J. Math. Biol. (2017) 75:199–237
DOI 10.1007/s00285-016-1084-3

Mathematical Biology

The mathematics of xenology: di-cographs, symbolic ultrametrics, 2-structures and tree-representable systems of binary relations

Marc Hellmuth[1,2] · **Peter F. Stadler**[3,4,5,6,7] · **Nicolas Wieseke**[8,9]

Received: 8 March 2016 / Revised: 20 November 2016 / Published online: 30 November 2016

Abstract The concepts of orthology, paralogy, and xenology play a key role in molecular evolution. Orthology and paralogy distinguish whether a pair of genes originated by speciation or duplication. The corresponding binary relations on a set of genes form complementary cographs. Allowing more than two types of ancestral event types leads to symmetric symbolic ultrametrics. Horizontal gene transfer, which leads to xenol-

✉ Marc Hellmuth
mhellmuth@mailbox.org

Peter F. Stadler
studla@bioinf.uni-leipzig.de

Nicolas Wieseke
wieseke@informatik.uni-leipzig.de

1 Department of Mathematics and Computer Science, University of Greifswald, Walther-Rathenau-Straße 47, 17487 Greifswald, Germany

2 Center for Bioinformatics, Saarland University, Building E 2.1, P.O. Box 151150, 66041 Saarbrücken, Germany

3 Bioinformatics Group, Department of Computer Science, University of Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany

4 Interdisciplinary Center of Bioinformatics, University of Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany

5 Max-Planck-Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany

6 Institute of Theoretical Chemistry, University of Vienna, Währingerstraße 17, 1090 Wien, Austria

7 Santa Fe Institute, 1399 Hyde Park Rd.,Santa Fe, NM 87501, USA

8 Parallel Computing and Complex Systems Group, Department of Computer Science, University of Leipzig, Johannisgasse 26, 04103 Leipzig, Germany

9 Interdisciplinary Center of Bioinformatics, University of Leipzig, Johannisgasse 26, 04103 Leipzig, Germany

ogous gene pairs, however, is inherent asymmetric since one offspring copy "jumps" into another genome, while the other continues to be inherited vertically. We therefore explore here the mathematical structure of the non-symmetric generalization of symbolic ultrametrics. Our main results tie non-symmetric ultrametrics together with di-cographs (the directed generalization of cographs), so-called uniformly non-prime (*unp*) 2-structures, and hierarchical structures on the set of strong modules. This yields a characterization of relation structures that can be explained in terms of trees and types of ancestral events. This framework accommodates a horizontal-transfer relation in terms of an ancestral event and thus, is slightly different from the the most commonly used definition of xenology. As a first step towards a practical use, we present a simple polynomial-time recognition algorithm of *unp* 2-structures and investigate the computational complexity of several types of editing problems for *unp* 2-structures. We show, finally that these NP-complete problems can be solved exactly as Integer Linear Programs.

Keywords Xenologs · Paralogs · Orthologs · Gene tree · 2-Structures · Uniformly non-prime decomposition · Di-cograph · Symbolic ultrametric · Recognition algorithm · NP-completeness · Integer Linear Program

Mathematics Subject Classification 05C05 · 92D15 · 68R05 · 68R10

1 Introduction

The current flood of genome sequencing data poses new challenges for comparative genomics and phylogenetics. An important topic in this context is the reconstruction of large families of homologous proteins, RNAs, and other genetic elements. The distinction between orthologs, paralogs, and xenologs is a key step in any research program of this type. The distinction between orthologous and paralogous gene pairs dates back to the 1970s: pairs of genes whose last common ancestor in the "gene tree" corresponds to a speciation are orthologs; if the last common ancestor was a duplication event, the genes are paralogs (Fitch 1970). The importance of this distinction is two-fold: first it is informative in genome annotation. Orthologs usually fulfill corresponding functions in related organism. Paralogs, in contrast, are expected to have similar but distinct functions (Koonin 2005). Secondly, the orthology (or paralogy) relation conveys information about the events corresponding to internal nodes of the gene tree (Hernandez-Rosales et al. 2012) and about the underlying species tree (Hellmuth et al. 2013, 2015).

Based on a theory of symbolic ultrametrics (Böcker and Dress 1998) is was shown in Hellmuth et al. (2013) that the orthology and paralogy relations are necessarily complementary cographs provided the genetic repertoire evolved only by means of speciation, gene duplication, and gene loss. However, horizontal gene transfer (HGT), i.e., the incorporation of genes or other DNA elements from a source different than the parent(s), cannot be neglected under many circumstances. In fact, HGT plays an important role not only in the evolution of prokaryotes (Koonin et al. 2001) but also in eukaryotes (Keeling and Palmer 2008). This begs the question whether the

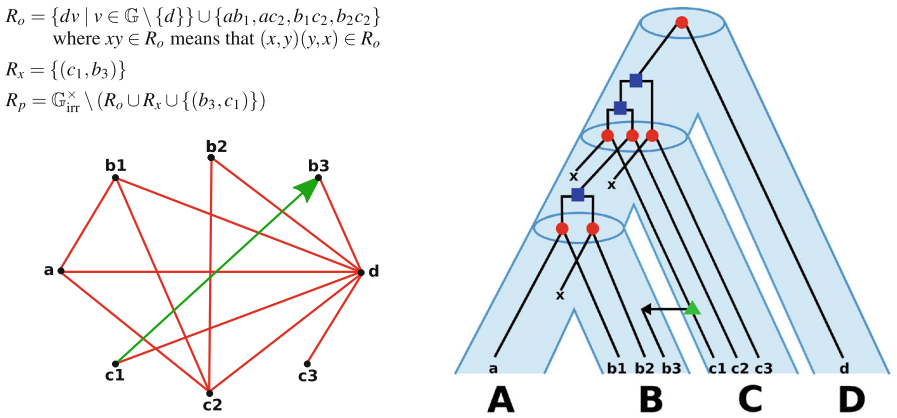

Fig. 1 Example of an evolutionary scenario showing the "true" evolution of a gene family evolving along the species tree (shown as *blue tube-like tree*). The corresponding gene tree T appears embedded in the species tree S. The speciation vertices in the gene tree (*red circuits*) appear on the vertices of the species tree (*blue ovals*), while the duplication vertices (*blue squares*) and the HGT-vertices (*green triangles*) are located on the edges of the species tree. Gene losses are represented with "x". The gene-tree T uniquely determines the relationships between the genes by means of the event at the least common ancestor $\text{lca}_T(x, y)$ of distinct genes $x, y \in \mathbb{G}$. There is a clear distinction between orthologs (comprised in R_o and indicated via *red edges*), paralogs (comprised in R_p and indicated via *non-drawn edges*), as well as xenologs, that are neither orthologs nor paralogs (comprised in R_x and indicated by *green directed arcs*) (color figure online)

combinatorial theory of orthology/paralogy can be extended to incorporate xenologs, i.e., pairs of genes that are separated in the gene tree by HGT events.

In contrast to orthology and paralogy, the definition of xenology is less well established and by no means consistent in the biological literature. The most commonly used definition stipulates that two genes are *xenologs* if their history since their common ancestor involves horizontal transfer of at least one of them (Fitch 2000; Jensen 2001). In this setting the HGT event itself is treated as gene duplication event. Every homolog is still either ortholog or a paralog. Both orthologs and paralogs may at them same time be xenologs (Jensen 2001).

The mathematical framework for orthology relations in terms of symbolic ultrametrics (Böcker and Dress 1998; Hellmuth et al. 2013), on the other hand, naturally accommodates more than two types of events associated with the internal nodes of the gene tree, see Hellmuth and Wieseke (2016) for an overview. It is appealing, therefore, to think of a HGT event as different from both speciation and duplication, in line with (Gray and Fitch 1983) where the term "xenologous" was originally introduced, see Fig. 1 for an illustrative example. The inherently asymmetric nature of HGT events, with their unambiguous distinction between the vertically transmitted "original" and horizontally transmitted "copy" furthermore suggests to relax the symmetry assumption and explore a generalization to directed graphs and symbolic "quasi-metrics". From the mathematical point of view it seems natural to ask which systems of binary relations on a set V (of genes) can be represented by a (phylogenetic) tree with leaf set V and a suitable labeling (of event types) on the internal nodes of T. To this end the theory of 2-structures (Ehrenfeucht and Rozenberg 1990a, b) provides an interesting starting point.

This contribution consists of two parts. In the first part (Sects. 2, 3), we will be concerned with the mathematical characterization of certain classes of 2-structures. To this end, we present the basic and relevant concepts used in this paper in Sect. 2. In particular, we briefly survey existing results concerning di-cographs, symbolic ultrametrics and 2-structures. In Sect. 3, we establish the characterization of 2-structures that have a particular tree-representation, so-called uniformly non-prime (*unp*) 2-structures, in terms of symbolic ultrametrics, di-cographs and so-called 1-clusters that are obtained from the tree-representation of the respective di-cographs. These results are summarized in Theorem 6. In the second part (Sect. 4), we use the characterization of *unp* 2-structure to design a conceptual quite simple quadratic-time algorithm to recognize whether a 2-structure is *unp*, and in the positive case, to construct the corresponding tree-representation (Algorithms 1–6 and Theorem 7). Furthermore, we are concerned with editing problems to obtain a *unp* 2-structure, showing the NP-completeness of the underlying decision problems (Theorem 9) and describe integer linear programming formulations to solve them.

2 Preliminaries

2.1 Basic notation

Throughout this contribution all sets are finite. We say that two sets A and B *overlap*, in symbols $A \between B$, if $A \cap B \neq \emptyset$ and neither $A \subseteq B$ nor $B \subseteq A$. Given a set V we identify binary relations $R \subseteq V \times V$ with the *directed graphs* (*di-graphs* for short) $G = (V, R)$ with vertex set V and *arc* set R. Throughout, we are concerned with irreflexive relations $R \subseteq V^2_{\text{irr}} := V \times V \backslash \{(v, v) | v \in V\}$ or, equivalently, loop-free digraphs. For an arc $e = (x, y) \in V^2_{\text{irr}}$ we write e^{-1} to designate the reverse arc (y, x). (*Undirected*) *graphs* are modeled by edge sets $E \subseteq \binom{V}{2}$ taken from the set of unordered pairs of vertices.

An undirected graph $G = (V, E)$ is *connected* if for any two vertices $x, y \in V$ there is a sequence of vertices $(x, v_1, \ldots, v_n, y)$, called *walk*, so that $\{x, v_1\}, \{v_n, y\}$ and $\{v_i, v_{i+1}\}$, $1 \leq i \leq n-1$ are contained in E. A di-graph $G = (V, E)$ is *(weakly) connected* if the undirected graph $G_u = (V, E_u)$ with $E_u = \{\{x, y\} \mid (x, y) \in E\}$ is connected. We say that a sequence of vertices $S = (x, v_1, \ldots, v_n, y)$ is a *walk* in the di-graph $G = (V, E)$, if S is a walk in the underlying undirected graph G_u. A graph $H = (W, F)$ is a *subgraph* of $G = (V, E)$ if $F \subseteq W \times W$ for di-graphs or $F \subseteq \binom{W}{2}$ for undirected graphs, $W \subseteq V$ and $F \subseteq E$. We will write $H \subseteq G$, if H is a subgraph of G. The subgraph $H = (W, F)$ is an *induced* di-graph if in addition $(x, y) \in W \times W$ and $(x, y) \in E$ implies $(x, y) \in F$. The corresponding condition in the undirected case reads $\{x, y\} \in \binom{W}{2}$ and $\{x, y\} \in E$ implies $\{x, y\} \in F$. A *connected component* of a (di-)graph is a connected subgraph that is maximal w.r.t. inclusion. A di-graph $G = (V, E)$ is *complete* if $E = V^2_{\text{irr}}$, and it is *arc-labeled* if there is a map $\varphi \colon E \to \Upsilon$ that assigns to each arc a label $i \in \Upsilon$.

A *tree* is a connected undirected graph that does not contain cycles. A *rooted* tree $T = (V, E)$ is a tree with one distinguished vertex $\rho \in V$ called *root*. The leaf set $L \subseteq V$ comprises all vertices that are distinct from the root and have degree 1. All

vertices that are contained in $V^0 := V \backslash L$ are called *inner* vertices. The first inner vertex $\mathrm{lca}(x, y)$ that lies on both unique paths from two vertices x, resp., y to the root, is called *lowest common ancestor* of x and y. We write $L(v)$ for the set of leaves in the subtree below a fixed vertex v, i.e., $L(v)$ is the set of all leaves for which v is located on the unique path from $x \in L(v)$ to the root of T. The *children* of an inner vertex v are its direct descendants. To be more precise, w is a child of v if $\{v, w\} \in E(T)$ and w is further away from the root than v. An *ordered* tree is a rooted tree in which an ordering is specified for the children of each vertex. Hence, ordered trees particularly imply a linear order $\leq$ of the leaves in L, and we say that x is left from y iff $x < y$.

Two rooted trees T_1 and T_2 on the same leaf set L are said to be *isomorphic* if there is a bijection $\psi : V(T_1) \to V(T_2)$ that induces a graph isomorphism from T_1 to T_2 which is the identity on L and maps the root of T_1 to the root of T_2.

It is well-known that there is a one-to-one correspondence between (isomorphism classes of) rooted trees on V and hierarchies on V. A *hierarchy on* V is a subset $\mathscr{C} \subseteq 2^V$ such that (i) $V \in \mathscr{C}$, (ii) $\{x\} \in \mathscr{C}$ for all $x \in V$, and (iii) $p \cap q \in \{p, q, \emptyset\}$ for all $p, q \in \mathscr{C}$. Condition (iii) states that no two members of $\mathscr{C}$ overlap. Members of $\mathscr{C}$ are called *clusters*. The number of clusters in a hierarchy is bounded (Hellmuth et al. 2015) and there is a well-known bijection between hierarchies and trees (Semple and Steel 2003):

Theorem 1 *Let $\mathscr{C}$ be a collection of non-empty subsets of V. Then, there is a rooted tree $T = (W, E)$ on V with $\mathscr{C} = \{L(v) \mid v \in W\}$ if and only if $\mathscr{C}$ is a hierarchy on V. Moreover, the number of clusters $|\mathscr{C}|$ in a hierarchy $\mathscr{C}$ on V is bounded by $2|V| - 1$.*

2.2 Di-cographs

Di-cographs are a generalization of the better-known undirected cographs. Cographs are obtained from single vertices by repeated application of disjoint union (*parallel* composition) and graph join (in this context often referred to as *series* composition) (Corneil et al. 1981; Brandstädt et al. 1999). In the case of di-cographs, the so-called *order* composition is added, which amounts to a directed variant of the join operation. More precisely, let $G_1, \ldots, G_k$ be a set of k disjoint digraphs. The disjoint union of the G_is is the digraph whose connected components are precisely the G_is. The series composition of the G_is is the union of these k graphs plus all possible arcs between vertices of different G_is. The order composition of the G_is is the union of these k graphs plus all possible arcs from G_i towards G_j, with $1 \leq i < j \leq k$.

We note that restricted to posets, di-cographs coincide with the series-parallel orders (Valdes et al. 1982). Di-cographs are characterized by the collection of forbidden induced subgraphs shown Fig. 2 (Crespelle and Paul 2006). An undirected graph is a cograph if and only if it does not contain P_4 as an induced subgraph (Corneil et al. 1981).

We emphasize that the results stated here are direct consequences of the results for 2-structure in Sect. 2.4. However, since di-cographs will play a central role for the characterization of certain 2-structures, we treat them here separately.

Given an arbitrary (di)graph $G = (V, E)$, a *(graph-)module* M is a subset $M \subset V$ such that for any $x \in M$ and $z \in V \backslash M$ it holds that $(x, z) \in E$ if and only if $(y, z) \in E$

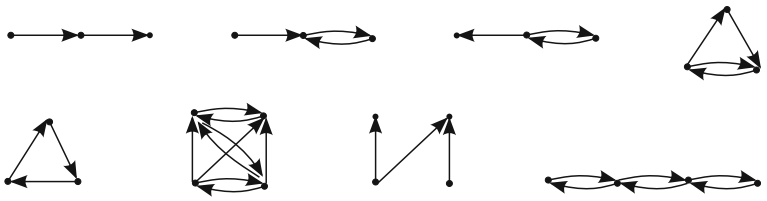

Fig. 2 Forbidden subgraphs for di-cographs. A di-graph G is a di-cograph if and only if it does not contain one of these graphs as an induced subgraph. Following Engelfriet et al. (1996), we denote them from *left to right* $D_3, A, B, \overline{D_3}$ (in the *1st line*), and $C_3, \overline{N}, N, P_4$ (in the *2nd line*). A similar picture appeared in Crespelle and Paul (2006)

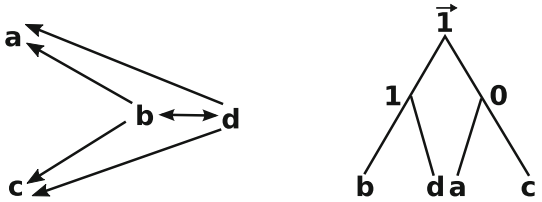

Fig. 3 Example of a di-cograph G together with its cotree. It can easily be seen that G that does not contain any of the forbidden subgraphs shown in Fig. 2

for all $y \in M$ and $(z, x) \in E$ if and only if $(z, y) \in E$ for all $y \in M$. The subfamily of so-called strong modules contains all modules that do not overlap other modules. The set of all strong modules forms a hierarchy and is called *modular decomposition* of G. For a given graph G we denote with $\mathbb{M}_{\text{str}}(G)$ the set of its strong modules. Since $\mathbb{M}_{\text{str}}(G)$ forms a hierarchy, there is an equivalent (ordered, rooted) tree, that is well known as the *modular decomposition tree* of G (Möhring and Radermacher 1984). The (unique) modular decomposition tree of a di-cograph is known as its *cotree*. Its leaves are identified with the vertices of the di-cograph and the inner vertices are labeled by the composition operations. Conversely, any ordered tree with internal vertices labeled by the operations *parallel*, *series*, or *order*, defines a unique di-cograph on its leaf-set.

Any inner vertex of the modular decomposition tree T of G corresponds to a strong module $L(v) \in \mathbb{M}_{\text{str}}(G)$. Moreover, each child u of v in T corresponds to a strong module $L(u) \subsetneq L(v)$ so that there is no other module $M \in \mathbb{M}_{\text{str}}(G)$ with $L(u) \subsetneq M \subsetneq L(v)$. Therefore, we refer to the module $L(u)$ as a *the child* of the module $L(v)$ if u is a child of v in the modular decomposition tree.

For simplicity, we use for a di-cograph G and its respective cotree T the labeling function $t : V^0(T) \to \{0, 1, \overrightarrow{1}\}$ defined by

$$t(\text{lca}(x, y)) = \begin{cases} 0, & \text{if } (x, y)(y, x) \notin E(G) \text{ (“parallel”)} \\ 1, & \text{if } (x, y)(y, x) \in E(G) \text{ (“series”)} \\ \overrightarrow{1}, & \text{else (“order”)} \end{cases}$$

throughout this contribution. Since the vertices in the cotree T are ordered, the label $\overrightarrow{1}$ on some $\text{lca}(x, y)$ of two distinct leaves $x, y \in L$ means that there is an arc $(x, y) \in E(G)$, while $(y, x) \notin E(G)$, whenever x is placed to the left of y in T, see Fig. 3 for an example. For a given cotree T and inner vertex v, we will also call the

strong module $L(v)$ of a di-cograph parallel, series, or order, if it is labeled 0, 1 and $\overrightarrow{1}$, respectively. The modular decomposition of a digraph that is not a cograph also contains strong modules that are neither parallel, nor series, nor order. Such modules are called *prime*.

2.3 Symbolic ultrametrics

Let V and Υ be non-empty sets and let $\delta\colon V \times V \to \Upsilon, (x, y) \mapsto \delta(xy)$ be a map that assigns to each pair $(x, y) \in V \times V$ the unique label $\delta(xy) \in \Upsilon$. To simplify the notation we write $\delta(xy)$ instead of $\delta((x, y))$.

Definition 1 A *symmetric symbolic ultrametric* is a map $\delta\colon V \times V \to \Upsilon$ that satisfies the following three axioms:

(U0′) $\delta(x, y) = \delta(y, x)$ for all $x, y \in X$
(U1′) there exists no subset $\{x, y, u, v\} \in \binom{V}{4}$ such that

$$\delta(xy) = \delta(yu) = \delta(uv) \neq \delta(yv) = \delta(xv) = \delta(xu).$$

(U2′) $|\{\delta(xy), \delta(xz), \delta(yz)\}| \leq 2 \quad$ for all $x, y, z \in V$;

Symmetric symbolic ultrametrics were first introduced by Böcker and Dress (1998) as a combinatorial generalization of ultrametrics. They play a central role in the reconstruction of event-labeled phylogenetic trees based on symmetric events (as the paralogy- or orthology-relation) (Hellmuth et al. 2013; Hellmuth and Wieseke 2015, 2016).

None of the axioms (U0′), (U1′), (U2′) refer to the "diagonal", i.e., the labels chosen for $\delta(x, x)$ for any $x \in V$. It will be convenient in the following to restrict ourselves to maps $\delta : V^2_{\text{irr}} \to \Upsilon$ because the 2-structures that will play a key role in this contribution are defined on V^2_{irr} only. Restricting the definition of symbolic ultrametrics to V^2_{irr} thus will simplify the presentation of the close relationships between 2-structures and symbolic ultrametrics without loss of generality. We note, finally, that the recent applications of symmetric symbolic ultrametrics to phylogenetic combinatorics (Hellmuth et al. 2013; Hellmuth and Wieseke 2015, 2016) assigned a special label to the diagonal, thereby effectively restricting the definition to V^2_{irr} as well.

For a given symmetric map $\delta\colon V^2_{\text{irr}} \to \Upsilon$ we define graphs $G_i(\delta) = (V, E_i), i \in \Upsilon$ with edge sets $E_i = \{\{x, y\} \mid x, y \in V, x \neq y, \delta(xy) = i\}$. Conversely, given a collection of graphs $G_i = (V, E_i)$ for $i \in \Upsilon$ there is a symmetric map $\delta\colon V^2_{\text{irr}} \to \Upsilon$ defined by $(x, y) \mapsto i$ iff $xy \in E_i$ provided the edge sets $\{E_i \mid i \in \Upsilon\}$ form a partition.

Symmetric symbolic ultrametrics can be characterized by means of their graph representations:

Theorem 2 (Hellmuth et al. 2013) *A map $\delta\colon V^2_{\text{irr}} \to \Upsilon$ is a symmetric symbolic ultrametric if and only if δ satisfies (U0′), (U2′) and the condition*

(U1″) the graph $G_i(\delta)$ is an undirected cograph for each $i \in \Upsilon$.

While the literature so far covers symmetric relations, we are interested in this contribution in non-symmetric relations. Therefore, we define symbolic ultrametrics here to include also the non-symmetric case.

Consider a map $\delta\colon V_{\text{irr}}^2 \to \Upsilon$ and two distinct vertices $x, y \in V$. We denote by $D_{xy} := \{\delta(xy), \delta(yx)\}$ the set of labels assigned to the pairs (x, y) and (y, x). By construction $D_{xy} = D_{yx}$ and $|D_{xy}| = 1$ iff $\delta(xy) = \delta(yx)$. We write $D_{xyz} := \{D_{xy}, D_{xz}, D_{yz}\}$ for any three vertices $x, y, z \in V$. Note that D_{xyz} is the set of distinct label pairs assigned to the constituent unordered pairs of the 3-tuple, not the set of distinct labels assigned to the six underlying ordered pairs. Furthermore, we define $G_i(\delta) = (V, E_i)$ as the digraphs with arc sets $E_i = \{(x, y) \in V_{\text{irr}}^2 \mid \delta(xy) = i\}$ for all $i \in \Upsilon$. As in the undirected case there is 1–1 correspondence maps $\delta\colon V_{\text{irr}}^2 \to \Upsilon$ and collections of digraphs $\{G_i | i \in \Upsilon\}$ whose arc sets form a partition of V_{irr}^2.

Definition 2 A *symbolic ultrametric* on V is map $\delta\colon V_{\text{irr}}^2 \to \Upsilon$ satisfying the following two conditions

(U1) $G_i(\delta)$ is a di-cograph for all $i \in \Upsilon$.
(U2) $|D_{xyz}| \leq 2$ for all $x, y, z \in V$.

We will refer to axiom (U2) as the "$\Delta(xyz)$-Condition" or simply "Triangle-Condition".

It is not hard to see that symmetric symbolic ultrametrics as defined above are indeed the symbolic ultrametrics in the sense of Def. 2 with a symmetric map δ: Axiom (U0') stipulates symmetry of δ. Moreover, if δ is a symmetric symbolic ultrametric in the sense of Def. 1, then (U2) states that $|D_{xyz}| = |\{\{\delta(xy), \delta(yx)\}, \{\delta(xz), \delta(zx)\}, \{\delta(zy), \delta(yz)\}\}| = |\{\delta(xy), \delta(xz), \delta(yz)\}| \leq 2$. Hence, (U2) and (U2′) a equivalent. Axiom (U1) simplifies for symmetric maps to the statement that $G_i(\delta)$ must be an undirected cograph for all $i \in \Upsilon$ due to condition (U1″) in Theorem 2. Hence, Definition 2 serves as a natural generalization of symbolic ultrametrics to maps that do not necessarily fulfill the symmetry condition (U0′).

Symbolic ultrametrics also generalize the usual notion of an ultrametric d on V (Böcker and Dress 1998). The latter are defined as a real-valued symmetric map $d\colon V \times V \to \mathbb{R}$ that (i) vanishes exactly on the diagonal and (ii) satisfies $d(x, z) \leq \max\{d(x, y), d(y, z)\}$ for all $x, y, z \in X$. A key property of ultrametrics is that the two larger distances of $d(x, y), d(x, z)$, and $d(y, z)$ necessarily coincide, see e.g. Semple and Steel (2003). For a quadruple of distinct vertices $x, y, u, v \in V$ with $d(x, y) = d(y, u) = d(u, v)$ and $d(y, v) = d(x, v) = d(x, u)$ this fact implies $d(x, y) = d(y, u) \geq d(x, u)$ and $d(x, v) = d(x, u) \geq d(u, v)$, whence all six distances must be equal. Reading the real-valued distances as labels, this observation rules out the existence of a quadruple forbidden by condition (U1′) and property (ii) directly translates to (U2′).

2.4 2-Structures

2-Structures were introduced in Ehrenfeucht and Rozenberg (1990a, b). We refer to Ehrenfeucht et al. (1995, 1999) and Engelfriet et al. (1996) for excellent additional

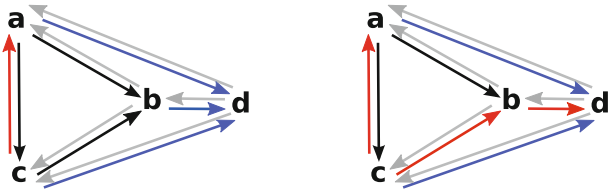

Fig. 4 Example of two different 2-structures that differ only in the respective label on the arc cb and bd. While the rhs. 2-structure is prime, the lhs. 2structure has as non-trivial modules $\{a, c\}$ and $\{a, b, c\}$

surveys. We follow the original terminology where possible. We will, however, deviate at times to remain consistent with the literature on co-graphs and symbolic ultrametrics.

Definition 3 A (labeled) *2-structure* is a triple $g = (V, \Upsilon, \varphi)$ where V and Υ are nonempty sets and $\varphi\colon V^2_{\text{irr}} \to \Upsilon$ is a map.

We refer to V as the vertices and Υ as the labels. The function φ maps each pair (x, y), called an *arc* of g to a unique label $\varphi(xy) := \varphi((x, y)) \in \Upsilon$. We will sometimes write V_g, Υ_g and φ_g to emphasize that the vertex set, label set, and the labeling function, resp., belong to the 2-structure g.

Isomorphic 2-structures $g = (V, \Upsilon, \varphi)$ and $h = (V, \Upsilon', \varphi')$, in symbols $g \simeq h$, differ only by a bijection $\alpha : \Upsilon \to \Upsilon'$ of their labels, i.e., $\varphi'(e) = \alpha(\varphi(e))$ and $\varphi(e) = \alpha^{-1}(\varphi'(e))$ for all $e \in V^2_{\text{irr}}$.

2-Structures can be considered as arc-labeled complete graphs, see Fig. 4 for examples. Conversely, every directed or undirected graph G with vertex set V has a representation as a 2-structure by labeling the edges of the complete graph by 0 or 1 depending on whether the arc is absent or present in G. Thus we can interpret 2-structures as a natural generalizations of (di-)graphs. Moreover, 2-structures are equivalent to a set of disjoint binary relations $R_1, \ldots, R_k$ where each tuple (x, y) has label i iff $(x, y) \in R_i$ or label 0 if (x, y) is not present in any of these relations.

Extending the definition for symbolic ultrametrics above, we define for a given 2-structure $g = (V, \Upsilon, \varphi)$ and each $i \in \Upsilon$ the graph $G_i(g) = (V, E_i)$ with arc set $E_i = \{(x, y) \in V^2_{\text{irr}} \mid \varphi(xy) = i\}$.

Given a subset $X \subseteq V$ the *substructure of* $g = (V, \Upsilon, \varphi)$ *induced by* X has vertex set X and all arcs $(a, b) \in X^2_{\text{irr}}$ retain the label $\varphi(ab) \in \Upsilon$, i.e., $g[X] := (X, \Upsilon, \varphi' = \varphi_{|X^2_{\text{irr}}})$. A 2-structure h is a substructure of the 2-structure g iff there is a subset $X \subseteq V_g$ so that $h \simeq g[X]$.

Definition 4 A *module* (or clan) of a 2-structure is a subset $M \subseteq V$, such that $\varphi(xz) = \varphi(yz)$ and $\varphi(zx) = \varphi(zy)$ holds for all $x, y \in M$ and $z \in V \setminus M$.

The empty set $\emptyset$, the complete vertex set V_g, and the singletons $\{v\}$ are always modules. They are called the *trivial* modules of g. We will assume from here on, that a module is non-empty unless otherwise indicated. The set of all modules of the 2-structure g will be denoted by $\mathbb{M}(g)$.

Lemma 1 *A subset $M \subseteq V$ is a module of a 2-structure $g = (V, \Upsilon, \varphi)$ if and only if M is also a graph-module of $G_i(g)$ for all $i \in \Upsilon$.*

Proof We denote for a fixed vertex $x \in M \subseteq V$ and an arbitrary vertex $z \in V \setminus M$ the labels as follows: $\varphi(xz) = i_z \in \Upsilon$ and $\varphi(zx) = j_z \in \Upsilon$.

M is a module in g if and only if $\varphi(yz) = i_z$ and $\varphi(zy) = j_z$ for all $y \in M, z \in V \setminus M$ if and only if $(y, z) \in E(G_{i_z}(g))$ and $(z, y) \in E(G_{j_z}(g))$ (and thus, $(y, z), (z, y) \notin E(G_k(g))$ for any $k \in \Upsilon$ with $k \neq i_z, j_z$) for all $y \in M, z \in V \setminus M$

Thus, M is a module in g if and only if M is a module in $G_l(g)$ for all $l \in \Upsilon$. □

A very useful property of modules is summarized by

Lemma 2 (Ehrenfeucht and Rozenberg (1990a), Lemma 4.11) *Let $X, Y \in \mathbb{M}(g)$ be two disjoint modules of $g = (V, \Upsilon, \varphi)$. Then there are labels $i, j \in \Upsilon$ such that $\varphi(xy) = i$ and $\varphi(yx) = j$ for all $x \in X$ and $y \in Y$.*

There are important subclasses of 2-structures g:

1. g is *prime* if $\mathbb{M}(g)$ consists of trivial modules only,
2. g is *complete* if for all $e, e' \in V^2_{\text{irr}}$, $\varphi(e) = \varphi(e')$,
3. g is *linear* if there are two distinct labels $i, j \in \Upsilon$ such that the relations $<_i, <_j$ defined by,

$$x <_i y \text{ iff } \varphi(xy) = i, \quad \text{and} \quad x <_j y \text{ iff } \varphi(xy) = j$$

are linear orders of the vertex set V_g.

In particular, if g is linear then there is a linear order $<$ of V such that $x < y$ if and only if $\varphi(xy) = i$ and $\varphi(yx) = j$. Clearly, if $|V_g| = 2$ all modules are trivial, and hence g is prime. On the other hand $|V_g| = 2$ also implies that g is either linear or complete. For $|V_g| \geq 3$, however, the tree types of 2-structures are disjoint.

Not all 2-structures necessarily fall into one of these three types. For example, the 2-structure g with $V_g = \{x, y, z\}$, $\varphi_g(xy) = \varphi_g(yx) = 1$, and $\varphi_g(xz) = \varphi_g(zx) = \varphi_g(zy) = \varphi_g(yz) = 2$ is neither prime, nor linear, nor complete. We finally note that our notion of prime is called "primitive" in Ehrenfeucht and Rozenberg (1990a); see also Ehrenfeucht et al. (1999); Engelfriet et al. (1996).

Definition 5 A 2-structure is *uniformly non-prime* (*unp*) if it does not have a prime substructure h of size $|V_h| \geq 3$.

Uniformly non-prime 2-structures were investigated already in Engelfriet et al. (1996). They are a key concept in the present contribution.

Definition 6 A module M of g is *strong* if M does not overlap with any other module of g.

This notion was termed "prime" in Ehrenfeucht and Rozenberg (1990a); see also Ehrenfeucht et al. (1999); Engelfriet et al. (1996). We write $\mathbb{M}_{\text{str}}(g) \subseteq \mathbb{M}(g)$ for the set of all strong modules of g.

While there may be exponentially many modules, the size of the set of strong modules is $O(|V|)$ (Ehrenfeucht et al. 1994). For example, the 2-structure $g = (V, \Upsilon, \varphi)$ with $\varphi(xy) = \varphi(ab)$ for all $(x, y), (a, b) \in V^2_{\text{irr}}$ has $2^{|V|}$ modules, however, the $|V|+1$ strong modules are V and the singletons $\{v\}$, $v \in V$.

Since V and the singletons $\{v\}$ are strong modules and strong modules do not overlap by definition, we see immediately that $\mathbb{M}_{\text{str}}(g)$ forms a hierarchy and by Thm. 1 gives rise to a unique tree representation T_g of g, also called *inclusion tree*. The vertices of T_g are (identified with) the elements of $\mathbb{M}_{\text{str}}(g)$. Adjacency in T_g is defined by the maximal proper inclusion relation, that is, there is an edge $\{M, M'\}$ between $M, M' \in \mathbb{M}_{\text{str}}(g)$ iff $M \subsetneq M'$ and there is no $M'' \in \mathbb{M}_{\text{str}}(g)$ such that $M \subsetneq M'' \subsetneq M'$. The root of T_g is V and the leaves are the singletons $\{v\}$, $v \in V$. Although $\mathbb{M}_{\text{str}}(g) \subseteq \mathbb{M}(g)$ does not represent all modules, any module $M \in \mathbb{M}(G)$ is the union of children of the strong modules in the tree T_g (Möhring and Radermacher 1984; Ehrenfeucht and Rozenberg 1990b). Thus, T_g represents at least implicitly all modules of g.

The hierarchical structure of $\mathbb{M}_{\text{str}}(g)$ implies that there is a unique partition $\mathbb{M}_{\max}(g) = \{M_1, \ldots, M_k\}$ of V_g into maximal (w.r.t. inclusion) strong modules $M_j \neq V_g$ of g (Ehrenfeucht and Rozenberg 1990a, b). Since $V_g \notin \mathbb{M}_{\max}(g)$ the set $\mathbb{M}_{\max}(g)$ consists of $k \geq 2$ strong modules, whenever $|V_g| > 1$.

In order to infer g from T_g we need to determine the label $\varphi(xy)$ of all pairs of distinct leaves x, y of T_g and thus of V_g. Hence, we need to define a labeling function t_g that assigns the "missing information" to the inner vertex of T_g. To this end, we will need to understand the quotient $g/\mathbb{M}_{\max}(g)$, i.e., the 2-structure $(\mathbb{M}_{\max}(g), \Upsilon, \varphi')$ with $\varphi'(M_i, M_j) = \varphi(xy)$ for some $x \in M_i$ and $y \in M_j$. Thus a quotient $g/\mathbb{M}_{\max}(g)$ is obtained from g by contracting each module in $M \in \mathbb{M}_{\max}(g)$ into a single node, and then inheriting the edge classes from g. By Lemma 2, the quotient $g/\mathbb{M}_{\max}(g)$ is well-defined. Although 2-structures are not necessarily prime, linear or complete, their quotients $g/\mathbb{M}_{\max}(g)$ are always of one of these types.

Lemma 3 (Ehrenfeucht and Rozenberg 1990b; Engelfriet et al. 1996) *Let g be a 2-structure. Then the quotient $g/\mathbb{M}_{\max}(g)$ is either linear, or complete, or prime. If g is unp, then $g/\mathbb{M}_{\max}(g)$ is either linear, or complete.*

We shall say that an inner vertex v of T_g (or, equivalently, the module $L(v)$) is linear, complete, or prime if the quotient $g[L(v)]/\mathbb{M}_{\max}(g[(L(v)])$ is linear, complete, or prime, respectively. In order to recover g from T_g one defines a labeling function σ that assigns the quotient of $g[L(v)]$ to each inner vertex v:

$$\sigma(v) = g[L(v)]/\mathbb{M}_{\max}(g[L(v)]).$$

This type of labeled tree-representation is called shape(g) or (strong) module decomposition of g (Ehrenfeucht and Rozenberg 1990a, b; Engelfriet et al. 1996). If we restrict ourselves to *unp* structures, the strong modules are "generic" in the sense that they are completely determined by the cardinalities of their domains and the ordering of the vertices in T_g.

We can therefore define a simplified labeling function t_g for *unp* structures g. If $M \in \mathbb{M}_{\max}(g)$ is a complete module and $M_1, \ldots, M_l$ are the children of M, then there is an $i \in \Upsilon$ such that for all vertices $x \in M_r$, $y \in M_s$, $r \neq s$ we have $\varphi(xy) = \varphi(yx) = i$. Therefore, we can set $t_g(M) = (i, i)$, implying that for all vertices x, y with $\text{lca}(x, y) = M$ it holds that $\varphi(xy) = \varphi(yx) = i$. If $M \in \mathbb{M}_{\max}(g)$ is a linear module, then we can assume that the children of M are ordered $M_1, \ldots, M_l$ such that

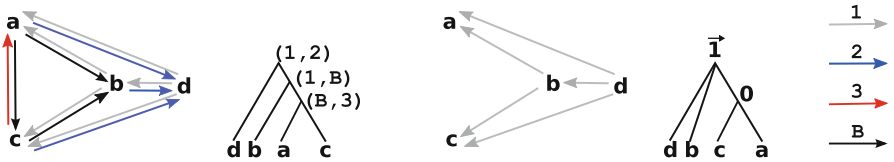

Fig. 5 Example of a *unp* 2-structure g with its tree-representation (T_g, t_g) (*1st* and *2nd from left*) and an underlying di-cograph $G_1(g)$ with respective cotree (*3rd* and *4th from left*). The *arc-colors* correspond to the respective labels `1`, `2`, `3`, `B` (*right-most*). In fact, all underlying di-graphs $G_2(g)$, $G_3(g)$ and $G_B(g)$ are di-cographs. To see that (U2) is satisfied for g observe that $D_{abc} = \{\{B, 1\}, \{B, 3\}\}$, $D_{acd} = \{\{B, 3\}, \{1, 2\}\}$, and $D_{abd} = D_{bcd} = \{\{B, 1\}, \{1, 2\}\}$ (color figure online)

$\varphi(xy) = i$ and $\varphi(yx) = j$ for some $i, j \in \Upsilon$ if and only if $x \in M_r$, $y \in M_s$ and $1 \leq r < s \leq l$. Therefore, we can set $t_g(M) = (i, j)$, implying that for all vertices x, y with $\mathrm{lca}(x, y) = M$ and x is left of y in T_g it holds that $\varphi(xy) = i$ and $\varphi(yx) = j$.

Definition 7 The *tree-representation* (T_g, t_g) of a *unp* 2-structure g is an ordered inclusion tree T_g of the hierarchy $\mathbb{M}_{\mathrm{str}}(g)$ together with a labeling $t_g : V^2_{\mathrm{irr}} \to \Upsilon$ such that for all vertices $x, y \in V_g$ of the *unp* 2-structure g it holds that

$$t_g(\mathrm{lca}(x, y)) = (i, j),$$

where $i = j$ if and only if $\varphi_g(xy) = \varphi_g(yx) = i$; and x is to the left of y in T_g if and only if $\varphi_g(xy) = i, \varphi_g(yx) = j$, and $i \neq j$.

Figure 5 shows an illustrative example.

We call two tree-representations (T, t) and (T', t') of a 2-structure g *isomorphic* if T and T' are isomorphic via a map $\psi : V(T) \to V(T')$ such that $t'(\psi(v)) = t(v)$ holds for all $v \in V(T)$. In the latter we write $(T, t) \simeq (T', t')$

Lemma 4 (Ehrenfeucht and Rozenberg 1990b; Engelfriet et al. 1996) *For any 2-structures h, g we have $(T_g, t_g) \simeq (T_h, t_h)$ if and only if $h \simeq g$.*

The tree-representation (T, t) of g contains no prime nodes if and only if g has no induced substructures of small size that are prime (Ehrenfeucht and Rozenberg 1990; Schmerl and Trotter 1993), see Ehrenfeucht et al. (1999) for a concise proof. The following theorem summarizes a couple of important properties of *unp* 2-structures Engelfriet et al. (1996, cf. Thm. 3.6).

Theorem 3 *If g is a 2-structure then the following statements are equivalent:*

1. *g is unp.*
2. *The tree-representation (T_g, t_g) of g has no inner vertex v labeled prime, that is, the quotient $g[L(v)]/\mathbb{M}_{\max}(g[(L(v)])$ is always linear or complete.*
3. *g has no prime substructure of size 3 or 4.*

In particular, if g is *unp*, then every substructure on a subset X with $|X| = 3$ or $|X| = 4$ has at least one non-trivial module, i.e, a module $M \subseteq X$ with $|M| \geq 2$.

We next examine a particular subclass of 2-structures, the so-called reversible 2-structures. As we shall see, they are simpler to handle than general 2-structures. Nevertheless, there is no loss of generality as far as modules are concerned.

Definition 8 A 2-structure $g = (V, \Upsilon, \varphi)$ is *reversible*, if for all $e, f \in V^2_{\text{irr}}$, $\varphi(e) = \varphi(f)$ implies that $\varphi(e^{-1}) = \varphi(f^{-1})$.

Equivalently, g is reversible, if for each label $i \in \Upsilon$ there is a unique label $j \in \Upsilon$ such that $\varphi(x, y) = i$ implies $\varphi(y, x) = j$.

The definition of modules simplifies for reversible 2-structures. It suffices to require that M satisfies $\varphi(xz) = \varphi(yz)$ for all $z \in V \backslash M$ and $x, y \in M$, because $\varphi(xz) = \varphi(yz)$ and reversibility implies $\varphi(zx) = \varphi(zy)$.

Definition 9 A 2-structure $\text{rev}(g)$ is the reversible refinement of the 2-structure $g = (V, \Upsilon, \varphi)$ if and only if (i) $V_{\text{rev}(g)} = V$ and (ii) for all $e, f \in V^2_{\text{irr}}$ it holds that $\varphi_{\text{rev}(g)}(e) = \varphi_{\text{rev}(g)}(f)$ if and only if $\varphi(e) = \varphi(f)$ and $\varphi(e^{-1}) = \varphi(f^{-1})$.

Let us recall some well-established results concerning (reversible) 2-structures.

Theorem 4 (Ehrenfeucht and Rozenberg (1990a)) *For every 2-structure g the following properties hold:*

1. $\text{rev}(g)$ *is reversible, i.e.,* $\text{rev}(\text{rev}(g)) = \text{rev}(g)$.
2. g *is reversible iff* $g = \text{rev}(g)$.
3. $\mathbb{M}(g) = \mathbb{M}(\text{rev}(g))$.
4. *A 2-structure h is a substructure of g iff* $\text{rev}(h)$ *is a substructure of* $\text{rev}(g)$.

According to Ehrenfeucht and Rozenberg (1990b, Remark 6.3), we can fix the labels of the reversible refinement of g and construct $\text{rev}(g)$ as follows.

Remark 1 For a given 2-structure $g = (V, \Upsilon, \varphi)$ set $\text{rev}(g) = (V, \Upsilon_{\text{rev}(g)} := \Upsilon \times \Upsilon, \varphi_{\text{rev}(g)})$ with $\varphi_{\text{rev}(g)}(e) = (\varphi(e), \varphi(e^{-1}))$ that maps each arc e to the ordered pairs of labels on e and its reverse e^{-1}.

Thus, for the 2-structures $g = (V, \Upsilon, \varphi)$ and the so-constructed 2-structure $\text{rev}(g)$ we have for all $e, f \in V^2_{\text{irr}}$: $\varphi_{\text{rev}(g)}(e) = \varphi_{\text{rev}(g)}(f) = (i, j)$ if and only if $\varphi(e) = \varphi(f) = i$ and $\varphi(e^{-1}) = \varphi(f^{-1}) = j$. Hence, $\text{rev}(g)$ is a reversible refinement of g.

Remark 2 Each symbolic ultrametric $\delta\colon V^2_{\text{irr}} \to \Upsilon$ gives rise to a 2-structure $g = (V, \Upsilon, \varphi)$ with $\varphi = \delta$. To simplify the language, we will say that g satisfies Condition (U1) and (U2) whenever φ satisfies (U1) and (U2).

3 Characterization of *unp* 2-structures

Let $\mathscr{R} = \{R_1, \ldots, R_n\}$ be a set of disjoint relations. Or goal is to characterize those $\mathscr{R}$ that are obtained from a common tree T and a suitable labeling t of the inner vertices of T. To this end we need to understand on the one hand the relationships by symbolic ultrametrics and event-labeled trees, and on the other hand the connection between symbolic ultrametrics and 2-structures. Moreover, we will establish the connection between 2-structures g and certain modules in the modular decomposition of the underlying graphs $G_i(g)$. The main results of this section are summarized in Theorem 6.

3.1 2-Structures and symbolic ultrametrics

We begin this section with the characterization of reversible 2-structures by means of symbolic ultrametrics, which will be generalized to arbitrary 2-structures at the end of this subsection.

Proposition 1 *For every reversible 2-structure $g = (V, \Upsilon, \varphi)$ the following two statements are equivalent:*

(1) *g is unp.*
(2) *φ is a symbolic ultrametric.*

Proof The straightforward but tedious case-by-case analysis is given in the appendix.

The next step is to generalize Proposition 1 to arbitrary 2-structures. To this end, we first prove two technical results:

Lemma 5 *Let $g = (V, \Upsilon, \varphi)$ be a 2-structure. Then condition (U2) is satisfied for g if and only if (U2) is satisfied in* $\mathrm{rev}(g)$.

Proof Condition (U2) is satisfied in g if and only if $|D_{abc}| \leq 2$ in g for all $a, b, c \in V$ if and only if there are two arcs e, f in this triangle induced by a, b, c such that $\varphi(e) = \varphi(f)$ and $\varphi(e^{-1}) = \varphi(f^{-1})$ if and only if $\varphi_{\mathrm{rev}(g)}(e) = \varphi_{\mathrm{rev}(g)}(f)$ (and thus, by reversibility of g, $\varphi_{\mathrm{rev}(g)}(e^{-1}) = \varphi_{\mathrm{rev}(g)}(f^{-1})$) if and only if $|D_{abc}| \leq 2$ in $\mathrm{rev}(g)$ for all $a, b, c \in V$. □

Lemma 6 *Let $g = (V, \Upsilon, \varphi)$ be a 2-structure satisfying (U2). Then $G_i(g)$ is a di-cograph for all $i \in \Upsilon$ if and only if $G_j(\mathrm{rev}(g))$ is a di-cograph for all $j \in \Upsilon_{\mathrm{rev}(g)}$.*

Proof The straightforward but tedious case-by-case analysis is given in the appendix.

It is now easy to establish our first main result:

Theorem 5 *The following two statements are equivalent for all 2-structures $g = (V, \Upsilon, \varphi)$:*

(1) *g is unp.*
(2) *φ is a symbolic ultrametric.*

Proof Lemmas 5 and 6 together imply that φ is a symbolic ultrametric if and only $\varphi_{\mathrm{rev}(g)}$ is a symbolic ultrametric. Proposition 1 implies that $\varphi_{\mathrm{rev}(g)}$ is a symbolic ultrametric if and only if $\mathrm{rev}(g)$ is *unp*. Now recall that *unp* 2-structures are defined in terms of their modules and that $\mathbb{M}(g) = \mathbb{M}(\mathrm{rev}(g))$ (cf. Theorem 4(4)). Therefore, $\mathrm{rev}(g)$ is a *unp* 2-structure if and only if g is a *unp* 2-structure. The theorem follows immediately. □

3.2 2-Structures and 1-clusters

Assume that $g = (V, \Upsilon, \varphi)$ is a 2-structure with the property that $G_i(g)$ is a di-cograph for all $i \in \Upsilon$. Each di-cograph $G_i(g)$ is represented by a unique ordered tree T_i, called

cotree (Corneil et al. 1981; Crespelle and Paul 2006). Recall, in our notation the label of an inner vertex in the cotree is always one of 0, 1, $\overrightarrow{1}$. We say that a leaf set $L(v)$ is a *1-cluster of* T_i if v has a label distinct from 0. The set $\mathscr{C}_i^1$ of 1-clusters of T_i therefore is a subset of the clusters that form the hierarchy equivalent to T_i. Consider the set

$$\mathscr{C}^1(g) := \cup_{i\in\Upsilon}\mathscr{C}_i^1 \cup \big\{\{v\} \mid v \in V_g\big\}$$

comprising the 1-clusters for each T_i and the singletons.

Remark 3 Any two disjoint (graph-)modules M, M' of a di-graph G are either adjacent or non-adjacent, i.e., for each vertex of $x \in M$ and each vertex of $y \in M'$ there is an arc (x, y) or (y, x) in G or there is no edge between any vertex of M and any vertex of M' (Möhring 1985; Engelfriet et al. 1996). Now, let $G = (V, E)$ be a di-cograph, $M \in \mathbb{M}_{\text{str}}(G)$ a strong module of G and $M_1, \ldots, M_l$ the children of M in the respective cotree, i.e., the inclusion-maximal elements of $\mathbb{M}_{\text{str}}(G[M])$. By construction, module M is a 1-cluster of G if and only if M is a series or order module. Moreover, M is a 1-cluster if and only if for two vertices $x \in M_i$ and $y \in M_j$, with $i \neq j$ there exists at least one of the arcs $(x, y) \in E$ or $(y, x) \in E$.

Lemma 7 *Let $g = (V, \Upsilon, \varphi)$ be a* reversible *unp 2-structure. Then $\mathscr{C}^1(g)$ is a hierarchy and, in particular, $\mathbb{M}_{\text{str}}(g) = \mathscr{C}^1(g)$.*

Moreover, for each cluster $M \in \mathscr{C}^1(g)$ there are at most two distinct di-cographs $G_i(g), G_j(g)$ such that $M \in \mathscr{C}_i^1$ and $M \in \mathscr{C}_j^1$. In other words, each cluster $M \in \mathscr{C}^1(g)$ appears as a 1-cluster in at most two different cotrees.

Proof We show that $\mathbb{M}_{\text{str}}(g) = \mathscr{C}^1(g)$. It then follows that $\mathscr{C}^1(g)$ is a hierarchy.

We start by proving that $\mathbb{M}_{\text{str}}(g) \subseteq \mathscr{C}^1(g)$. Since g is *unp* there is a tree-representation (T_g, t_g) of g. Let v be an inner vertex in T labeled with (i, j). By construction each vertex v of this tree T_g represents a strong module $L(v)$ of g. Hence, $L(v) \in \mathbb{M}_{\text{str}}(g)$ and thus, we can apply Lemma 1 and conclude that $L(v)$ is a module of $G_i(g)$.

We continue to prove that $L(v)$ is a strong module of $G_i(g)$. This is then used to show that $L(v)$ is contained in $\mathscr{C}_i^1 \subseteq \mathscr{C}^1(g)$ and hence, $\mathbb{M}_{\text{str}}(g) \subseteq \mathscr{C}^1(g)$. To this end, we show first that all $a, b \in L(v)$ are contained in the same connected subgraph of both $G_i(g)$ and $G_j(g)$.

Claim 1 All $a, b \in L(v)$ are contained in the same connected subgraph of both $G_i(g)$ and $G_j(g)$. Moreover, $\varphi(ab) = i$ if and only if $\varphi(ba) = j$ for all $a, b \in V$.

Proof of Claim 1 Let $v_1, \ldots, v_k$ be the children of v, ordered from left to right. If $i = j$, an ordering is not necessary. For leaves $x \in L(v_r)$ and $y \in L(v_s)$ we have now $\varphi(xy) = i$ and $\varphi(yx) = j$ if $r < s$. Hence, $x \in L(v_r)$ and $y \in L(v_s)$ and $r < s$ implies that $(x, y) \in E(G_i(g))$ and thus, all vertices $L(v)$ are contained in one connected subgraph of $G_i(g)$. Analogously, all vertices $L(v)$ are contained in one connected subgraph of $G_j(g)$. Since $\varphi(xy) = i, \varphi(yx) = j$ if $r < s$ and g is reversible, we observe that $\varphi(ab) = i$ if and only if $\varphi(ba) = j$ for all $a, b \in V$. □

Claim 2 $L(v)$ is a strong in $G_i(g)$ and $G_j(g)$.

Proof of Claim 2 Assume, for contradiction, that the module $L(v)$ is not strong in $G_i(g)$. Hence there is a further module M in $G_i(g)$ such that $M \between L(v)$. Since $L(v)$ is a strong module in g, M cannot be a module in g. Since $M \between L(v)$, we have $M \cap L(v) \neq \emptyset$, $M \not\subseteq L(v)$ and $L(v) \not\subseteq M$. By the latter, and since all $a, b \in L(v)$ are contained in one connected subgraph of $G_i(g)$ there must be an arc (u, x) or (x, u) in $G_i(g)$, for some $x \in L(v) \backslash M$ and $u \in M \cap L(v)$. Wlog. assume that (u, x) is an arc in $G_i(g)$, since the following arguments can be applied analogously for the case that (x, u) is an arc in $G_i(g)$. Thus, we have $\varphi(ux) = i$ and since g is reversible, $\varphi(xu) = j$.

Since M and $L(v)$ are modules in $G_i(g)$ and there is an arc (u, x) in $G_i(g)$, where particularly $x \in L(v)$ and $u \in M$, we can conclude that for all vertices $x' \in L(v)$, there is an arc (y, x') in $G_i(g)$. This implies additionally that all $y \in M$ form an arc (y, u), since u is also contained in $L(v)$. However, since M is not a module in g and g is reversible, there must be a vertex $y \in M$ such that $\varphi(yz) \neq \varphi(uz)$ and $\varphi(zy) \neq \varphi(zu)$ for some $z \in V \backslash M$. Since $\varphi(yx') = \varphi(ux')$ for all $x' \in L(v)$, we can conclude that for the latter chosen vertex z we have $z \in V \backslash (M \cup L(v))$.

Since $\varphi(yz) \neq \varphi(uz), \varphi(zy) \neq \varphi(zu)$, g is reversible, and M is a module in $G_i(g)$ with $u, y \in M$, we conclude that $\varphi(yz), \varphi(zy), \varphi(uz)$, and $\varphi(zu)$ must all be different from i. To see this, assume for contradiction that for some $e \in \{(y, z), (u, z)\}$ it holds that $\varphi(e) = i$. Since M is a module in $G_i(g)$ it follows that $\varphi(f) = i$ for $f \in \{(y, z), (u, z)\} \backslash \{e\}$; a contradiction to $\varphi(yz) \neq \varphi(uz)$. The same argument applies for $e, f \in \{(z, y), (z, u)\}$.

Thus, assume that $\varphi(yz) = l \neq \varphi(uz) = k$ for some $l, k \in \Upsilon$ distinct from i. Since all $\varphi(yz), \varphi(zy), \varphi(uz), \varphi(zu)$ are distinct from i while $\varphi(yu) = i$, and since Condition (U2) must be fulfilled for the set D_{yuz}, we obtain $\varphi(zy) = k$, and $\varphi(zu) = l$. Since g is reversible, neither $\varphi(uy) = k$, nor $\varphi(uy) = l$. But then there is $D_3^k(yuz)$ and $D_3^l(yuz)$ in $G_k(g)$ and $G_l(g)$; a contradiction to (U1).

Thus, $L(v)$ is a strong module in $G_i(g)$. By analogous arguments, $L(v)$ is a strong module in $G_j(g)$. □

Claim 3 $L(v)$ is contained in $\mathscr{C}_i^1 \subseteq \mathscr{C}^1(g)$. Therefore, $\mathbb{M}_{\text{str}}(g) \subseteq \mathscr{C}^1(g)$.

Proof of Claim 3 All strong modules of $G_i(g)$ are represented in the respective cotree T_i. As already observed, since $t_g(v) = (i, j)$ for all leaves $x \in L(v_r)$ and $y \in L(v_s)$ we have $\varphi(xy) = i$ if $r < s$. Hence, (x, y) is an arc in $G_i(g)$ for all $x \in L(v_r)$, $y \in L(v_s)$, $r < s$. If $i = j$ then even (y, x) is an arc in $G_i(g)$. Hence, $L(v)$ cannot be labeled with "0" in the cotree, because otherwise it is not possible to have all arcs (x, y) with $x \in L(v_1)$ and $y \in L(v_i)$, $1 < i \leq k$. In particular, if $i \neq j$, then $L(v)$ must labeled "$\overrightarrow{1}$" in $G_i(g)$; if $i = j$, then the strong module $L(v)$ must labeled "1" in $G_i(g)$. Hence, $L(v)$ is contained in $\mathscr{C}_i^1 \subseteq \mathscr{C}^1(g)$. Therefore, $\mathbb{M}_{\text{str}}(g) \subseteq \mathscr{C}^1(g)$. □

We proceed to show that $\mathscr{C}^1(g) \subseteq \mathbb{M}_{\text{str}}(g)$. Let $L(v) \in \mathscr{C}_i^1 \subseteq \mathscr{C}^1(g)$ be a strong module with label different from "0" obtained from the cotree T_i. Clearly, $|L(v)| > 1$, since the singletons $\{v\}$ are by definition not contained in $\mathscr{C}_i^1$, albeit they are by construction contained in $\mathscr{C}^1(g)$.

Claim 4 $L(v) \in \mathscr{C}_i^1 \subseteq \mathscr{C}^1(g)$ is a module of g.

Proof of Claim 4 Assume for contradiction that $L(v)$ is not a module in g. Since g is reversible, there must be two vertices $a, b \in L(v)$ and $c \in V \backslash L(v)$ such that $\varphi(ac) = j \neq \varphi(bc) = k$. In particular, j and k must both be distinct from i, as otherwise $L(v)$ would not be a module in $G_i(g)$. Since g is reversible, $\varphi(ca) \neq \varphi(cb)$ and by analogous arguments as before, neither $\varphi(ca) = i$ nor $\varphi(cb) = i$. The latter arguments and reversibility of g imply that $\varphi(xc) \neq i$ and $\varphi(cx) \neq i$ for all $x \in L(v)$, as otherwise $L(v)$ would not be a module in $G_i(g)$ or $L(v)$ would be a module in g; a contradiction. Since $L(v) \in \mathscr{C}_i^1$ the vertex v has either label 1 or $\overrightarrow{1}$ in T_i. Thus, the subgraph in $G_i(g)$ induced by the vertices in $L(v)$ is connected. Let $(a, v_1, \ldots, v_n, b)$ be a walk in $G_i(g)$ with $v_i \in L(v)$ for $1 \leq i \leq n$, that connects the vertices a and b. Since none of the labels $\varphi(ac), \varphi(ca), \varphi(v_1c), \varphi(cv_1)$ is i, and since Condition (U2) must be fulfilled for the set D_{av_1c}, we can conclude that the label $\varphi(ac) = j$ must occur on at least one of the arcs (v_1c) or (cv_1). We distinguish the two cases (i) $\varphi(v_1c) = j \neq \varphi(cv_1)$ and (ii) $\varphi(v_1c) = \varphi(cv_1) = j$. □

Case (i) Since none of the labels $\varphi(v_2c), \varphi(cv_2)$ is i, and Condition (U2) must be fulfilled for the set $D_{v_1v_2c}$, we obtain that $\varphi(v_2c) = j$ or $\varphi(cv_2) = j$. Repeating the latter, we obtain $\varphi(v_nc) = j$ or $\varphi(cv_n) = j$. Since Condition (U2) must be fulfilled for the set D_{v_nbc}, and none of the labels of (v_nc) and (cv_n) is i, but $\varphi(bc) = k$, we can conclude that $\varphi(cb) = j$ and the labels j and k must occur on the two arcs (v_n, c) and (c, v_n). The case $\varphi(cv_n) = k$ cannot occur, since then there is $D_3^k(bcv_n)$. Thus, $\varphi(v_nc) = k$ and by reversibility of g, $\varphi(cv_n) = j$. By analogous arguments, $\varphi(cv_{n-1}) = j$ and $\varphi(v_{n-1}c) = k$, and, iterative, $\varphi(cv_1) = j$, $\varphi(v_1c) = k$ and $\varphi(ca) = k$. Since g is reversible and $\varphi(v_1a) = i$ or $\varphi(av_1) = i$ we can conclude that there are forbidden subgraphs $D_3^j(acv_1)$ and $D_3^k(acv_1)$; a contradiction.

Case (ii) If $\varphi(cv_1) = \varphi(v_1c) = j$, then by analogous arguments as in Case (i), $\varphi(cv_n) = \varphi(v_nc) = j$. But then $|D_{bcv_n}| = 3$, violating (U2) and thus, g is not *unp*.

In summary, $L(v)$ is a module of g. □

It remains to show that $L(v) \in \mathscr{C}^1(g)$ is also strong in g. Assume for contradiction that $L(v)$ is not strong in g. Hence, there is a module M in g with $|M| > 1$ and $L(v) \between M$. Lemma 1 implies that M is also a module in $G_i(g)$ and hence, M overlaps $L(v)$ in $G_i(g)$; a contradiction, since $L(v)$ is strong in $G_i(g)$. Therefore, $\mathscr{C}^1(g) \subseteq \mathbb{M}_{\text{str}}(g)$.

Since $\mathbb{M}_{\text{str}}(g) \subseteq \mathscr{C}^1(g)$ and $\mathscr{C}^1(g) \subseteq \mathbb{M}_{\text{str}}(g)$ we can conclude that $\mathscr{C}^1(g) = \mathbb{M}_{\text{str}}(g)$ and thus, $\mathscr{C}^1(g)$ is a hierarchy.

Finally, we show that each cluster $M \in \mathscr{C}^1(g)$ appears at most in two different cotrees. Let $M \in \mathscr{C}^1(g) = \mathbb{M}_{\text{str}}(g)$ and assume that $t_g(M) = (i, j)$. By Claims 1, 2, and 3, M is a strong module in $G_i(g)$ with label $\overrightarrow{1}$, if $i \neq j$, and label 1, if $i = j$ in the cotree of $G_i(g)$. Thus, $M \in \mathscr{C}_i^1$ in all cases and additionally, $M \in \mathscr{C}_j^1$, if $i \neq j$.

It remains to show that there is no further $k \in \Upsilon, k \neq i, j$ with $M \in \mathscr{C}_k^1$. Assume that this is not the case, and thus there is a di-cograph $G_k(g)$ such that M is labeled 1 or $\overrightarrow{1}$, in the respective cotree T_k. Let $M_1, \ldots, M_r$ be the children of M in T_g

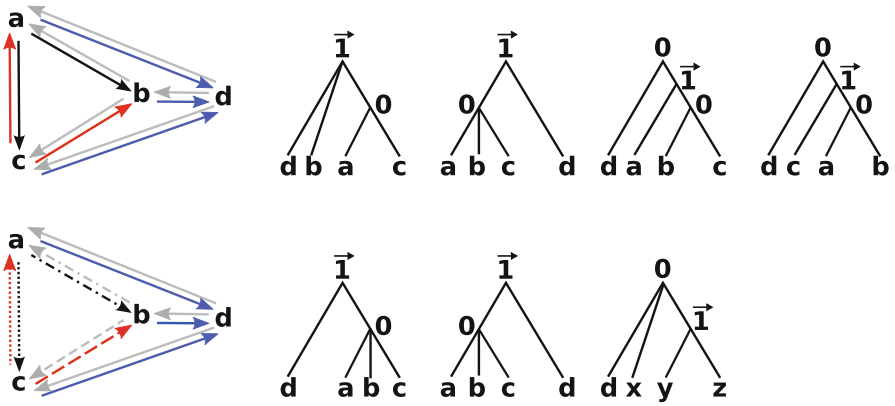

Fig. 6 Example of a non-reversible 2-structure g (*upper left part*) and its reversible refinement $\mathrm{rev}(g)$ (*lower left part*). Labels are indicated by *colored arcs*. The triangle-condition (U2) is violated in both g and $\mathrm{rev}(g)$. Hence, neither 2-structure has a tree representation despite the fact that all $G_i(g)$ and $G_i(\mathrm{rev}(g))$ are cographs. There are four different cographs, and hence cotrees, for g and eight for $\mathrm{rev}(g)$. From these we obtain $\mathscr{C}^1(g) = \{\{a,b,c,d\},\{a,b,c\},\{a\},\{b\},\{c\},\{d\}\}$. Note that $\mathscr{C}^1(g)$ is a hierarchy even though g is not *unp*. Moreover, the clusters $\{a,b,c,d\}, \{a,b,c\}$ are both contained in two different cotrees. Thus the converse of Lemma 7 does not hold in general for non-reversible 2-structures. The leaves x, y, z of the right-most cotree of $\mathrm{rev}(g)$ can be chosen arbitrarily, as long as $\{x,y,z\} = \{a,b,c\}$. Thus, $C^1(\mathrm{rev}(g))$ contains the cluster $\{a,b\}$ and $\{b,c\}$ and therefore does not form a hierarchy, cf. Lemma 8 and Theorem 6 (color figure online)

and $N_1, \ldots, N_s$ be the children of M in T_k. Let x_l be a vertex contained in M_l for $1 \leq l \leq r$. Since every arc between distinct $x_l, x_{l'}$ is labeled i or j, and in particular, not k, and since $x_1, \ldots, x_r$ are contained in M, we can conclude that $x_1, \ldots, x_r \in N_m$ for some $m \in \{1, \ldots, s\}$. Otherwise, there would be a label k on some arc between some $x_l, x_{l'}$. Now take a further vertex $y \in M_l$. By analogous arguments for the vertices $x_1, \ldots, x_{l-1}, y, x_{l+1}, \ldots, x_r$ and since $x_1, \ldots, x_r \in N_m$, we obtain that $y \in N_m$. By induction, all vertices in $\cup_{i=1}^r M_i$ must be contained in N_m. Thus, $M = \cup_{i=1}^r M_i \subseteq N_m \subsetneq M$; a contradiction. Therefore, $M \in \mathscr{C}_i^1$, $M \in \mathscr{C}_j^1$ but $M \notin \mathscr{C}_k^1$ for any $k \neq i, j$, from what the statement follows. □

Although it might be possible to derive a result similar to Lemma 7 for non-reversible 2-structures g (with some more elaborated technical arguments), the fact that $\mathscr{C}^1(g)$ is a hierarchy and that each 1-cluster appears in at most 2 cotrees is not sufficient to conclude that g is *unp*. Figure 6 gives a counterexample. Surprisingly, however, the triangle-condition (U2) and the property that $\mathscr{C}^1(g)$ is a hierarchy, are equivalent for reversible 2-structures that fulfill (U1).

Lemma 8 *Let $g = (V, \Upsilon, \varphi)$ be a* reversible *2-structure such that $G_i(g)$ is a di-cograph for all $i \in \Upsilon$. Then the following statements are equivalent*

1. *The Triangle-Condition (U2) is satisfied for g,*
2. *$\mathscr{C}^1(g)$ is a hierarchy. In particular, $\mathbb{M}_{\mathrm{str}}(g) = \mathscr{C}^1(g)$.*

Proof If g satisfies (U1) and (U2), then, by Theorem 5, g is *unp*. Now apply Lemma 7.

Suppose g does not satisfy (U2), i.e., there are three vertices $a, b, c \in V$ such that $\varphi(ab) = i$ and $|D_{abc}| = 3$. Since g is reversible we conclude that $\varphi(ac), \varphi(ca), \varphi(bc)$,

and $\varphi(cb)$ are all distinct from i. In the cotree T_i of $G_i(g)$, $\mathrm{lca}(ab)$ must be labeled either 1 or $\overrightarrow{1}$. Next we observe that c cannot be descendant of $\mathrm{lca}(ab)$ in T_i, since otherwise we would have $\mathrm{lca}(ac) \in \{1, \overrightarrow{1}\}$ or $\mathrm{lca}(bc) \in \{1, \overrightarrow{1}\}$. This implies that at least one of the arcs (ac), (ca), (bc), (cb) must be present in $G_i(g)$, which is only possible iff φ mapped one of those arcs to the label i; a contradiction. Therefore, there is a cluster in T_i that contains a and b but not c, and this cluster is also contained in $\mathscr{C}^1(g)$. Now, let $\varphi(ac) = j \neq i$. Since g is reversible, we know that $\varphi(ab)$, $\varphi(ba)$, $\varphi(bc)$, and $\varphi(cb)$ are all distinct from j. Using the same argument as above, one can show that in the cotree T_j there is a cluster containing a and c but not b, which is contained in $\mathscr{C}^1(g)$. But then these two particular clusters overlap, and hence, $\mathscr{C}^1(g)$ is not a hierarchy. Since $\mathbb{M}_{\mathrm{str}}(g)$ is a hierarchy, we can conclude that $\mathscr{C}^1(g) \neq \mathbb{M}_{\mathrm{str}}(g)$. □

Corollary 1 *Let $g = (V, \Upsilon, \varphi)$ be a 2-structure such that $G_i(g)$ is a di-cograph for all $i \in \Upsilon$. The following statements are equivalent*

1. *g satisfies the Triangle-Condition (U2).*
2. *$\mathscr{C}^1(\mathrm{rev}(g))$ is a hierarchy.*
3. *$\mathscr{C}^1(\mathrm{rev}(g)) = \mathbb{M}_{\mathrm{str}}(g)$.*

Proof By Lemma 5, g satisfies (U2) if and only if $\mathrm{rev}(g)$ satisfies (U2). Together with Lemma 6 this implies that g satisfies (U1) if and only if $\mathrm{rev}(g)$ satisfies (U1). Therefore $\mathrm{rev}(g)$ satisfies (U1) and (U2), which, by Lemma 8, is equivalent to $\mathrm{rev}(g)$ satisfying (U1) and $\mathscr{C}^1(\mathrm{rev}(g))$ being a hierarchy. In this case we have in particular $\mathscr{C}^1(\mathrm{rev}(g)) = \mathbb{M}_{\mathrm{str}}(\mathrm{rev}(g))$. By Theorem 4(4), $\mathbb{M}(\mathrm{rev}(g)) = \mathbb{M}(g)$ and hence, $\mathbb{M}_{\mathrm{str}}(\mathrm{rev}(g)) = \mathbb{M}_{\mathrm{str}}(g)$. Therefore, the statement is true. □

Collecting the results derived above we obtain the main result of this contribution (Fig. 7):

Theorem 6 (Characterization of *unp* 2-structures) *The following statements are equivalent for every 2-structure $g = (V, \Upsilon, \varphi)$:*

1. *g is unp.*
2. *φ is a symbolic ultrametric.*
3. *g fulfills the following two properties:*
   (a) *$G_i(g)$ is a di-cograph for all $i \in \Upsilon$.*
   (b) *$\mathscr{C}^1(\mathrm{rev}(g))$ is a hierarchy. In particular, $\mathscr{C}^1(\mathrm{rev}(g)) = \mathbb{M}_{\mathrm{str}}(g)$.*
4. *$\mathrm{rev}(g)$ fulfills the following two properties:*
   (a) *$G_i(rev(g))$ is a di-cograph for all $i \in \Upsilon_{\mathrm{rev}(g)}$.*
   (b) *$\mathscr{C}^1(\mathrm{rev}(g))$ is a hierarchy.*
5. *$\mathrm{rev}(g)$ is unp.*

Proof The equivalence of Item (1.) and (2.) are already given in Theorem 5. Hence, g satisfies (U1) and (U2). By Corollary 1, φ is a symbolic ultrametric if and only if (U1) and the condition that $\mathscr{C}^1(\mathrm{rev}(g))$ is a hierarchy (with $\mathscr{C}^1(\mathrm{rev}(g)) = \mathbb{M}_{\mathrm{str}}(g)$) is satisfied. Thus, Item (2.) and (3.) are equivalent. By Lemma 6 and since g satisfies (U2), g satisfies (U1) if and only if $\mathrm{rev}(g)$ satisfies (U1), and thus Item (3.) and (4.) are equivalent. Lemma 8 implies that the statement that $\mathrm{rev}(g)$ satisfies (U1) so that

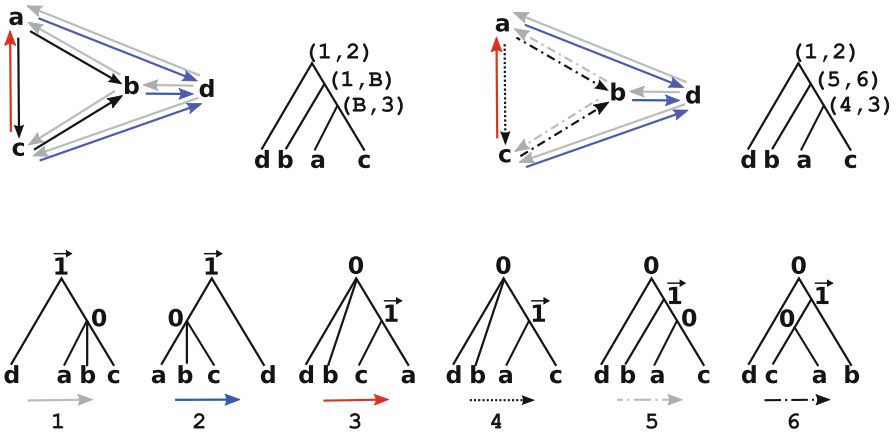

Fig. 7 Example of a non-reversible *unp* 2-structure g with tree-representation (T_g, t_g) (*upper left*) and its reversible refinement $\mathrm{rev}(g)$ with $(T_{\mathrm{rev}(g)}, t_{\mathrm{rev}(g)})$ (*upper right*). Labels are indicated by *colors* and *line-styles*. The label B in the right-most tree corresponds to the *solid black arc* in g. All $G_i(g)$ and $G_i(\mathrm{rev}(g))$ are di-cographs and the triangle condition is satisfied. The cotrees corresponding to the $G_i(\mathrm{rev}(g))$ are shown in the *second row*. They generate $\mathscr{C}^1(\mathrm{rev}(g)) = \{\{a,b,c,d\},\{a,b,c\},\{a,c\},\{a\},\{b\},\{c\},\{d\}\}$, i.e., a hierarchy. By Lemma 8 and Theorem 6, $\mathscr{C}^1(\mathrm{rev}(g)) = \mathbb{M}_{\mathrm{str}}(\mathrm{rev}(g)) = \mathbb{M}_{\mathrm{str}}(g)$. The trees T_g and $T_{\mathrm{rev}(g)}$ are isomorphic and differ only in the labels (color figure online)

in addition $\mathscr{C}^1(\mathrm{rev}(g))$ is a hierarchy is equivalent to the property that $\mathrm{rev}(g)$ satisfies (U1) and (U2) and thus, $\varphi_{\mathrm{rev}(g)}$ is a symbolic ultrametric. By Theorem 5, $\mathrm{rev}(g)$ is *unp*. Hence, Item (4.) and (5.) are equivalent. □

4 Algorithms and complexity results

4.1 Recognition algorithm

We first consider the problem of recognizing whether a 2-structure $g = (V, \Upsilon, \varphi)$ has a tree-representation (T_g, t_g) and thus, whether g is *unp*. In what follows, the integer n will always denote $|V|$.

There are $O(n^2)$ time algorithms described in Ehrenfeucht et al. (1994) and McConnell (1995) to compute the modular decomposition of 2-structures. Ehrenfeucht et al. (1994) proposed a divide-and-conquer algorithm based on particular partitions of V defined by maximal modules that do not contain a certain vertex v. These partitions together with connected components of a graph that reflects whether vertex v is "distinguished" by other vertices are then be used to stepwisely compute the modular decomposition tree of a 2-structure. The "incremental" algorithm by McConnell (1995) stepwisely extends small substructures h of a given 2-structure g by one vertex while computing an updated modular decomposition tree. The modular decomposition for an arbitrary 2-structure is then obtained by a series of such incremental steps. Clearly, these algorithms can be used in order to verify whether a 2-structure is *unp* or not: Simply check whether the modular decomposition tree has an inner vertex labeled "prime".

Here, we propose an alternative $O(n^2)$ time algorithm to recognize *unp* 2-structures that is based on their characterization via di-cographs and 1-clusters (cf. Theorem 6(4)). On the one hand, our established results allow to design a conceptual simple algorithm by means of Theorem 6 and, on the other hand, the method developed here is an interesting starting point for novel heuristics for corresponding NP-complete editing problems as we shall discuss later.

For a given 2-structure $g = (V, \Upsilon, \varphi)$ the approach works as follows:

Step 1 Compute the reversible refinement $rev(g)$ of g.
Step 2 Check whether for all $i \in \Upsilon$ the graph $G_i(rev(g))$ is a di-cograph or not.
Step 3 If they are all di-cographs then compute the 1-clusters $\mathscr{C}_i^1$ for all $i \in \Upsilon$.
Step 4 If $\mathscr{C}^1(\mathrm{rev}(g))$ is a hierarchy, then a tree can be constructed with the method described in McConnell and Montgolfier (2005).

Complete pseudocode for the recognition procedure and all necessary subroutines is given in the Appendix (Algorithm 1– 6). The following Lemma implies that a tree can be can be constructed from a hierarchy in linear time w.r.t. the number of elements in the hierarchy, which is bounded by $2|V| - 1$ due to Theorem 1.

Lemma 9 (McConnell and Montgolfier 2005) *Given a hierarchy $\mathscr{C}$, it takes $O(|\mathscr{C}|)$ time to construct its inclusion tree.*

Steps 1–4 can be implemented in such a way that the designed recognition algorithm runs in $O(n^2)$-time. For full details on the algorithmic aspects we refer to the appendix.

We continue to show how the tree-representation of a *unp* 2-structure $g = (V, \Upsilon, \varphi)$ can be computed. As shown in the appendix, we can compute the hierarchy $\mathscr{C}^1(\mathrm{rev}(g))$ in $O(|V|^2)$ time. By Lemma 7, $\mathscr{C}^1(\mathrm{rev}(g)) = \mathbb{M}_{\mathrm{str}}(\mathrm{rev}(g))$. By Theorem 4(4), $\mathbb{M}(g) = \mathbb{M}(\mathrm{rev}(g))$ and hence, $\mathbb{M}_{\mathrm{str}}(\mathrm{rev}(g)) = \mathbb{M}_{\mathrm{str}}(g)$, which implies that $\mathscr{C}^1(\mathrm{rev}(g)) = \mathscr{C}^1(g)$. By Theorem 1, the number of clusters contained in $\mathscr{C}^1(g)$ is bounded by $2|V| - 1$. Thus, we can compute the inclusion tree T_g of $\mathscr{C}^1(g)$ in $O(|V|)$ time by means of Lemma 9. In order to get the correct labeling t_g we proceed as follows. We traverse T_g via breadth-first search, starting with the root v of T_g that represents V. Take any two children u_1, u_2 of v and any two vertices $x \in L(u_1)$, $y \in L(u_2)$ and check the labeling of the arcs (xy) and (yx). Assume that $\varphi(xy) = i$ and $\varphi(yx) = j$, where $i \leq j$. Then set $t_g(v) = (i, j)$ and place u_1 left of u_2 in T_g. An ordering of the children of v is not necessary if $i = j$. This step has to be repeated for all pairs of children of v and thus has total time complexity of $O(\deg(v)^2)$, where $\deg(v)$ denotes the number of children of v. After this we proceed with a child of v, playing now the role of v. Hence, the ordering of the tree T_g and the labeling t_g can be computed in $\sum_{v \in V(T_g)} \deg(v)^2 \leq \sum_{v \in V(T_g)} |V| \deg(v) = |V| \cdot 2(|V| - 1)$ and thus, in $O(|V|^2)$ time.

Taken the latter together with the preceding results, we obtain the following result.

Theorem 7 *For a given 2-structure $g = (V, \Upsilon, \varphi)$ it can be verified in $O(|V|^2)$ time whether g is unp or not, and in the positive case, the tree-representation (T_g, t_g) can be computed in $O(|V|^2)$ time.*

4.2 Tree-representable sets of relations, complexity results and ILP

From the practical point of view, 2-structures $g = (V, \Upsilon, \varphi)$ with $\Upsilon = \{0, 1, \ldots, k\}$ can be used to represent sets of disjoint relations $R_1, \ldots, R_k$, that is, φ is chosen such that

$$\varphi(xy) = \begin{cases} i \neq 0, & \text{if } (x, y) \in R_i \\ 0 & \text{else, i.e., } (x, y) \notin R_i,\ 1 \leq i \leq k \end{cases}$$

Moreover, 2-structures can even be used to represent *non-disjoint* relations as follows. Assume that we have arbitrary binary relations $R_1, \ldots, R_k$ over some set V. For two vertices $x, y \in V$ let $I_{xy} \subseteq \{1, \ldots, k\}$ be the inclusion-maximal subset such that $(x, y) \in R_i$ for all $i \in I_{xy}$. Furthermore, let b_{xy} be the (unique) integer encoded by the bit vector $b_1 b_2 \ldots b_k$ with $b_i = 1$ iff $i \in I_{xy}$. Clearly, $g = (V, \mathbb{N}, \varphi)$ with $\varphi(xy) = b_{xy}$ is a 2-structure, and the disjoint relations can be represented in a tree if and only if g is *unp*.

Those relations might represent the evolutionary relationships between genes, i.e., genes x, y are in relation R_i if the lowest common ancestor $\mathrm{lca}(x, y)$ in the corresponding gene tree was labeled with a particular event i, as e.g. speciation, duplication, horizontal-gene transfer, retro-transposition, and others. By way of example, methods as `ProteinOrtho` (Lechner et al. 2011, 2014) allow to estimate pairs of orthologs without inferring a gene or species tree. Hence, in practice such relations represent often only estimates of the true relationship between genes. Thus, in general the 2-structure of such estimates $R_1 \ldots, R_k$ will not be *unp* and hence, there is no tree-representation of such estimates. One possibility to attack this problem is to optimally edit the estimate $g = \{V, \Upsilon, \varphi\}$ to a *unp* 2-structure $g^* = \{V, \Upsilon, \varphi^*\}$ by changing the minimum number of labels assigned by φ.

We first consider the problem to rearrange a symmetric map d to obtain a symmetric symbolic ultrametric δ.

Problem 1 Symmetric Symbolic Ultrametric Editing/Deletion/Completion [SymSU- E/D/C]

Input: Given a symmetric map $d : V_{\text{irr}}^2 \to \Upsilon$, a fixed symbol $* \in \Upsilon$ and an integer k.

Question Is there a symmetric symbolic ultrametric $\delta : V_{\text{irr}}^2 \to \Upsilon$, such that

- $|D| \leq k$ *(Editing)*
- if $d(x, y) \neq *$, then $\delta(x, y) = d(x, y)$; and $|D| \leq k$ *(Completion)*
- $\delta(x, y) = d(x, y)$ or $\delta(x, y) = *$; and $|D| \leq k$ *(Deletion)*

where $D = \{(x, y) \in X \times X \mid d(x, y) \neq \delta(x, y)\}$.

The editing problem is clear. The completion problem is motivated by assuming that we might have an reliable assignment d on a subset W of V_{irr}^2, however, the assignment for the pairs $(x, y) \in \overline{W} = V_{\text{irr}}^2 \backslash W$ is unreliable or even unknown. For those pairs we use an extra symbol $*$ and set $d(x, y) = *$ for all $(x, y) \in \overline{W}$. Since we trust in the assignment d for all elements in W we aim at changing the least number of non-reliable estimates only, that is, only pairs (x, y) with $d(x, y) = *$ are allowed

to be changed so that the resulting map becomes a symmetric symbolic ultrametric. Conversely, the deletion problem asks to change a minimum number of assignments $d(x, y) \neq * \text{ to } \delta(x, y) = *$.

The following result was given in Hellmuth and Wieseke (2016).

Theorem 8 SymSU- E/D/C *is NP-complete.*

We can use this result in order to show that the analogous problems for 2-structures are NP-complete, as well.

Problem 2 *unp* 2- Structure Editing/Deletion/Completion [u2s- E/D/C]

Input: Given a 2-structure $g = (V, \Upsilon, \varphi_g)$, a fixed symbol $* \in \Upsilon$ and an integer k.

Question Is there a *unp* 2-structure $h = (V, \Upsilon, \varphi_h)$, such that

- $|D| \leq k$ *(Editing)*
- if $\varphi_g(x, y) \neq *$, then $\varphi_h(x, y) = \varphi_g(x, y)$; and $|D| \leq k$ *(Completion)*
- $\varphi_h(x, y) = \varphi_g(x, y)$ or $\varphi_h(x, y) = *$; and $|D| \leq k$ *(Deletion)*

where $D = \{(x, y) \in X \times X \mid \varphi_g(x, y) \neq \varphi_h(x, y)\}$.

Theorem 9 u2s- E/D/C *is NP-complete.*

Proof Since we can test whether a 2-structure is *unp* with Algorithm 1 in polynomial time, u2s- E/D/C $\in NP$.

To show NP-hardness, we simply reduce the instance $d : V_{\text{irr}}^2 \to \Upsilon$ of SymSU to the instance $g = (V, \Upsilon, d)$ of u2s. By Theorem 5, δ is a symbolic ultrametric if and only if $h = (V, \Upsilon, \delta)$ is *unp* and thus, we obtain the NP-hardness of u2s- E/D/C. □

The latter proof in particular implies that u2s- E/D/C is even NP-complete in the case that $\varphi_g(xy) = \varphi_g(yx)$ for all distinct $x, y \in V$.

We showed in Hellmuth et al. (2015) that the cograph editing problem and in Hellmuth and Wieseke (2016) that the symmetric symbolic ultrametric editing/completion/deletion is amenable to formulations as Integer Linear Program (ILP). We will extend these results here to solve the symbolic ultrametric editing/completion/deletion problem.

Let $d : V_{\text{irr}}^2 \to \Upsilon$ be an arbitrary map with $\Upsilon = \{*, 1, \ldots, n\}$ and $K_{|V|} = (V, E = V_{\text{irr}}^2)$ be the corresponding complete di-graph with arc-labeling such that each arc $(x, y) \in E$ obtains label $d(x, y)$.

For each of the three problems and hence, a given symmetric map d we define for each distinct $x, y \in V$ and $i \in \Upsilon$ the binary constants $\mathfrak{d}_{x,y}^i$ with $\mathfrak{d}_{x,y}^i = 1$ if and only if $d(x, y) = i$. Moreover, we define the binary variables E_{xy}^{ij} for all $i, j \in \Upsilon$ and $x, y \in V$ that reflect the labeling of the arcs in $K_{|V|}$ of the final symbolic ultrametric δ, i.e., E_{xy}^{ij} is set to 1 if and only if $\delta(x, y) = i$ and $\delta(y, x) = j$. In the following we will write E_{xy}^i as a shortcut for $\sum_{j \in \Upsilon} E_{xy}^{ij}$. Note that $E_{xy}^i \in \{0, 1\}$ and $E_{xy}^i = 1$ if and only if $\delta(x, y) = i$.

In order to find the closest symbolic ultrametric δ, the objective function is to minimize the symmetric difference of d and δ among all different symbols $i \in \Upsilon$:

$$\min \sum_{i \in \Upsilon} \left(\sum_{(x,y) \in V_{\text{irr}}^2} (1 - \mathfrak{d}_{xy}^i) E_{xy}^i + \sum_{(x,y) \in V_{\text{irr}}^2} \mathfrak{d}_{xy}^i (1 - E_{xy}^i) \right) \tag{1}$$

The same objective function can be used for the symbolic ultrametric completion and deletion problem.

For the symbolic ultrametric completion we must ensure that $\delta(x, y) = d(x, y)$ for all $d(x, y) \neq *$. Hence we set for all x, y with $d(x, y) = i \neq *$:

$$E_{xy}^i = 1. \tag{2}$$

For the symbolic ultrametric deletion we must ensure that $\delta(x, y) = d(x, y)$ or $\delta(x, y) = *$. In other words, for all $d(x, y) = i \neq *$ it must hold that for some $j \in \Upsilon$ either $E_{xy}^{ij} = 1$ or $E_{xy}^{*j} = 1$. Hence, we set for all $(x, y) \in V_{\text{irr}}^2$:

$$E_{xy}^* = 1, \text{ if } d(x, y) = *, \text{ and } E_{xy}^i + E_{xy}^* = 1, \text{ else.} \tag{2'}$$

For the cograph editing problem we neither need Constraint 2 nor 2'. However, for all three problems we need the following.

Each tuple (x, y) with $x \neq y$ has exactly one pair of values $(i, j) \in \Upsilon \times \Upsilon$ assigned to it, such that $E_{xy}^{ij} = E_{yx}^{ji}$. Hence, we add the following constraints for all distinct $(x, y) \in V_{\text{irr}}^2$ and $(i, j) \in \Upsilon \times \Upsilon$.

$$\sum_{i,j \in \Upsilon} E_{xy}^{ij} = 1 \text{ and } E_{xy}^{ij} = E_{yx}^{ji}. \tag{3}$$

In order to satisfy Condition (U2) and thus, that all induced triangles have at most two label-pairs we need to add the following constraints:

$$E_{xy}^{ij} + E_{yz}^{kl} + E_{zx}^{rs} \leq 2 \tag{4}$$

for all (not necessarily distinct) labels $i, j, k, l, r, s \in \Upsilon$ with pairwise distinct $\{i, j\}, \{k, l\}$, and $\{r, s\}$ and for all distinct $x, y, z \in V$.

Finally, in order to satisfy Condition (U1) and thus, that each mono-chromatic subgraph comprising all arcs with fixed label i is a di-cograph, we need the a couple of constraints that encode forbidden subgraphs. Since these conditions are straightforward to derive, we just give two example constraints to avoid induced P_4's and $\overline{N}$'s.

$$E_{ab}^i + E_{ba}^i + E_{bc}^i + E_{cb}^i + E_{cd}^i + E_{dc}^i - E_{ac}^i - E_{ca}^i - E_{ad}^i - E_{da}^i - E_{bd}^i - E_{db}^i \leq 5 \tag{5}$$

$$E^i_{ac} + E^i_{ca} + E^i_{ad} + E^i_{da} + E^i_{bd} + E^i_{db} + E^i_{ba} + E^i_{bc} + E^i_{dc} - E^i_{ab} - E^i_{cb} - E^i_{bd} \leq 8 \quad (5')$$

for all $i \in \Upsilon$ and all ordered tuple (a, b, c, d) of distinct $a, b, c, d \in V$.

It is easy to verify that the latter ILP formulation needs $O(|\Upsilon|^2|V|^2)$ variables and $O(|\Upsilon|^6|V|^3 + |\Upsilon||V|^4)$ constraints.

5 Concluding remarks

From an applications point of view, the main result of this contribution is a classification of those relationships between genes that can be derived from a gene phylogeny and the knowledge of event types assigned to interior *nodes* of phylogenetic tree. All such relations necessarily have co-graph structure. Fitch's version of the orthology and paralogy relations are the most important special cases. However, this class of relations is substantially more general and also include non-symmetric relations. These can account in particular for pairs of genes that are related by an ancestral horizontal transfer event and keep track of the directionality of the transfer.

It remains an open question to what extent this "lca-xenology" relation (Hellmuth and Wieseke 2016) can be inferred directly from sequence similarity data similar to the orthology and paralogy relations. If, however, the (estimates of) the relations R_o, R_p, and R_x (as in Fig. 1) are given, one can use the simplified map $\varphi_g(xy) = 1$ if $(x, y) \in R_o \cup R_x$ and $\varphi_g(xy) = 0$ otherwise, in order to determine whether R_o, R_p, and R_x are tree-representable. That is, it suffices to check whether $G_1(g)$ is a di-cograph or not. If so, the corresponding cotree with its vertex labels "0" (paralogs), "1" (orthologs), and "$\overrightarrow{1}$" (HGT) faithfully represents R_o, R_p, and R_x. In biological terms, this co-tree is (a not necessarily fully resolved) event-labeled gene tree, that in extension of Hellmuth et al. (2015) also provides information on HGT events.

The most commonly used definition of the xenology relation, however, is based on the presence of one or more horizontal transfer events along the unique path in the gene tree that connects two genes. It cannot be expressed in terms labels at the lowest common ancestor only. This raises the question whether edge labeled phylogenetic trees given rise to similar systems of relations on the leaf set.

The mathematical results presented in this contribution do not yet provide a practically applicable workflow to solve the gene-tree/species-tree reconciliation problem for real-world data. Nevertheless they outline quite clearly how such an approach could look like and which practical problems have to be overcome to make it work. The estimation of orthology directly from sequence data via pairwise similarities is necessarily imperfect (due to complementary deletion of paralogs and due to sometimes large variations is evolutionary rates) but yields pairwise orthology assignments on par with phylogenetics-based methods (Altenhoff et al. 2016). In Hellmuth et al. (2015), we have shown that pairwise sequence similarities are indeed sufficient to reconstruct event-labeled gene trees in the absence of HGT.

The second key problem is the identification of horizontal transfer events. In principle, likely xenologs (i.e., genes that have been introduced into a genome by horizontal transfer) can be identified directly from sequence data (Soucy et al. 2015). Sequence

composition often identifies a gene as a recent addition to a genome. In the absence of horizontal transfer, the similarities of pairs of true orthologs in the species pairs AB and AC are expected to be linearly correlated. Outliers are likely candidates for HGT events and thus can be "relabeled". A much more detailed analysis of the relational properties of horizontally transferred genes will be discussed in a forthcoming contribution.

A key issue for the practical purpose of a relation-based approach is the correction of noise and unavoidable errors in the input data. In the absence of horizontal transfer this boils down to the cograph editing problem. In the current setting this generalizes to the need for noise reduction in the input data by editing estimated input relations to *unp* 2-structures. Most likely, heuristics for cograph editing can be adapted to this task, see e.g. Dondi et al. (2016), Hellmuth and Wieseke (2016), Hellmuth et al. (2015), Lafond et al. (2016) and Lafond and El-Mabrouk (2015). Given the unavoidable noise in the input data it appears promising to consider a probabilistic inference of *unp* 2-structures depending on weighted or explicitly probabilistic estimates of orthology and/or horizontal transfer. While at present probabilistic models are based on explicit reconciliation of gene-trees and species trees (Sennblad and Lagergren 2009), there is no fundamental reason that would preclude e.g. a maximum likelihood or Bayesian formulation of cograph or 2-structure editing using similarities among small subsets of genes to derive propensities for a focal gene pair to be orthologs, paralogs, or xenologs.

Acknowledgements We thanks Maribel Hernández-Rosales for discussions. This work was funded by the German Research Foundation (DFG) (Proj. Nos. MI439/14-1 to P.F.S. and N.W.).

Appendix

Proofs of Proposition 1 and Lemma 6

Let $g = (V, \Upsilon, \varphi)$ be a 2-structure. In the proof of Proposition 1 we write $\Delta(xyz)$ as a shorthand for "*the Condition (U2) must be fulfilled for the set* D_{xyz}", where $x, y, z \in V$. Moreover, for any forbidden subgraph K that might occur in the graph $G_j(g)$ of some 2-structure g, we use the symbols $K^j(abc)$ and $K^j(abcd)$, resp., to designate the fact that $G_j(g)$ contains the forbidden subgraph K induced by the vertices a, b, c, resp., a, b, c, d in $G_j(g)$.

Proof of Proposition 1

In order to prove Proposition 1, we have to show that *unp* 2-structures are characterized by the conditions

(U1) $G_i(g)$ is a di-cograph for all $i \in \Upsilon$ and
(U2) for all vertices $x, y, z \in V$ it holds $|\{D_{xy}, D_{xz}, D_{yz}\}| \leq 2$.

We will frequently apply the following argument without explicitly stating it every time: By definition, if g is reversible then $\varphi(e) = \varphi(f)$ iff $\varphi(e^{-1}) = \varphi(f^{-1})$. Hence, for reversible g, $D_{ab} \neq D_{xy}$ implies that $\varphi(ab) \neq \varphi(xy), \varphi(yx)$ and $\varphi(ba) \neq \varphi(xy), \varphi(yx)$.

$\Rightarrow$: Let $g = (V, \Upsilon, \varphi)$ be a reversible *unp* 2-structure. If $|V| < 3$ then (U1) and (U2) are trivially satisfied. Thus we assume w.l.o.g. that $|V| \geq 3$. Furthermore, suppose there is a label $i \in \Upsilon$ such that $G_i(g)$ is not a di-cograph, i.e., $G_i(g)$ contains one of the forbidden subgraphs. Since g is reversible, the forbidden subgraphs A, B, $\overline{D_3}$, and $\overline{N}$ cannot occur.

Now let h be a substructure of g with $|V_h| = 3$ containing D_3 or C_3, or $|V_h| = 4$ containing P_4 or N, respectively. It is not hard to check that for each of these four graphs and any two distinct vertices $a, b \in V_h$ there is always a vertex $v \in V_h \backslash \{a, b\}$ so that $\varphi(av) \neq \varphi(bv)$. Therefore, $\{a, b\}$ cannot form a module in h. For P_4 and N one checks that for any three distinct vertices $a, b, c \in V_h$ and $v \in V_h \backslash \{a, b, c\}$ we always have $\varphi(av) \neq \varphi(bv)$, or $\varphi(av) \neq \varphi(cv)$, or $\varphi(bv) \neq \varphi(cv)$, so that $\{a, b, c\}$ cannot form a module in h. Thus, h contains only trivial modules and, hence, is prime. This contradiction implies that (U1) must be fulfilled.

Since $g = (V, \Upsilon, \varphi)$ has a tree-representation without prime nodes, and since three distinct leaves can have at most two distinct least common ancestors, Condition (U2) must hold as well.

$\Leftarrow$: Now assume that φ is a symbolic ultrametric, i.e., condition (U1) and (U2) are fulfilled for a reversible 2-structure g. In order to show that g is *unp* we have to demonstrate that all substructures h of g with $|V_h| = 3$ and $|V_h| = 4$ are non-prime (cf. Theorem 3).

Claim 1 If h is a substructure of g with $V_h = \{a, b, c\}$, then h is non-prime.

Proof of Claim 1 Since $\Delta(abc)$ we may assume that $D_{ab} = D_{ac}$, otherwise we simply relabel the vertices. If $|D_{ab}| = 1$, then $\{b, c\}$ forms a module in h. Assume that $|D_{ab}| = 2$. There are two cases, either $\varphi(ab) = \varphi(ac)$, then $\{b, c\}$ is a module in h, or $\varphi(ba) = \varphi(ac) = i$. In the latter case, $\varphi(bc) = i$ since otherwise either or $D_3^i(abc)$ or $C_3^i(abc)$ would occur. Therefore, $\{a, c\}$ forms a module in h.

Hence, in all cases, a substructure h of g with $V_h = \{a, b, c\}$ forms a non-prime structure. □

Claim 2 If h is a substructure of g with $V_h = \{a, b, c, d\}$, then h is non-prime.

Proof of Claim 2 There are two cases, either $|D_{ab}| = 1$ or $|D_{ab}| = 2$. For both cases, we will examine numerous sub-cases that might occur, and show that for each of these cases h contains non-trivial modules and thus, is non-prime. □

Case $|D_{ab}| = 1$:

Since $\Delta(abc)$ we can assume that $D_{ab} = D_{ac}$, otherwise relabel the vertices. Thus, $\varphi(ab) = \varphi(ba) = \varphi(ac) = \varphi(ca) = i$ for some $i \in \Upsilon$. Since $\Delta(acd)$ we have the three distinct cases

(i) $\varphi(ad) = \varphi(cd) = i$,
(ii) either (A) $\varphi(ad) = i$ or (B) $\varphi(cd) = i$
(iii) neither $\varphi(ad) = i$ nor $\varphi(cd) = i$.

In Case (i) and (iiA), $\{b, c, d\}$ is a module in h. In Case (iiB), the arc (bc) or (bd) must be labeled with i as otherwise there is $P_4^i(abcd)$. If $\varphi(bc) = i$, then $\{a, b, d\}$ is a module in h. If $\varphi(bd) = i$, then $\{a, d\}$ is a module in h.

Consider now Case (iii). Since $\Delta(acd)$, it follows that $D_{ad} = D_{cd}$ and in particular, $i \notin D_{ad} = D_{cd}$, since g is reversible. Let first $|D_{ad}| = |\{j\}| = 1$. Since $\Delta(abd)$, we have that either $\varphi(bd) = j$, in which case $\{a, b, c\}$ is a module in h or $\varphi(bd) = i$, which implies that $\varphi(bc) = i$, since otherwise $P_4^i(abcd)$. In the latter case, $\{a, c, d\}$ forms a module in h. If $|D_{ad}| = 2$, we have only the case that $\varphi(ad) = \varphi(cd) = j$ for some $j \in \Upsilon$. In the two other cases $\varphi(ad) = \varphi(dc) = j$ or $\varphi(da) = \varphi(cd) = j$ we would obtain $D_3^j(adc)$. Since $\Delta(abd)$, we obtain that either (I) $\varphi(bd) = i$, (II) $\varphi(bd) = j$ or (III) $\varphi(db) = j$. Case (I) implies that $\varphi(bc) = i$ as otherwise there is $P_4^i(abcd)$. Hence, $\{a, c, d\}$ form a module in h. In Case (II) $\{a, b, c\}$ is a module in h and Case (III) cannot occur, as otherwise there is $D_3^j(abd)$.

Case $|D_{ab}| = 2$: Since $\Delta(abc)$, we can assume wlog. that $D_{ab} = D_{ac}$, otherwise we relabel the vertices. Hence, we have either (I) $\varphi(ab) = \varphi(ac) = i$ or (II) $\varphi(ba) = \varphi(ac) = i$. Note that in Case (I), $\varphi(ba) = \varphi(ca) = i' \neq i$ and in Case (II) $\varphi(ab) = \varphi(ca) = i' \neq i$.

Consider Case (I). Since $\Delta(acd)$, we have one of the four distinct cases

(i) $D_{ac} = D_{ad} = D_{cd}$
(ii) $D_{ac} = D_{ad} \neq D_{cd}$
(iii) $D_{ac} = D_{cd} \neq D_{ad}$
(iv) $D_{ac} \neq D_{cd}$ and $D_{ac} \neq D_{ad}$

In Case (Ii) it is not possible to have $\varphi(cd) = \varphi(da) = i$ as otherwise there is $C_3^i(acd)$. If $\varphi(ad) = i$, then $\{b, c, d\}$ is module in h. If $\varphi(da) = i$, then $\varphi(db) = \varphi(dc) = i$, since otherwise there is $D_3^i(abd)$, $D_3^i(acd)$, $C_3^i(abd)$ or $C_3^i(acd)$. In that case, $\{a, b, c\}$ is a module in h.

In Case (Iii) it is not possible to have $\varphi(da) = i$, since otherwise there is $D_3^i(acd)$. Thus, $\varphi(ad) = i$ and therefore, $\{b, c, d\}$ forms a module in h.

In Case (Iiii) it is not possible to have $\varphi(cd) = i$, since otherwise there is $D_3^i(acd)$. Hence, $\varphi(dc) = i$. But then, at least one of the remaining arcs (bc), (cb), (bd), (db) must have label i, since otherwise there is $N^i(abcd)$. If $\varphi(bc) = i$, then $\{a, b, d\}$ is a module in h. If $\varphi(cb) = i$, then $\varphi(db) = i$ as otherwise there is $D_3^i(bcd)$ or $C_3^i(bcd)$. Hence, $\{a, c, d\}$ is a module in h. The case $\varphi(bd) = i$ is not possible, since then there is $D_3^i(abd)$. If $\varphi(db) = i$, then $\{a, d\}$ is a module in h.

In Case (Iiv) and since $\Delta(acd)$, we have $D_{ad} = D_{cd}$. If $D_{ad} = \{j\}$ and thus, $|D_{ad}| = 1$, then $\Delta(abd)$ implies that either $\varphi(bd) = \varphi(db) = j \neq i$, or $\varphi(bd) = i$, or $\varphi(db) = i$. If $\varphi(bd) = j \neq i$, then $\{a, b, c\}$ is a module in h. The case $\varphi(bd) = i$ cannot happen, since otherwise there is $D_3^i(abd)$. If $\varphi(db) = i$, then either $\varphi(bc) = i$ or $\varphi(cb) = i$, otherwise there is $N^i(abcd)$. The case $\varphi(bc) = i$ is not possible, otherwise there is $D_3^i(bcd)$. If $\varphi(cb) = i$, then $\{a, c, d\}$ is a module in h.

Assume now that in Case (Iiv) we have $|D_{ad}| = 2$. Again, since $\Delta(acd)$, we have $D_{ad} = D_{cd}$. Assume that $j \in D_{ad}$. There are two case, either $\varphi(ad) = \varphi(cd) = j \neq i$ or $\varphi(ad) = \varphi(dc) = j \neq i$. However, the latter case is not possible, otherwise there is $D_3^j(acd)$. Hence, let $\varphi(ad) = \varphi(cd) = j \neq i$. Since $\Delta(abd)$ we can conclude that either $\varphi(bd) = i$, or $\varphi(db) = i$, or $\varphi(bd) = j$, or $\varphi(db) = j$. The cases $\varphi(bd) = i$ and $\varphi(db) = j$ are not possible, otherwise there is $D_3^i(abd)$ and $D_3^j(abd)$, respectively. If $\varphi(db) = i$, then $\varphi(bc) = i$ or $\varphi(cb) = i$, otherwise there is $N^i(abcd)$. This case can

be treated as in the previous step and we obtain the module $\{a, c, d\}$ in h. If $\varphi(bd) = j$, then $\{a, b, c\}$ is a module in h.

Consider now Case (II) $\varphi(ba) = \varphi(ac) = i$, and $\varphi(ab) = \varphi(ca) = i' \neq i$. Hence, $\varphi(bc) = i$, otherwise there is $D_3^i(abc)$ or $C_3^i(abc)$. Again, since $\Delta(acd)$, we have one of the four distinct cases (i), (ii), (iii) or (iv), as in Case (I).

Consider the Case (IIi). If $\varphi(dc) = i$, then $\{a, b, d\}$ is a module in h. Thus, assume $\varphi(cd) = i$. The case $\varphi(da) = i$ is not possible, since then there is $C_3^i(acd)$. If $\varphi(ad) = i$, then $\varphi(bd) = i$, otherwise there is $D_3^i(abd)$ or $C_3^i(abd)$. Now, $\{a, c, d\}$ is a module in h.

Now, Case (IIii). The case $\varphi(da) = i$ is not possible, otherwise there is $D_3^i(acd)$ and thus, $\varphi(ad) = i$. Then $\varphi(bd) = i$, otherwise there is $D_3^i(abd)$ or $C_3^i(abd)$. Therefore, $\{a, c, d\}$ is a module in h.

Consider the Case (IIiii). The case $\varphi(cd) = i$ is not possible, otherwise there is $D_3^i(acd)$. Thus, $\varphi(dc) = i$ and therefore, $\{a, b, d\}$ is a module in h.

In Case (IIiv) and since $\Delta(acd)$, we have $D_{ad} = D_{cd}$. If $D_{ad} = \{j\}$ and thus, $|D_{ad}| = 1$, then $\Delta(abd)$ implies that either $\varphi(bd) = j \neq i$, or $\varphi(bd) = i$, or $\varphi(db) = i$. If $\varphi(bd) = j \neq i$, then $\{a, b, c\}$ is a module in h. If $\varphi(bd) = i$, then $\{a, c, d\}$ is a module in h. The case $\varphi(db) = i$ cannot happen, otherwise there is $D_3^i(bcd)$.

If $|D_{ad}| = 2$ and $j \in D_{ad}$, then there are two cases either $\varphi(ad) = \varphi(cd) = j \neq i$ or $\varphi(ad) = \varphi(dc) = j \neq i$. However, the latter case is not possible, otherwise there is $D_3^j(acd)$. Hence, let $\varphi(ad) = \varphi(cd) = j \neq i$. Since $\Delta(abd)$ we can conclude that either $\varphi(bd) = i$, or $\varphi(db) = i$, or $\varphi(bd) = j$, or $\varphi(db) = j$. The cases $\varphi(db) = i$ and $\varphi(db) = j$ are not possible, otherwise there is $D_3^i(abd)$ and $D_3^j(abd)$, respectively. If $\varphi(bd) = i$ or $\varphi(bd) = j$, then $\{a, c, d\}$, resp., $\{a, b, c\}$ is a module in h. □

In summary, in each of the cases a substructure h of g with 3 or 4 vertices is non-prime whenever (U1) and (U2) holds. Thus g is *unp*. □

Proof of Lemma 6

⇒: Let $G_i(g)$ be a di-cograph for all $i \in \Upsilon$. Moreover, assume for contradiction that there is a label $j \in \Upsilon_{\mathrm{rev}(g)}$ such that $G_j(\mathrm{rev}(g))$ is not a di-cograph. Then $G_j(\mathrm{rev}(g))$ contains a forbidden subgraph. Since $\mathrm{rev}(g)$ is reversible, only the subgraphs D_3, C_3, N, and P_4 are possible. Moreover, by construction of $\mathrm{rev}(g)$ and because $\varphi_{\mathrm{rev}(g)}(e) = \varphi_{\mathrm{rev}(g)}(f)$ implies $\varphi(e) = \varphi(f)$, we have $G_j(\mathrm{rev}(g)) \subseteq G_k(g)$ for some $k \in \Upsilon$.

In the following we will show that the existence of one of the forbidden subgraphs D_3, C_3, N, and P_4 in any $G_j(\mathrm{rev}(g))$ leads to a contradiction. We proceed case by case.

Case: $G_j(\mathrm{rev}(g))$ contains D_3 for some $j \in \Upsilon_{\mathrm{rev}(g)}$.

If $G_j(\mathrm{rev}(g))$ contains D_3 induced by the vertices x, y, z, we can wlog. assume that the vertices are labeled so that $\varphi_{\mathrm{rev}(g)}(xy) = \varphi_{\mathrm{rev}(g)}(yz) = j \neq \varphi_{\mathrm{rev}(g)}(xz)$, $\varphi_{\mathrm{rev}(g)}(zy) = \varphi_{\mathrm{rev}(g)}(yx) = k \neq \varphi_{\mathrm{rev}(g)}(zx)$ and $j \neq k$. By construction of $\mathrm{rev}(g)$ we obtain $\varphi(xy) = \varphi(yz) = j'$, $\varphi(zy) = \varphi(yx) = k'$ for some distinct $j', k' \in \Upsilon$.

However, since g does not contain forbidden subgraphs in $G_{j'}(g)$, there must be an arc connecting x and z with label j'. The possibilities $\varphi(zx) = j' \neq \varphi(xz)$ and $\varphi(zx) = \varphi(xz) = j'$ cannot occur, since then $G_{j'}(g)$ would contain a C_3 or $\overline{D_3}$ as forbidden subgraph. Hence, it must hold that $\varphi(xz) = j'$. Analogously, one shows that $\varphi(zx) = k'$. By construction of $\mathrm{rev}(g)$, we obtain $\varphi_{\mathrm{rev}(g)}(xz) = j$, and $\varphi_{\mathrm{rev}(g)}(zx) = k$; a contradiction.

Case: $G_j(\mathrm{rev}(g))$ *contains* C_3 *for some* $j \in \Upsilon_{\mathrm{rev}(g)}$.

If $G_j(\mathrm{rev}(g))$ contains a C_3 induced by the vertices x, y, z, we can wlog. assume that the vertices are labeled so that $\varphi_{\mathrm{rev}(g)}(xy) = \varphi_{\mathrm{rev}(g)}(yz) = \varphi_{\mathrm{rev}(g)}(zx) \neq \varphi_{\mathrm{rev}(g)}(yx) = \varphi_{\mathrm{rev}(g)}(xz) = \varphi_{\mathrm{rev}(g)}(zy)$. Thus, $\varphi(xy) = \varphi(yz) = \varphi(zx) = j'$ and $\varphi(yx) = \varphi(xz) = \varphi(zy) = k'$. We have $j' \neq k'$ as otherwise $\varphi_{\mathrm{rev}(g)}(xy) = \varphi_{\mathrm{rev}(g)}(yx)$. Therefore, $G_{j'}(g)$ contains the forbidden subgraph C_3; a contradiction.

Case: $G_j(\mathrm{rev}(g))$ *contains* P_4 *for some* $j \in \Upsilon_{\mathrm{rev}(g)}$.

If $G_j(\mathrm{rev}(g))$ contains a P_4 induced by the vertices a, b, c, d, we can wlog. assume that the vertices are labeled so that $\varphi_{\mathrm{rev}(g)}(e) = \varphi_{\mathrm{rev}(g)}(f) = j$ for all $e, f \in E' = \{(a, b), (b, a), (b, c), (c, b), (c, d), (d, c)\}$. For all these arcs $e, f \in E'$ it additionally holds that $\varphi(e) = \varphi(f) = j'$. Moreover, for all other arcs $e \in \{a, b, c, d\}^{\times}_{\mathrm{irr}} \backslash E'$ it is not possible that $\varphi(e) = \varphi(e^{-1}) = j'$, as otherwise, $\varphi_{\mathrm{rev}(g)}(e) = \varphi_{\mathrm{rev}(g)}(e^{-1}) = j$ and the P_4 would not be an induced subgraph of $G_j(\mathrm{rev}(g))$. By the latter argument and since $G_{j'}(g)$ does not contain an induced P_4 there must be at least one arc $e \in \{a, b, c, d\}^{\times}_{\mathrm{irr}} \backslash E'$ with $\varphi(e) = j'$, but $\varphi(e^{-1}) \neq j'$. Now full enumeration of all possibilities (which we leave to the reader) to set one, two, or three of these arcs to the label j' yields one of the forbidden subgraphs $\overline{D_3}$, A, B or $\overline{N}$ in $G_{j'}(g)$; a contradiction.

Case: $G_j(\mathrm{rev}(g))$ *contains* N *for some* $j \in \Upsilon_{\mathrm{rev}(g)}$.

If $G_j(\mathrm{rev}(g))$ contains an N induced by the vertices a, b, c, d, we can wlog. assume that the vertices are labeled so that $\varphi_{\mathrm{rev}(g)}(ba) = \varphi_{\mathrm{rev}(g)}(bc) = \varphi_{\mathrm{rev}(g)}(dc) = j \neq \varphi_{\mathrm{rev}(g)}(ab) = \varphi_{\mathrm{rev}(g)}(cb) = \varphi_{\mathrm{rev}(g)}(cd) = k$. Thus, $\varphi(ba) = \varphi(bc) = \varphi(dc) = j' \neq \varphi(ab) = \varphi(cb) = \varphi(cd) = k'$. Since $G_{j'}(g)$ is cograph, there must be an arc $e \in E' = \{(a, c), (c, a), (a, d), (d, a), (b, d), (d, b)\}$ with $\varphi(e) = j'$. Moreover, for this arc e it must hold that $\varphi(e^{-1}) \neq k'$ as otherwise, $\varphi_{\mathrm{rev}(g)}(e) = j$. The graph $G_k(\mathrm{rev}(g))$ also contains an N induced by the vertices a, b, c, d. Hence, by analogous arguments there is an $f \in E', e \neq f$ with $\varphi(f) = k'$ with $\varphi(f^{-1}) \neq j'$. Assume first that e is (a, c) or (c, a) and thus, $D_{ac} = \{j', j''\}$ where $j'' = j'$ is allowed. If f is (a, d) or (d, a), then $D_{ad} = \{k', k''\}$ where $k'' = k'$ is allowed. But then $D_{acd} = \{\{k', j'\}, \{j', j''\}, \{k', k''\}\}$ with $j' \neq k'$ and thus $|D_{acd}| = 3$ violating Condition (U2) in g; a contradiction. If f is (b, d) or (d, b), then $D_{bd} = \{k', k''\}$ where $k'' = k'$ is allowed. Thus, $\{j', j''\}, \{k', j'\} \in D_{acd}$ and $\{k', k''\}, \{k', j'\} \in D_{abd}$. The only way to satisfy $|D_{acd}| = 2$ and $|D_{abd}| = 2$ is achieved by $D_{ad} = \{k', j'\}$. However, the case $\varphi(e) = j'$ and $\varphi(e^{-1}) = k'$ with $e \in E'$ is not allowed. All other cases, starting with $e \in E' \backslash \{(a, c), (c, a)\}$ can be treated analogously.

$\Leftarrow$: Let $G_j(\mathrm{rev}(g))$ be a di-cograph for all $j \in \Upsilon_{\mathrm{rev}(g)}$. Moreover, assume for contradiction that there is a label $i \in \Upsilon$ such that $G_i(g)$ is not a di-cograph. Hence, $G_i(g)$ contains a forbidden subgraph.

In the following we will show that the existence of one of the forbidden subgraphs in any $G_i(g)$ leads to a contradiction. Again we analyze the possible forbidden subgraph separately.

Case: $G_i(g)$ contains D_3, A or B for some $i \in \Upsilon$.

If $G_i(g)$ contains a forbidden subgraph D_3, A, B then there are arcs (a, b), (b, c) contained in these forbidden subgraphs with $\varphi(ab) = \varphi(bc) = i$ but $\varphi(ac) \neq i$ and $\varphi(ca) \neq i$. Moreover, since $G_j(\mathrm{rev}(g))$ does not contain these forbidden subgraphs for any $j \in \Upsilon_{\mathrm{rev}(g)}$, we also obtain that $\varphi(ab) = \varphi(bc) = i$ but $\varphi(ba) \neq \varphi(cb)$. But this implies that $|D_{abc}| = 3$ in g; a contradiction to (U2).

Case: $G_i(g)$ contains $\overline{D_3}$ or C_3 for some $i \in \Upsilon$.

If $G_i(g)$ contains a forbidden subgraph $\overline{D_3}$ or C_3, then there are arcs (a, b), (b, c) contained in these forbidden subgraphs with $\varphi(ab) = \varphi(bc) = i$ and $\varphi(ba) \neq i, \varphi(cb) \neq i$. If $\varphi(ba) = \varphi(cb)$ and the case $\overline{D_3}$ is contained $G_i(g)$, then $G_j(\mathrm{rev}(g))$ contains the D_3 as forbidden subgraph. If $\varphi(ba) = \varphi(cb)$ and the case C_3 is contained $G_i(g)$, then $G_j(\mathrm{rev}(g))$ contains the D_3 or C_3 as forbidden subgraph. Hence, $\varphi(ba) \neq \varphi(cb)$. For the case $\overline{D_3}$, we observe that $|D_{abc}| = 3$ in g; a contradiction to (U2). For the case C_3, we can conclude by analogous arguments, $\varphi(ba) \neq \varphi(ca)$ and $\varphi(cb) \neq \varphi(ca)$ and again, $|D_{abc}| = 3$ in g; a contradiction.

Case: $G_i(g)$ contains N for some $i \in \Upsilon$.

Similarly, if N is contained in $G_i(g)$ then there are arcs (b, a), (b, c), (d, c) contained in N with $\varphi(ba) = \varphi(bc) = \varphi(dc) = i$ and $\varphi(e) \neq i$ for all $e \in \{(a, c), (c, a), (b, d), (d, b)\}$. Since $G_j(\mathrm{rev}(g))$ does not contain N it holds that $\varphi(ab) \neq \varphi(cb)$ or $\varphi(cb) \neq \varphi(cd)$. If $\varphi(ab) \neq \varphi(cb)$ then $|D_{abc}| = 3$, as $\varphi(ac) \neq i$ and $\varphi(ca) \neq i$; a contradiction to (U2). On the other hand, if $\varphi(cb) \neq \varphi(cd)$ then $|D_{bcd}| = 3$, as $\varphi(bd) \neq i$ and $\varphi(db) \neq i$; again a contradiction to (U2).

Case: $G_i(g)$ contains P_4 for some $i \in \Upsilon$.

The P_4 on four vertices a, b, c, d cannot be contained in any $G_i(g)$, since for any two arcs $e, f \in E' = \{(a, b), (b, a), (b, c), (c, b), (c, d), (d, c)\}$ of this P_4 it still holds $\varphi_{\mathrm{rev}(g)}(e) = \varphi_{\mathrm{rev}(g)}(f) = i'$ and for any arc e not in E', $\varphi_{\mathrm{rev}(g)}(e) \neq i'$. Hence, if $G_i(g)$ contains a P_4, then $G_{i'}(\mathrm{rev}(g))$ contains a P_4 as forbidden subgraph; a contradiction.

Case: $G_i(g)$ contains $\overline{N}$ for some $i \in \Upsilon$.

If $G_i(g)$ contains the forbidden subgraph $\overline{N}$ on four vertices a, b, c, d, then for the three arcs e_1, e_2, e_3 with $\varphi(e_j) = \varphi(e_j^{-1}) = i$, it still holds, that $\varphi_{\mathrm{rev}(g)}(e_j) = \varphi_{\mathrm{rev}(g)}(e_j^{-1}) = i'$, $1 \leq j \leq 3$. However, for the other arcs f_1, f_2, f_3 with $\varphi(f_j) = i \neq \varphi(f_j^{-1})$, we can infer that $\varphi_{\mathrm{rev}(g)}(f_j) \neq i'$ and $\varphi_{\mathrm{rev}(g)}(f_j^{-1}) \neq i'$. Thus, $G_{i'}(\mathrm{rev}(g))$ contains a P_4 on the three edges e_1, e_2, e_3 as forbidden subgraph; a contradiction. □

Algorithmic considerations

We show that the characterization of *unp* 2-structures in terms of di-cographs and 1-clusters (cf. Theorem 6(4)) can be used to derive a simple algorithm for the recognition of *unp* 2-structures. In the following the integer n will always denote $|V|$ as a measure of the input size.

Pseudocode for the recognition procedure is given in Algorithm 1. Furthermore, we give pseudocode for all necessary subroutines (Algorithms 2 to 6). We omit the procedure for computing the modular decomposition $\mathbb{M}_{\text{str}}(G)$ of a digraph $G = (V, E)$, as McConnell and de Montgolfier McConnell and Montgolfier (2005) already presented an $O(|V| + |E|)$ time algorithm for this problem.

We first prove the correctness of Algorithms 4, 5, and 6.

Lemma 10 *Given a digraph G and its modular decomposition $\mathbb{M}_{\text{str}}(G)$, Algorithm 4 recognizes whether G is a di-cograph or not.*

Proof At first, Algorithm 4 computes the inclusion tree T of $\mathbb{M}_{\text{str}}(G)$ and then iterates over all strong modules $M \in \mathbb{M}_{\text{str}}(G)$. For each strong module M two arbitrary but distinct children $M', M'' \in \mathbb{M}_{\text{str}}(G)$ of M in T are selected and it is checked if there is an arc between two vertices $x \in M'$ and $y \in M''$. If G is a di-cograph and there is an arc $(x, y) \in E$ or $(y, x) \in E$, then by Remark 3, M must be either series or order. In other words, if we have found an $(x, y) \in E$ or $(y, x) \in E$, but M is neither series nor order, it must be prime which implies that G was not a di-cograph. However, it might be possible, that the chosen elements x and y do not form an arc $(x, y) \in E$ or $(y, x) \in E$, but then M is either prime or parallel. If M is prime there must be arcs (x', y') or (y', x'), that we might have not observed in the preceding step, where $x' \in M'$, $y' \in M''$ for some children M', M'' of M, otherwise M would be parallel. However, this case is covered by counting the numbers of all arcs between the vertices of maximal strong submodules contained in series or order modules M. If the accumulated number e of all counted arcs is equal to the number of arcs $|E|$ in G, then all modules $M' \in \mathbb{M}_{\text{str}}(G)$ which are neither series nor order must be parallel. Hence, no prime modules exists and therefore G is a di-cograph. □

Lemma 11 *Given a di-cograph G_i and its modular decomposition $\mathbb{M}_{\text{str}}(G_i)$, Algorithm 5 computes the 1-clusters $\mathscr{C}_i^1$ of G_i.*

Proof At first, Algorithm 5 computes the inclusion tree T of $\mathbb{M}_{\text{str}}(G_i)$. Then, for each strong module M two arbitrary vertices from distinct children $M', M'' \in \mathbb{M}_{\text{str}}(G)$ of M in T are selected. If there is an arc $(x, y) \in E$ or $(y, x) \in E$, then by Remark 3, M cannot be parallel and hence, M is a 1-cluster and therefore, has to be added to the set of 1-clusters $\mathscr{C}_i^1$. □

The next lemma shows that Algorithm 6 correctly recognizes, whether $\mathscr{C}^1(\text{rev}(g))$ is a hierarchy or not. However, due to efficiency and also simplicity of the algorithm, we deal here with multisets, $\mathscr{C} = \biguplus_{i \in \Upsilon_{rev(g)}} \mathscr{C}_i^1$. The symbol "$\biguplus$" denotes the multiset-union of sets where the multiplicity of an element M in $\mathscr{C}$ is given by the number of sets that contain M.

Algorithm 1 Recognition of *unp* 2-Structures

INPUT: 2-structure $g = (V, \Upsilon, \varphi)$ with $n = |V|$ vertices and $k = |\Upsilon|$ labels;
$g' = (V, \Upsilon', \varphi') \leftarrow$ `Compute` rev(g)
if $|\Upsilon'| > 2(n-1)$ **then**
  return *FALSE*
end if
set of digraphs $\mathscr{G} \leftarrow$ `Compute monochromatic subgraphs` $G_i(g')$
multiset of clusters $\mathscr{C} \leftarrow \emptyset$
for G_i in $\mathscr{G}$ **do**
  set of strong modules $\mathbb{M}_{\text{str}} \leftarrow$ `Compute the modular decomposition of` G_i (cf. McConnell and Montgolfier (2005))
  if `Check di-cograph property for` G_i `with modular decomposition` $\mathbb{M}_{\text{str}}(G_i)$ **then**
    $\mathscr{C}_i^1 \leftarrow$ `Get 1-clusters from di-cograph` G_i `with modular decomposition` $\mathbb{M}_{\text{str}}(G_i)$
  else
    return *FALSE*
  end if
  $\mathscr{C} \leftarrow \biguplus_{i \in \Upsilon'} \mathscr{C}_i^1$
end for
if $|\mathscr{C}| > 2(n-1)$ **then**
  return *FALSE*
end if
if `Check hierarchy property for` $\mathscr{C}$ **then**
  return *TRUE*
else
  return *FALSE*
end if

Algorithm 2 Compute rev(g)

INPUT: 2-structure $g = (V, \Upsilon, \varphi)$ with $n = |V|$ vertices;
$\Upsilon' \leftarrow \emptyset$
for $i = 1, \ldots, n$ **do**
  for $j = 1, \ldots, n$ **do**
    $\varphi'(i, j) \leftarrow (\varphi(i, j), \varphi(j, i))$
    $\Upsilon' \leftarrow \Upsilon' \cup \{\varphi'(i, j)\}$
  end for
end for
return $g' = (V, \Upsilon', \varphi')$

Lemma 12 *Given a multiset* $\mathscr{C} = \biguplus_{i \in \Upsilon_{rev(g)}} \mathscr{C}_i^1$ *of the 1-clusters of a set of di-cographs* $G_i = (V, E_i)$*, Algorithm* 6 *recognizes whether* $\mathscr{C}^1 = \bigcup_{i \in \Upsilon_{rev(g)}} \mathscr{C}_i \cup \{v | v \in V\}$ *is a hierarchy or not.*

Proof Note that the multiset $\mathscr{C}$ may contain a cluster C more than once, as C can be part of different 1-clusters $\mathscr{C}_i^1$. Furthermore, $\mathscr{C}$ does not contain the singletons. However, it is easy to see that $\mathscr{C}^1$ is a hierarchy if and only if the singletons are contained in $\mathscr{C}^1$ (which is satisfied by construction), there is a 1-cluster equal to V and for all $C', C'' \in \mathscr{C}$ it holds that $C' \cap C'' \in \{C', C'', \emptyset\}$. The latter is equivalent to the following statement. For all $C', C'' \in \mathscr{C}, |C'| \leq |C''|$ it holds that either $C' \cap C'' = \emptyset$ or $C' \subseteq C''$.

Algorithm 3 Compute monochromatic subgraphs $G_i(g')$

INPUT: 2-structure $g' = (V, \Upsilon', \varphi')$ with $n = |V|$ vertices and $k' = |\Upsilon'|$ labels;
define bijection $\mu : \Upsilon' \to 1 \dots k'$
$\mathscr{G} \leftarrow \emptyset$
for $i = 1, \dots, k'$ **do**
  $E_i \leftarrow \emptyset$
  $G_i = (V, E_i)$
end for
for $i = 1, \dots, n$ **do**
  for $j = 1, \dots, n$ **do**
    $E_{\mu(\varphi'(i,j))} \leftarrow E_{\mu(\varphi'(i,j))} \cup \{(i, j)\}$
  end for
end for
return $\mathscr{G} = \bigcup_{i=1}^{k'} \{G_i\}$

Algorithm 4 Check di-cograph property for G_i with modular decomposition $\mathbb{M}_{\text{str}}(G_i)$

INPUT: digraph G_i and its modular decomposition $\mathbb{M}_{\text{str}}(G_i)$;
tree $T \leftarrow$ inclusion tree of $\mathbb{M}_{\text{str}}$
arc counter $e \leftarrow 0$
for $M \in V(T)$ **do**
  $M', M'' \leftarrow$ two arbitrary but distinct child vertices of M in T
  $x, y \leftarrow$ two arbitrary elements $x \in M'$ and $y \in M''$
  if $(x, y) \in E(G_i)$ or $(y, x) \in E(G_i)$ **then**
    if M is not series or order **then**
      return *FALSE*
    else
      increase e by the number of arcs between all elements from distinct children of M in T
    end if
  end if
end for
if $e \neq |E(G_i)|$ **then**
  return *FALSE*
end if
return *TRUE*

In Line 4, a list $\mathscr{C}_{\leq}$ is created with all $C \in \mathscr{C}$ being sorted ascending by cardinality. Hence, $\mathscr{C}_{\leq}(|\mathscr{C}_{\leq}|)$ is one of the largest clusters. In Line 6, it is checked if this largest cluster contains all elements from the ground set $V = \{1, \dots, n\}$. If not then $V \notin \mathscr{C}$ and therefore $\mathscr{C}^1$ is not a hierarchy. In Lines 9 to 14, lists $\mathscr{L}_i$ are created, containing all clusters $C \in \mathscr{C}$ with $i \in C$. The relative order of clusters in $\mathscr{L}_i$ is identical to the relative order of clusters in $\mathscr{C}_{\leq}$. In each iteration of Lines 16 to 28 the smallest cluster L is selected among all remaining clusters $\bigcup_{i=1}^{n} \mathscr{L}_i$. For each $i \in L$ obviously $L \in \mathscr{L}_i$. If $s, t \in L$ then it is checked if $\mathscr{L}_s = \mathscr{L}_t$. This can be done, as $\mathscr{L}_s$ and $\mathscr{L}_t$ have the same relative order of clusters. If $s, t \in L$ and $\mathscr{L}_s = \mathscr{L}_t$ then it follows that $s, t \in L'$ for all $L' \in \mathscr{L}_s \cup \mathscr{L}_t$. As this holds for all pairwise distinct $s, t \in L$ and $|L| \leq |L'|$ for all $L' \in \bigcup_{i=1}^{n} \mathscr{L}_i$ it follows that $L \subseteq L'$ for all $L' \in \bigcup_{i=1}^{n} \mathscr{L}_i$ with $L \cap L' \neq \emptyset$. As $\mathscr{L}_s = \mathscr{L}_t$ it is sufficient to keep only one of the lists, e.g., $\mathscr{L}_s$ (Line 24). Finally, L is removed from $\mathscr{L}_s$ (Line 27) and the while-loop is repeated with the next smallest cluster. □

We now show the correctness of Algorithm 1.

Algorithm 5 Get 1-clusters from di-cograph G_i with modular decomposition $\mathbb{M}_{\text{str}}(G_i)$

INPUT: di-cograph G_i and its modular decomposition $\mathbb{M}_{\text{str}}(G_i)$;
$\mathscr{C}_i^1 \leftarrow \emptyset$
tree $T \leftarrow$ inclusion tree of $\mathbb{M}_{\text{str}}(G_i)$
for $M \in V(T)$ **do**
  $M', M'' \leftarrow$ two arbitrary but distinct child vertices of M in T
  $x, y \leftarrow$ two arbitrary elements $x \in M'$ and $y \in M''$
  if $(x, y) \in E(G_i)$ or $(y, x) \in E(G_i)$ **then**
    $\mathscr{C}_i^1 \leftarrow \mathscr{C}_i^1 \cup M$
  end if
end for
return $\mathscr{C}_i^1$

Algorithm 6 Check hierarchy property for $\mathscr{C}$

INPUT: multiset of clusters $\mathscr{C}$, on the ground set $\{1, \ldots, n\}$;
for each element in $\mathscr{C}$ compute a unique identifier $id : \mathscr{C} \to \{1, \ldots, |\mathscr{C}|\}$
for each element in $\mathscr{C}$ compute its bit string representation $bsr : \mathscr{C} \to \{0, 1\}^n$ with $bsr(C_j)[i] = 1$ iff $i \in C_j, C_j \in \mathscr{C}$
sorted list $\mathscr{C}_{\leq} \leftarrow$ sort $\mathscr{C}$ ascending by cardinality of its elements
for i=1, …, n **do**
  if $bsr(\mathscr{C}_{\leq}(|\mathscr{C}_{\leq}|))[i] \neq 1$ **then**
    return *FALSE*
  end if
  $\mathscr{L}_i \leftarrow \mathscr{C}_{\leq}$
  for $C_j \in \mathscr{L}_i$ **do**
    if $bsr(C_j)[i] = 0$ **then**
      remove C_j from $\mathscr{L}_i$
    end if
  end for
end for
while $\mathscr{L}_1 \neq \emptyset$ **do**
  $s \leftarrow$ the smallest i such that $|\mathscr{L}_i(1)| \leq |\mathscr{L}_j(1)|$ for all $i \neq j$
  $L \leftarrow \mathscr{L}_s(1)$
  for $t \in L$ with $t \neq s, \mathscr{L}_t \neq \emptyset$ **do**
    for $r = 1, \ldots, |\mathscr{L}_s|$ **do**
      if $id(\mathscr{L}_s(r)) \neq id(\mathscr{L}_t(r))$ **then**
        return *FALSE*
      end if
      $\mathscr{L}_t \leftarrow \emptyset$
    end for
  end for
  remove L from $\mathscr{L}_s$
end while
return *TRUE*

Lemma 13 *Given a 2-structure $g = (V, \Upsilon, \varphi)$, Algorithm 1 recognizes whether g is unp or not.*

Proof In fact, Algorithm 1 recognizes, for the reversible refinement $rev(g)$, whether all monochromatic subgraphs $G_i(rev(g))$ are di-cographs and whether in addition the 1-clusters in $\mathscr{C}^1(rev(g))$ form a hierarchy. By Theorem 6, this suffices to decide whether g is *unp* or not.

It is easy to see that Algorithm 2 computes the reversible refinement of g by means of Definition 9 and Remark 1 with $\varphi_{\mathrm{rev}(g)}(e) = (\varphi_g(e), \varphi_g(e^{-1}))$. Hence, in Line 2 the reversible refinement $g' = \mathrm{rev}(g)$ of g is computed.

If $\mathrm{rev}(g)$ is *unp*, then there exists a tree-representation $(T_{\mathrm{rev}(g)}, t_{\mathrm{rev}(g)})$. As $T_{\mathrm{rev}(g)}$ has at most $n-1$ inner vertices there can be at most $n-1$ different labels $t_{\mathrm{rev}(g)}(lca(x, y)) = (i, j)$, each composed of at most two distinct labels $i, j \in \Upsilon_{rev(g)}$. Assuming that all labels are pairwise distinct leads to $2(n-1) \leq |\Upsilon_{rev(g)}|$ distinct labels in total. Hence, if $|\Upsilon_{rev(g)}| > 2(n-1)$ then $rev(g)$ is not *unp*. It is easy to see that, given the 2-structure $\mathrm{rev}(g)$, Algorithm 3 (which is called in Line 6) computes the respective monochromatic subgraphs $G_i(\mathrm{rev}(g))$. By Lemma 10, for each G_i Algorithm 4 (which is called in Line 10) checks whether G_i is a di-cograph or not, and by Lemma 11 in Line 11 the corresponding 1-clusters C_i^1 are returned. In Line 15, the 1-clusters $\mathscr{C}_i^1$ of all di-cographs G_i are collectively stored in the multiset $\mathscr{C}$, without removing duplicated entries.

Since $T_{\mathrm{rev}(g)}$ has at most most $n-1$ inner vertices and since each 1-cluster appears in at most 2 distinct cotrees whenever $\mathrm{rev}(g)$ is *unp* (cf. Lemma 7), we can conclude that $\mathscr{C}$ can contain at most $2(n-1)$ elements. Hence, if $|\mathscr{C}| > 2(n-1)$ then $\mathrm{rev}(g)$ is not *unp*, and therefore, g is not *unp* (Line 17).

Finally, by Lemma 12 it is checked in Line 20, if the set of 1-clusters $\mathscr{C}^1$ is a hierarchy. Hence, *TRUE* is returned if g is *unp* and *FALSE* else. □

Before we show the time complexity of Algorithm 1 we first show the time complexity of the two subroutines Algorithms 4 and 6.

Lemma 14 *For a given digraph $G = (V, E)$ and its modular decomposition $\mathbb{M}_{\mathrm{str}}$, Algorithm 4 runs in time $O(n + m)$ with $n = |V|$ and $m = |E|$.*

Proof By Lemma 9 computing the inclusion tree T of $\mathbb{M}_{\mathrm{str}}(G)$ in Line 2 takes time $O(n)$ as there are at most $O(n)$ strong modules. In the for-loop from Line 4 to 14 for each strong module M it is checked, whether or not there is an arc between two arbitrary vertices from two distinct children of M in T. This has to be done for all $O(n)$ strong modules $M \in \mathbb{M}_{\mathrm{str}}(G)$. Only if there is an arc it is further checked whether M is series or order. This can be done by checking all the arcs between vertices x and y from distinct children of M in T. In both cases (M being series and order) there is at least one arc $(x, y) \in E$ or $(y, x) \in E$, between any pair of vertices x and y. Furthermore, as only vertices from distinct children of M in T are considered, every pair (x, y) is checked at most once once. Hence, the number of all pairwise checks is bounded by $O(m)$. For the same reason, counting the arcs (Line 11) can also be done in $O(m)$ time. This accounts to a running time of $O(n + m)$ in total. □

Lemma 15 *For a given multiset of clusters $\mathscr{C}$ of size N on the ground set $\{1, \ldots, n\}$, Algorithm 6 runs in time $O(nN)$.*

Proof Computing the identifier id for each cluster in $\mathscr{C}$ (Line 2) takes time $O(N)$, computing the bit string representation for each cluster in $\mathscr{C}$ (Line 3) takes time $O(nN)$, and sorting the clusters of $\mathscr{C}$ (Line 4) using bucket sort with n buckets takes time $O(N + n)$. The for-loop from Line 5 to Line 15 runs in time $O(nN)$, as there are $O(N)$ clusters in $\mathscr{C}$ which possibly have to be removed in Line 12 from the respective

lists $\mathscr{L}_i$. The while-loop (Lines 16 to 28) is executed at most $O(N)$ times, as in each iteration one of the N clusters is removed from all the lists $\mathscr{L}_i$ that contain it (Line 24 and 27). The for-loop from Line 19 to Line 26 is executed for all of the $O(n)$ many elements $t \in L$. However, as in each execution of the inner loop (Lines 20 to 25) one of the n lists $\mathscr{L}_i$ gets empty, Lines 20 to 25 are executed n times in total and each execution takes $O(N)$ time. Hence, the time that Algorithm 6 spends on computing Lines 20 to 25 is bounded by $O(nN)$. This sums up to a total running time of $O(nN)$ for Algorithm 6. □

Finally, we show the time complexity of $O(n^2)$ for Algorithm 1.

Lemma 16 *For a given 2-structure $g = (V, \Upsilon, \varphi)$ with $n = |V|$, Algorithm 1 runs in time $O(n^2)$.*

Proof Computing the reversible refinement of g in Line 2 takes $O(n^2)$ time using Algorithm 2. In Line 3 it is assured that there are at most $2(n-1)$ labels and hence $N = |\Upsilon_{rev(g)}| < 2(n-1)$ monochromatic subgraphs $G_i(rev(g))$. Computing those $O(n)$ subgraphs at once using Algorithm 3 in Line 6 takes $O(n^2)$ time. The for-loop from Line 8 to Line 16 runs for each of the $O(n)$ many digraphs $G_i(rev(g))$. As already stated, there is an $O(n+m)$ time complexity algorithm for computing the modular decomposition of a digraph (Line 9) given in McConnell and Montgolfier (2005). By Lemma 14, Algorithm 4 (Line 10) has also a time complexity of $O(n+m)$. Algorithm 5 (Line 11) has a time complexity of $O(n)$, as by Lemma 9 constructing the inclusion tree within Line 3 of Algorithm 5 takes time $O(n)$ as there are at most $O(n)$ strong modules within G_i. Hence, all procedures within the for-loop (Lines 8 to 16) have a time complexity of $O(n+m)$. Precisely, the time complexity is $O(n+m_i)$ with $m_i = |E(G_i(rev(g)))|$ the number of arcs of $G_i(rev(g))$. The total running time of the for-loop therefore is $O(n+m_1)+O(n+m_2)+\ldots O(n+m_N) = O(n^2+\sum_{i=1}^{N} m_i)$. As each arc (x, y) occurs in exactly one of the digraphs $G_i(rev(g))$ it follows that $\sum_{i=1}^{N} m_i = n(n-1)$, which leads to a running time of $O(n^2)$ for Line 8 to 16. Line 17 assures that the multiset $\mathscr{C}$ contains at most $2(n-1)$ clusters. Hence, $|\mathscr{C}| \in O(n)$. Therefore, and by Lemma 15 Algorithm 6 runs in time $O(n^2)$. This leads to a time complexity of $O(n^2)$ for Algorithm 1. □

References

Altenhoff AM, Boeckmann B, Capella-Gutierrez S, Dalquen DA, DeLuca T, Forslund K, Huerta-Cepas J, Linard B, Pereira C, Pryszcz LP, Schreiber F, da Silva AS, Szklarczyk D, Train CM, Bork P, Lecompte O, von Mering C, Xenarios I, Sjölander K, Jensen LJ, Martin MJ, Muffato M, Gabaldón T, Lewis SE, Thomas PD, Sonnhammer E, Dessimoz C (2016) Standardized benchmarking in the quest for orthologs. Nat Methods 13(5):425–430

Böcker S, Dress AWM (1998) Recovering symbolically dated, rooted trees from symbolic ultrametrics. Adv Math 138:105–125

Brandstädt A, Le VB, Spinrad JP (1999) Graph classes: a survey. Society for Industrial and Applied Mathematics, Philadelphia

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

Crespelle C, Paul C (2006) Fully dynamic recognition algorithm and certificate for directed cographs. Discr. Appl. Math. 154:1722–1741

Dondi R, El-Mabrouk N, Lafond M (2016) Correction of weighted orthology and paralogy relations-complexity and algorithmic results. In: International workshop on algorithms in bioinformatics. Springer, pp 121–136

Ehrenfeucht A, Gabow HN, Mcconnell RM, Sullivan SJ (1994) An O(n^2) divide-and-conquer algorithm for the prime tree decomposition of two-structures and modular decomposition of graphs. J Algorithms 16(2):283–294

Ehrenfeucht A, Harju T, Rozenberg G (1995) Theory of 2-structures. In: Fülöp Z, Gécseg F (eds) Automata, languages and programming: proceedings of the 22nd international colloquium, ICALP 95 Szeged, Hungary, July 10–14, 1995. Springer, Berlin, pp 1–14

Ehrenfeucht A, Harju T, Rozenberg G (1999) The theory of 2-structures: a framework for decomposition and transformation of graphs. World Scientific, Singapore

Ehrenfeucht A, Rozenberg G (1990) Primitivity is hereditary for 2-structures. Theor Comput Sci 70(3):343–358

Ehrenfeucht A, Rozenberg G (1990) Theory of 2-structures, part I: clans, basic subclasses, and morphisms. Theor Comput Sci 70:277–303

Ehrenfeucht A, Rozenberg G (1990) Theory of 2-structures, part II: representation through labeled tree families. Theor Comput Sci 70:305–342

Engelfriet J, Harju T, Proskurowski A, Rozenberg G (1996) Characterization and complexity of uniformly nonprimitive labeled 2-structures. Theor Comput Sci 154:247–282

Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst Zool 19:99–113

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

Gray GS, Fitch WM (1983) Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from Staphylococcus aureus. Mol Biol Evol 1:57–66

Hellmuth M, Hernandez-Rosales M, Huber KT, Moulton V, Stadler PF, Wieseke N (2013) Orthology relations, symbolic ultrametrics, and cographs. J Math Biol 66:399–420

Hellmuth M, Wieseke N (2015) On symbolic ultrametrics, cotree representations, and cograph edge decompositions and partitions. In: Xu D (ed) Computing and combinatorics, lecture notes in computer science, vol 9198. Springer International Publishing, Cham, pp 609–623

Hellmuth M, Wieseke N (2016) From sequence data including orthologs, paralogs, and xenologs to gene and species trees. Springer International Publishing, Cham

Hellmuth M, Wieseke N (2016) On tree representations of relations and graphs: Symbolic ultrametrics and cograph edge decompositions. arXiv:1509.05069 (preprint)

Hellmuth M, Wieseke N, Lechner M, Lenhof HP, Middendorf M, Stadler PF (2015) Phylogenomics with paralogs. Proc Natl Acad Sci USA 112:2058–2063

Hernandez-Rosales M, Hellmuth M, Wieseke N, Huber KT, Moulton V, Stadler PF (2012) From event-labeled gene trees to species trees. BMC Bioinf 13(Suppl. 19):S6

Jensen RA (2001) Orthologs and paralogs—we need to get it right. Genome Biol 2:8

Keeling PJ, Palmer JD (2008) Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet 9:605–618

Koonin E (2005) Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39:309–338

Koonin EV, Makarova KS, Aravind L (2001) Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol 55:709–742

Lafond M, Dondi R, El-Mabrouk N (2016) The link between orthology relations and gene trees: a correction perspective. Algorithms Mol Biol 11(1):1

Lafond M, El-Mabrouk N (2015) Orthology relation and gene tree correction: complexity results. In: International workshop on algorithms in bioinformatics. Springer, pp 66–79

Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ (2011) `Proteinortho`: detection of (co-)orthologs in large-scale analysis. BMC Bioinf 12:124

Lechner M, Hernandez-Rosales M, Doerr D, Wiesecke N, Thevenin A, Stoye J, Hartmann RK, Prohaska SJ, Stadler PF (2014) Orthology detection combining clustering and synteny for very large datasets. PLoS One 9(8):e105,015

McConnell RM (1995) An $o(n^2)$ incremental algorithm for modular decomposition of graphs and 2-structures. Algorithmica 14(3):229–248

McConnell RM, de Montgolfier F (2005) Linear-time modular decomposition of directed graphs. Discr Appl Math 145(2):198–209

Möhring RH (1985) Algorithmic aspects of the substitution decomposition in optimization over relations, set systems and boolean functions. Ann Oper Res 4(1):195–225

Möhring RH, Radermacher FJ (1984) Substitution decomposition for discrete structures and connections with combinatorial optimization. Ann Discr Math 19:257–356
Schmerl JH, Trotter WT (1993) Critically indecomposable partially ordered sets, graphs, tournaments and other binary relational structures. Discr Math 113(1):191–205
Semple C, Steel M (2003) Phylogenetics, Oxford lecture series in mathematics and its applications, vol 24. Oxford University Press, Oxford
Sennblad B, Lagergren J (2009) Probabilistic orthology analysis. Syst Biol 58:411–424
Soucy SM, Huang J, Gogarten JP (2015) Horizontal gene transfer: building the web of life. Nat Rev Genet 16:472–482
Valdes J, Tarjan RE, Lawler EL (1982) The recognition of series parallel digraphs. SIAM J Comput 11:298–313