Phylogenetics from Paralogs

Marc Hellmuth
Nicolas Wieseke
Markus Lechner
Hans-Peter Lenhof
Martin Middledorf

SFI WORKING PAPER: 2014-04-008

SANTA FE INSTITUTE

Phylogenetics from 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

Motivation: Sequence-based phylogenetic approaches heavily rely on initial data sets to be composed of orthologous sequences only. Paralogs are treated as a dangerous nuisance that has to be detected and removed. Recent advances in mathematical phylogenetics, however, have indicated that gene duplications can also convey meaningful phylogenetic information provided orthologs and paralogs can be distinguished with a degree of certainty.

Results: We demonstrate that plausible phylogenetic trees can be inferred from paralogy information only. To this end, tree-free estimates of orthology, the complement of paralogy, are first corrected to conform cographs and then translated into equivalent event-labeled gene phylogenies. A certain subset of the triples displayed by these trees translates into constraints on the species trees. While the resolution is very poor for individual gene families, we observe that genome-wide data sets are sufficient to generate fully resolved phylogenetic trees of several groups of eubacteria. The novel method introduced here relies on solving three intertwined NP-hard optimization problems: the cograph editing problem, the maximum consistent triple set problem, and the least resolved tree problem. Implemented as Integer Linear Program, paralogy-based phylogenies can be computed exactly for up to some twenty species and their complete protein complements.

Availability: The ILP formulation is implemented in the Software `ParaPhylo` using IBM ILOG CPLEX™ Optimizer 12.6 and is freely available from `http://pacosy.informatik.uni-leipzig.de/paraphylo`.

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` (Ebersberger et al., 2009), are used to filter genome-wide data sets to at least deplete sequences with detectable paralogs (homologs in the same species).

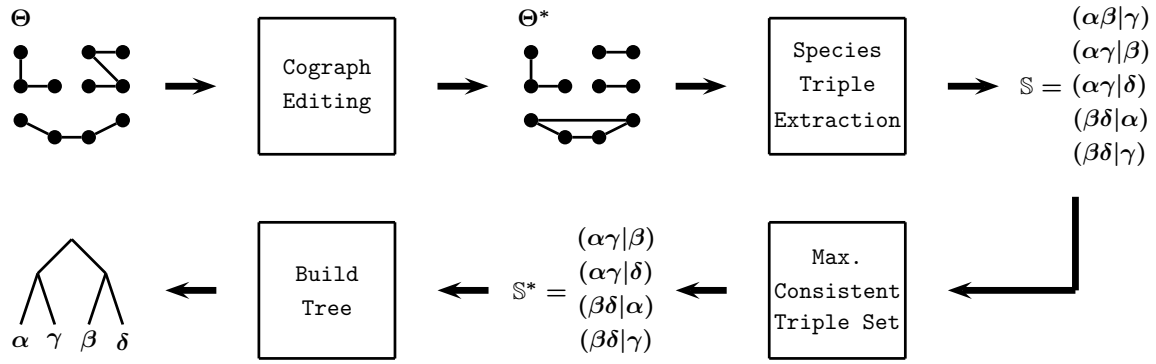


Figure 1: Outline of computational framework. Starting from an estimated orthology relation Θ, its graph representation G_Θ is edited to obtain the closest cograph G_{Θ^*}, which in turn is equivalent to a (not necessarily fully resolved) gene tree T and an event labeling t. From (T, t) we extract the 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 $\mathbb{S}^*$.

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 Fitch (2000) and Gabaldón and Koonin (2013) 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 (Altenhoff and Dessimoz, 2009). Many tools of this type have become available over the last decade, see Kristensen et al. (2011) for a recent review. The output of such methods is an estimate Θ of the 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), since if all genes are pairwise orthologs, then the corresponding minimally resolved gene tree is a star. Building upon the theory of symbolic ultrametrics (Böcker and Dress, 1998) we showed that a symmetric relation R on a set of genes is an orthology relation if and only if R yields a cograph (Hellmuth et al., 2013). Furthermore, the corresponding cotree, which can be efficiently computed 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). Hernandez-Rosales et al. (2012) then showed that certain triples of genes from three different species must also be displayed in the species tree, and thus provide direct information on the structure of the species tree. Estimates of 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 orthology relation Θ^*. Work on the cograph-editing problem, which asks for the cograph most similar to an arbitrary input graph (Liu et al., 2011, 2012), 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. A conceptually elegant solution is provided by the theory of supertrees in the form of the largest set of consistent triples (Jansson et al., 2005). 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 (Liu et al., 2012), maximal consistent triple set (Wu, 2004; Jansson, 2001; Jansson et al., 2005), and least resolved supertree (Jansson et al., 2012)) 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.

2 Theory

2.1 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 $T(\rho_T) = T$. There is a well-known one-to-one correspondence between phylogenetic trees and *hierarchies* on L, i.e., systems $\mathcal{C}$ of subsets of L satisfying (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. More precisely, 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, see (Semple and Steel, 2003).

Rooted Triples: Rooted triples (Dress et al., 2012) are a key concept in the theory of supertrees (Semple and Steel, 2003; Bininda-Emonds, 2004). 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` (Aho et al., 1981) 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 (Jansson et al., 2012). 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 (Byrka et al., 2010a; van Iersel et al., 2009). Polynomial-time approximation algorithms for this problem and further theoretical results are reviewed by (Byrka et al., 2010b).

2.2 Triple Closure Operations and Inference Rules

If R is consistent it is often possibe 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 $\mathrm{cl}(R) = \cap_{T \in \langle R \rangle} \mathfrak{R}(T)$.

This operation, which has been extensively studied in the last decades (Bryant and Steel, 1995; Grünewald et al., 2007; Bryant, 1997; Huber et al., 2005; Böcker et al., 2000), satisfies the usual three properties of a closure operator: (i) $R \subseteq \mathrm{cl}(R)$; (ii) $\mathrm{cl}(\mathrm{cl}(R)) = \mathrm{cl}(R)$ and (iii) if $R' \subseteq R$, then $\mathrm{cl}(R') \subseteq \mathrm{cl}(R)$. We say R is *closed* if $R = \mathrm{cl}(R)$. Obviously, $\mathfrak{R}(T)$ is closed. We write $R \vdash (\mathsf{xy}|\mathsf{z})$ iff $(\mathsf{xy}|\mathsf{z}) \in \mathrm{cl}(R)$.

A brute force computation of the closure of a given consistent set R runs in $O(|R|^5)$ time (Bryant and Steel, 1995): For any three leaves in L_R test whether $R \cup \{r\}$ is consistent for exactly one of the three possible triples; if so, r is added to the closure. Extending earlier work of Dekker (1986), Bryant and Steel (1995) 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:

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

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. In the Supplemental Material we prove the following two key results that will play an important role in 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.*

2.3 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. Two genes $x, y \in \mathfrak{G}$ are paralogs if and only if $x \neq y$ and $(x, y) \notin \Theta$. 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. 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 (Hellmuth et al., 2013), 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 Brandstädt et al. (1999) for a survey of cographs and many other equivalent characterizations. Cographs can be recognized in linear time (Corneil et al., 1985; Habib and Paul, 2005). 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 a NP-complete problem (Liu et al., 2011, 2012). 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 7 in the Supplemental Material, it suffices to solve the cograph editing for each connected component separately.

2.4 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-Rosales et al., 2012 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 Hernandez-Rosales et al. (2012) 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-complete.

3 ILP Formulation, Implementation & Data

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 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 more detailed description proving 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 $p \in \{1, \ldots, \|\mathfrak{S}\| - 2\}$.
$N_{\alpha\beta,p}$	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.

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 with the additional constraint that $(x, y) \notin E_{\Theta^*}$ whenever $\sigma(x) = \sigma(y)$, i.e., no pair of orthologs is found in the same species. While, we use binary variables E_{xy} to express whether or not $(x, y) \in E_{\Theta^*}$, the input relation Θ is represented by the binary constants $\Theta_{ab} = 1$ iff $(a, b) \in \Theta$. In the weighted variant of the problem, $\Theta \in [0, 1]$ is interpreted as a confidence in the orthology assignment. The minimization of the edge edit distance between Θ and Θ^* can be expressed as

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

Since $E_{xy} \equiv E_{yx}$ we use these variables interchangeably. Consistency with σ is enforced by

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

The condition that G_{Θ^*} is a cograph is readily expressed by forbidding P_4 as induced subgraph on all quadruples. This amounts to the constraints

$$E_{wx} + E_{xy} + E_{yz} - E_{xz} - E_{wy} - E_{wz} \leq 2 \qquad \texttt{(ILP 3)}$$

for all ordered tuples (w, x, y, z) of four distinct indices $w, x, y, z \in \mathfrak{G}$. In summary, $O(|\mathfrak{G}|^2)$ binary variables are required and Equations (ILP 2) and (ILP 3) establish $O(|\mathfrak{G}|^4)$ constraints. In practice, the effort is not dominated by the number of edges, since the connected components of G_Θ can be treated independently.

3.2 Extraction of All Species Triples

The construction of the species tree S is based upon the set $\mathbb{S}$ of species triples, which we encode 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 $\alpha, \beta, \gamma \in \mathfrak{S}$.

The key observation is that $(xy|z)$ has a speciation vertex at its root node iff (x, z) and (y, z) are orthologs, i.e., if $E_{xz} = 1$ and $E_{yz} = 1$. We show in the Supplemental Material that the following constraints for all

pairwise distinct species $\alpha, \beta, \gamma, \delta \in \mathfrak{S}$ and all $\sigma(x) = \alpha$, $\sigma(y) = \beta$, and $\sigma(z) = \gamma$ restrict $\mathbb{S}$ to the triples derived from $\mathbb{G}$:

$$\begin{aligned} (1 - E_{xy}) + E_{xz} + E_{yz} - T_{(\alpha\beta|\gamma)} &\leq 2 \\ E_{xy} + (1 - E_{xz}) + E_{yz} - T_{(\alpha\gamma|\beta)} &\leq 2 \\ E_{xy} + E_{xz} + (1 - E_{yz}) - T_{(\beta\gamma|\alpha)} &\leq 2 \\ T_{(\alpha\delta|\gamma)} + T_{(\beta\delta|\gamma)} - T_{(\alpha\beta|\gamma)} &\leq 1 \end{aligned} \tag{ILP 4}$$

In order to remove the remaining degrees of freedom in the choice of the binary value $T_{(\alpha\beta|\gamma)}$ for the triples $(\alpha\beta|\gamma) \notin \mathbb{G}$ we add the objective function

$$\min \sum_{\{\alpha,\beta,\gamma\} \in \binom{S}{3}} T_{(\alpha\beta|\gamma)} + T_{(\alpha\gamma|\beta)} + T_{(\beta\gamma|\alpha)} \tag{ILP 5}$$

This ILP formulation requires $O(|\mathfrak{S}|^3)$ variables and $O(|\mathfrak{G}|^3 + |\mathfrak{S}|^4)$ constraints.

3.3 Find Maximal Consistent Triple Set

Chang et al. (2011) proposed an ILP approach to find maximal consistent triple sets. 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 6 (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 $|\mathbb{S}' \cap \mathbb{S}|$.

In complete analogy to the previous subsection we define variables $T'_{(\alpha\beta|\gamma)} = 1$ iff $(\alpha\beta|\gamma) \in \mathbb{S}'$. For any three pairwise distinct $\alpha, \beta, \gamma \in \mathfrak{S}$ there are three variables $T'_{(\alpha\beta|\gamma)}$, $T'_{(\alpha\gamma|\beta)}$, and $T'_{(\beta\gamma|\alpha)}$. Strict density of $\mathbb{S}'$ implies that it contains exactly one of the possible three triples for any three species, i.e.,

$$T'_{(\alpha\beta|\gamma)} + T'_{(\alpha\gamma|\beta)} + T'_{(\beta\gamma|\alpha)} = 1. \tag{ILP 6}$$

As a consequence of Theorem 1, we can use the 2-order inference rules (i)-(iii) to ensure that $\mathbb{S}'$ is consistent. These can be expressed in the language of ILP in the following form. For all pairwise distinct $\alpha, \beta, \gamma, \delta \in \mathfrak{S}$ we set:

$$\begin{aligned} T'_{(\alpha\beta|\gamma)} + T'_{(\alpha\delta|\gamma)} - T'_{(\beta\delta|\gamma)} &\leq 1. \\ 2T'_{(\alpha\beta|\gamma)} + 2T'_{(\alpha\delta|\beta)} - T'_{(\beta\delta|\gamma)} - T'_{(\alpha\delta|\gamma)} &\leq 2 \\ 2T'_{(\alpha\beta|\gamma)} + 2T'_{(\gamma\delta|\beta)} - T'_{(\alpha\beta|\delta)} - T'_{(\gamma\delta|\alpha)} &\leq 2 \end{aligned} \tag{ILP 7}$$

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

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

The intersection $\mathbb{S}^* = \mathbb{S}' \cap \mathbb{S}$ is expressed by another set of binary variables $T^*_{(\alpha\beta|\gamma)} = 1$ iff $(\alpha\beta|\gamma) \in \mathbb{S}^*$ restricted by the following constraints.

$$0 \leq T'_{(\alpha\beta|\gamma)} + T_{(\alpha\beta|\gamma)} - 2T^*_{(\alpha\beta|\gamma)} \leq 1 \tag{ILP 9}$$

Here, we require $O(|\mathfrak{S}|^3)$ variables and $O(|\mathfrak{S}|^4)$ constraints.

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 8)

$$\max \sum_{(\alpha\beta|\gamma) \in \mathbb{S}} T'_{(\alpha\beta|\gamma)} * w(\alpha\beta|\gamma). \tag{ILP 10}$$

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

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 Chang et al. (2011).

To build an arbitrary tree for the consistent triple set $\mathbb{S}^*$, one can use one of the fast implementations of `BUILD` (Semple and Steel, 2003). 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.

Since a phylogenetic tree S is equivalently specified by its hierarchy $\mathcal{C} = \{L(v) \mid v \in V(S)\}$ we construct the clusters induced by all triples of $\mathbb{S}^*$ and check whether they form a hierarchy on $\mathfrak{S}$. Following (Chang et al., 2011), 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 Chang et al. (2011), 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. In other words $N_{\alpha\beta,p} = 1$ iff $M_{\alpha p} = 1$ and $M_{\beta p} = 1$, which can be expressed as

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

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

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

In the Supplemental Material we proof that Eq. (`ILP 12`) ensures the existence of at least one cluster p that contains α and β but not γ, for each triple $(\alpha\beta|\gamma) \in \mathbb{S}^*$.

We do not insist on the existence of a cluster q that contains γ but not α and β for every triple $(\alpha\beta|\gamma)$. Thus we do not necessarily construct singleton clusters. Moreover, there is no constraint that decodes for a complete cluster $p = \{\mathfrak{S}\}$ in M. Instead, we only need to capture that M defines a "partial" hierarchy, i.e., any two clusters satisfy $p \cap q \in \{p, q, \emptyset\}$. We use the "three-gamete condition" (Chang et al., 2011) for this purpose. For each gamete $\Gamma\Lambda \in \{01, 10, 11\}$ 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 require

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

These constraints capture that $C_{p,q,\Gamma\Lambda} = 1$ as long as $M_{\alpha p} = \Gamma$ and $M_{\alpha q} = \Lambda$ for some $\alpha \in \mathfrak{S}$. To ensure compatibility of the clusters the constraints

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

are enforced for all p, q. 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. The corresponding constraints are

$$0 \leq Y_p|\mathfrak{S}| - \sum_{\alpha\in\mathfrak{S}} M_{\alpha p} \leq |\mathfrak{S}| - 1. \tag{ILP 15}$$

Finally, in order to minimize the number of non-trivial columns in M, and thus the number of inner vertices in the respective tree, we add the objective function

$$\min \sum_p Y_p \tag{ILP 16}$$

This ILP uses $O(|\mathfrak{S}|^3)$ variables and constraints.

3.5 Implementation Details and Test Data

ILP Solver: The ILP approach is implemented using IBM ILOG CPLEX$^{\text{TM}}$ Optimizer 12.6 in the weighted version of the maximum consistent triple set problem. 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.

Simulated Data: To evaluate our approach we use simulated and real-life data sets. Artificial data is created with the method described in (Hernandez-Rosales et al., 2014) to generate explicit species/gene tree histories from which the orthology relation is directly accessible. We simulate 100 orthology data sets for five, ten, and 15 species and ten to 100 gene families, each. All simulations are performed with parameters 1 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.

The reconstructed trees are compared with the binary trees generated by the simulation. Therefore, we use the software `TreeCmp` (Bogdanowicz et al., 2012) 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 (Yule, 1925).

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. Noise is added with a probability $p \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$.

Real-life Data: As real-life applications we consider two sets of eubacterial genomes: the set of eleven *Aquificales* species studied in (Lechner et al., 2014), and 19 *Enterobacteriales* species from `RefSeq`, see Supplemental Material for accession numbers. An initial estimate of the orthology relation is computed with `Proteinortho` (Lechner et al., 2011) from all the annotated proteins using an E-value threshold of $1e-10$. 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.

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 10`). Precisely,

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

Equivalently, support values s_v for each subtree rooted at v can be computed by considering only those triples $(\alpha\beta|\gamma)$ with the two closer related species $\alpha, \beta \in L(v)$ and $\gamma \notin L(v)$.

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

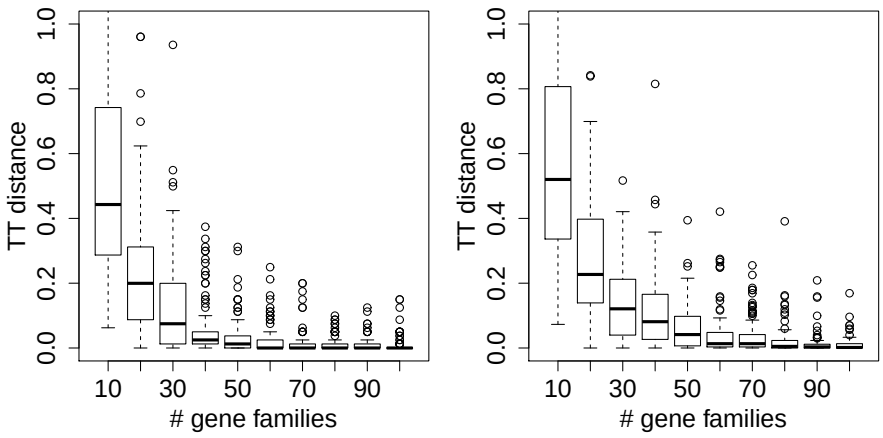


Figure 2: Accuracy of reconstructed species trees as function of number of independent gene families. Tree distance is measured by the triple metric (TT) for 100 reconstructed phylogenetic trees with ten (l.h.s.) and 15 (r.h.s.) species.

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.

4 Results and Discussion

The theoretical considerations of Section 2 show that empirical estimates of orthology implicitly contain information on the species phylogeny which can be extracted, e.g., by the ILP formulation outlined in Section 3. We first used simulated data to demonstrate that the workflow of Fig. 1 is indeed feasible and leads to correct trees. To obtain fully resolved species trees, a sufficient number of gene duplications must have occurred, since the phylogenetic information utilized by our approach is entirely contained in the duplication events. Note, if no paralogs exist, then G_Θ is a clique, the corresponding minimally resolved gene tree is a star and no species triples can be obtained. In small sets with five species 95% of the trees could be exactly reconstructed from at least 50 gene families. For ten and 15 species with 100 gene families 80%, resp. 50%, of the trees could be properly reconstructed, see Fig. 2.

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 levels. 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) and against horizontal gene transfer (noise model iv), while missing edges in Θ have a larger impact, see Fig. 3 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 alike. From such components fewer species triples can be extracted and therefore, introducing false species triples is unlikely, while missing species triples can be supplemented by other components. On the other hand, if there are many false paralogs (underpredicting orthology) more false species triples are introduced, resulting in inaccurate trees.

With the *Aquificales* data set `Proteinortho` predicts 2887 gene families, from which 823 contain duplications. The reconstructed species tree (see Fig. 4) is almost identical to the tree presented in (Lechner et al., 2014). It only differs in the two *Sulfurihydrogenibium* species not being 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 the tree correctly. Additionally, *Hydrogenivirga sp.* is misplaced next to *Persephonella marina*. This does not come as a surprise: Lechner et al. (2014) already suspected that the data from this species was contaminated with material from *Hydrogenothermaceae*. The reconstructed tree has a support of 0.6.

The second data set comprises the genomes of 19 *Enterobacteriales* with 8308 gene families of which 10 consists of more than 50 genes and 1301 containing duplications. Our orthology-based tree shows the expected groupings of *Escherichia* and *Shigella* species and identifies the monophyletic groups comprising *Salmonella*,

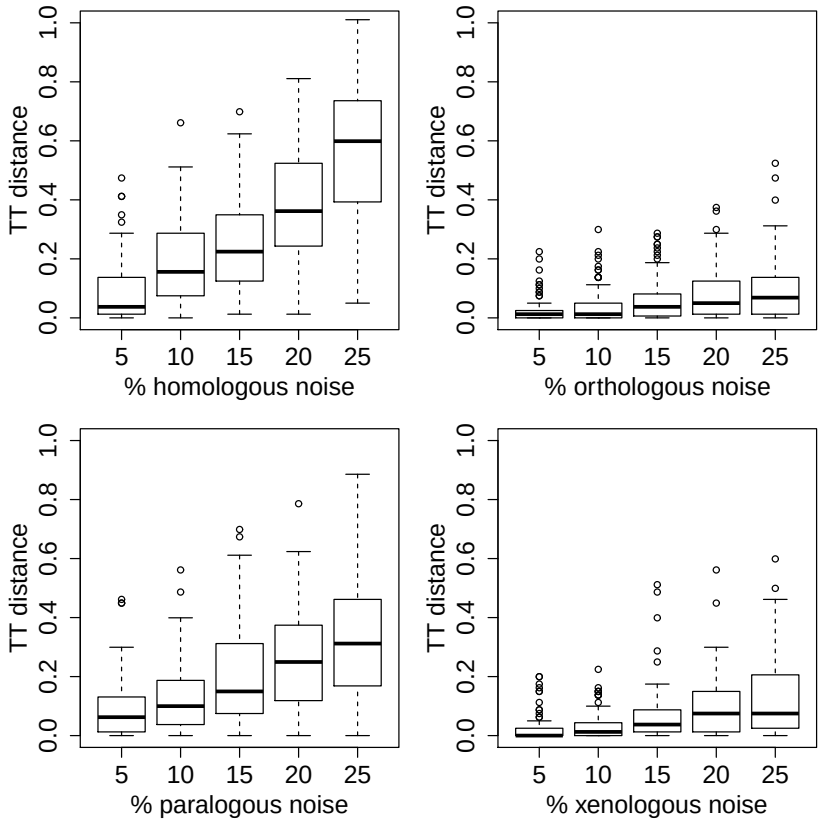


Figure 3: Accuracy of reconstructed species trees as function of noise level ($p = 5 - 25\%$) and noise type in the raw orthology data Θ. Tree distance is measured by the triple metric (TT) for 100 reconstructed phylogenetic trees with ten species.

Data	**CE**	**TE**	**MCS**	**LRT**	**Total**
Simulations[1]	45	5	< 1	< 1	51
Aquificales	32	64	< 1	< 1[2]	102
Enterobacteriales	442	1008	9	< 1[2]	1639

Table 2: Running time in seconds on an Intel® Core™2 Duo CPU with 2.4GHz for individual sub-tasks: **CE** cograph editing, **TE** triple extraction, **MCS** minimal consistent subset of triples, **LRT** least resolved tree. See Supplement for more details.

Klebsiella, and *Yersinia* species. The topology of the deeper nodes agrees only in part with the reference tree from `PATRIC` database (Wattam et al., 2013), see Supplemental Material for additional information. The resulting tree has a support of 0.48, 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 on a regular desktop computer, see Table 4. 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.

5 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.

[1] Average of 2000 simulations, 10 species,100 gene families.

[2] A unique tree was obtained using `BUILD` .

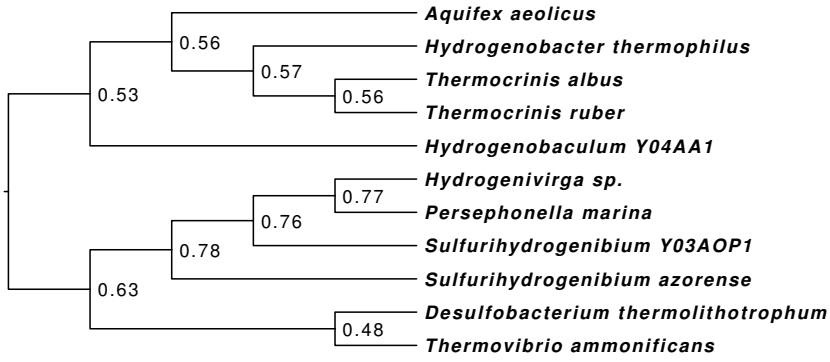


Figure 4: Phylogenetic tree of eleven *Aquificales* species inferred from paralogy. Internal node labels indicate support of subtrees.

In fact, 1:1 orthologs – the preferred data in sequence-based phylogenetics – do not contribute *at all* to the phylogenetic reconstruction in our approach. Access to the phylogenetic information implicit in (co-)orthology data, however, requires the solution of three NP-complete combinatorial optimization problems. Here we solve these tasks exactly for moderate-size problems by means of an ILP formulation. Using phylogenomic data for *Aquificiales* and *Enterobacteriales* we demonstrated that non-trivial phylogenies can indeed be re-constructed from tree-free orthology estimates alone. Simulated data, furthermore, indicate that the method is rather robust and can tolerate surprisingly large levels of noise such as mispredicted orthology and horizontal gene transfer, provided enough independent gene families are present in the data. Lack of duplications, however, limits our resolution at very short time scales, a regime in which sequence-based approaches work very accurately.

The current implementation does not easily scale to very large data sets. We suspect that substantial improvements will come from sophisticated ILP formulations requiring deeper insights into strict dense triple sets. Paralleling the developments in sequence-based phylogenetics, where the problems of finding a good input alignment and finding the tree(s) maximizing the parsimony score, likelihood or Bayesian posterior probability are also NP-complete, it may be advantageous to settle for heuristic solutions. Within decades of development these have improved to the point where they are no longer a limiting factor in phylogenetic reconstruction. 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 further research into 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 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, 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).

References

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

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

Bininda-Emonds, O. (2004). *Phylogenetic Supertrees*. Kluwer Academic Press, Dordrecht, The Netherlands.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Lechner, M., Nickel, A. I., Wehner, S., Riege, K., Wieseke, N., Beckmann, B. M., Hartmann, R. K., and Marz, M. (2014). Genomewide comparison and novel ncRNAs of aquificales. *PLoS ONE*. submitted.

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

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

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

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

Wattam et al. (2013). Patric, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Research*.

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

Yule, G. U. (1925). 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.

Phylogenetics from 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 Diestel (2012) 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 (Semple and Steel, 2003)**.** *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 (Dress et al., 2012), constitute an important concept in the context of supertree reconstruction (Semple and Steel, 2003; Bininda-Emonds, 2004) 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 (Semple and Steel, 2003) as `BUILD`, that either constructs a phylogenetic tree T displaying R or recognizes that R is not consistent or *inconsistent* (Aho et al., 1981). Various practical implementations have been described starting with (Aho et al., 1981), improved variants are discussed in (Rauch Henzinger et al., 1999; Jansson et al., 2005). The problem of determining a maximum consistent subset of an inconsistent set of triples, however, is NP-hard and also APX-hard, see (Byrka et al., 2010a; van Iersel et al., 2009) and the references therein. We refer to (Byrka et al., 2010b) 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 (Jansson et al., 2012).

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 (Aho et al., 1981; Huson et al., 2010; Bryant and Steel, 1995). 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 (Aho et al., 1981; Bryant and Steel, 1995)**.** *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 (Huson et al., 2010)**.** *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 (Bryant and Steel, 1995)**.** *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 (Bryant and Steel, 1995; Grünewald et al., 2007; Bryant, 1997; Huber et al., 2005; Böcker et al., 2000). 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 (Bryant and Steel, 1995): For any three leaves $x, y, z \in L_R$ test whether exactly one of the sets $R \cup \{(\mathsf{xy|z})\}$, $R \cup \{(\mathsf{xz|y})\}$, $R \cup \{(\mathsf{zy|x})\}$ 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 $R \vdash (\mathsf{xy|z})$ if any phylogenetic tree that displays all triples of R also displays $(\mathsf{xy|z})$. In other words, $R \vdash (\mathsf{xy|z})$ iff $(\mathsf{xy|z}) \in \mathrm{cl}(R)$. In a work of Bryant and Steel (Bryant and Steel, 1995), in which the authors extend and generalize the work of Dekker (Dekker, 1986), it was shown under which conditions it is possible to infer triples by using only subsets $R' \subseteq R$, i.e., under which conditions $R \vdash (\mathsf{xy|z}) \implies R' \vdash (\mathsf{xy|z})$ for some $R' \subseteq R$. In particular, we will make frequent use of the following inference rules:

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

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

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

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.*

We are now in the position to prove the following important and helpful lemmas and theorem. The final theorem basically states, consistent strict dense triple sets can be characterized by the closure of any two element subset of R.

Lemma 4. *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 5. *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 5 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 4 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 6. *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 (Böcker and Dress, 1998).

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. (Brandstädt et al., 1999) 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. (Corneil et al., 1981, p. 166)). The next theorem summarizes the results, that rely on the theory of so-called symbolic ultrametrics developed in (Böcker and Dress, 1998) and have been established in a more general context in (Hellmuth et al., 2013).

Theorem 5 (Hellmuth et al., 2013)**.** *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 (Corneil et al., 1985; Habib and Paul, 2005). 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 (Liu et al., 2011, 2012).

Lemma 7. *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. □ □

Thus it suffices to solve the cograph editing problem separately for the connected components of G.

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 Hernandez-Rosales et al. (2012).

S 1.5.1 Reconciliation Maps

We start with a formal definition of reconciliation maps.

Definition 1 (Hernandez-Rosales et al., 2012)**.** *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 (Doyon et al., 2009).

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 Guigó et al. (1996); Page and Charleston (1997); Arvestad et al. (2003); Bonizzoni et al. (2005); Górecki and J. (2006); Hahn (2007); Bansal and Eulenstein (2008); Chauve et al. (2008); Burleigh et al. (2009); Larget et al. (2010). 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. Hernandez-Rosales et al. (2012) 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

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

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 6 (Hernandez-Rosales et al., 2012)**.** *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 (Semple and Steel, 2003). 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 6.

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 7 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 Chang et al. (2011). 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 6). 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 Chang et al. (2011) 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 6, 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 8. *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 8 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.

$$(1 - E_{xy}) + E_{xz} + E_{yz} - T_{(\alpha\beta|\gamma)} \leq 2 \tag{ILP 4}$$

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

$$E_{xy} + (1 - E_{xz}) + E_{yz} - T_{(\alpha\gamma|\beta)} \leq 2 \qquad \text{(ILP 4)}$$
$$E_{xy} + E_{xz} + (1 - E_{yz}) - T_{(\beta\gamma|\alpha)} \leq 2$$

Case (2) $t(\mathrm{lca}(x,y)) = \bullet = t(\mathrm{lca}(x,y,z))$*:* Lemma 8 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:

$$T_{(\alpha\delta|\gamma)} + T_{(\beta\delta|\gamma)} - T_{(\alpha\beta|\gamma)} \leq 1 \qquad \text{(ILP 4)}$$

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 4) 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 4) is added for all $\{\alpha, \beta, \gamma, \delta\} \in \binom{\mathfrak{S}}{4}$.

In particular, these constraints ensure, that for each triple (xy|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 (xy|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 (xy|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:

$$\min \sum_{\{\alpha,\beta,\gamma\} \in \binom{S}{3}} T_{(\alpha\beta|\gamma)} + T_{(\alpha\gamma|\beta)} + T_{(\beta\gamma|\alpha)} \qquad \text{(ILP 5)}$$

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 Jansson (2001); Wu (2004). In an earlier ILP approach, Chang et al. (2011) 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:

$$T'_{(\alpha\beta|\gamma)} + T'_{(\alpha\gamma|\beta)} + T'_{(\beta\gamma|\alpha)} = 1. \qquad \text{(ILP 6)}$$

We now apply the inference rules in Eq. (i)-(iii) and the results of Theorem 1. We ensure consistency of $\mathbb{S}'$ by adding the following constraints for all ordered tuples $(\alpha, \beta, \gamma, \delta)$ for all $\{\alpha, \beta, \gamma, \delta\} \in \binom{\mathfrak{S}}{4}$:

$$T'_{(\alpha\beta|\gamma)} + T'_{(\alpha\delta|\gamma)} - T'_{(\beta\delta|\gamma)} \leq 1. \qquad \text{(ILP 7)}$$
$$2T'_{(\alpha\beta|\gamma)} + 2T'_{(\alpha\delta|\beta)} - T'_{(\beta\delta|\gamma)} - T'_{(\alpha\delta|\gamma)} \leq 2$$
$$2T'_{(\alpha\beta|\gamma)} + 2T'_{(\gamma\delta|\beta)} - T'_{(\alpha\beta|\delta)} - T'_{(\gamma\delta|\alpha)} \leq 2$$

The constraints in Eq. (ILP 7) are a direct translation of the inference rules in Eqns.((i)-(iii)). By Remark 3, these three inference rules are the only ones that imply new triples for pairs of triples for any dense triple set. Moreover, by Theorem 1 we know that testing pairs of triples is sufficient for verifying consistency.

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

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

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.

$$\max \sum_{(\alpha\beta|\gamma)\in\mathbb{S}} T'_{(\alpha\beta|\gamma)} * w(\alpha\beta|\gamma). \tag{ILP 10}$$

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

$$0 \leq T'_{(\alpha\beta|\gamma)} + T_{(\alpha\beta|\gamma)} - 2T^*_{(\alpha\beta|\gamma)} \leq 1 \tag{ILP 9}$$

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 Chang et al. (2011), 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 Chang et al. (2011), to solve the NP-hard problem Jansson et al. (2012) 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 (Semple and Steel, 2003). 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 Chang et al. (2011) 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 Chang et al. (2011). 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

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

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 Chang et al. (2011), 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

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

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

$$C_{p,q,01} + C_{p,q,10} + C_{p,q,11} \leq 2. \qquad \texttt{(ILP 14)}$$

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

$$0 \leq Y_p|\mathfrak{S}| - \sum_{\alpha\in\mathfrak{S}} M_{\alpha p} \leq |\mathfrak{S}| - 1 \qquad \texttt{(ILP 15)}$$

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

$$\min \sum_p Y_p \qquad \texttt{(ILP 16)}$$

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 Semple and Steel (2003); Dress et al. (2012).

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

S 2.5 Implementation Details

The ILP approach was 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 contains any paralogous genes at all. If not, then the resulting gene tree would be a star, not containing any species triple information. Hence, the ILP for extracting the triples is skipped.

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 Bogdanowicz et al. (2012).

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 devided 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 $lca(x, y)$ 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 Bogdanowicz et al. (2012), together with an option to compute normalized distances. Therefore, average distances between random Yule trees Yule (1925) 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.

As stated in the main text, we defined a support value $s \in [0, 1]$ for the reconstructed trees. This value utilizing the triple weights $w(\alpha\beta|\gamma)$ from Eq. `ILP 10`. 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 supportorting this subtree in relation to the frequency of all triples supporting the existence or non-existence of this subtree.

S 3 Data Sets

Beside simulated data sets two real-life data sets of eubacterial genomes are analyzed in this study.

For the first set we took eleven species from the three *Aquificales* 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).

For the second set we used the following 19 species from the *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* (NC_021231.1, NC_021232.1), *Proteus mirabilis* BB2000 (NC_022000.1), *Salmonella bongori* Sbon 167 (NC_021870.1, NC_021871.1), *Salmonella enterica* serovar Agona 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 4 Additional Results

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. Figure S1 shows boxplots for the tree distance as a funcion of the number of independent gene families.

The complete results for the 2000 simulated data sets of 10 species and 100 gene families with a varying amount of noise are depicted in Figure S2.

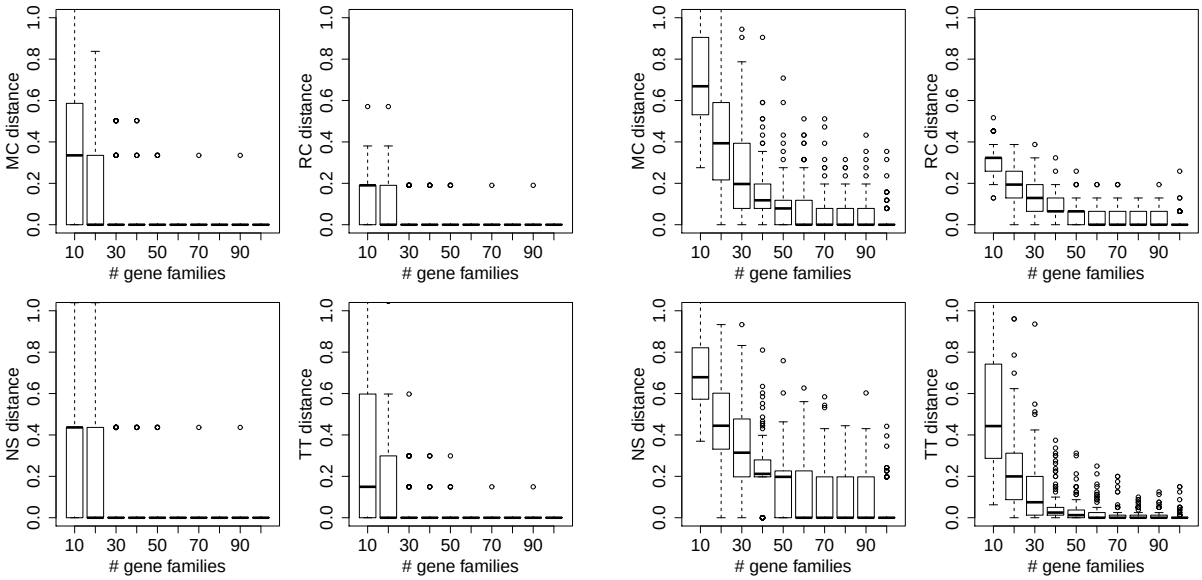
Real-life Data: Figure S3 depicts the phylogenetic tree of *Aquificales* species obtained from paralogy data in comparison to the tree suggested by Lechner et al. (2014). The trees obtained from bootstrapping experiments are given in Figure S4. The majority-rule consensus trees for both bootstrapping 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 S5 depicts the phylogenetic tree of *Enterobacteriales* species obtained from paralogy data in comparison to the tree from `PATRIC` database (Wattam et al., 2013). The trees obtained from bootstrapping experiments are given in Figure S6. 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.

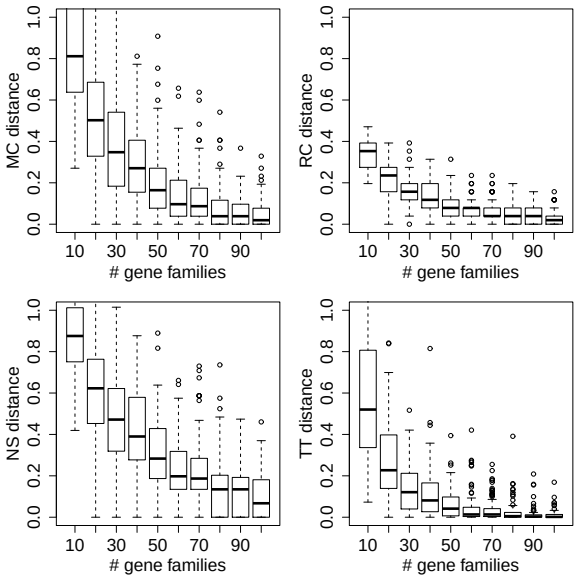
S 4.1 Additional Comments on Running Time

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 4.1 shows the runtimes for simulated and real-life data sets for each individual sub-task. Note that the time used for triple extraction is quite high, compared to cograph editing. This is due to the fact, that in both cases initializing the ILP solver is the dominating factor. In the implementation we first perform a check, if for a given gene family cograph editing and/or triple extraction has to be performed. In the real-life data



(A) (B)



(C)

Figure S1: Matching Cluster (MC), Robinson-Foulds (RC), Nodal Splitted (NS) and Triple metric (TT) tree distances of 100 reconstructed phylogenetic trees with five (A), ten (B), and 15 (C) species and ten to 100 gene families, each.

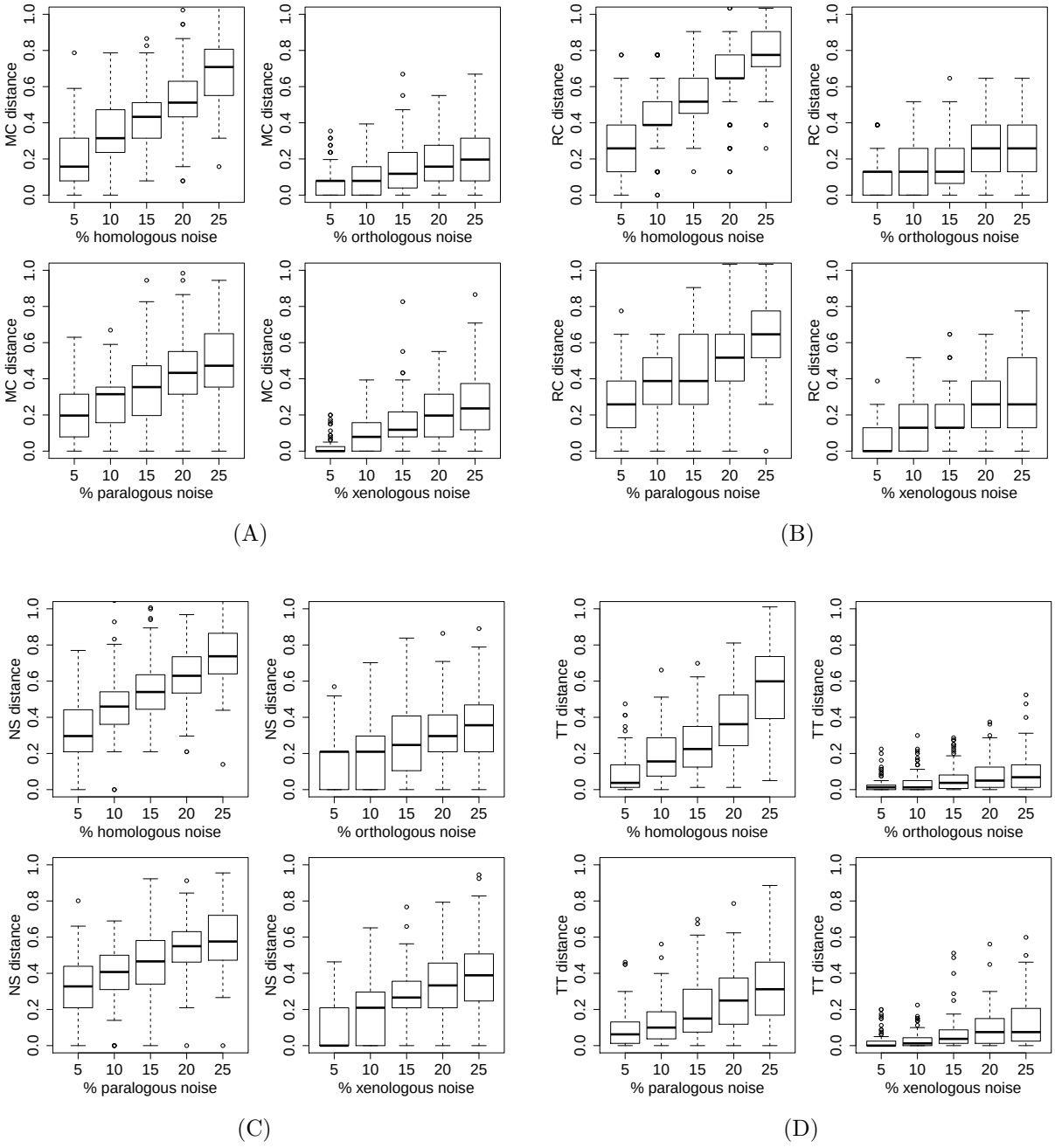


Figure S2: Matching Cluster (A), Robinson-Foulds (B), Nodal Splitted (C) and Triple metric (D) tree distances of 100 reconstructed phylogenetic trees with ten species. For each model noise was added with a probability of 0.05 to 0.25.

sets many connected components of G_{Θ}, containing duplications, are cographs already. Thus, cograph editing was skipped more often than triple extraction. Another oddity is the extraordinary short runtime for triple extraction in the *Enterobacteriales* data set. During the bootstrapping experiments for this set much longer times were observed, dominating the total runtime.

Data	**CE**	**TE**	**MCS**	**LRT**	**Total** [1]
Simulations[2]	45 [3]	5	< 1	< 1 (2) [4]	51
Aquificales	32	64	< 1	< 1 (5)[5]	102
Enterobacteriales	442	1008	9 [6]	< 1 (140758)[5]	1639

Table S1: Running time in seconds on an Intel® Core™2 Duo CPU with 2.4GHz for individual sub-tasks: **CE** cograph editing, **TE** triple extraction, **MCS** minimal consistent subset of triples, **LRT** least resolved tree.

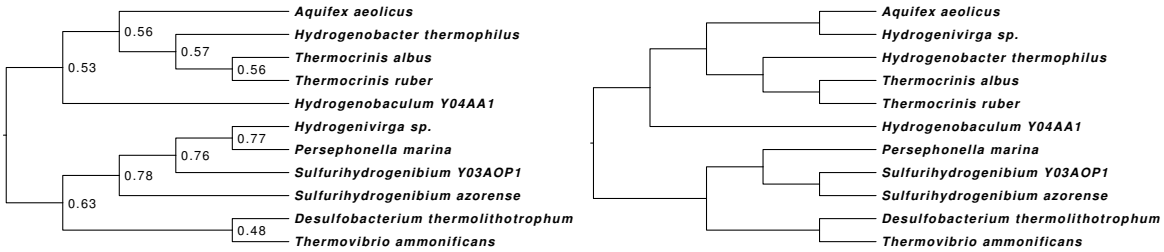


Figure S3: Phylogenetic tree of eleven *Aquificales* species. L.h.s.: tree computed from paralogy data. Internal node labels indicate support of subtrees. R.h.s.: reference tree from Lechner et al. (2014).

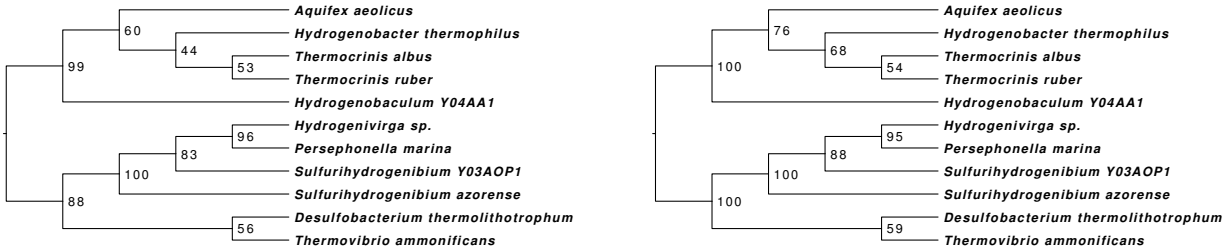


Figure S4: Cograph-based (l.h.s.) and triple-based (r.h.s.) bootstrapping trees of eleven *Aquificales* species.

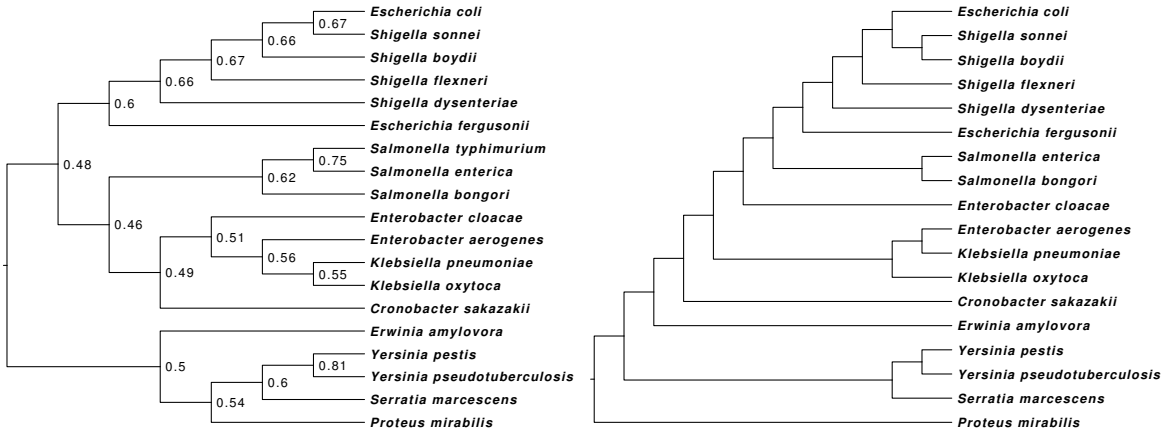


Figure S5: Phylogenetic trees of 19 *Enterobacteriales* species. L.h.s.: tree computed from paralogy data. Internal node labels indicate support of subtrees. R.h.s.: reference tree from `PATRIC` database, projected to the 19 considered species. *Salmonella typhimurium* is missing in `PATRIC` tree.

References

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

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

[1]Total time includes parsing input and writhing output files.

[2]Average of 2000 simulations, 10 species, 100 gene families.

[3]200,000 cographs, 21 not optimally solved within time limit of 30 min.

[4]In 99.7% of the simulations the least resolved tree could be found using `BUILD`. Second value indicates average runtime for cases where ILP formulation was used.

[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 (2688 sec. on average).

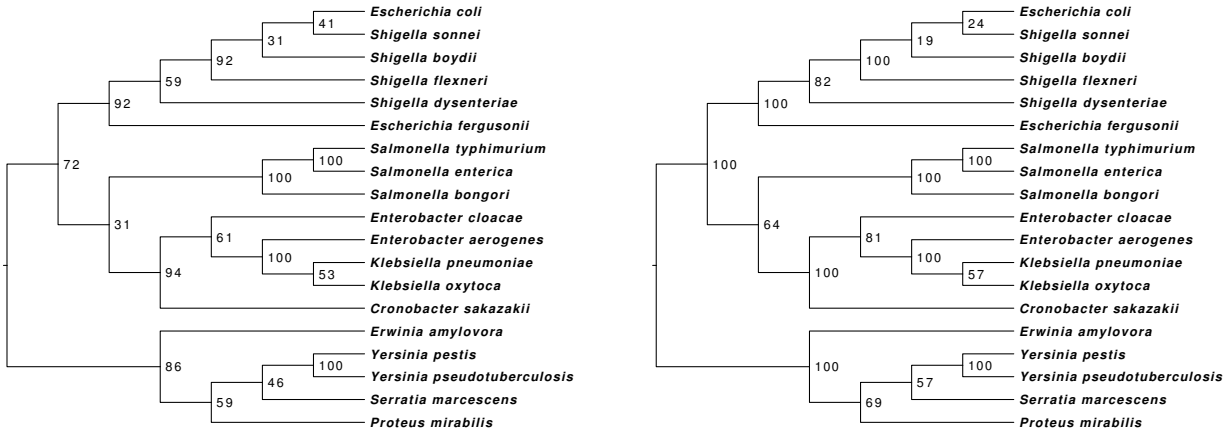


Figure S6: Cograph-based (l.h.s.) and triple-based (r.h.s.) bootstrapping trees of 19 *Enterobacteriales* species.

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

Bininda-Emonds, O. (2004). *Phylogenetic Supertrees*. Kluwer Academic Press, Dordrecht, The Netherlands.

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

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

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

Bonizzoni, P., Della Vedova, G., and Dondi, R. (2005). Reconciling a gene tree to a species tree under the duplication cost model. *Theor. Comp. Sci.*, 347:36–53.

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

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

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

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

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

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

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

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

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

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

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

Diestel, R. (2012). *Graph Theory, 4th Edition*, volume 173 of *Graduate texts in mathematics*. Springer.

Doyon, J.-P., Chauve, C., and Hamel, S. (2009). Space of gene/species trees reconciliations and parsimonious models. *J. Comp. Biol.*, 16:1399–1418.

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

Górecki, P. and J., T. (2006). DSL-trees: A model of evolutionary scenarios. *Theor. Comp. Sci.*, 359:378–399.

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

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

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

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

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

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

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

Huson, D. H., Rupp, R., and Scornavacca, C. (2010). *Phylogenetic networks: concepts, algorithms and applications*. Cambridge University Press.

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

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

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

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

Lechner, M., Nickel, A. I., Wehner, S., Riege, K., Wieseke, N., Beckmann, B. M., Hartmann, R. K., and Marz, M. (2014). Genomewide comparison and novel ncRNAs of aquificales. *PLoS ONE*. submitted.

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

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

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

Rauch Henzinger, M., King, V., and Warnow, T. (1999). Constructing a tree from homeomorphic subtrees, with applications to computational evolutionary biology. *Algorithmica*, 24:1–13.

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

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

Wattam et al. (2013). Patric, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Research*.

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

Yule, G. U. (1925). 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.