathfrak{R}(T) = \mathfrak{R}(\mathrm{Aho}(\mathfrak{R}(T))$. Thus $\mathrm{Aho}(\mathfrak{R}(T)) = T$.*

S 1.4 Orthology Relations, Symbolic Representations, and Cographs

For a gene tree $T = (V, E)$ on $\mathfrak{G}$ we define $t : V^0 \to M$ as a map that assigns to each inner vertex an arbitrary symbol $m \in M$. Such a map t is called a *symbolic dating map* or *event-labeling* for T; it is *discriminating* if $t(u) \neq t(v)$, for all inner edges $\{u, v\}$, see [7].

In the rest of this paper we are interested only in event-labelings t that map inner vertices into the set $M = \{\bullet, \square\}$, where the symbol "$\bullet$" denotes a speciation event and "$\square$" a duplication event. We denote with (T, t) a gene tree T with corresponding event labeling t. If in addition the map σ is given, we write this as $(T, t; \sigma)$.

An orthology relation $\Theta \subset \mathfrak{G} \times \mathfrak{G}$ is a symmetric relation that contains all pairs (x, y) of orthologous genes. Note, this implies that $(x, x) \notin \Theta$ for all $x \in \mathfrak{G}$. Hence, its complement $\overline{\Theta}$ contains all leaf pairs (x, x) and pairs (x, y) of non-orthologous genes and thus, in this context all paralogous genes.

For a given orthology relation Θ we want to find an event-labeled phylogenetic tree T on $\mathfrak{G}$, with $t : V^0 \to \{\bullet, \square\}$ such that

1. $t(\mathrm{lca}_T(x, y)) = \bullet$ for all $(x, y) \in \Theta$
2. $t(\mathrm{lca}_T(x, y)) = \square$ for all $(x, y) \in \overline{\Theta} \setminus \{(x, x) \mid x \in \mathfrak{G}\}$.

In other words, we want to find an event-labeled tree T on $\mathfrak{G}$ such that the event on the most recent common ancestor of the orthologous genes is a speciation event and of paralogous genes a duplication event. If such a tree T with (discriminating) event-labeling t exists for Θ, we call the pair (T, t) a *(discriminating) symbolic representation* of Θ.

S 1.4.1 Symbolic Representations and Cographs

Empirical orthology estimations will in general contain false-positives. In addition orthologous pairs of genes may have been missed due to the scoring function and the selected threshold. Hence, not for all estimated orthology relations there is such a tree. In order to characterize orthology relations we define for an arbitrary symmetric relation $R \subseteq \mathfrak{G} \times \mathfrak{G}$ the underlying graph $G_R = (\mathfrak{G}, E_R)$ with edge set $E_R = \left\{\{x, y\} \in \binom{\mathfrak{G}}{2} \mid (x, y) \in R\right\}$.

As we shall see, orthology relations Θ and cographs are closely related. A cograph is a P_4-free graph (i.e. a graph such that no four vertices induce a subgraph that is a path on 4 vertices), although there are a number of equivalent characterizations of such graphs (see e.g. [10] for a survey).

It is well-known in the literature concerning cographs that, to any cograph $G = (V, E)$, one can associate a canonical *cotree* $\mathrm{CoT}(G) = (W \cup V, F)$ with leaf set V together with a labeling map $\lambda_G : W \to \{0, 1\}$ defined on the inner vertices of $\mathrm{CoT}(G)$. The key observation is that, given a cograph $G = (V, E)$, a pair $\{x, y\} \in \binom{V}{2}$ is an edge in G if and only if $\lambda_G(\mathrm{lca}_{\mathrm{CoT}(G)}(x, y)) = 1$ (cf. [18, p. 166]). The next theorem summarizes the results, that rely on the theory of so-called symbolic ultrametrics developed in [7] and have been established in a more general context in [33].

Theorem 5 ([33])**.** *Suppose that Θ is an (estimated) orthology relation and denote by $\overline{\Theta}^{\neq} := \overline{\Theta} \setminus \{(x, x) \mid x \in \mathfrak{G}\}$ the complement of Θ without pairs (x, x). Then the following statements are equivalent:*

(i) Θ has a symbolic representation.

(ii) Θ has a discriminating symbolic representation.

(iii) $G_\Theta = \overline{G}_{\overline{\Theta}^{\neq}}$ is a cograph.

This result enables us to find the corresponding discriminating symbolic representation (T, t) for Θ (if one exists) by identifying T with the respective cotree $\mathrm{CoT}(G_\Theta)$ of the cograph G_Θ and setting $t(v) = \bullet$ if $\{x, y\} \in E(G_\Theta)$ and thus, $\lambda_{G_\Theta}(v) = 1$ and $t(v) = \square$ if $\{x, y\} \notin E(G_\Theta)$ and thus $\lambda_{G_\Theta}(v) = 0$

We identify the discriminating symbolic representation (T, t) for Θ (if one exists) with the cotree $\mathrm{CoT}(G_\Theta)$ as explained above.

S 1.4.2 Cograph Editing

It is well-known that cographs can be recognized in linear time [19, 31]. However, the cograph editing problem, that is given a graph $G = (V, E)$ one aims to convert G into a cograph $G^* = (V, E^*)$ such that the number $|E \triangle E^*|$ of inserted or deleted edges is minimized is an NP-complete problem [45, 46]. In view of the above results, this implies the following:

Theorem 6. *Let $\Theta \subset \mathfrak{G} \times \mathfrak{G}$ be an (estimated) orthology relation. It can be recognized in linear time whether Θ has a (discriminating) symbolic representation.*

For a given positive integer K the problem of deciding if there is an orthology relation Θ^* that has a (discriminating) symbolic representation s.t. $|\Theta \triangle \Theta^*| \leq K$ is NP-complete.

As the next result shows, it suffices to solve the cograph editing problem separately for the connected components of G.

Lemma 8. *For any graph $G(V, E)$ let $F \in \binom{V}{2}$ be a minimal set of edges so that $G' = (V, E \triangle F)$ is a cograph. Then $(x, y) \in F \setminus E$ implies that x and y are located in the same connected component of G.*

Proof. Suppose, for contradiction, that there is a minimal set F connecting two distinct connected components of G, resulting in a cograph G'. W.l.o.g., we may assume that G has only two connected components C_1, C_2. Denote by G'' the graph obtained from G' by removing all edges $\{x, y\}$ with $x \in V(C_1)$ and $y \in V(C_2)$. If G'' is not a cograph, then there is an induced P_4, which must be contained in one of the connected components of G''. By construction this induced P_4 is also contained in G'. Since G' is a cograph no such P_4 exists and hence G'' is also a cograph, contradicting the minimality of F. □

S 1.5 From Gene Triples to Species Triples and Reconciliation Maps

A gene tree T on $\mathfrak{G}$ arises in evolution by means of a series of events along a species tree S on $\mathfrak{S}$. In our setting these may be duplications of genes within a single species and speciation events, in which the parent's gene content is transmitted to both offsprings. The connection between gene and species tree is encoded in the reconciliation map, which associates speciation vertices in the gene tree with the interior vertex in the species tree representing the same speciation event. We consider the problem of finding a species tree for a given gene tree. In this subsection We follow the presentation of [35].

S 1.5.1 Reconciliation Maps

We start with a formal definition of reconciliation maps.

Definition 1 ([35])**.** *Let $S = (W, F)$ be a species tree on $\mathfrak{S}$, let $T = (V, E)$ be a gene tree on $\mathfrak{G}$ with corresponding event labeling $t : V^0 \to \{\bullet, \square\}$ and suppose there is a surjective map σ that assigns to each gene the respective species it is contained in. Then we say that S is a* species tree for $(T, t; \sigma)$ *if there is a map $\mu : V \to W \cup F$ such that, for all $x \in V$:*

(i) If $x \in \mathfrak{G}$ then $\mu(x) = \sigma(x)$.

(ii) If $t(x) = \bullet$ then $\mu(x) \in W \setminus \mathfrak{S}$.

(iii) If $t(x) = \square$ then $\mu(x) \in F$.

(iv) Let $x, y \in V$ with $x \prec_T y$. We distinguish two cases:

1. *If $t(x) = t(y) = \square$ then $\mu(x) \preceq_S \mu(y)$ in S.*
2. *If $t(x) = t(y) = \bullet$ or $t(x) \neq t(y)$ then $\mu(x) \prec_S \mu(y)$ in S.*

(v) If $t(x) = \bullet$ then $\mu(x) = \mathrm{lca}_S(\sigma(L(x)))$

We call μ the reconciliation map from (T, t, σ) to S.

A reconciliation map μ maps leaves $x \in \mathfrak{G}$ to leaves $\mu(x) := \sigma(x)$ in S and inner vertices $x \in V^0$ to inner vertices $w \in W \setminus \mathfrak{S}$ in S if $t(x) = \bullet$ and to edges $f \in F$ in S if $t(x) = \square$, such that the ancestor relation $\preceq_S$ is implied by the ancestor relation $\preceq_T$. Definition 1 is consistent with the definition of reconciliation maps for the case when the event labeling t on T is not known, see [24].

S 1.5.2 Existence of a Reconciliation Map

The reconciliation of gene and species trees is usually studied in the situation that only S, T, and σ are known and both μ and t and must be determined [29, 47, 3, 9, 26, 32, 4, 17, 13, 42]. In this form, there is always a solution (μ, t), which however is not unique in general. A variety of different optimality criteria have been used in the literature to obtain biologically plausible reconciliations. The situation changes when not just the gene tree T but a symbolic representation (T, t) is given. Then a species tree need not exists. [35] derived necessary and sufficient conditions for the existence of a species tree S so that there exists a reconciliation map from (T, t) to S. We briefly summarize the key results.

For $(T, t; \sigma)$ we define the triple set

$$\mathbb{G} = \big\{r \in \mathfrak{R}(T) \big| t(\mathrm{lca}_T(L_r)) = \bullet \text{ and } \sigma(x) \neq \sigma(y), \\ \text{for all } x, y \in L_r \text{ pairwise distinct}\big\}$$

In other words, the set $\mathbb{G}$ contains all triples $r = (\mathsf{ab}|\mathsf{c})$ of $\mathfrak{R}(T)$ where all three genes in $a, b, c \in L_r$ are contained in different species and the event at the most recent common ancestor of L_r is a speciation event, i.e., $t(\mathrm{lca}_T(a, b, c)) = \bullet$. It is easy to see that in this case S must display $(\sigma(\mathsf{a})\sigma(\mathsf{b})|\sigma(\mathsf{c}))$, i.e., it is a necessary condition that the triple set

$$\mathbb{S} = \{(\alpha\beta|\gamma) | \, \exists (\mathsf{ab}|\mathsf{c}) \in \mathbb{G} \text{ with } \sigma(a) = \alpha, \sigma(b) = \beta, \sigma(c) = \gamma\}$$

is consistent. This condition is also sufficient:

Theorem 7 ([35]). *There is a species tree on $\sigma(\mathfrak{G})$ for (T, t, σ) if and only if the triple set $\mathbb{S}$ is consistent. A reconciliation map can then be found in polynomial time.*

S 1.5.3 Maximal Consistent Triple Sets

In general, however, $\mathbb{S}$ may not be consistent. In this case it is impossible to find a valid reconciliation map. However, for each consistent subset $\mathbb{S}^* \subset \mathbb{S}$, its corresponding species tree S^*, and a suitably chosen homeomorphic image of T one can find the reconciliation. For a phylogenetic tree T on L, the *restriction* $T|_{L'}$ of T to $L' \subseteq L$ is the phylogenetic tree with leaf set L' obtained from T by first forming the minimal spanning tree in T with leaf set L' and then by suppressing all vertices of degree two with the exception of ρ_T if ρ_T is a vertex of that tree, see [49]. For a consistent subset $\mathbb{S}^* \subset \mathbb{S}$ let $L' = \{x \in \mathfrak{G} \mid \exists r \in \mathbb{S}^* \text{ with } \sigma(x) \in L_r\}$ be the set of genes (leaves of $T|_{L'}$) for which a species $\sigma(x)$ exits that is also contained in some triple $r \in \mathbb{S}^*$. Clearly, the reconciliation map of $T|_{L'}$ and the species tree S^* that displays $\mathbb{S}^*$ can then be found in polynomial time by means of Theorem 7.

S 2 ILP Formulation

The workflow outline in the main text consists of three stages, each of which requires the solution of hard combinatorial optimization problem. Our input data consist of an Θ or of a weighted version thereof. In the weighted case we assume the edge weights $w(x, y)$ have values in the unit interval that measures the confidence in the statement "$(x, y) \in \Theta$". Because of measurement errors, our first task is to correct Θ to an irreflexive, symmetric relation Θ^* that is a valid orthology relation. As outlined in section S 1.4.1, G_{Θ^*} must be cograph so that $(x, y) \in \Theta^*$ implies $\sigma(x) \neq \sigma(y)$. By Lemma 8 this problem has to be solved independently for every connected component of G_Θ. The resulting relation Θ^* has the symbolic representation (T, t).

In the second step we identify the best approximation of the species tree induced by (T, t). To this end, we determine the maximum consistent subset $\mathbb{S}^*$ in the set $\mathbb{S}$ of species triples induced by those triples of (T, t) that have a speciation vertex as their root. The hard part in the ILP formulation for this problem is to enforce consistency of a set of triples [16]. This step can be simplified considerably using the fact that for every consistent triple set $\mathbb{S}^*$ there is a strict dense consistent triple set $\mathbb{S}'$ that contains $\mathbb{S}^*$ (Lemma 7). This allows us to write $\mathbb{S}^* = \mathbb{S}' \cap \mathbb{S}$. The gain in efficiency in the corresponding ILP formulation comes from the fact that a strict dense set of triples is consistent if and only if all its two-element subsets are consistent (Theorem 1), allowing for a much faster check of consistency.

In the third step we determine the least resolved species tree S from the triple set $\mathbb{S}^*$ since this tree makes least assumptions of the topology and thus, of the evolutionary history. In particular, it displays only those triples that are either directly derived from the data or that are logically implied by them. Thus S is the tree with the minimal number of (inner) vertices that displays $\mathbb{S}^*$. Our ILP formulation uses ideas from the work of [16] to construct S in the form of an equivalent partial hierarchy.

S 2.1 Cograph Editing

Given the edge set of an input graph, in our case the pairs $(x, y) \in \Theta$, our task is to determine a modified edge set so that the resulting graph is a cograph. The input is conveniently represented by binary constants $\Theta_{ab} = 1$ iff $(a, b) \in \Theta$. The edges of the adjusted cograph G_{Θ^*} are represented by binary variables $E_{xy} = E_{yx} = 1$ if and only if $\{x, y\} \in E(G_{\Theta^*})$. Since $E_{xy} \equiv E_{yx}$ we use these variables interchangeably, without distinguishing the indices. Since genes residing in the same organism cannot be orthologs, we exclude edges $\{x, y\}$ whenever $\sigma(x) = \sigma(y)$ (which also forbids loops $x = y$. This is expressed by setting

$$E_{xy} = 0 \text{ for all } \{x, y\} \in \binom{\mathfrak{G}}{2} \text{ with } \sigma(x) = \sigma(y). \tag{ILP 2}$$

To constrain the edge set of G_{Θ^*} to cographs, we use the fact that cographs are characterized by P_4 as forbidden subgraph. This can be expressed as follows. For every ordered four-tuple $(w, x, y, z) \in \mathfrak{G}^4$ with pairwise distinct w, x, y, z we require

$$E_{wx} + E_{xy} + E_{yz} - E_{xz} - E_{wy} - E_{wz} \leq 2 \tag{ILP 3}$$

Constraint (ILP 3) ensures that for each ordered tuple (w, x, y, z) it is not the case that there are edges $\{w, x\}$, $\{x, y\}$, $\{y, z\}$ and at the same time no edges $\{x, z\}$, $\{w, y\}$, $\{w, z\}$ that is, w, x, y and z induce the path $w - x - y - z$ on four vertices. Enforcing this constraint for all orderings of w, x, y, z ensures that the subgraph induced by $\{w, x, y, z\}$ is P_4-free.

In order to find the closest orthology cograph G_{Θ^*} we minimize the symmetric difference of the estimated and adjusted orthology relation. Thus the objective function is

$$\min \sum_{(x,y)\in\mathfrak{G}\times\mathfrak{G}} (1 - \Theta_{xy})E_{xy} + \sum_{(x,y)\in\mathfrak{G}\times\mathfrak{G}} \Theta_{xy}(1 - E_{xy}) \tag{ILP 1}$$

Remark 5. *We have defined Θ above as a binary relation. The problem can be generalized to a weighted version in which the input Θ is a real valued function $\Theta : \mathfrak{G} \times \mathfrak{G} \to [0, 1]$ measuring the confidence with which a pair (x, y) is orthologous. The ILP formulation remains unchanged.*

The latter ILP formulation makes use of $O(|\mathfrak{G}|^2)$ variables and Equations (ILP 2) and (ILP 3) impose $O(|\mathfrak{G}|^4)$ constraints.

S 2.2 Extraction of All Species Triples

Let Θ be an orthology relation with symbolic representation $(T, t; \sigma)$ so that $\sigma(x) = \sigma(y)$ implies $(x, y) \notin \Theta$. By Theorem 7, the species tree S displays all triples $(\alpha\beta|\gamma)$ with a corresponding gene triple $(\mathsf{xy}|\mathsf{z}) \in \mathbb{G} \subseteq \mathfrak{R}(T)$, i.e., a triple $(\mathsf{xy}|\mathsf{z})$ with speciation event at the root of $t(\mathrm{lca}_T(x, y, z) = \bullet$ and $\sigma(x) = \alpha$, $\sigma(y) = \beta$, $\sigma(z) = \gamma$ are pairwise distinct species. We denote the set of these triples by $\mathbb{S}$. Although all species triples can be extracted in polynomial time, e.g. by using the BUILD algorithm, we give here an ILP formulation to complete the entire ILP pipeline. It will also be useful as a starting point for the final step, which consists in finding a minimally resolved trees that displays $\mathbb{S}$. Instead of using the symbolic representation $(T, t; \sigma)$ we will directly make use of the information stored in Θ using the following simple observation.

Lemma 9. *Let Θ be an orthology relation with discriminating symbolic representation $(T, t; \sigma)$ that is identified with the cotree of the corresponding cograph $G_\Theta = (\mathfrak{G}, E_\Theta)$. Assume that $(\mathsf{xy}|\mathsf{z}) \in \mathfrak{R}(T)$ is a triple where all genes x, y, z are contained in pairwise different species. Then it holds: $t(\mathrm{lca}(x, y)) = \square$ if and only if $\{x, y\} \notin E_\Theta$ and $t(\mathrm{lca}(x, y, z)) = \bullet$ if and only if $\{x, z\}, \{y, z\} \in E_\Theta$*

Proof. Assume there is a triple $(\mathsf{xy}|\mathsf{z}) \in \mathfrak{R}(T)$ where all genes x, y, z are contained in pairwise different species. Clearly, $t(\mathrm{lca}(x, y)) = \square$ iff $(x, y) \notin \Theta$ iff $\{x, y\} \notin E_\Theta$. Since, $\mathrm{lca}(x, y) \neq \mathrm{lca}(x, z) = \mathrm{lca}(y, z) = \mathrm{lca}(x, y, z)$ we have $t(\mathrm{lca}(x, z)) = t(\mathrm{lca}(y, z)) = \bullet$, which is iff $(x, z), (y, z) \in \Theta$ and thus, iff $\{x, z\}, \{y, z\} \in E_\Theta$. □

The set $\mathbb{S}$ of species triples is encoded by the binary variables $T_{(\alpha\beta|\gamma)} = 1$ iff $(\alpha\beta|\gamma) \in \mathbb{S}$. Note that $(\beta\alpha|\gamma) \equiv (\alpha\beta|\gamma)$. In order to avoid superfluous variables and symmetry conditions connecting them we assume that the first two indices in triple variables are ordered. Thus there are three triple variables $T_{(\alpha\beta|\gamma)}$, $T_{(\alpha\gamma|\beta)}$, and $T_{(\beta\gamma|\alpha)}$ for any three distinct $\alpha, \beta, \gamma \in \mathfrak{S}$.

Assume that $(\mathsf{xy}|\mathsf{z}) \in \mathfrak{R}(T)$ is an arbitrary triple displayed by T. In the remainder of this section, we assume that these genes x, y and z are from pairwise different species $\sigma(x) = \alpha$, $\sigma(y) = \beta$ and $\sigma(z) = \gamma$. Given that in addition $t(\mathrm{lca}(x, y, z)) = \bullet$, we need to ensure that $T_{(\alpha\beta|\gamma)} = 1$. If $t(\mathrm{lca}(x, y, z)) = \bullet$ then there are two cases: *(1)* $t(\mathrm{lca}(x, y)) = \square$ or *(2)* $t(\mathrm{lca}(x, y)) = \bullet$. These two cases needs to be considered separately for the ILP formulation.

Case (1) $t(\mathrm{lca}(x, y)) = \square \neq t(\mathrm{lca}(x, y, z))$*:* Lemma 9 implies that $E_{xy} = 0$ and $E_{xz} = E_{yz} = 1$. This yields, $(1 - E_{xy}) + E_{xz} + E_{yz} = 3$. To infer that in this case $T_{(\alpha\beta|\gamma)} = 1$ we add the next constraint.

$$(1 - E_{xy}) + E_{xz} + E_{yz} - T_{(\alpha\beta|\gamma)} \leq 2 \qquad \texttt{(ILP 15)}$$

These constraints need, by symmetry, also be added for the possible triples $(\mathsf{xz}|\mathsf{y})$, resp., $(\mathsf{yz}|\mathsf{x})$ and the corresponding species triples $(\alpha\gamma|\beta)$, resp., $(\beta\gamma|\alpha)$:

$$\begin{aligned} E_{xy} + (1 - E_{xz}) + E_{yz} - T_{(\alpha\gamma|\beta)} &\leq 2 \qquad \texttt{(ILP 15)} \\ E_{xy} + E_{xz} + (1 - E_{yz}) - T_{(\beta\gamma|\alpha)} &\leq 2 \end{aligned}$$

Case (2) $t(\mathrm{lca}(x, y)) = \bullet = t(\mathrm{lca}(x, y, z))$*:* Lemma 9 implies that $E_{xy} = E_{xz} = E_{yz} = 1$. Since $\mathrm{lca}(x, y) \neq \mathrm{lca}(x, y, z)$ and the gene tree we obtained the triple from is a discriminating representation, that is consecutive event labels are different, there must be an inner vertex $v \notin \{\mathrm{lca}(x, y), \mathrm{lca}(x, y, z)\}$ on the path from $\mathrm{lca}(x, y)$ to $\mathrm{lca}(x, y, z)$ with $t(v) = \square$. Since T is a phylogenetic tree, there must be a leaf $w \in L(v)$ with $w \neq x, y$ and $\mathrm{lca}(x, y, w) = v$ which implies $t(\mathrm{lca}(x, y, w)) = t(v) = \square$. For this vertex w we derive that $(xw|z), (yw|z) \in \mathfrak{R}(T)$ and in particular, $\mathrm{lca}(y, w, z) = \mathrm{lca}(x, y, z) = \mathrm{lca}(w, z)$. Therefore, $t(\mathrm{lca}(y, w, z)) = t(\mathrm{lca}(w, z)) = \bullet$.

Now we have to distinguish two subcases; either *Case (2a)* $\sigma(x) = \alpha = \sigma(w)$ (analogously one treats the case $\sigma(y) = \beta = \sigma(w)$ by interchanging the role of x and y) or *Case (2b)* $\sigma(x) = \alpha \neq \sigma(w) = \delta \notin \{\alpha, \beta, \gamma\}$. Note, the case $\sigma(w) = \sigma(z) = \gamma$ cannot occur, since we obtained (T, t) from the cotree of G_Θ and in particular, we have $t(\mathrm{lca}(w, z)) = \bullet$. Therefore, $E_{wz} = 1$ and hence, by Constraint `ILP 2` it must hold $\sigma(w) \neq \sigma(z)$.

(2a) Since $t(\mathrm{lca}(y, w, z)) = \bullet$ and $v = \mathrm{lca}(y, w)$ with $t(v) = \square$ it follows that the triple $(yw|z)$ fulfills the conditions of *Case 1*, and hence $T_{(\alpha\beta|\gamma)} = 1$ and we are done.

(2b) Analogously as in Case (2a), the triples $(xw|z)$ and $(yw|z)$ fulfill the conditions of Case (1), and hence we get $T_{(\alpha\delta|\gamma)} = 1$ and $T_{(\beta\delta|\gamma)} = 1$. However, we must ensure that also the triple $(\alpha\beta|\gamma)$ will be determined as observed species triple. Thus we add the constraint:

$$T_{(\alpha\delta|\gamma)} + T_{(\beta\delta|\gamma)} - T_{(\alpha\beta|\gamma)} \leq 1 \qquad \texttt{(ILP 15)}$$

which ensures that $T_{(\alpha\beta|\gamma)} = 1$ whenever $T_{(\alpha\delta|\gamma)} = T_{(\beta\delta|\gamma)} = 1$.

The first three constraints in Eq. (`ILP 15`) are added for all $\{x, y, z\} \in \binom{\mathfrak{G}}{3}$ and where all three genes are contained in pairwise different species $\sigma(x) = \alpha$, $\sigma(y) = \beta$ and $\sigma(z) = \gamma$ and the fourth constraint in Eq. (`ILP 15`) is added for all $\{\alpha, \beta, \gamma, \delta\} \in \binom{\mathfrak{S}}{4}$.

In particular, these constraints ensure, that for each triple $(\mathsf{xy}|\mathsf{z}) \in \mathbb{G}$ with speciation event on top and corresponding species triple $(\alpha\beta|\gamma)$ the variable $T_{(\alpha\beta|\gamma)}$ is set to 1.

However, the latter ILP constraints allow some degree of freedom for the choice of the binary value $T_{(\alpha\beta|\gamma)}$, where for all respective triples $(\mathsf{xy}|\mathsf{z}) \in \mathfrak{R}(T)$ holds $t(\mathrm{lca}(x, y, z)) = \square$. To ensure, that only those variables $T_{(\alpha\beta|\gamma)}$ are set to 1, where at least one triple $(\mathsf{xy}|\mathsf{z}) \in \mathfrak{R}(T)$ with $t(\mathrm{lca}(x, y, z)) = \bullet$ and $\sigma(x) = \alpha$, $\sigma(y) = \beta$, $\sigma(z) = \gamma$ exists, we add the following objective function that minimizes the number of variables $T_{(\alpha\beta|\gamma)}$ that are set to 1:

$$\min \sum_{\{\alpha,\beta,\gamma\} \in \binom{S}{3}} T_{(\alpha\beta|\gamma)} + T_{(\alpha\gamma|\beta)} + T_{(\beta\gamma|\alpha)} \qquad \texttt{(ILP 16)}$$

For the latter ILP formulation $O(|\mathfrak{S}|^3)$ variables and $O(|\mathfrak{G}|^3 + |\mathfrak{S}|^4)$ constraints are required.

S 2.3 Find Maximal Consistent Triple Set

Given the set of species triple $\mathbb{S}$ the next step is to extract a maximal subset $\mathbb{S}^* \subseteq \mathbb{S}$ that is consistent. This combinatorial optimization problem is known to be NP-complete [39, 53]. In an earlier ILP approach, [16] explicitly constructed a tree that displays $\mathbb{S}^*$. In order to improve the running time of the ILP we focus here instead on constructing a consistent, strict dense triple set $\mathbb{S}$' containing the desired solution $\mathbb{S}^*$ because the consistency check involves two-element subsets in this case (Theorem 1). From $\mathbb{S}'$ obtain the desired solution as $\mathbb{S}^* = \mathbb{S}' \cap \mathbb{S}$. We therefore introduce binary variables $T'_{(\alpha\beta|\gamma)} = 1$ iff $(\alpha\beta|\gamma) \in \mathbb{S}'$.

To ensure, that $\mathbb{S}'$ is strict dense we add for all $\{\alpha, \beta, \gamma\} \in \binom{\mathcal{S}}{3}$ the constraints:

$$T'_{(\alpha\beta|\gamma)} + T'_{(\alpha\gamma|\beta)} + T'_{(\beta\gamma|\alpha)} = 1. \tag{ILP 5}$$

We can now apply the inference rules in Eq. (ii) and the results of Theorem 1 and Lemma 4. Therefore, we add the following constraint for all ordered tuples $(\alpha, \beta, \gamma, \delta)$ for all $\{\alpha, \beta, \gamma, \delta\} \in \binom{\mathfrak{S}}{4}$:

$$2T'_{(\alpha\beta|\gamma)} + 2T'_{(\alpha\delta|\beta)} - T'_{(\beta\delta|\gamma)} - T'_{(\alpha\delta|\gamma)} \leq 2 \tag{ILP 6}$$

The constraint in Eq. (ILP 6) is a direct translation of the inference rule in Eqn. (ii). Moreover, by Theorem 1 and Lemma 4, we know that testing pairs of triples with Eq. (ii) is sufficient for verifying consistency.

To ensure maximal cardinality of $\mathbb{S}^* = \mathbb{S}' \cap \mathbb{S}$ we use the objective function

$$\max \sum_{(\alpha\beta|\gamma)\in\mathbb{S}} T'_{(\alpha\beta|\gamma)} \tag{ILP 4}$$

This ILP formulation can easily be adapted to solve a "weighted" maximum consistent subset problem: With $w(\alpha\beta|\gamma)$ we denote for every species triple $(\alpha\beta|\gamma) \in \mathbb{S}$ the number of connected components in G_{Θ^*} that contains three vertices $a, b, c \in \mathfrak{G}$ with $(\mathsf{ab}|\mathsf{c}) \in \mathbb{G}$ and $\sigma(a) = \alpha, \sigma(b) = \beta, \sigma(c) = \gamma$. In this way, we increase the significance of species triples in $\mathbb{S}$ that have been observed more times, when applying the following objective function.

$$\max \sum_{(\alpha\beta|\gamma)\in\mathbb{S}} T'_{(\alpha\beta|\gamma)} * w(\alpha\beta|\gamma). \tag{ILP 8}$$

Finally, we define binary variables $T^*_{(\alpha\beta|\gamma)}$ that indicate whether a triple $(\alpha\beta|\gamma) \in \mathbb{S}$ is contained in a maximal consistent triples set $\mathbb{S}^* \subseteq \mathbb{S}$, i.e., $T^*_{(\alpha\beta|\gamma)} = 1$ iff $(\alpha\beta|\gamma) \in \mathbb{S}^*$ and thus, iff $T_{(\alpha\beta|\gamma)} = 1$ and $T'_{(\alpha\beta|\gamma)} = 1$. Therefore, we add for all $\{\alpha, \beta, \gamma\} \in \binom{\mathcal{S}}{3}$ the binary variables $T^*_{(\alpha\beta|\gamma)}$ and add the constraints

$$0 \leq T'_{(\alpha\beta|\gamma)} + T_{(\alpha\beta|\gamma)} - 2T^*_{(\alpha\beta|\gamma)} \leq 1 \tag{ILP 7}$$

It is easy to verify, that in the latter ILP formulation $O(|\mathfrak{S}|^3)$ variables and $O(|\mathfrak{S}|^4)$ constraints are required.

S 2.4 Least Resolved Species Tree

The final step consists in finding a minimally resolved tree that displays all triples of $\mathbb{S}^*$. The variables $T^*_{(\alpha\beta|\gamma)}$ defined in the previous step take on the role of constants here.

There is an ILP approach by [16], for determining a maximal consistent triple sets. However, this approach relies on determining consistency by checking and building up a binary tree, a very time consuming task. As we showed, this can be improved and simplified by the latter ILP formulation. However, we will adapt now some of the ideas established by [16], to solve the NP-hard problem [40] of finding a least resolved tree.

To build an arbitrary tree for the consistent triple set $\mathbb{S}^*$, one can use the fast algorithm BUILD [49]. Moreover, if the tree obtained by BUILD for $\mathbb{S}^*$ is a binary tree, then Theorem 2 implies that the closure $\mathrm{cl}(\mathbb{S}^*)$ is strict dense and that this tree is a unique and least resolved tree for $\mathbb{S}^*$. Hence, as a preprocessing step one could use BUILD first, to test whether the tree for $\mathbb{S}^*$ is already binary and if not, proceed with the following ILP approach.

A phylogenetic tree S is uniquely determined by hierarchy $\mathcal{C} = \{L(v) \mid v \in V(S)\}$ according to Theorem 3. Thus it is possible to construct S by building the clusters induced by the triples of $\mathbb{S}^*$. Thus we need to translate the condition for $\mathcal{C}$ to be a hierarchy into the language of ILPs.

Following [16] we use a binary $|\mathfrak{S}| \times N$ matrix M, with entries $M_{\alpha p} = 1$ iff species α is contained in cluster p. By Lemma 1, it is clear that we need at most $2|\mathfrak{S}| - 1$ columns. As we shall see later, we exclude (implicitly) the trivial singleton clusters $\{x\} \in \mathfrak{S}$ and the cluster $\mathfrak{S}$. Hence, it suffices to use $N = 2|\mathfrak{S}| - 1 - |\mathfrak{S}| - 1 = |\mathfrak{S}| - 2$

clusters. Each cluster p, which is represented by the p-th column of M, corresponds to an inner vertex v_p in the species tree S so that $p = (L(v_p))$.

Since we are interested in finding a least resolved tree rather than a fully resolved one, we allow that number of clusters is smaller than $N-2$, i.e., we allow that some columns of M have no non-zero entries. Here, we deviate from the approach of [16]. Columns p with $\sum_{\alpha\in\mathfrak{S}} M_{\alpha p} = 0$ containing only 0 entries and thus, clusters $L(v_p) = \emptyset$, are called *trivial*, all other columns and clusters are called *non-trivial*. Clearly, the non-trivial clusters correspond to the internal vertices of S, hence we have to maximize the number of trivial columns of M. This condition also suffices to remove redundancy, i.e., non-trivial columns with the same entries.

We first give the ILP formulation that captures that all triples $(\alpha\beta|\gamma)$ contained in $\mathbb{S}^* \subseteq \mathbb{S}$ are displayed by a tree. A triple $(\alpha\beta|\gamma)$ is displayed by a tree if and only if there is an inner vertex v_p such that $\alpha, \beta \in L(v_p)$ and $\gamma \notin L(v_p)$ and hence, iff $M_{\alpha p} = M_{\beta p} = 1 \neq M_{\gamma p} = 0$ for this cluster p.

To this end, we define binary variables $N_{\alpha\beta,p}$ so that $N_{\alpha\beta,p} = 1$ iff $\alpha, \beta \in L(v_p)$ for all $\{\alpha, \beta\} \in \binom{\mathfrak{S}}{2}$ and $p = 1, \ldots, |\mathfrak{S}| - 2$. This condition is captured by the constraint:

$$0 \leq M_{\alpha p} + M_{\beta p} - 2N_{\alpha\beta,p} \leq 1. \tag{ILP 11}$$

We still need to ensure that for each triple $(\alpha\beta|\gamma) \in \mathbb{S}^*$ there is at least one cluster p that contains α and β but not γ, i.e., $N_{\alpha\beta,p} = 1$ and $N_{\alpha\gamma,p} = 0$ and $N_{\beta\gamma,p} = 0$. For each possible triple $(\alpha\beta|\gamma)$ we therefore add the constraint

$$1 - |\mathfrak{S}|(1 - T^*_{(\alpha\beta|\gamma)}) \leq \sum_p N_{\alpha\beta,p} - \frac{1}{2}N_{\alpha\gamma,p} - \frac{1}{2}N_{\beta\gamma,p}. \tag{ILP 12}$$

To see that (ILP 12) ensures $\alpha, \beta \in L(v_p)$ and $\gamma \notin L(v_p)$ for each $(\alpha\beta|\gamma) \in \mathbb{S}^*$ and some p, assume first that $(\alpha\beta|\gamma) \notin \mathbb{S}^*$ and hence, $T^*_{(\alpha\beta|\gamma)} = 0$. Then, $1 - |\mathfrak{S}|(1 - T^*_{(\alpha\beta|\gamma)}) = 1 - |\mathfrak{S}|$ and we are free in the choice of the variables $N_{\alpha\beta,p}$, $N_{\alpha\gamma,p}$, and $N_{\beta\gamma,p}$. Now assume that $(\alpha\beta|\gamma) \in \mathbb{S}^*$ and hence, $T^*_{(\alpha\beta|\gamma)} = 1$. Then, $1 - |\mathfrak{S}|(1 - T^*_{(\alpha\beta|\gamma)}) = 1$. This implies that at least one variable $N_{\alpha\beta,p}$ must be set to 1 for some p. If $N_{\alpha\beta,p} = 1$ and $N_{\alpha\gamma,p} = 1$, then constraint (ILP 11) implies that $M_{\alpha p} = M_{\beta p} = M_{\gamma p} = 1$ and thus $N_{\beta\gamma,p} = 1$. Analogously, if $N_{\alpha\beta,p} = 1$ and $N_{\beta\gamma,p} = 1$, then $N_{\alpha\gamma,p} = 1$. It remains to show that there is some cluster p with $N_{\alpha\beta,p} = 1$ and $N_{\alpha\gamma,p} = N_{\beta\gamma,p} = 0$. Assume, for contradiction, that for none of the clusters p with $N_{\alpha\beta,p} = 1$ holds that $N_{\alpha\gamma,p} = N_{\beta\gamma,p} = 0$. Then, by the latter arguments all of these clusters p satisfy: $N_{\alpha\gamma,p} = N_{\beta\gamma,p} = 1$. However, this implies that $N_{\alpha\beta,p} - \frac{1}{2}N_{\alpha\gamma,p} - \frac{1}{2}N_{\beta\gamma,p} = 0$ for all p, which contradicts the constraint (ILP 12). Therefore, if $T^*_{(\alpha\beta|\gamma)} = 1$, there must be at least one cluster p with $N_{\alpha\beta,p} = 1$ and $N_{\alpha\gamma,p} = N_{\beta\gamma,p} = 0$ and hence, $M_{\alpha p} = M_{\beta p} = 1$ and $M_{\gamma p} = 0$.

In summary the constraints above ensure that for the maximal consistent triple set $\mathbb{S}^*$ of $\mathbb{S}$ and for each triple $(\alpha\beta|\gamma) \in \mathbb{S}^*$ exists at least one column p in the matrix M that contains α and β, but not γ. Note that for a triple $(\alpha\beta|\gamma)$ we do not insist on having a cluster q that contains γ but not α and β and therefore, we do not insist on constructing singleton clusters. Moreover, there is no constraint that claims that the set $\mathfrak{S}$ is decoded by M. In particular, since we later maximize the number of trivial columns in M and since we do not gave ILP constraints that insist on finding clusters $\mathfrak{S}$ and $\{x\}$, $x \in \mathfrak{S}$, these clusters will not be defined by M. However, these latter clusters are clearly known, and thus, to decode the desired tree, we only require that M is a "partial" hierarchy, that is for every pair of clusters p and q holds $p \cap q \in \{p, q, \emptyset\}$. In such case the clusters p and q are said to be compatible. Two clusters p and q are incompatible if there are (not necessarily distinct) species $\alpha, \beta, \gamma \in \mathfrak{S}$ with $\alpha \in p \setminus q$ and $\beta \in q \setminus p$, and $\gamma \in p \cap q$. In the latter case we would have $(M_{\alpha p}, M_{\alpha q}) = (1, 0)$, $(M_{\beta p}, M_{\beta q}) = (0, 1)$, $(M_{\gamma p}, M_{\gamma q}) = (1, 1)$. Here we follow the idea of [16], and use the so-called three-gamete condition. For each gamete $(\Gamma, \Lambda) \in \{(0, 1), (1, 0)(1, 1)\}$ and each column p and q we define a set of binary variables $C_{p,q,\Gamma\Lambda}$. For all $\alpha \in \mathfrak{S}$ and $p, q = 1, \ldots, |\mathfrak{S}| - 2$ with $p \neq q$ we add

$$\begin{aligned} C_{p,q,01} &\geq -M_{\alpha p} + M_{\alpha q} \\ C_{p,q,10} &\geq M_{\alpha p} - M_{\alpha q} \\ C_{p,q,11} &\geq M_{\alpha p} + M_{\alpha q} - 1 \end{aligned} \tag{ILP 13}$$

These constraints capture that $C_{p,q,\Gamma\Lambda} = 1$ as long as if $M_{\alpha p} = \Gamma$ and $M_{\alpha q} = \Lambda$ for some $\alpha \in \mathfrak{S}$. To ensure that only compatible clusters are contained, we add for each of the latter defined variable

$$C_{p,q,01} + C_{p,q,10} + C_{p,q,11} \leq 2. \tag{ILP 14}$$

Hence the latter Equations (ILP 11)-(ILP 14) ensure we get a "partial" hierarchy M, where only the singleton clusters and the set $\mathfrak{S}$ is missing,

Finally we want to have for the maximal consistent triple sets $\mathbb{S}^*$ of $\mathbb{S}$ the one that determines the least resolved tree, i.e, a tree that displays all triples of $\mathbb{S}^*$ and has a minimal number of inner vertices and makes therefore, the fewest assumptions on the tree topology. Since the number of leaves $|\mathfrak{S}|$ in the species tree S is fixed and therefore the number of clusters is determined by the number of inner vertices, as shown in the proof of Lemma 1, we can conclude that a minimal number of clusters results in tree with a minimal number of inner vertices. In other words, to find a least resolved tree determined by the hierarchy matrix M, we need to maximize the number of trivial columns in M, i.e., the number of columns p with $\sum_{\alpha\in\mathfrak{S}} M_{\alpha p} = 0$.

For this, we require in addition to the constraints (ILP 11)-(ILP 14) for each $p = 1, \ldots, |\mathfrak{S}| - 2$ a binary variable Y_p that indicates whether there are entries in column p equal to 1 or not. To infer that $Y_p = 1$ whenever column p is non-trivial we add for each $p = 1, \ldots, |\mathfrak{S}| - 2$ the constraint

$$0 \leq Y_p|\mathfrak{S}| - \sum_{\alpha\in\mathfrak{S}} M_{\alpha p} \leq |\mathfrak{S}| - 1 \qquad \text{(ILP 10)}$$

If there is a "1" entry in column p and $Y_p = 0$ then, $Y_p|\mathfrak{S}| - \sum_{\alpha\in\mathfrak{S}} M_{\alpha p} < 0$, a contradiction. If column p is trivial and $Y_p = 1$ then, $Y_p|\mathfrak{S}| - \sum_{\alpha\in\mathfrak{S}} M_{\alpha p} = |\mathfrak{S}|$, again a contradiction. Finally, in order to minimize the number of non-trivial columns in M and thus, to obtain a least resolved tree for $\mathbb{S}^*$ we add the objective function

$$\min \sum_p Y_p \qquad \text{(ILP 9)}$$

Therefore, we obtain for the maximal consistent subset $\mathbb{S}^* \subseteq \mathbb{S}$ of species triples a "partial" hierarchy defined by M, that is, for all clusters $L(v_p)$ and $L(v_q)$ defined by columns p and q in M holds $L(v_p) \cap L(v_q) \in \{L(v_p), L(v_q), \emptyset\}$. The clusters $\mathfrak{S}$ and $\{x\}$, $x \in \mathfrak{S}$ will not be defined by M. However, from these clusters and the clusters determined by the columns of M it is easily build the corresponding tree, which by construction displays all triples in $\mathbb{S}^*$, see [49, 25].

The latter ILP formulation requires $O(|\mathfrak{S}|^3)$ variables and constraints.

S 3 Implementation and Data Sets

S 3.1 ILP Solver

The ILP approach has been implemented using IBM ILOG CPLEX$^{\text{TM}}$ Optimizer 12.6 in the weighted version of the maximum consistent triple set problem. For each component of G_Θ we check in advance if it is already a cograph. If this is not the case then an ILP instance is executed, finding the closest cograph. In a similar manner, we check for each resulting cograph whether it contains any paralogous genes at all. If not, then the cograph is a complete graph and the resulting gene tree would be a star, not containing any species triple information. Hence, extracting the species triples is skipped. Triple extraction is done using an polynomial time algorithm instead of the ILP formulation. Although the connected components of G_Θ are treated separately, some instances of the cograph editing problem have exceptionally long computation times. We therefore exclude components of G_Θ with more than 50 genes. In addition, we limit the running time for finding the closest cograph for one disconnected component to 30 minutes. If an optimal solution for this component is not found within this time limit, we use the best solution found so far. The other ILP computations are not restricted by a time limit.

S 3.2 Simulated Data

To evaluate the ILP approach we use simulated and real-life data sets. Artificial data is created with the the method described in [34] as well as the `Artificial Life Framework` (`ALF`) [20]. The first method generates explicit species/gene tree histories, from which the orthology relation is directly accessible. All simulations are performed with parameters 1.0 for gene duplication, 0.5 for gene loss and 0.1 for the loss rate, respectively increasing loss rate, after gene duplication. We do not consider cluster or genome duplications. `ALF` simulates the evolution of sequences along a branch length-annotated species tree, explicitly taking into account gene duplication, gene loss, and horizontal transfer events. To obtain bacteria-like data sets we adopted the procedure from [21]: a tree of *γ-proteobacteria* from the OMA project [2] was randomly pruned to obtain trees of moderate size, while conserving the original branch lengths. All simulations are performed with parameters 0.005 for gene duplication/loss rate. We do not consider cluster duplications/loss.

The presented method heavily depends on the amount of duplicated genes, which, in turn, is depending on the number of analyzed genes per species. Naturally, the question arose, how many genes, respectively gene

families, are needed, to provide enough information to reconstruct accurate species trees, assuming a certain gene duplication rate. Therefore, we evaluate the precision of reconstructed trees with respect to the number of species and gene families. 100 species trees of size 5, 10, 15, and 20 (`ALF` only) leaves are generated. For each tree, the evolution of ten to 100 (first simulation method) and 100 to 500 (`ALF`) gene families is simulated. This corresponds for the first simulation method to 32.6% (five species), 19.0% (ten species), and 13.5% (15 species) and for `ALF` simulations 11.2% (five species), 8.1% (ten species), and 7.5% (15 and 20 species) of all homologous pairs being paralogs (values determined from the simulations). Horizontal gene transfer and cluster duplication/loss were not considered.

The reconstructed trees are compared with the generated (binary) species trees. Therefore, we use the software `TreeCmp` [8] to compute distances for rooted trees based on Matching Cluster (MC), Robinson-Foulds (RC), Nodal Splitted (NS) and Triple metric (TT). The distances are normalized by the average distance between random Yule trees [54].

In order to estimate the effects of noise in the empirical orthology relation we consider several forms of perturbations (i) insertion and deletion of edges in the orthology graph (homologous noise), (ii) insertion of edges (orthologous noise), (iii) deletion of edges (paralogous noise), and (iv) modification of gene/species assignments (xenologous noise). In the first three models each possible edge is modified with probability p. Model (ii) simulates overprediction of orthology, while model (iii) simulates underprediction. Model (iv) retains the original orthology information but changes the associations between genes and their respective species with probability p. This simulates noise as expected in case of horizontal gene transfer. For each model we reconstruct the species trees of 100 simulated data sets with ten species and 100 gene families (first simulation method), respectively 1000 gene families (`ALF`). As before, no horizontal gene transfer or cluster duplications/losses were simulated. Noise is added with a probability $p \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$.

Horizontal transfer is an abundant process in particular in procaryotes that may lead to particular types of errors in each step of our approach, see the theoretical discussion below. We therefore investigated the robustness of our approach against HGT as a specific type of perturbation in some detail. To this end, we simulate data sets of 1000 gene families, using `ALF`, with a duplication/loss rate of 0.005 and evolutionary rates $r \in \{0.0, 0.0025, 0.005, 0.0075\}$ for horizontal transfer. Cluster duplications/losses, or horizontal transfers of groups of genes are not considered. The simulation is repeated 100 times for each combination of parameters. From the simulated sequences, orthologous pairs of genes are predicted with `Proteinortho` [43], using an E-value threshold of $1e-10$ and similarity parameter of 0.9. From this estimate of the orthology relation species trees are reconstructed.

The authors of [21] observed that increasing HGT rates have only a minor impact on the recall of orthology prediction, while the precision drops significantly, i.e., orthology prediction tools tend to mis-predict xenology as orthology. To evaluate the impact of noise solely coming from mis-predicting xenology as orthology, a second orthology relation is constructed from the same simulations. This orthology relation only differs from the simulated orthology relation by all simulated xenologs being predicted as orthologs, i.e., all paralogs are correctly detected (*perfect paralogy knowledge*), see S6 (B). Analogously, we evaluated the impact of noise solely coming from mis-predicting xenology as paralogy, i.e., all orthologs are correctly detected (*perfect orthology knowledge*), see S6 (C). From these orthology relations, species trees are reconstructed with the ILP approach, and compared with the generated species trees, used for the simulation.

S 3.3 Real-Life Data Sets

As real-life applications we consider two sets of eubacterial genomes. The set of eleven *Aquificales* species studied in [44] covers the three families *Aquificaceae*, *Hydrogenothermaceae*, and *Desulfurobacteriaceae*. The species considered are the *Aquificaceae*: *Aquifex aeolicus* VF5 (NC_000918.1, NC_001880.1), *Hydrogenivirga sp.* 128-5-R1-1 (ABHJ00000000.1), *Hydrogenobacter thermophilus* TK-6 (NC_013799.1), *Hydrogenobaculum sp.* Y04AAS1 (NC_011126.1), *Thermocrinis albus* DSM 14484 (NC_013894.1), *Thermocrinis ruber* DSM 12173 (CP007028.1), the *Hydrogenothermaceae*: *Persephonella marina* EX-H1 (NC_012439.1, NC_012440.1), *Sulfurihydrogenibium sp.* YO3AOP1 (NC_010730.1) *Sulfurihydrogenibium azorense* Az-Fu1 (NC_012438.1), and the *Desulfurobacteriaceae*: *Desulfobacterium thermolithotrophum* DSM 11699 (NC_015185.1), and *Thermovibrio ammonificans* HB-1 (NC_014917.1, NC_014926.1).

A larger set of 19 *Enterobacteriales* was taken from RefSeq: *Enterobacteriaceae* family: *Cronobacter sakazakii* ATCC BAA-894 (NC_009778.1, NC_009779.1, NC_009780.1), *Enterobacter aerogenes* KCTC 2190 (NC_015663.1), *Enterobacter cloacae* ATCC 13047 (NC_014107.1, NC_014108.1, NC_014121.1), *Erwinia amylovora* ATCC 49946 (NC_013971.1, NC_013972.1, NC_013973.1), *Escherichia coli* K-12 substr DH10B (NC_010473.1), *Escherichia fergusonii* ATCC 35469 (NC_011740.1, NC_011743.1), *Klebsiella oxytoca* KCTC 1686 (NC_016612.1), *Klebsiella pneumoniae* 1084 (NC_018522.1), *Proteus mirabilis* BB2000 (NC_022000.1), *Salmonella bongori* Sbon 167 (NC_021870.1, NC_021871.1), *Salmonella enterica* serovar Ag-

ona SL483 (NC_011148.1, NC_011149.1), *Salmonella typhimurium* DT104 (NC_022569.1, NC_022570.1), *Serratia marcescens* FGI94 (NC_020064.1), *Shigella boydii* Sb227 (NC_007608.1, NC_007613. 1), *Shigella dysenteriae* Sd197 (NC_007606.1, NC_007607.1, NC_009344.1), *Shigella flexneri* 5 str 8401 (NC_008258.1), *Shigella sonnei* Ss046 (NC_007384.1, NC_007385.1, NC_009345.1, NC_009346.1, NC_009347.1), *Yersinia pestis* Angola (NC_010157. 1, NC_010158.1, NC_010159.1), and *Yersinia pseudotuberculosis* IP 32953 (NC_006153.2, NC_006154.1, NC_006155.1).

S 3.4 Estimation of the Input Orthology Relation

An initial estimate of the orthology relation is computed with `Proteinortho` [43] from all the annotated proteins using an E-value threshold of $1e-10$ and similarity parameter of 0.9. Additionally, the genomes of all species were re-`blast`ed to detect homologous genes not annotated in the `RefSeq`. In brief, `Proteinortho` implements a modified pair-wise best hit strategy starting from `blast` comparisons. It first creates a graph consisting of all genes as nodes and an edge for every blast hit with an E-value above a certain threshold. In a second step edges between two genes a and b from different species are removed if a much better blast hit is found between a and a duplicated gene b' from the same species as b. Finally, the graph is filtered with spectral partitioning to result in disconnected components with a certain minimum algebraic connectivity.

The resulting orthology graph usually consists of several pairwise disconnected components, which can be interpreted as individual gene families. Within these components there may exist pairs of genes having `blast` E-values worse than the threshold so that these nodes are not connected in the initial estimate of Θ. Thus, the input data have a tendency towards underprediction of orthology in particular for distant species. Our simulation results suggest that the ILP approach handles overprediction of orthology much better. We therefore re-add edges that were excluded because of the E-value cut-off only within connected components of the raw Θ relation.

S 3.5 Evaluation of Phylogenies

For the analysis of simulated data we compare the reconstructed trees with the trees generated by the simulation. To this end we computed the four commonly used distances measures for rooted trees, Matching Cluster (MC), Robinson-Foulds (RC), Nodal Splitted (NS) and Triple metric (TT), as described by [8].

The MC metric asks for a minimum-weight one-to-one matching between the internal nodes of both trees, i.e., the clusters C_1 from tree T_1 with the clusters C_2 from tree T_2. For a given one-to-one matching the MC tree distance d_{MC} is defined as the sum of all weights $h_C(p_1, p_2) = |L(p_1) \setminus L(p_2) \cup L(p_2) \setminus L(p_1)|$ with $p_1 \in C_1$ and $p_2 \in C_2$. For all unmatched clusters p a weight $|L(p)|$ is added. The RC tree distance d_{RC} is equal to the number of different clusters in both trees divided by 2. The NS metric computes for each tree T_i a matrix $l(T_i) = (l_{xy})$ with $x, y \in L(T_i)$ and l_{xy} the length of the path from $\mathrm{lca}(x, y)$ to x. The NS tree distance d_{NS} is defined as the L^2 norm of these matrices, i.e., $d_{NS} = \|l(T_1) - l(T_2)\|_2$. The TT metric is based on the set of triples $\mathfrak{R}(T_i)$ displayed by tree T_i. For two trees T_1 and T_2 the TT tree distance is equal to the number of different triples in respective sets $\mathfrak{R}(T_1)$ and $\mathfrak{R}(T_2)$.

The four types of tree distances are implemented in the software `TreeCmp` [8], together with an option to compute normalized distances. Therefore, average distances between random Yule trees [54] are provided for each metric and each tree size from 4 to 1000 leaves. These average distances are used for normalization, resulting in a value of 0 for identical trees and a value of approximately 1 for two random trees. Note, however, distances greater 1 are also possible.

For the trees reconstructed from the real-life data sets we compute a support value $s \in [0, 1]$, utilizing the triple weights $w(\alpha\beta|\gamma)$ from Eq. (`ILP 8`). Precisely,

$$s = \frac{\sum_{(\alpha\beta|\gamma)\in\mathbb{S}^*} w(\alpha\beta|\gamma)}{\sum_{(\alpha\beta|\gamma)\in\mathbb{S}^*} w(\alpha\beta|\gamma) + w(\alpha\gamma|\beta) + w(\beta\gamma|\alpha)} \tag{2}$$

The support value of a reconstructed tree indicates how often the triples from the computed maximal consistent subset $\mathbb{S}^*$ were obtained from the data in relation to the frequency of all obtained triples. It is equal to 1 if there was no ambiguity in the data. Values around 0.33 indicate randomness.

In a similar way, we define support values for each subtree $T(v)$ of the resulting species tree T. Therefore, let $S^v = \{(\alpha\beta|\gamma) \in \mathfrak{R}(T) | \alpha, \beta \in L(v), \gamma \notin L(v)\}$ be the subset of the triples displayed by T with the two closer related species being leaves in the subtree $T(v)$ and the third species not from this subtree. Then, the subtree support is defined as:

$$s_v = \frac{\sum_{(\alpha\beta|\gamma)\in\mathbb{S}^v} w(\alpha\beta|\gamma)}{\sum_{(\alpha\beta|\gamma)\in\mathbb{S}^v} w(\alpha\beta|\gamma) + w(\alpha\gamma|\beta) + w(\beta\gamma|\alpha)} \tag{3}$$

Note that S^v only contains triples that support a subtree with leaf set $L(v)$. Therefore, the subtree support indicates how often triples are obtained supporting this subtree in relation to the frequency of all triples supporting the existence or non-existence of this subtree.

In addition, bootstrap trees are constructed for each data set, using two different bootstrapping approaches. (i) bootstrapping based on components, and (ii) bootstrapping based on triples. Let m be the number of pairwise disconnected components from the orthology graph G_{Θ^*}, n_i the number of species triples extracted from component i, and $n = \sum_{i=1}^{m} n_i$. In the first approach we randomly select m components with repetition from G_{Θ^*}. Then we extract the respective species triples and compute the maximal consistent subset and least resolved tree. In the second approach we randomly select n triples with repetition from $\mathbb{S}$. Each triple $(\alpha\beta|\gamma)$ is chosen with a probability according to its relative frequency $w(\alpha\beta|\gamma)/n$. From this set the maximal consistent subset and least resolved tree is computed. Bootstrapping is repeated 100 times. Majority-rule consensus trees are computed with the software `CONSENSE` from the `PHYLIP` package.

S 4 Additional Results

S 4.1 Robustness against Noise from Horizontal Gene Transfer

Horizontal gene transfer (HGT) is by far the most common deviation from vertical inheritance, e.g. [27]. The key problem with HGT in the context of orthology prediction is that pairs of genes that derive from a speciation rather than a duplication event may end up in the same genome. Since pairs of genes in the same genome are classified as paralogs by the initial orthology detection heuristics and subsequently by ILP constraint `ILP 2` during cograph editing. Such *pseudo-orthologous* pairs can lead to a misplaced node with an incorrect event label in the cotree. This may, under some circumstances, lead to the inference of false species triples, see Figure S1. Note, the latter problem still remains even if we would have detected all events on the gene tree correctly but use the triple sets $\mathbb{G}$ and $\mathbb{S}$ without any additional restrictions. Again Fig. S1 serves as an example. Therefore, it is of central interest to understand in more detail the relation between symbolic representations, reconciliation maps and triple sets that take also HGT into account, which might solve this problem.

When all true paralogs are known, we obtain surprisingly accurate species tree, see Figure S6 (B). The species trees reconstructed from a perfect orthology relation are somewhat less accurate, see Figure S6 (C). The most pressing practical issue, therefore, is to identify true paralogs, i.e., pairs of genes that originate from a duplication event and subsequently are inherited only vertically. In addition, phylogeny-free methods to identify xenologs e.g. based on sequence composition [41, 50] are a promising avenue for future work to improve the initial estimates of orthology and paralogy.

S 4.2 Simulated Data

The results for simulated data sets with a varying number of independent gene families suggest, that a few hundred gene families are sufficient to contain enough information for reconstructing proper phylogenetic species trees. The reconstructions for data sets generated with `ALF` need much more gene families to obtain a similar accuracy, as compared to simulations with the first simulation method. This can be explained by the fact that the simulations of the first method resulted in a higher amount of paralogs, ranging from 13.5% to 32.6%, compared to the ALF simulations (7.5% to 11.2%). Another reason is that due to the construction of the gene trees, used for `ALF` simulations, the distribution of branch lengths, and hence, the distribution of duplications among the species tree, is very heterogeneous. The average percentage of short branches (for which less than 1 duplication is expected, using a duplication rate of 0.005 and n gene families) is ranging from 11.3% (5 species, 500 gene families) to 33.6% (20 species, 100 gene families). Note, that the lack of duplications leads to species trees that are not fully resolved, and hence have a larger distance to the generated trees used for the simulation. Figures S2 (first simulation method) and S3 (`ALF` simulations) show boxplots for the four tree distances as a function of the number of independent gene families.

The complete results for the 2000 simulated data sets of 10 species and 100, resp. 1000 gene families with a varying amount of noise are depicted in Figures S4 (first simulation method) and S5 (`ALF`).

The results for simulated data sets with horizontal gene transfer show that our method is very robust against noise introduced by horizontal gene transfer, which appears as mis-predicted orthology. Even xenologous noise of up to 39.5% of the homologous pairs had only a minor impact on the obtained tree distances. The triple support values s for the reconstructed species trees, which ranges between 0.978 (HGT rate 0.0025) and 0.943 (HGT rate 0.0075). This shows that only very few false species triples have been inferred. However, these triples could be excluded during the computation of the maximal consistent subset, as they are usually dominated by the amount of correctly identified species triples. The small differences between generated and reconstructed

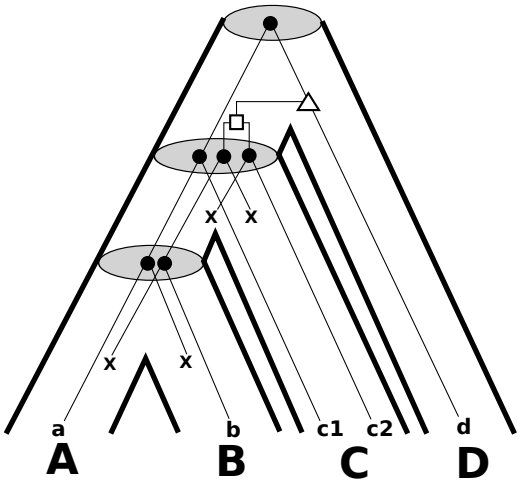


Figure S1: Shown is a gene tree T on $\mathfrak{G} = \{a, b, c1, c2, d\}$ evolving along species tree S on $\mathfrak{S} = \{A, B, C, D\}$. In this scenario false gene triples in $\mathbb{G}$ and thus, false species triples in $\mathbb{S}$ are introduced, due to the HGT-event ($\triangle$) followed by a duplication event ($\square$) and certain losses ($\mathbf{x}$). Here, we obtain that $\mathsf{bc2|a} \in \mathbb{G}$ and thus $\mathsf{BC|A} \in \mathbb{S}$, contradicting that $\mathsf{AB|C} \in \mathfrak{R}(S)$.

species trees can be explained by the fact that the method forces homologous genes within the same species to be paralogous, although, due to horizontal gene transfer their lowest common ancestor can be a speciation event. This leads to the estimated orthology not being a cograph, introducing errors during the cograph editing step. Figure S6 shows boxplots for the tree distance as a function of the percentage of xenologous noise.

S 4.3 Real-life Data

Figure S7 depicts the phylogenetic tree of *Aquificales* species obtained from paralogy data in comparison to the tree suggested by [44]. The trees obtained from bootstrapping experiments are given in Figure S8. The majority-rule consensus trees for both bootstrapping approaches are identical to the previously computed tree. However, the bootstrap support appears to be smaller next to the leaves. This is in particular the case for closely related species with only a few duplicated genes exclusively found in one of the species.

Figure S9 depicts the phylogenetic tree of *Enterobacteriales* species obtained from paralogy data in comparison to the tree from `PATRIC` database [52]. The trees obtained from bootstrapping experiments are given in Figure S10. When assuming the `PATRIC` to be correct, then the subtree support values appear to be a much more reliable indicator, compared to the bootstrap values.

S 4.4 Additional Comments on Running Time

The CPLEX Optimizer is capable of solving instances with approximately a few thousand variables. As the ILP formulation for cograph editing requires $O(|\mathfrak{G}|^2)$ many variables, we can solve instances with up to 100 genes per connected component in G_Θ. However, for our computations we limit the size of each component to 50 genes. Furthermore, the ILP formulations for finding the maximal consistent triple set and least resolved species tree requires $O(|\mathfrak{S}|^3)$ many variables. Hence, problem instances of up to about 20 species can be processed.

Table S 4.4 shows the runtimes for simulated and real-life data sets for each individual sub-task. Note that the time used for cograph editing is quite high, compared to the other sub-tasks. This is due to the fact, that cograph editing if performed for each connected component in G_Θ individually, and initializing the ILP solver is a relevant factor. In the implementation we first perform a check, if for a given gene family cograph editing has to be performed. Triple extraction is performed with a polynomial time algorithm. Another oddity is the extraordinary short runtime for the computation of the maximal consistent subset of species triples in the *Enterobacteriales* data set. During the bootstrapping experiments for this set much longer times were observed.

[1] Total time includes triple extraction, parsing input, and writhing output files.
[2] Average of 2000 simulations generated with `ALF`, 10 species, 1000 gene families.
[3] 2,000,000 cographs, 41 not optimally solved within time limit of 30 min.
[4] In 95.95% of the simulations the least resolved tree could be found using `BUILD`.
[5] A unique tree was obtained using `BUILD`. Second value indicates runtime with ILP solving enforced.
[6] Note that the bootstrap computations had a much longer runtime (125 sec. on average).

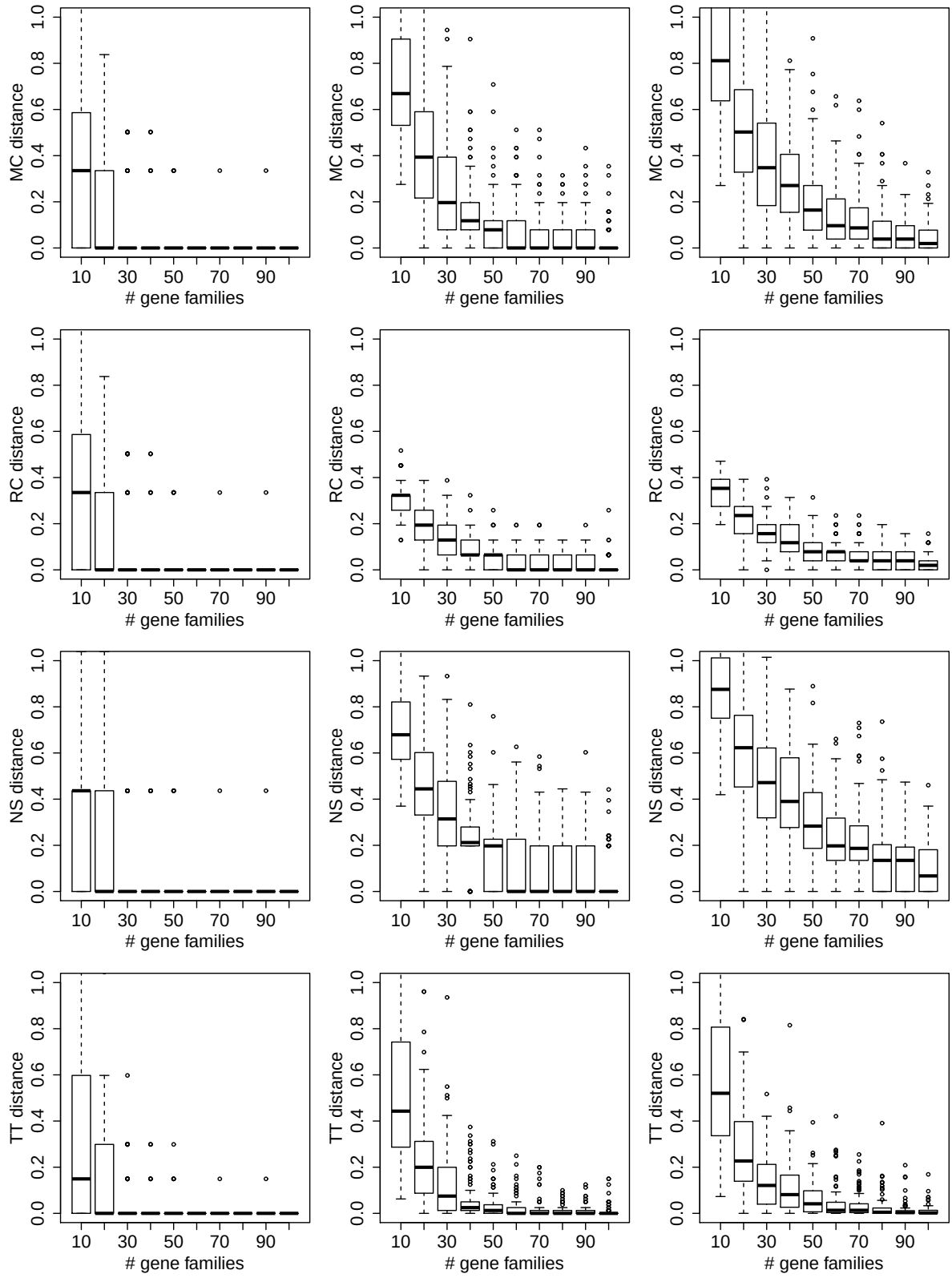


Figure S2: Matching Cluster (MC), Robinson-Foulds (RC), Nodal Splitted (NS) and Triple metric (TT) tree distances of 100 reconstructed phylogenetic trees with (from left to right) five, ten, and 15 species and 10 to 100 gene families, each. Simulations are generated with first simulation method.

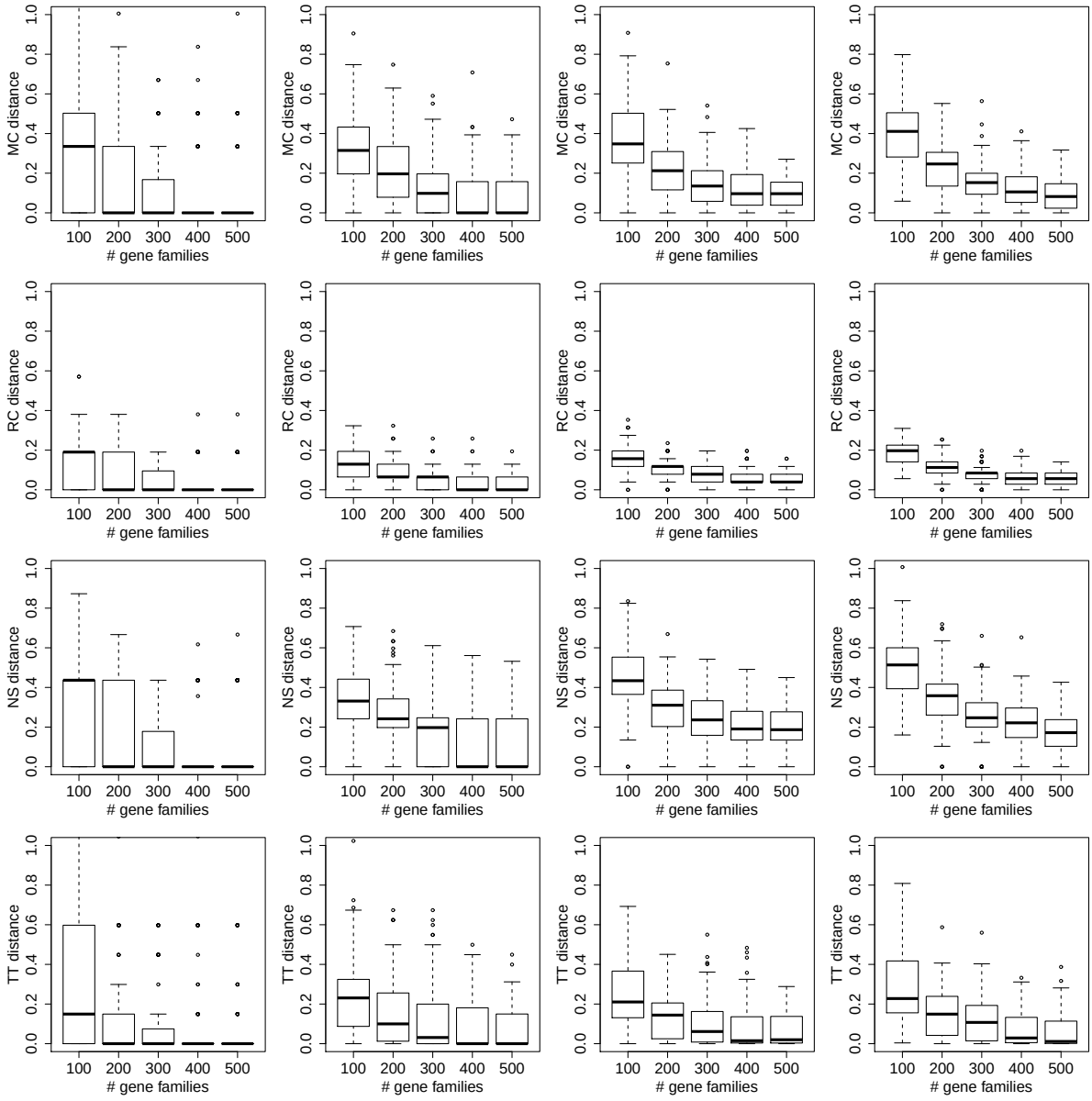


Figure S3: Matching Cluster (MC), Robinson-Foulds (RC), Nodal Splitted (NS) and Triple metric (TT) tree distances of 100 reconstructed phylogenetic trees with (from left to right) five, ten, 15, and 20 species and 100 to 500 gene families, each. Simulations are generated with `ALF`.

Data	CE	MCS	LRT	Total[1]
Simulations[2]	125[3]	< 1	< 1[4]	126
Aquificales	34	< 1	< 1 (6)[5]	34
Enterobacteriales	2673	2[6]	< 1 (1749)[5]	2676

Table S1: Running time in seconds on 2 Six-Core AMD Opteron™ Processors with 2.6GHz for individual sub-tasks: **CE** cograph editing, **MCS** maximal consistent subset of triples, **LRT** least resolved tree.

References

[1] A. V. Aho, Y. Sagiv, T. G. Szymanski, and J. D. Ullman. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM J. Comput.*, 10:405–421, 1981.

[2] A. M. Altenhoff, A. Schneider, G. H. Gonnet, and C. Dessimoz. OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res.*, 39(Database issue):D289–294, Jan 2011.

[3] L. Arvestad, A. C. Berglund, J. Lagergren, and B. Sennblad. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics*, 19:i7–i15, 2003.

[4] M. S. Bansal and O. Eulenstein. The multiple gene duplication problem revisited. *Bioinformatics*, 24:i132–i138, 2008.

[5] O.R.P Bininda-Emonds. *Phylogenetic Supertrees.* Kluwer Academic Press, Dordrecht, The Netherlands, 2004.

[6] Sebastian Böcker, David Bryant, Andreas W.M. Dress, and Mike A. Steel. Algorithmic aspects of tree amalgamation. *Journal of Algorithms*, 37(2):522 – 537, 2000.

[7] Sebastian Böcker and Andreas W. M. Dress. Recovering symbolically dated, rooted trees from symbolic ultrametrics. *Adv. Math.*, 138:105–125, 1998.

[8] Damian Bogdanowicz, Krzysztof Giaro, and Borys Wróbel. Treecmp: Comparison of trees in polynomial time. *Evolutionary Bioinformatics Online*, 8:475, 2012.

[9] Paola Bonizzoni, Gianluca Della Vedova, and Riccardo Dondi. Reconciling a gene tree to a species tree under the duplication cost model. *Theor. Comp. Sci.*, 347:36–53, 2005.

[10] Andreas Brandstädt, Van Bang Le, and Jeremy P Spinrad. *Graph Classes: A Survey.* SIAM Monographs on Discrete Mathematics and Applications. Soc. Ind. Appl. Math., Philadephia, 1999.

[11] D. Bryant. *Building trees, hunting for trees, and comparing trees: theory and methods in phylogenetic analysis.* PhD thesis, University of Canterbury, 1997.

[12] D. Bryant and M. Steel. Extension operations on sets of leaf-labelled trees. *Adv. Appl. Math.*, 16(4):425–453, 1995.

[13] J. G. Burleigh, M. S. Bansal, A. Wehe, and O. Eulenstein. Locating large-scale gene duplication events through reconciled trees: implications for identifying ancient polyploidy events in plants. *J. Comput. Biol.*, 16:1071–1083, 2009.

[14] J. Byrka, P. Gawrychowski, K. T. Huber, and S. Kelk. Worst-case optimal approximation algorithms for maximizing triplet consistency within phylogenetic networks. *J. Discr. Alg.*, 8:65–75, 2010.

[15] Jaroslaw Byrka, Sylvain Guillemot, and Jesper Jansson. New results on optimizing rooted triplets consistency. *Discr. Appl. Math.*, 158:1136–1147, 2010.

[16] Wen-Chieh Chang, Gordon J Burleigh, David F Fernández-Baca, and Oliver Eulenstein. An ilp solution for the gene duplication problem. *BMC bioinformatics*, 12(Suppl 1):S14, 2011.

[17] C. Chauve, J. P. Doyon, and N. El-Mabrouk. Gene family evolution by duplication, speciation, and loss. *J. Comput. Biol.*, 15:1043–1062, 2008.

[18] D. G. Corneil, H. Lerchs, and L. Steward Burlingham. Complement reducible graphs. *Discr. Appl. Math.*, 3:163–174, 1981.

[19] D. G. Corneil, Y. Perl, and L. K. Stewart. A linear recognition algorithm for cographs. *SIAM J. Computing*, 14:926–934, 1985.

[20] D. A. Dalquen, M. Anisimova, G. H. Gonnet, and C. Dessimoz. ALF–a simulation framework for genome evolution. *Mol. Biol. Evol.*, 29(4):1115–1123, Apr 2012.

[21] Daniel A. Dalquen, Adrian M. Altenhoff, Gaston H. Gonnet, and Christophe Dessimoz. The impact of gene duplication, insertion, deletion, lateral gene transfer and sequencing error on orthology inference: A simulation study. *PLoS ONE*, 8(2):e56925, 02 2013.

[22] M. C. H. Dekker. Reconstruction methods for derivation trees. Master's thesis, Vrije Universiteit, Amsterdam, Netherlands, 1986.

[23] Reinhard Diestel. *Graph Theory, 4th Edition*, volume 173 of *Graduate texts in mathematics*. Springer, 2012.

[24] Jean-Philippe Doyon, Cedric Chauve, and Sylvie Hamel. Space of gene/species trees reconciliations and parsimonious models. *J. Comp. Biol.*, 16:1399–1418, 2009.

[25] Andreas W. M. Dress, Katharina T. Huber, Jacobus Koolen, Vincent Moulton, and Andreas Spillner. *Basic phylogenetic combinatorics*. Cambridge University Press, 2012.

[26] P. Górecki and Tiuryn J. DSL-trees: A model of evolutionary scenarios. *Theor. Comp. Sci.*, 359:378–399, 2006.

[27] J Grilli, M Romano, F Bassetti, and M Cosentino Lagomarsino. Cross-species gene-family fluctuations reveal the dynamics of horizontal transfers. *Nucleic Acids Res.*, 42:6850–6860, 2014.

[28] Stefan Grünewald, Mike Steel, and M. Shel Swenson. Closure operations in phylogenetics. *Mathematical Biosciences*, 208(2):521 – 537, 2007.

[29] R. Guigó, I. Muchnik, and T. F. Smith. Reconstruction of ancient molecular phylogeny. *Mol. Phylogenet. Evol.*, 6:189–213, 1996.

[30] Sylvain Guillemot and Matthias Mnich. Kernel and fast algorithm for dense triplet inconsistency. *Theoretical Computer Science*, 494(0):134 – 143, 2013. Theory and Applications of Models of Computation (TAMC 2010).

[31] Michel Habib and Christophe Paul. A simple linear time algorithm for cograph recognition. *Discrete Applied Mathematics*, 145(2):183–197, 2005.

[32] M. W. Hahn. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol.*, 8:R141, 2007.

[33] Marc Hellmuth, Maribel Hernandez-Rosales, Katharina T. Huber, Vincent Moulton, Peter F. Stadler, and Nicolas Wieseke. Orthology relations, symbolic ultrametrics, and cographs. *Journal of Mathematical Biology*, 66(1-2):399–420, 2013.

[34] M. Hernandez-Rosales, M. Hellmuth, N. Wieseke, and P. F. Stadler. Simulation of gene family histories. *BMC Bioinformatics*, 15(Suppl 3):A8, 2014.

[35] Maribel Hernandez-Rosales, Marc Hellmuth, Nicolas Wieseke, Katharina T Huber, Vincent Moulton, and Peter F Stadler. From event-labeled gene trees to species trees. *BMC Bioinformatics*, 13(Suppl 19):S6, 2012.

[36] Katharina T Huber, Vincent Moulton, Charles Semple, and M Steel. Recovering a phylogenetic tree using pairwise closure operations. *Applied mathematics letters*, 18(3):361–366, 2005.

[37] Daniel H Huson, Regula Rupp, and Celine Scornavacca. *Phylogenetic networks: concepts, algorithms and applications*. Cambridge University Press, 2010.

[38] J Jansson, J. H.-K. Ng, K. Sadakane, and W.-K. Sung. Rooted maximum agreement supertrees. *Algorithmica*, 43:293–307, 2005.

[39] Jesper Jansson. On the complexity of inferring rooted evolutionary trees. *Electronic Notes Discr. Math.*, 7:50–53, 2001.

[40] Jesper Jansson, Richard S. Lemence, and Andrzej Lingas. The complexity of inferring a minimally resolved phylogenetic supertree. *SIAM J. Comput.*, 41:272–291, 2012.

[41] K S Jaron, J C Moravec, and N Martıınková. `SigHunt`: horizontal gene transfer finder optimized for eukaryotic genomes. *Bioinformatics*, 30:1081–1086, 2014.

[42] B. R. Larget, S. K. Kotha, C. N. Dewey, and C. Ane. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics*, 26:2910–2911, 2010.

[43] Marcus Lechner, Sven Findeiß, Lydia Steiner, Manja Marz, Peter F. Stadler, and Sonja J. Prohaska. Proteinortho: Detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics*, 12:124, 2011.

[44] Marcus Lechner, Astrid Nickel, Stefanie Wehner, Konstantin Riege, Nicolas Wieseke, Benedikt Beckmann, Roland Hartmann, and Manja Marz. Genomewide comparison and novel ncrnas of aquificales. *BMC Genomics*, 15(1):522, 2014.

[45] Yunlong Liu, Jianxin Wang, Jiong Guo, and Jianer Chen. Cograph editing: Complexity and parametrized algorithms. In B. Fu and D. Z. Du, editors, *COCOON 2011*, volume 6842 of *Lect. Notes Comp. Sci.*, pages 110–121, Berlin, Heidelberg, 2011. Springer-Verlag.

[46] Yunlong Liu, Jianxin Wang, Jiong Guo, and Jianer Chen. Complexity and parameterized algorithms for cograph editing. *Theoretical Computer Science*, 461(0):45 – 54, 2012.

[47] R. D. Page and M. A. Charleston. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol. Phylogenet. Evol.*, 7:231–240, 1997.

[48] Monika Rauch Henzinger, Valerie King, and Tandy Warnow. Constructing a tree from homeomorphic subtrees, with applications to computational evolutionary biology. *Algorithmica*, 24:1–13, 1999.

[49] Charles Semple and Mike Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, UK, 2003.

[50] A Tsirigos and I Rigoutsos. A new computational method for the detection of horizontal gene transfer events. *Nucleic Acids Res.*, 33:922–933, 2005.

[51] L. van Iersel, S. Kelk, and M. Mnich. Uniqueness, intractability and exact algorithms: reflections on level-k phylogenetic networks. *J. Bioinf. Comp. Biol.*, 7:597–623, 2009.

[52] Wattam et al. Patric, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Research*, 2013.

[53] Bang Ye Wu. Constructing the maximum consensus tree from rooted triples. *J. Comb. Optimization*, 8:29–39, 2004.

[54] G. U. Yule. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Phil. Trans. R. Soc. B*, 213:21–87, 1925.

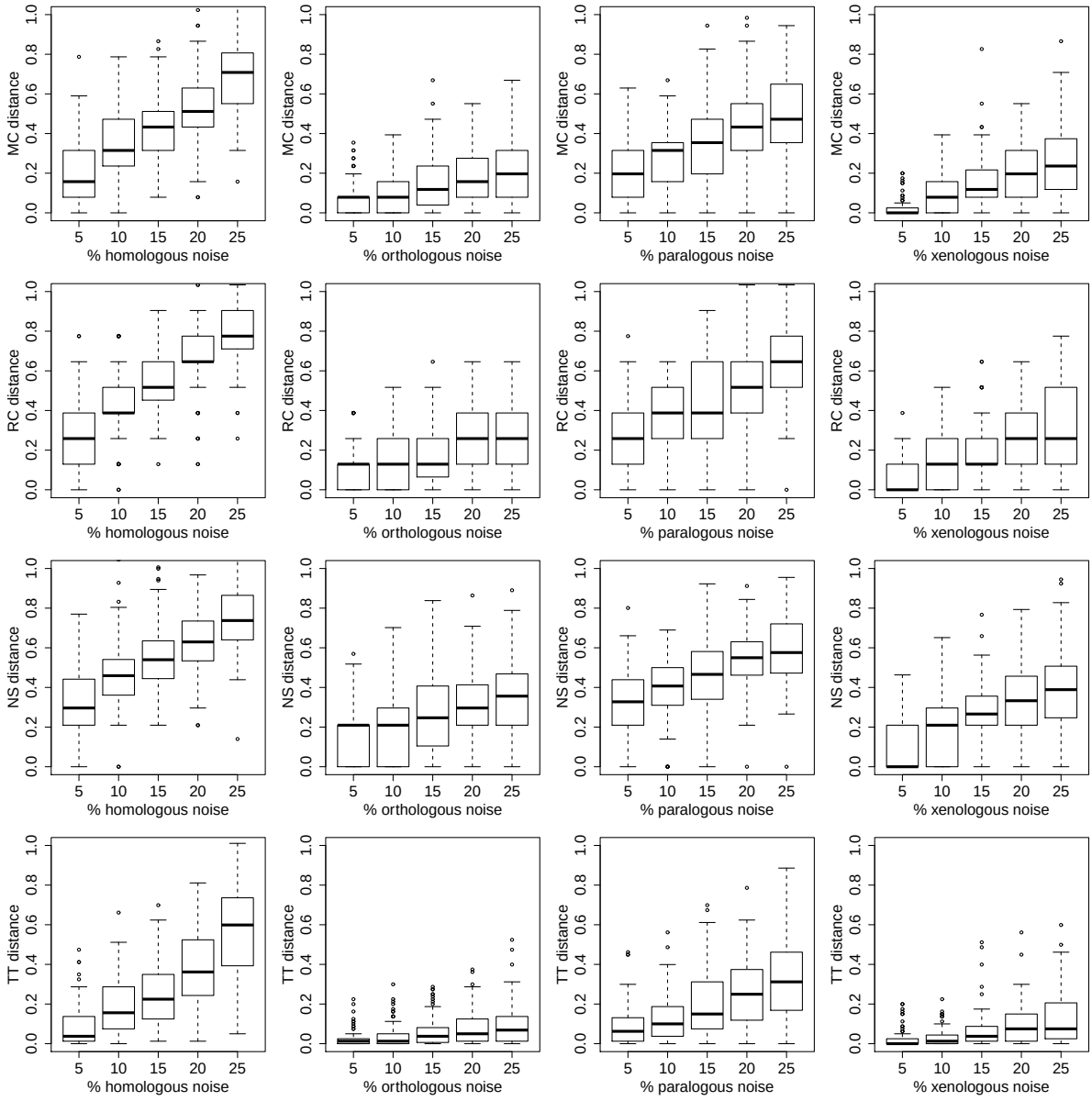


Figure S4: Matching Cluster (MC), Robinson-Foulds (RC), Nodal Splitted (NS) and Triple metric (TT) tree distances of 100 reconstructed phylogenetic trees with ten species and 100 gene families generated with first simulation method. For each model noise was added with a probability of 0.05 to 0.25.

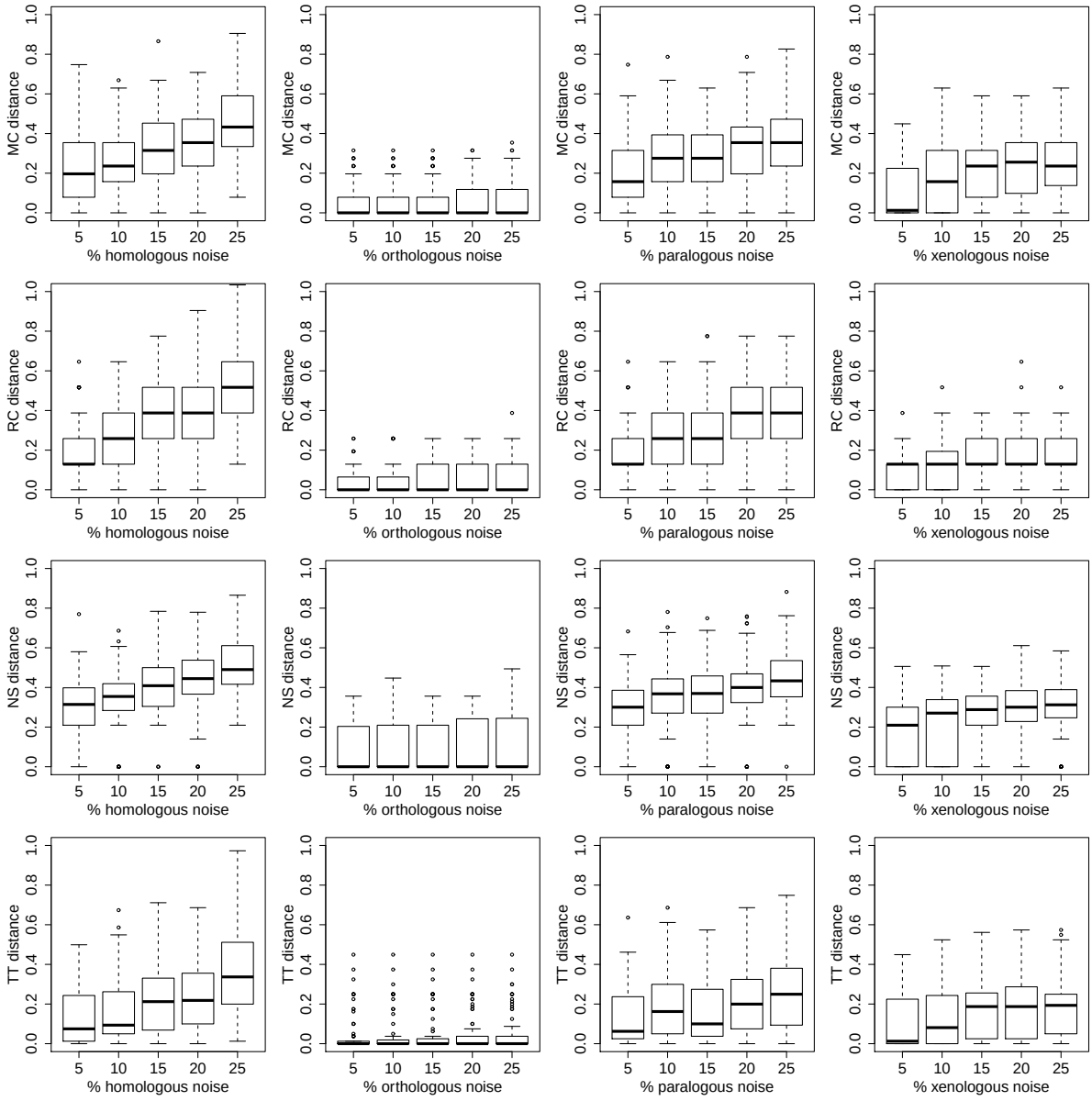
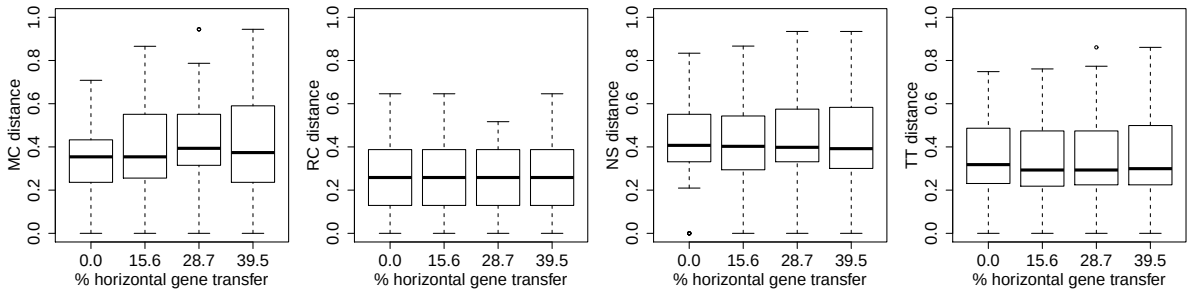
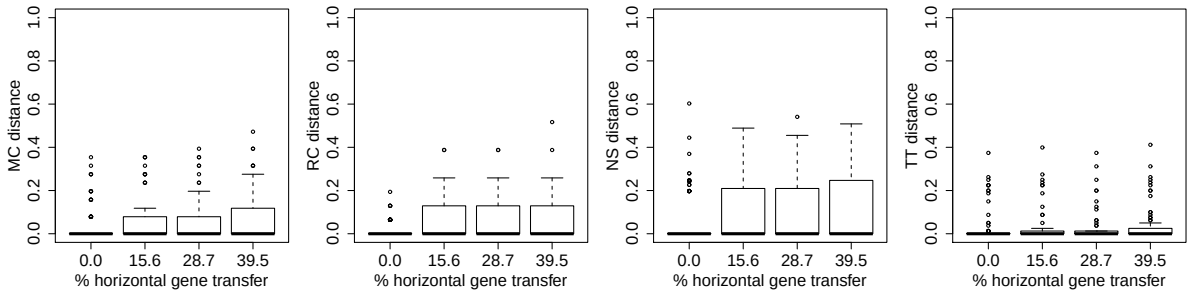


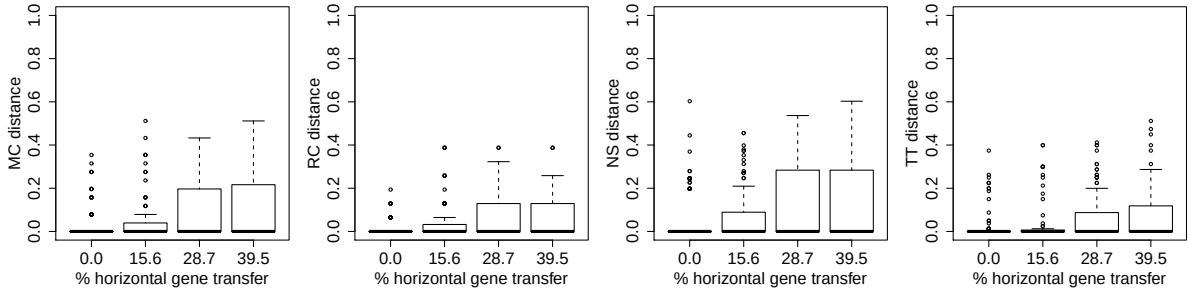
Figure S5: Matching Cluster (MC), Robinson-Foulds (RC), Nodal Splitted (NS) and Triple metric (TT) tree distances of 100 reconstructed phylogenetic trees with ten species and 1000 gene families generated with `ALF`. For each model noise was added with a probability of 0.05 to 0.25.



(A)



(B)



(C)

Figure S6: Matching Cluster (MC), Robinson-Foulds (RC), Nodal Splitted (NS) and Triple metric (TT) tree distances of 100 reconstructed phylogenetic trees with ten species. ALF simulations are performed with duplication/loss rates of 0.005 $\cong$ 6.1% and hgt rates of 0.0025 to 0.0075, resulting in xenologous noise between 0.0% to 39.5%. Reconstructions are based on (A) `Proteinortho` orthology estimation, (B) perfect paralogy knowledge, and (C) perfect orthology knowledge.

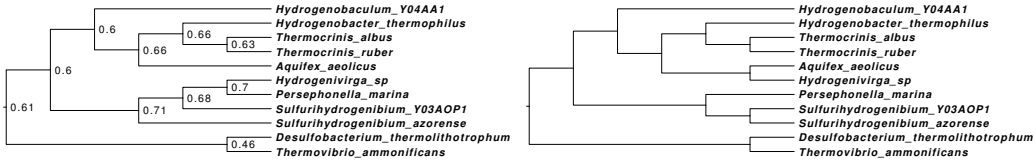


Figure S7: Phylogenetic tree of eleven *Aquificales* species. L.h.s.: tree computed from paralogy data. Internal node labels indicate support of subtrees. R.h.s.: reference tree from [44].

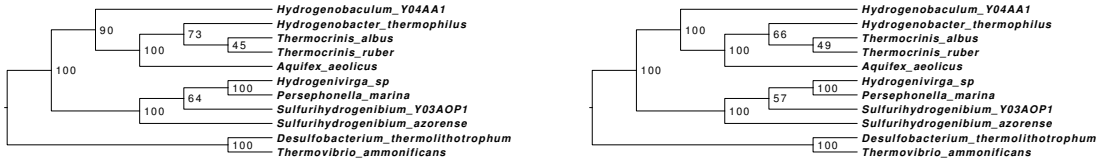


Figure S8: Cograph-based (l.h.s.) and triple-based (r.h.s.) bootstrapping trees of eleven *Aquificales* species.

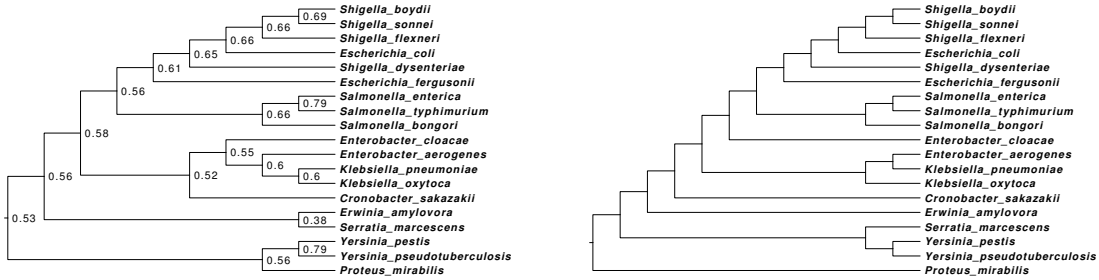


Figure S9: Phylogenetic trees of 19 *Enterobacteriales* species. L.h.s.: tree computed from paralogy data. Internal node labels indicate support of subtrees. R.h.s.: reference tree from `PATRIC` database, projected to the 19 considered species. *Salmonella typhimurium* is missing in `PATRIC` tree.

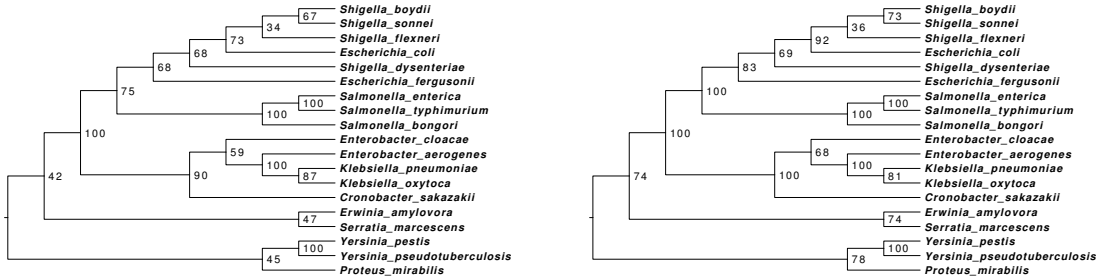


Figure S10: Cograph-based (l.h.s.) and triple-based (r.h.s.) bootstrapping trees of 19 *Enterobacteriales* species.