Preprint manuscript No.
(will be inserted by the editor)

Orthology Relations, Symbolic Ultrametrics, and Cographs

Marc Hellmuth · Maribel Hernandez-Rosales · Katharina T. Huber · Vincent Moulton · Peter F. Stadler · Nicolas Wieseke

November 11, 2011

Abstract Orthology detection is an important problem in comparative and evolutionary genomics and, consequently, a variety of orthology detection methods have been devised in recent years. Although many of these methods are dependent on generating gene and/or species trees, it has been shown that orthology can be estimated at acceptable levels of accuracy without having to infer gene trees and/or reconciling gene trees with species trees. Thus, it is of interest to understand how much information about the gene tree, the species tree, and their reconciliation is already contained in the orthology relation on the underlying set of genes. Here we shall show that a result by Böcker and Dress concerning symbolic ultrametrics, and subsequent algorithmic results by Semple and Steel for processing these structures can throw a considerable amount of light on this problem. More specifically, building upon these authors' results, we present some new characterizations for symbolic ultrametrics and new algorithms for recovering the associated trees, with an emphasis on how these

Marc Hellmuth
Center for Bioinformatics, Saarland University, Building E 2.1, D-66041 Saarbrücken, Germany, E-mail: marc.hellmuth@bioinf.uni-sb.de

Maribel Hernandez-Rosales
Max-Planck-Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, and Interdisciplinary Center of Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany, E-mail: maribel@bioinf.uni-leipzig.de

Katharina T. Huber and Vincent Moulton
School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK, E-mail: katharina.huber@cmp.uea.ac.uk, E-mail: vincent.moulton@cmp.uea.ac.uk

Peter F. Stadler
Bioinformatics Group, Department of Computer Science; and Interdisciplinary Center of Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, and Max-Planck-Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, and Inst. f. Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria, and Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA, Tel: **49 341 97 16690, E-mail: studla@bioinf.uni-leipzig.de

Nicolas Wieseke
Parallel Computing and Complex Systems Group, Department of Computer Science; and Interdisciplinary Center of Bioinformatics, University of Leipzig, Johannisgasse 26, D04103 Leipzig, Germany

algorithms could be potentially extended to deal with arbitrary orthology relations. In so doing we shall also show that, somewhat surprisingly, symbolic ultrametrics are very closely related to cographs, graphs that do not contain an induced path on any subset of four vertices. We conclude with a discussion on how our results might be applied in practice to orthology detection.

Keywords Orthology · Symbolic Ultrametric · Cograph · Cotree · Rooted triples

1 Introduction

With the current deluge of DNA sequencing data, orthology detection has become an important task in bioinformatics, and it lies at the heart of many comparative and evolutionary genomic studies. A variety of orthology detection methods have been devised in recent years (see e.g. Kristensen et al (2011) for an overview). Many of these methods are tree-based and typically rely on the reconciliation of a gene tree with a species tree (cf. e.g. `TreeFam` (Li et al, 2006), `PhyOP` (Goodstadt and Ponting, 2006), `PHOG` (Datta et al, 2009) and `EnsemblCompara GeneTrees` (Hubbard et al, 2007), `MetaPhOrs` (Pryszcz et al, 2011)). Even so, computing gene trees from sequence data is not only computationally demanding, but it is also a rather error-prone task especially for large data sets. Moreover, in a recent benchmark study it was shown that orthology can be estimated at acceptable levels of accuracy without even having to infer gene trees and/or to reconcile gene trees with species trees (Altenhoff and Dessimoz, 2009). It makes sense, therefore, to look at the connection of trees and orthology from a different angle: *How much information about the gene tree, the species tree, and their reconciliation is already contained in the orthology relation between genes?*

In this paper, we shall explore the following model for shedding light on this question. Suppose that X is a set of genes having a common origin, and that their evolutionary history is given by a gene tree, i.e. a (graph-theoretical) tree $T=(V,E)$ with vertex set V, edge set E and leaf set X. Typically one can think of T as being derived from a species tree, in which case the interior vertices of T will correspond to speciation or duplication events.[1] Note that two genes x,y in X are orthologs if the event corresponding to the (unique) least common ancestor $\mathrm{lca}_T(x,y)$ of x and y in T is a speciation; if x and y are not orthologs then $\mathrm{lca}_T(x,y)$ will correspond to some other events such as a duplication. In particular, we obtain a map t from the set of interior vertices of T to some set M of events, and, consequently, a map $d_{(T;t)}$ from distinct pairs x,y in X to M given by putting $d_{(T;t)}(x,y)=t(\mathrm{lca}_T(x,y))$. These concepts are illustrated in Fig. 1. Note that in practice, we do not necessarily know the pair $(T;t)$, but that there are bioinformatics methods that allow us to estimate the values $d_{(T;t)}(x,y)$ for $x,y\in X$ (Altenhoff and Dessimoz, 2009; Lechner et al, 2011). Hence, in this set-up, the above question can be rephrased as follows: *Given an arbitrary symmetric map $\delta : X\times X \rightarrow M$, i.e. an orthology relation, can we determine if there is a pair $(T;t)$ with for which $d_{(T;t)}(x,y)=\delta(x,y)$ holds for $x,y\in X$ distinct and, if not, can we at least find some pair $(T;t)$ where this is almost true?*

[1] In reality, other events such as horizontal gene transfer might also occur, although we will not consider these explicitly here.

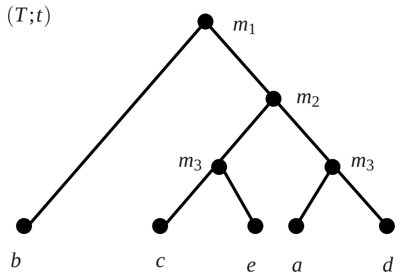

Fig. 1 A phylogenetic tree $T = (V,E)$ on the set $X = \{a,\ldots,e\}$, together with a map t from the set of interior vertices of T to the set of events $M = \{m_1,m_2,m_3\}$, as indicated by the labels on the interior vertices of T. The vertex in V that is the least common ancestor of c and a has label m_2 and so $d_{(T;t)}(c,a) = m_2$.

Intriguingly, a solution to the first part of this question has already been given by Böcker and Dress (1998) in a different context. In particular, they completely characterized maps of the form $d_{(T;t)}$, with $(T;t)$ as above, maps which they called *symbolic ultrametrics*. Moreover, in subsequent work Semple and Steel (2003) presented an algorithm that can be used to reconstruct T and t from any given symbolic ultrametric. In this paper, we shall build upon these results, presenting some new characterizations for symbolic ultrametrics and novel algorithms for recovering the associated trees, with an emphasis on how these results and algorithms could be potentially used to cope with arbitrary orthology relations. In so doing we shall also show that, somewhat surprisingly, symbolic ultrametrics are very closely related to a well-studied class of graphs called cographs, which is precisely the class of graphs that do not contain induced paths on any subset of four vertices (Corneil et al, 1981).

The rest of this paper is organized as follows. In Section 2, we present the basic and relevant concepts used in this paper. In Section 3, we recall the aforementioned results concerning symbolic ultrametrics from (Böcker and Dress, 1998) and (Semple and Steel, 2003), and prove some mild generalizations of these results that are relevant to the question above concerning orthology relations. In Section 4, we show that symbolic ultrametrics can also be characterized in terms of cographs (see Proposition 3) and that the tree corresponding to a symbolic ultrametric can also be recovered using cotrees, trees that can be canonically associated to cographs. In Section 5, we present a connection between symbolic ultrametrics and a certain collection of partitions that can be associated to the corresponding tree (see Corollary 3). We use this result in the following section to help obtain a new algorithm for deciding whether or not a map is a symbolic ultrametric and, if this is the case, for constructing its corresponding tree representation. We conclude in Section 7 with a discussion on how our results might be applied in practice to orthology detection.

2 Preliminaries: Phylogenetic Trees and Rooted Triples

In this section, we present the relevant basic concepts and notation. Unless stated otherwise, we will follow Semple and Steel (2003).

In the remainder of this paper, X will always denote a finite set of size at least three.

A tree $T=(V,E)$ is a connected cycle-free graph with vertex set $V(T)=V$ and edge set $E(T)=E$. A vertex of T of degree one is a called a *leaf* of T and all other vertices of T are called *interior*. A *star* is a tree that has at most one interior vertex. An edge of T is *interior* if both of its end vertices are interior vertices. The sets of interior vertices and interior edges of T are denoted by V^0 and E^0, respectively.

A *rooted tree* $T=(V,E)$ is a tree that contains a distinguished vertex $\rho_T \in V$ called the *root*. Without explicitly stating it we will always assume that a rooted tree is directed in that all edges of T are directed away from ρ_T. For ease of representation we will always draw rooted trees with the root at the top. A rooted tree T is called *binary* if every interior vertex of T has outdegree two. We define a partial order $\preceq_T$ on V by setting $v \preceq_T w$ for any two vertices $v,w \in V$ for which v is a vertex on the path from ρ_T to w. In particular, if $v \preceq_T w$ we call v an *ancestor* of w.

A *phylogenetic tree T (on X)* is a rooted tree with leaf set X that does not contain any vertices with in- and outdegree one and whose root ρ_T has indegree zero. For $A \subseteq X$ a non-empty subset, we define $\mathrm{lca}_T(A)$, or the *most recent common ancestor of A*, to be the unique vertex in T that is the greatest lower bound of A under the partial order $\preceq_T$. In case $A=\{x,y\}$ we put $\mathrm{lca}_T(x,y)=\mathrm{lca}_T(\{x,y\})$. We denote by $T(W)$ the (rooted) subtree of T with root $\mathrm{lca}_T(W)$. For convenience, we will sometimes denote the root of $T(W)$ by ρ_W. Two phylogenetic trees T_1 and T_2 on X are said to be *isomorphic* if there is a bijection $\psi : V(T_1) \to V(T_2)$ that induces a (directed) graph isomorphism from T_1 to T_2 which is the identity on X and maps the root of T_1 to the root of T_2.

Suppose T is a phylogenetic tree on X with root ρ_T and a non-empty subset $Y \subseteq X$ with $|Y| \geq 2$. Then the *restriction* $T|Y$ of T to Y is the phylogenetic tree obtained from $T(Y)$ by suppressing all vertices of degree two with the exception of ρ_T if $\rho_T \in V(T(Y))$. For every vertex $v \in V(T)$ we denote by $C(v)$ the subset of X such that $v=\mathrm{lca}_T(C(v))$ and put $\mathscr{C}(T)=\bigcup_{v\in V(T)}\{C(v)\}$. We say that a phylogenetic tree S on X *refines* T, in symbols $T \leq S$, if $\mathscr{C}(T) \subseteq \mathscr{C}(S)$. In addition, we say that T *displays* a phylogenetic tree S on Y if S can be obtained from the restriction $T|Y$ of T to Y by contracting interior edges. Note that contraction of non-interior edges would not result in a valid phylogenetic tree as such a tree could e. g. have an interior vertex contained in Y. We say that a set $\mathscr{R}$ of phylogenetic trees all having leaves in X is *compatible* if $\mathscr{R}=\emptyset$ or if there is an phylogenetic tree T on X that displays every tree contained in $\mathscr{R}$.

A *(rooted) triple* is a binary phylogenetic tree on a set Y with $|Y|=3$. For $Y := \{x,y,z\} \in \binom{X}{3}$, we denote by $xy|z$ the unique triple t on Y with root ρ_t for which $\mathrm{lca}_t(x,y) \neq \rho_t$ holds. Given an phylogenetic tree T on X we denote by

$$\mathscr{R}_T := \left\{ T|Y \,:\, Y \in \binom{X}{3} \text{ and } T|Y \text{ is binary} \right\} \tag{1}$$

its set of rooted triples. Note that, for any phylogenetic tree T on X, we have $|\mathscr{R}_T| \leq \binom{|X|}{3}$ and that the maximum is attained precisely if T is binary.

The importance of sets of rooted triples stems from the fact that the set $\mathscr{R}_T$ of rooted triples displayed by a phylogenetic tree T uniquely determines T up to isomorphism, i.e. if T' is a phylogenetic tree for which $\mathscr{R}_T = \mathscr{R}_{T'}$ holds then T and T' must be isomorphic. In fact, a more general result of this nature is presented by Semple and Steel (2003, p. 119-120):

Theorem 1 *Let $\mathscr{R}$ be a collection of triples so that the union of their leaf sets is X. Then there is a polynomial-time algorithm — called* `BUILD` *— that, when applied to $\mathscr{R}$, either:*

(i) outputs a phylogenetic tree on X that displays $\mathscr{R}$ if $\mathscr{R}$ is compatible; or
(ii) outputs the statement "$\mathscr{R}$ is not compatible".

Note that the original version of the `BUILD` algorithm is due to Aho et al (1981). A more efficient solution of the same problem has been described e.g. by Rauch Henzinger et al (1999).

3 Symbolic Ultrametrics

In this section, we recall some results from Böcker and Dress (1998) and Semple and Steel (2003, Section 7) concerning symbolic ultrametrics. We also prove some mild generalizations of these results with the view to their possible application to orthology relations. More details on such applications will be discussed in the last section.

From now on, M will always denote a non-empty finite set, the symbol $\odot$ will always denote a special element not contained in M, and $M^{\odot} := M \cup \{\odot\}$. The symbol $\odot$ corresponds to a "non-event" and is introduced for purely technical reasons. It will always correspond only to the leaves of a phylogenetic tree since these will not usually correspond to events such as speciation and duplication.

Now, suppose $\delta : X \times X \to M^{\odot}$ is a map. We call δ a *symbolic ultrametric*[2] if it satisfies the following conditions:

(U0) $\delta(x,y) = \odot$ if and only if $x = y$;
(U1) $\delta(x,y) = \delta(y,x)$ for all $x,y \in X$, i.e. δ is symmetric;
(U2) $|\{\delta(x,y), \delta(x,z), \delta(y,z)\}| \leq 2$ for all $x,y,z \in X$; and
(U3) there exists no subset $\{x,y,u,v\} \in \binom{X}{4}$ such that

$$\delta(x,y) = \delta(y,u) = \delta(u,v) \neq \delta(y,v) = \delta(x,v) = \delta(x,u). \tag{2}$$

Note that every symmetric map δ on X with $|X| = 3$ that also satisfies Properties (U0) and (U2) is a symbolic ultrametric on X. Also note that every *ultrametric d* on X (that is, a symmetric map d from $X \times X$ to the real numbers which vanishes on the diagonal and that satisfies the additional property that $d(x,z) \leq \max\{d(x,y), d(y,z)\}$ holds for all $x,y,z \in X$) is also symbolic ultrametric if the special symbol $\odot$ is identified with 0. Ultrametrics are well-studied in phylogenetics as they correspond to weighted, rooted trees (cf. e.g. Semple and Steel (2003)).

[2] Note that in Böcker and Dress (1998) a symbolic ultrametric is defined without the requirement (U0), which we have introduced for technical reasons.

Now, suppose that $T=(V,E)$ is a phylogenetic tree on X and that $t:V\to M^{\odot}$ is a map such that $t(x)=\odot$ for all $x\in X$. We call such a map t a *symbolic dating map* for T; it is *discriminating* if $t(u)\neq t(v)$, for all edges $\{u,v\}\in E$. To the pair $(T;t)$ we associate the map $d_{(T;t)}$ on $X\times X$ by setting, for all $x,y\in X$,

$$d_{(T;t)}:X\times X\to M^{\odot};d_{(T;t)}(x,y)=t(\mathrm{lca}_T(x,y)). \tag{3}$$

Clearly this map is symmetric and satisfies (U0). We call the pair $(T;t)$ a *symbolic representation* of a map $\delta:X\times X\to M^{\odot}$ if $\delta(x,y)=d_{(T;t)}(x,y)$ holds for all $x,y\in X$; it is called discriminating if t is discriminating (see Fig 1 for an example of a discriminating symbolic representation where we have omitted the assignment of the value $\odot$ to the leaves). Note that we call two symbolic representations $(T;t)$ and $(T';t')$ of δ *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 Böcker and Dress (1998), the following fundamental result concerning the relationship between symbolic ultrametrics and symbolic representations is proven:

Theorem 2 *Suppose $\delta:X\times X\to M^{\odot}$ is a map. Then there is a discriminating symbolic representation of δ if and only if δ is a symbolic ultrametric. Furthermore, up to isomorphism, this representation is unique.*

Given any symbolic ultrametric δ on X, we denote the unique discriminating symbolic representation of δ given by this theorem by $(T_\delta;t_\delta)$.

Note that the symbolic tree representation of an orthology relation on a set of genes need not necessarily be discriminating, since duplication events do not necessarily have to come directly after speciation events and vice versa. To help deal with this, we shall now prove a simple result concerning the relationship between symbolic ultrametrics and arbitrary symbolic representations. To this end, suppose that $T=(V,E)$ is a phylogenetic tree on X and that $t:V\to M^{\odot}$ is a symbolic dating map that is not discriminating. Then there must exist some $e=\{u,v\}\in E^0$ such that $t(u)=t(v)$. Let v_e denote the vertex in T obtained by collapsing the edge e. Then the tree $T_e=(V_e,E_e)$ with vertex set $V_e=V\setminus\{u,v\}\cup\{v_e\}$, edge set $E_e=E\setminus\{e\}\cup\{\{e_v,w\}:\{w,u\}\text{ or }\{w,v\}\in E\}$ is clearly a phylogenetic tree on X. Furthermore the map $t_e:V_e\to M^{\odot}$ defined by putting, for all $w\in V_e$,

$$t_e(w)=t(w)\text{ if }w\neq v_e\text{ and }t(v_e)=t(u) \tag{4}$$

is again a symbolic dating map for T_e. Clearly, this construction can be repeated, with $(T_e;t_e)$ now playing the role of $(T;t)$, until a phylogenetic tree $\hat{T}=(\hat{V},\hat{E})$ on X is obtained together with a discriminating symbolic dating map $\hat{t}$ on $\hat{T}$.

Proposition 1 *Let $\delta:X\times X\to M^{\odot}$ be a map. Then the following are equivalent:*

(i) *δ is a symbolic ultrametric.*
(ii) *there is a discriminating symbolic representation of δ.*
(iii) *there is a symbolic representation of δ.*

Moreover, if δ is a symbolic ultrametric, and $(T;t)$ is any symbolic representation of δ, then $(\hat{T};\hat{t})$ is isomorphic to $(T_\delta;t_\delta)$.

Proof (i) ⇒ (ii): Apply Theorem 2.

(ii) ⇒ (iii): This is obvious.

(iii) ⇒ (i): It is straight-forward to check that if there is a symbolic representation $(T;t)$ of δ, then δ must satisfy (U0)–(U3). Then apply Theorem 2.

To see that the final statement holds, note that if $\delta : X \times X \to M^{\odot}$ has a symbolic representation $(T;t)$, and $e = \{u,v\} \in E(T)$ with $t(u) = t(v)$, then $d_{(T_e;t_e)} = d_{(T;t)}$. Therefore, $d_{(\hat{T};\hat{t})} = d_{(T';t')}$ must also hold. Moreover, $\hat{t}$ is discriminating by construction and thus, by Theorem 2, the proposition follows. □

We conclude this section by recalling a practical approach for constructing the discriminating symbolic representation $(T_\delta;t_\delta)$ from a given symbolic ultrametric $\delta : X \times X \to M^{\odot}$ based on the BUILD algorithm, that was described by Semple and Steel (2003, Section 7.6).

Let $\delta : X \times X \to M^{\odot}$ be a symbolic ultrametric on X and let $\mathscr{R}(\delta)$ be the set of triples $xy|z$, $\{x,y,z\} \in \binom{X}{3}$ satisfying one of the following two conditions:

(R1) $\delta(x,y) \neq \delta(x,z) = \delta(y,z)$, or
(R2) $\delta(x,y) = \delta(x,z) = \delta(y,z)$, and there is some $w \in X$ such that $\delta(x,w) = \delta(y,w) \neq \delta(z,w) = \delta(x,y)$.

Furthermore, denote by $\mathscr{R}_\delta \subseteq \mathscr{R}(\delta)$ the subset of $\mathscr{R}(\delta)$ consisting only of the triples satisfying condition (R1). If δ is a symbolic ultrametric then $\mathscr{R}(\delta) = \mathscr{R}_{T_\delta}$ (Böcker and Dress, 1998, Lemma 2). Moreover, we have the following result, which is a mild generalization of (Semple and Steel, 2003, p. 167-8):

Proposition 2 *Let $\delta : X \times X \to M^{\odot}$ be a map that satisfies Properties (U0)–(U2). Then the following are equivalent:*

(i) δ is a symbolic ultrametric.
(ii) $\mathscr{R}(\delta)$ is compatible.
(iii) $\mathscr{R}_\delta$ is compatible.

In particular, δ is a symbolic ultrametric if and only if the BUILD algorithm applied to $\mathscr{R}_\delta$ or $\mathscr{R}(\delta)$ returns a phylogenetic tree T, in which case the map $t : V(T) \to M^{\odot}$, $v \mapsto \delta(x,y)$ with $v = \mathrm{lca}_T(x,y)$, $x,y \in X$, is well-defined and $(T;t)$ is isomorphic to the discriminating symbolic representation for δ.

Proof Clearly all 3 assertions are equivalent if $|X| = 3$. So assume $|X| \geq 4$. The implications (i) ⇒ (ii) and (ii) ⇒ (iii) are trivial in view of the observation preceding Proposition 2.

(iii) ⇒ (i): Suppose for contradiction that $\mathscr{R}_\delta$ is compatible but that δ is not a symbolic ultrametric. Then δ does not satisfy Property (U3) and so there exists some $\{x,y,u,v\} \in \binom{X}{4}$ such that $\delta(x,y) = \delta(y,u) = \delta(u,v) \neq \delta(y,v) = \delta(x,v) = \delta(x,u)$. But then $\mathscr{R} := \{xy|v, xu|y, uv|x\} \subseteq \mathscr{R}_\delta$ must hold which is impossible as $\mathscr{R}$ is not compatible and thus $\mathscr{R}_\delta$ cannot be compatible. □

It follows from this result and Theorem 2 that we can decide in polynomial time whether or not δ is a symbolic ultrametric by applying the BUILD algorithm to the set $\mathscr{R}_\delta$, which will also construct a symbolic representation of δ in case it is. The following additional consequence, which will not be used later, is also worth noting:

Corollary 1 *Suppose δ is a symbolic ultrametric on X. Then δ has a unique symbolic representation if and only if $|\mathscr{R}(\delta)| = \binom{|X|}{3}$.*

Proof Suppose first that $|\mathscr{R}(\delta)| = \binom{|X|}{3}$. Then $|\mathscr{R}_{T_\delta}| = \binom{|X|}{3}$ in view of (Böcker and Dress, 1998, Lemma 2) recalled above as δ is a symbolic ultrametric. Since only a binary phylogenetic tree can display $\binom{|X|}{3}$ triples, it follows that T_δ must be binary. But this implies immediately that $(T_\delta;t_\delta)$ is the unique symbolic representation for δ because any symbolic representation for δ can be obtained from $(T_\delta;t_\delta)$ by resolving interior vertices of T_δ.

Conversely, assume that δ has a unique symbolic representation $(T;t)$. Then T must be binary as otherwise, by Proposition 1, there would exist an interior vertex of T that could be resolved to obtain a new symbolic representation $(T';t')$ for δ contradicting the uniqueness of $(T;t)$. But then $(T;t)$ is isomorphic to $(T_\delta;t_\delta)$ and so $|\mathscr{R}_{T_\delta}| = \binom{|X|}{3}$. Since δ is a symbolic ultrametric on X, Lemma 2 of Böcker and Dress (1998) implies $\mathscr{R}_{T_\delta} = \mathscr{R}(\delta)$ and so the corollary follows. □

4 Cographs and Cotrees

In this section, we shall investigate a connection between symbolic ultrametrics and *complement-reducible graphs* or *cographs*. As mentioned in the introduction, a cograph is a P_4-free graph (i.e. a graph such that no four vertices induce a subgraph that is a path of length 3), although there are a number of equivalent characterizations of such graphs (see e.g. (Brandstädt et al, 1999) for a survey).

Let $\delta : X \times X \to M^\odot$ be a map satisfying Properties (U0) and (U1). For $x \in X$ and $m \in M$, we define the *neighborhood* $N_m(x)$ of x with respect to m and δ as

$$N_m(x) = N_{m,\delta}(x) := \{y \in X \,:\, \delta(x,y) = m\}\,. \tag{5}$$

Note that, in view of Property (U0), $x \notin N_m(x)$ and that, in view of Property (U1), $y \in N_m(x)$ if and only if $x \in N_m(y)$. We also define, for each fixed $m \in M$, an undirected graph $G_m(\delta) = (V_m, E_m)$ with vertex set $V_m = V_m(\delta) = X$ and edge set

$$E_m = E_m(\delta) := \left\{ \{x,y\} \in \binom{X}{2} \,:\, y \in N_m(x),\ x \in X \right\}. \tag{6}$$

For example, if $\delta = d_{(T;t)}$ for the pair $(T;t)$ depicted in Fig. 1, then $G_{m_1}(\delta)$ is the graph with vertex set $\{a,\ldots,e\}$ and edge set $\{\{a,b\},\{d,b\},\{e,b\},\{c,b\}\}$, and $G_{m_3}(\delta)$ is the graph with the same vertex set as $G_{m_1}(\delta)$ and edge set $\{\{a,d\},\{c,e\}\}$.

The following result gives the aforementioned connection between symbolic ultrametrics and cographs:

Proposition 3 *Let $\delta : X \times X \to M^\odot$ be a map satisfying Properties (U0) and (U1). Then δ is a symbolic ultrametric if and only if*

(U2') For all $\{x,y,z\} \in \binom{X}{3}$ there is an $m \in M$ such that $E_m(\delta)$ contains two of the three edges $\{x,y\}$, $\{x,z\}$, and $\{y,z\}$.

(U3') $G_m(\delta)$ is a cograph for all $m \in M$.

Proof Suppose that δ is a map as in the statement of the proposition. Note that we may assume $|X| \geq 4$.

Clearly, δ satisfies (U2) if and only if it satisfies Property (U2'). Moreover, it is easy to see that (U3') implies (U3). Thus if (U3') and (U2') hold, then δ is a symbolic ultrametric on X. Thus, it only remains to show that if δ satisfies (U2) and (U3) (i.e. δ is a symbolic ultrametric), then it must satisfy (U3').

Suppose this is not the case, i.e. (U3') does not hold. Then there exists $\{x,y,u,v\} \in \binom{X}{4}$ and some $m \in M$ such that the subgraph of $G(\delta)$ induced on $\{x,y,u,v\}$ is a path of length three. Suppose that this path is x,y,u,v. Then $\delta(x,y) = \delta(y,u) = \delta(u,v) = m$ and $m \notin \{\delta(x,u), \delta(x,v), \delta(y,v)\}$. But (U2) implies $\delta(x,u) = \delta(x,v) = \delta(y,v)$, and so (U3) does not hold. This contradiction completes the proof. □

Intriguingly, it is well-known in the literature concerning cographs that, to any cograph G, one can associate a canonical *cotree* $T(G) = (V,E)$. This is a rooted tree with root[3] ρ, leaf set equal to the vertex set $V(G)$ of G and inner vertices that represent so-called "join" and "union" operations together with a labeling map $\lambda_G : V^0 \to \{0,1\}$ such that $\lambda_G(\rho) = 1$ and, if $v \in V^0$ and $w_1, \ldots, w_k \in V^0$, $k \geq 2$, are the children of v, then $|\lambda_G(v) - \lambda_G(w_i)| = 1$, for all $i = 1, \ldots, k$ (cf. (Corneil et al, 1981)). For example, if $\delta = d_{(T;t)}$ for the pair $(T;t)$ depicted in Fig. 1, then the cotrees associated to the cographs $G_{m_1}(\delta)$, $G_{m_2}(\delta)$, and $G_{m_3}(\delta)$, respectively, are depicted in Fig. 2. Note that the cotree associated to a cograph has root labeled with 0 if and only if the cograph is disconnected.

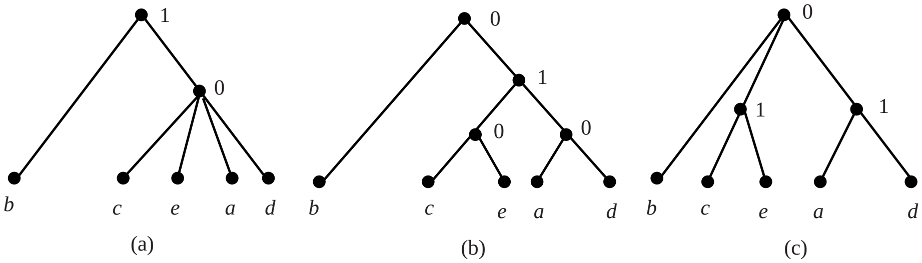

Fig. 2 For the symbolic ultrametric $\delta = d_{(T;t)}$, with $(T;t)$ pictured in Fig. 1, the three cotrees $(T(G_{m_i}(\delta)), \lambda_{G_{m_i}(\delta)})$, $i = 1,2,3$, pictured in that order from left to right. Note that the tree T depicted in Fig. 1 refines each of the cotrees.

The key observation about cographs that concerns us here is that, given a cograph G, a pair $\{x,y\} \in \binom{V(G)}{2}$ is an edge in G if and only if $\lambda_G(\mathrm{lca}_{T(G)}(x,y)) = 1$ (cf. (Corneil et al, 1981, p. 166)). It is therefore natural to ask what the relationship is between the discriminating representation of a symbolic ultrametric δ and the cotrees associated to the cographs coming from δ given by Proposition 3. We shall now show that there is a very close connection between these structures.

[3] Note that in cotrees the root might have outdegree one; in such cases we will simply suppress this vertex and its outgoing edge.

To this end, suppose $\delta : X \times X \to M^{\odot}$ is a map satisfying Properties (U0) and (U1) and $m \in M$. Consider the map $\delta_m : X \times X \to \{0,1,\odot\}$ defined, for all $x,y \in X$, by putting

$$\delta_m(x,y) = \begin{cases} \odot & \text{if} \quad x=y, \\ 1 & \text{if} \quad \{x,y\} \in E_m(\delta), \\ 0 & \text{if} \quad \text{else}. \end{cases} \tag{7}$$

Note if δ is a symbolic ultrametric on X, then it is easy to see that δ_m is also a symbolic ultrametric on X, $m \in M$ (essentially because $G(\delta_m)$ is a cograph).

Lemma 1 *Let $\delta : X \times X \to M^{\odot}$ be a symbolic ultrametric. Then, for all $m \in M$, $(T(G_m(\delta)); \lambda_{G_m(\delta)})$ is the discriminating symbolic representation for δ_m.*

Proof Suppose $m \in M$, and let $T' = T(G_m(\delta))$ and $t' = \lambda_{G_m(\delta)}$. In view of Theorem 2 it suffices to show that $\delta_m(x,y) = d_{(T';t')}(x,y)$ holds for all $x,y \in X$. Let $x,y \in X$. Then, by the aforementioned properties of the cotree associated to a cograph and Proposition 3, it follows that $d_{(T';t')}(x,y) = t'(\mathrm{lca}_{T'}(x,y)) = 1$ if and only if $\{x,y\} \in E_m(\delta)$ if and only if $\delta_m(x,y) = 1$, as required. □

Using this lemma, we now prove a technical result which, given a symbolic ultrametric δ, relates triples in $\mathscr{R}(\delta)$ and, for $m \in M$, triples in $\mathscr{R}_{\delta_m}$.

Theorem 3 *Let $\delta : X \times X \to M^{\odot}$ be a symbolic ultrametric. Then the following hold:*

(i) *For all $m \in M$, $\mathscr{R}_{\delta_m} \subseteq \mathscr{R}_{\delta}$.*
(ii) *For all $m \in M$, $\mathscr{R}(\delta_m) \subseteq \mathscr{R}(\delta)$.*
(iii) *$\mathscr{R}_{\delta} = \bigcup_{m \in M} \mathscr{R}_{\delta_m}$.*

Proof (i) Suppose $m \in M$ and $xy|z \in \mathscr{R}_{\delta_m}$. Then $\delta_m(x,y) \neq \delta_m(x,z) = \delta_m(y,z)$ and so either (a) $\delta_m(x,y) = 1$ and $\delta_m(x,z) = \delta_m(y,z) = 0$ or (b) $\delta_m(x,y) = 0$ and $\delta_m(x,z) = \delta_m(y,z) = 1$.

If Case (a) holds then $\{x,y\} \in E_m(\delta)$ and $\{x,z\},\{y,z\} \notin E_m(\delta)$. Hence $\delta(x,y) = m$ and $\delta(x,z), \delta(y,z) \neq m$. Since δ is an ultrametric and so satisfies Property (U2) it follows that $\delta(x,z) = \delta(y,z)$. Consequently, $xy|z \in R_{\delta}$ in this case.

If Case (b) holds then $\{x,y\} \notin E_m(\delta)$ and $\{x,z\},\{y,z\} \in E_m(\delta)$. But then $\delta(x,z) = \delta(y,z) = m \neq \delta(x,y)$ and so $xy|z \in R_{\delta}$ must hold in this case, too.

(ii) Let $m \in M$. Suppose $xy|z \in \mathscr{R}(\delta_m)$. Assume first that $xy|z$ satisfies Property (R1). Then Assertion (i) implies $xy|z \in \mathscr{R}_{\delta} \subseteq \mathscr{R}(\delta)$. So assume that $xy|z$ does not satisfy Property (R1). Then $xy|z \notin \mathscr{R}_{\delta_m}$ and $xy|z$ must satisfy Property (R2), that is, $\delta_m(x,y) = \delta_m(x,z) = \delta_m(y,z)$ and there must exist some $w \in X$ such that $\delta_m(x,w) = \delta_m(y,w) \neq \delta_m(z,w) = \delta_m(x,y)$. We distinguish the cases $\delta_m(x,y) = \delta_m(x,z) = \delta_m(y,z) = 1$ and $\delta_m(x,y) = \delta_m(x,z) = \delta_m(y,z) = 0$.

Assume first that $\delta_m(x,y) = \delta_m(x,z) = \delta_m(y,z) = 1$ holds. Then $m = \delta(x,y) = \delta(x,z) = \delta(y,z)$ and so $\delta(z,w) = m$ and $m \notin \{\delta(x,w), \delta(y,w)\}$ must hold. But then Property (U2) implies that $\delta(x,w) = \delta(y,w) \neq m$ and so (R2) holds. Thus, $xy|z \in \mathscr{R}(\delta)$ in this case.

Now, assume that $\delta_m(x,y) = \delta_m(x,z) = \delta_m(y,z) = 0$ holds. Then $m \notin \{\delta(x,y), \delta(x,z), \delta(y,z), \delta(z,w)\}$ and so $m = \delta(x,w) = \delta(y,w)$. By Property (U2) it follows that $m_1 := \delta(y,z) = \delta(z,w) = \delta(z,x) \neq m$. If $m_2 := \delta(x,y) = m_1$ then $xy|z$ satisfies Property (R2) for δ and so $xy|z \in \mathscr{R}(\delta)$. If $m_2 \neq m_1$ then $xy|z \in \mathscr{R}_{\delta_{m_2}} \subseteq \mathscr{R}_\delta \subseteq \mathscr{R}(\delta)$ in view of Assertion (i). This completes the proof of (ii).

(iii) Statement (i) clearly implies $\bigcup_{m\in M} \mathscr{R}_{\delta_m} \subseteq \mathscr{R}_\delta$. To see that the converse set inclusion holds, let $xy|z \in \mathscr{R}_\delta$. Then there exists some $m \in M$ such that $m = \delta(x,y) \neq \delta(x,z) = \delta(y,z)$ and thus $\{x,y\} \in E_m(\delta)$ and $\{x,z\},\{y,z\} \notin E_m(\delta)$. Hence, $\delta_m(x,y) = 1 \neq 0 = \delta_m(x,z) = \delta_m(y,z)$ and so $xy|z \in \mathscr{R}_{\delta_m}$, as required. □

Using this theorem, we now see how the discriminating symbolic representation T_δ of a symbolic ultrametric δ can be constructed from the cotrees $T(G_m(\delta))$, $m \in M$ (or, equivalently, the discriminating symbolic representations of the maps δ_m, $m \in M$). The first statement of the following corollary is illustrated in Fig. 2.

Corollary 2 *Let $\delta : X \times X \to M^\odot$ be a symbolic ultrametric. Then, for each $m \in M$, $T(G_m(\delta)) \leq T_\delta$. Moreover, T_δ is isomorphic to the tree obtained by applying `BUILD` to the set $\bigcup_{m\in M} \mathscr{R}_{\delta_m}$.*

Proof The second statement follows immediately from Theorem 3(ii) and Proposition 2.

To see that $T(G_m(\delta)) \leq T_\delta$ holds for all $m \in M$, note that since δ_m is a symbolic ultrametric $\mathscr{R}(\delta_m) = \mathscr{R}_{T_{\delta_m}}$ holds by Lemma 2 of (Böcker and Dress, 1998) recalled above. Hence by Theorem 3 (ii), $\mathscr{R}_{T_{\delta_m}} \subseteq \mathscr{R}_{T_\delta}$. By Theorem 6.4.1 of Semple and Steel (2003) this last statement holds if and only if $T_{\delta_m} \leq T_\delta$. Now apply Lemma 1. □

Remark 1 By modifying the argument in the proof of part (iii) of Theorem 3, it is straight-forward to show, under the same assumptions given in the theorem plus the additional assumption $|M| \geq 3$, that T_δ is isomorphic to the tree obtained by applying `BUILD` to the set $\bigcup_{m\in M'} \mathscr{R}_{\delta_m}$, for any $M' \subseteq M$ with $|M'| = |M| - 1$. However, in general it is not possible to obtain T_δ using `BUILD` in this way by using subsets of M with size less than $|M| - 1$ (see Fig. 3).

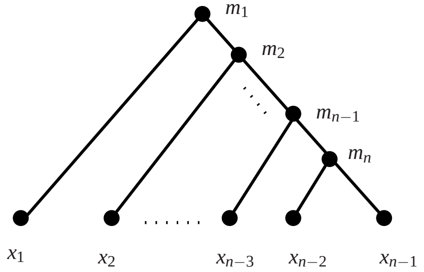

Fig. 3 A symbolic representation of a symbolic ultrametric δ on the set $X = \{x_1, \ldots, x_{n-1}\}$ with values in the set $M = \{m_1, \ldots, m_n\}$. It can be shown that it is not possible to reconstruct T_δ by applying `BUILD` to the set $\bigcup_{m\in M'} R_{\delta_m}$, for any $M' \subseteq M$ with $|M'| \leq n-2$.

5 Pseudo-Cherries, Cliques, and Partitions

In this section, we will show that the cliques[4] in a certain graph $G(\delta)$ that can be associated to a symbolic ultrametric $\delta : X \times X \to M^{\odot}$ are closely related to the structure of the discriminating symbolic representation of δ (see Proposition 4). We use this result in the next section to help derive a new algorithm for determining whether a map is a symbolic ultrametric or not. We shall also show that cliques in $G(\delta)$ can be characterized in terms of cliques in the graphs $G_m(\delta)$, $m \in M$, defined in the previous section (see Corollary 3).

Let $\delta : X \times X \to M^{\odot}$ be a symmetric map that satisfies (U0). For $\delta(x,y) = m \in M$, $x \neq y$, we have $\{x,y\} \subseteq N_m(x) \triangle N_m(y)$, where $\triangle$ denotes the usual symmetric difference of sets. For future reference note that, with $N_m[x] := N_m(x) \cup \{x\}$, $x \in X$, we have

$$N_m(x) \triangle N_m(y) = \{x,y\} \text{ if and only if } N_m[x] = N_m[y], \tag{8}$$

for all $m \in M$ and all $y,z \in X$. Also note that this condition is satisfied for at most one $m \in M$ for any given pair $\{x,y\} \in \binom{X}{2}$.

Now, define $G(\delta)$ to be the graph with vertex set X and edge set

$$E(\delta) := \left\{ \{x,y\} \in \binom{X}{2} : N_m[x] = N_m[y] \text{ for some } m \in M \right\}. \tag{9}$$

For example, if $\delta = d_{(T;t)}$ for the pair $(T;t)$ depicted in Fig. 1, then the graph $G(\delta)$ is the graph with vertex set $\{a,\ldots,e\}$ and edge set $\{\{c,e\},\{a,d\}\}$. We denote by $\mathfrak{C}(G)$ the (set-inclusion) maximal cliques of a graph G, and for brevity we let $\mathfrak{C}(\delta)$ denote $\mathfrak{C}(G(\delta))$, for $\delta : X \times X \to M^{\odot}$ a symmetric map that satisfies (U0). Note that $\delta(x,y) = \delta(u,v)$ holds for any clique $C \in \mathfrak{C}(\delta)$ with $|C| \geq 2$ and any two $\{x,y\},\{u,v\} \in \binom{C}{2}$ in case δ is a symbolic ultrametric. Also note that there exists a $C \in \mathfrak{C}(\delta)$ with $|C| \geq 2$ in this case because the tree T_δ has a vertex such that all of its children (of which there must be at least two) are leaves.

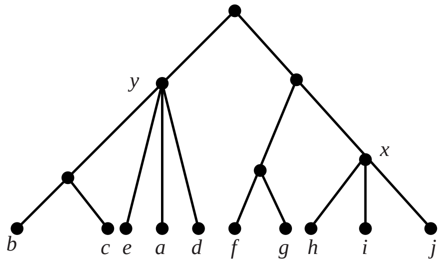

Fig. 4 A phylogenetic tree T on $X = \{a,b,c,\cdots,j\}$. The vertices $x = \mathrm{lca}_T(C')$ and $y = \mathrm{lca}_T(C)$ are the most recent common ancestors of the sets $C = \{a,d,e\}$ and $C' = \{h,i,j\}$. Both C and C' are pseudo-cherries of T. However, C' is also a cherry of T whereas C is not.

[4] A clique in a graph is a subset of its vertices such that every two vertices in the subset are connected by an edge.

Suppose that T is a phylogenetic tree on X with root ρ. Let $C \subseteq X$ be a non-empty subset of X and put $v_C = \mathrm{lca}_T(C)$. We call C a *pseudo-cherry* of T if a leaf x of T is adjacent to v_C if and only if $x \in C$. If, in addition, every vertex $v \in V(T)$ adjacent to v_C that does not lie on a path from ρ to v_C is contained in X, then we call C a *cherry* of T. Note that a pseudo-cherry must contain at least one element and that the definition of a cherry C reduces to the usual definition of a cherry (as given e.g. by Semple and Steel (2003)) in case $|C| = 2$. We illustrate these two definitions in Fig. 4.

Now, let $t : V(T) \to M^{\odot}$ be a symbolic dating map for T. For each $m \in M$, we define a relation $\sim_m$ on X by putting, for all $x, y \in X$, $x \sim_m y$ if $x = y$ or, in case x and y are distinct, $t(u) = m$ holds for every interior vertex u of T that lies on the unique path from x to y. Clearly $\sim_m$ is an equivalence relation on X. We write $\widetilde{\Pi}_m$ for the corresponding partition of X. Note that the $\sim_m$-equivalence classes can in some cases be estimated directly from data without having to construct a tree (e.g. for inparalogs, that is, paralogs which all arise from duplication events after a speciation event).

We now show that if δ is a symbolic ultrametric, then the cliques in the graph $G(\delta)$ correspond to pseudo-cherries in the discriminating symbolic representation of δ.

Proposition 4 *Let T be phylogenetic tree on X, $t : V(T) \to M^{\odot}$ be a symbolic dating map, and $\delta = d_{(T;t)}$ be the associated symbolic ultrametric on X. Then:*

(i) $x \sim_m y$ if and only if $N_m[x] = N_m[y]$, for all $\{x, y\} \in \binom{X}{2}$ and all $m \in M$.
(ii) The graph $G(\delta)$ is the disjoint union of its maximal cliques.
(iii) If the map t is discriminating, then a non-empty subset C of X is a maximal clique of $G(\delta)$ if and only if C is a pseudo-cherry of T.

Proof (i) Suppose first that $\{x, y\} \in \binom{X}{2}$ such that $x \sim_m y$ for some $m \in M$. Assume for contradiction that $N_m[x] \neq N_m[y]$, that is, $(N_m(x) \bigtriangleup N_m(y)) \setminus \{x, y\} \neq \emptyset$, in view of Equ. (8). Choose some element z in that set. Then the restriction $T' := T|\{x, y, z\}$ of T to $\{x, y, z\}$ is either the star with leaf set $\{x, y, z\}$ or isomorphic to one of the triples in $\mathscr{R} := \{xy|z, yz|x, xz|y\}$. If T' were the star on $\{x, y, z\}$ then $\mathrm{lca}_T(x, z) = \mathrm{lca}_T(z, y) = \mathrm{lca}_T(x, y)$ would follow. But then $\delta(x, z) = \delta(z, y) = \delta(x, y) = m$ so that $z \in N_m(x) \cap N_m(y)$, contradicting $z \in N_m(x) \bigtriangleup N_m(y) - \{x, y\}$. Thus T' must be isomorphic to one of the triples in $\mathscr{R}$.

If T' were isomorphic to the triple $xy|z$ then $\mathrm{lca}_T(x, z) = \mathrm{lca}_T(y, z)$ and so $\delta(x, z) = \delta(y, z)$. But the choice of z implies that we may assume without loss of generality that $z \in N_m(x) \setminus N_m(y)$, so that $\delta(x, z) = m \neq \delta(y, z)$, a contradiction. If T' were isomorphic to the triple $xz|y$, then $\mathrm{lca}_T(x, y) = \mathrm{lca}_T(z, y)$ and so $\delta(z, y) = \delta(x, y) = m$ would follow as, by assumption, $x \sim_m y$ holds. But then $z \notin N_m(x) \setminus N_m(y)$, a contradiction. By symmetry, T' cannot be isomorphic to the remaining triple $yz|x$ either which yields the final contradiction.

Conversely, suppose that $\{x, y\} \in \binom{X}{2}$ such that $N_m[x] = N_m[y]$, some $m \in M$. We need to show that $t(w) = m$ holds for every interior vertex $w \in V(T)^0$ on the path P from x to y. Assume, for contradiction, that there exists some interior vertex $u \in V(T)^0$ on P with $t(u) \neq m$. Then $u \neq \mathrm{lca}_T(x, y)$ since $t(\mathrm{lca}_T(x, y)) = \delta(x, y) = m$ as, by assumption, $N_m[x] = N_m[y]$. Starting at x and traversing P, let $u' \in V(P)$ and $u'' \in V(P)$ denote the predecessor and successor of u, respectively. Since T is an phylogenetic tree and so has no vertex with in- and outdegree one, there must exist a leaf $z \in V(T)$ such that the path from u to z does not cross the edges $\{u',u\},\{u'',u\} \in E(P)$. Thus $z \notin \{x,y\}$ and either $\operatorname{lca}_T(x,z) = u$ or $\operatorname{lca}_T(y,z) = u$ must hold. By symmetry, we may assume without loss of generality that $\operatorname{lca}_T(x,z) = u$. Then $\delta(x,z) = t(u) \neq m$ and so $z \notin N_m(x)$. By construction of z, we have $\operatorname{lca}_T(y,z) = \operatorname{lca}_T(y,x)$ and so $\delta(y,z) = \delta(y,x) = m$. Hence, $z \in N_m(y)$ and so $z \in N_m(y) \setminus N_m(x) \subseteq N_m(y) \vartriangle N_m(x) = \{x,y\}$. This is a contradiction in view of Equ. (8). Thus, $t(w) = m$ for every interior vertex $w \in V(T)^0$ on P, as required.

(ii) The observation that $G(\delta)$ is the disjoint union of its maximal cliques is a trivial consequence of (i) and the fact that $\sim_m$ is an equivalence relation on X, for all $m \in M$.

(iii) Suppose that t is discriminating. Then the definition of a pseudo-cherry immediately implies that any pseudo-cherry of T must be a maximal clique of $G(\delta)$.

Conversely, assume that C is a maximal clique of $G(\delta)$. Put $v_C := \operatorname{lca}_T(C)$. We show first that every leaf of T adjacent with v_C must be contained in C. To see this, note that if there is a leaf $z \in X - C$ of T adjacent to v_C then $\operatorname{lca}_T(z,x) = v_C$ would hold for all $x \in C$ and, so, $\delta(z,x) = t(v_C)$ would follow for all such x. But then $C \cup \{z\}$ would be a clique in $G(\delta)$ that contains C which is impossible as C is a maximal clique in $G(\delta)$.

Now, for contradiction, assume that there exists some leaf $z \in V(T)$ of T that is contained in C but is not adjacent to v_C. Then, by the definition of v_C, we must have $|C| \geq 2$. Put $m = t(v_C)$ and note that $\delta(x,y) = m$ holds for all $\{x,y\} \in \binom{C}{2}$. Also note that the path P from v_C to z must be of length at least two. Let $w \in V(T)^0$ denote the child of v_C on P. Since t is discriminating, it follows that $t(w) \neq m$. Let $y \in X$ be a leaf of T for which there exists a directed path from w to y and this path does not have an edge in common with the path from w to z. Note that $y \in C$ cannot hold since $\operatorname{lca}_T(z,y) = w$ and, so, $\delta(z,y) = t(w) \neq m$. Thus, $y \in X - C$. Since $y \notin N_m(z)$ and $y \in N_m(x)$ clearly holds for all $x \in C$, we obtain $y \in N_m(x) \Delta N_m(z) = \{x,z\}$ in view of the fact that C is a clique, $x,z \in C$, and Equ. (8), a contradiction. Thus, C is a pseudo-cherry of T. □

We now give an alternative description of the maximal cliques of $G(\delta)$ for δ a symbolic ultrametric, in terms of the graphs $G_m(\delta)$ defined in the last section. To this end, we first describe a general way of constructing a partition from a collection of subsets of a non-empty, finite set. Denote the power set of a non-empty, finite set Y by $\mathscr{P}(Y)$ and assume that Z is a finite, non-empty set. We say that a collection $\mathfrak{A} \in \mathscr{P}(Z)$ is a *cover for* Z if $\bigcup_{A \in \mathfrak{A}} A = Z$ holds. Now, suppose $\mathfrak{A} \in \mathscr{P}(Z)$ is a cover for Z. Then we associate to $\mathfrak{A}$ a collection $\Pi(\mathfrak{A})$ of subsets $B \subseteq Z$ that satisfy the following three conditions:

(P1) there exists some $A \in \mathfrak{A}$ such that $B \subseteq A$,
(P2) there are no two distinct elements $x,y \in B$ such that there exists some $A \in \mathfrak{A}$ with $x \in A$ and $y \notin A$, and
(P3) B is (set-inclusion) maximal with respect to satisfying Property (P2).

The proof of the following lemma is routine.

Lemma 2 *Suppose Z is a non-empty, finite set. If a collection $\mathfrak{A}$ of subsets of Z is a cover for Z then $\Pi(\mathfrak{A})$ is a partition of Z.*

Now, suppose $\delta : X \times X \to M^{\odot}$ is map that satisfies Properties (U0) and (U1). Then, for all $m \in M$, Lemma 2 implies that $\Pi(\mathfrak{C}(G_m(\delta)))$ is a partition of X, since any vertex of a graph must be a vertex in a maximal clique of that graph. For example, consider again the symbolic ultrametric $\delta = d_{(T;t)}$ associated to the pair $(T;t)$ depicted in Fig. 1. Then $\Pi(\mathfrak{C}(G_{m_3}(\delta))) = \{\{b\},\{a,d\},\{c,e\}\}$ and $\Pi(\mathfrak{C}(G_{m_1}(\delta)))$ and $\Pi(\mathfrak{C}(G_{m_2}(\delta)))$ are the partitions that consist of all singletons of $\{a,\dots,e\}$.

We now show that for all $m \in M$ the partition $\widetilde{\Pi}_m$ corresponding to the equivalence relation $\sim_m$ defined above can be given in terms of the cliques of $G_m(\delta)$.

Theorem 4 *Let T be a phylogenetic tree on X and let $t : V(T) \to M^{\odot}$ be a symbolic dating map. Then $\Pi(\mathfrak{C}(G_m(d_{(T;t)}))) = \widetilde{\Pi}_m$ holds for all $m \in M$.*

Proof Suppose $m \in M$ and put $\delta = d_{(T;t)}$ and $\Pi_m = \Pi(\mathfrak{C}(G_m(\delta)))$. Since both Π_m and $\widetilde{\Pi}_m$ are partitions of X it suffices to show that a subset $A \subseteq X$ with $|A| \geq 2$ is an element in Π_m if and only if it is an element in $\widetilde{\Pi}_m$.

Suppose first that $A \in \widetilde{\Pi}_m$. Let $\{x,y\} \in \binom{A}{2}$. Then $x \sim_m y$ and so $t(\mathrm{lca}_T(x,y)) = m$. Thus $\{x,y\} \in E(G_m(\delta))$. Consequently, there must exist a maximal clique $C \in \mathfrak{C}(G_m(\delta))$ such that $x,y \in C$. Without loss of generality, we may assume that C is of minimal size with this property. Since A is a maximal clique in $\mathfrak{C}(\delta)$ it follows that $A \subseteq C$ and that there cannot exist some $C' \in \mathfrak{C}(G_m(\delta))$ and distinct $x,y \in A$ such that $x \in C'$ and $y \notin C'$. But then A satisfies Properties (P1) – (P3) with regards to $\mathfrak{C}(G_m(\delta))$ and so $A \in \Pi_m$ must hold.

Conversely, suppose $A \in \Pi_m$ and assume for contradiction that A is not an equivalence class in $\widetilde{\Pi}_m$. Let $\{x,y\} \in \binom{A}{2}$. Then there must exist some interior vertex v on the path P from x to y in T such that $t(v) \neq m$. Since $A \in \Pi_m$ we cannot have $v = \mathrm{lca}_T(x,y)$. Assume without loss of generality that v lies on the path from $\mathrm{lca}_T(x,y)$ to the leaf x. Also assume without loss of generality that v is such that $t(w) = m$ holds for all interior vertices w on the path P' from v to x. Since T does not have degree two vertices (except possibly the root of T) there must exist a child w of v that is not a vertex of P'. Let $z \in V(T)$ denote a leaf of T such that w lies on the path from v to z. Since X is the vertex set of $G_m(\delta)$ and $t(\mathrm{lca}_T(z,y)) = m \neq t(v) = t(\mathrm{lca}_T(x,z))$ there must exist some $D \in \mathfrak{C}(G_m(\delta))$ such that $z,y \in D$ and $x \notin D$. But this is impossible as $x,y \in A$ and $A \in \Pi_m$. □

As a consequence we now immediately obtain the aforementioned relationship:

Corollary 3 *Suppose $\delta : X \times X \to M^{\odot}$ is a symbolic ultrametric. Then the maximal cliques of $G(\delta)$ are the set-inclusion maximal subsets in $\bigcup_{m\in M} \Pi(\mathfrak{C}(G_m(\delta)))$.*

Proof The statement follows from Proposition 4(ii) and (iii), the fact that a non-empty subset C of X is a pseudo-cherry of T_δ if and only if $A \in \widetilde{\Pi}_m$ holds for some $m \in M$, and Theorem 4. □

6 A Bottom-Up Construction of Symbolic Representations

We have seen in Proposition 2 that the `BUILD` algorithm can be used to determine whether a map is a symbolic ultrametric or not, and if so, constructs its discriminating symbolic representation. `BUILD` can be thought of as a “top-down” algorithm as, in essence, it starts at the root of the tree (if it exists) and ends when it reaches the leaves. In this section, we present an alternative “bottom-up” algorithm, called `BOTTOM-UP`, which will use our clique-based analysis of symbolic representations in the last section. Such an algorithm could provide a potentially useful alternative to `BUILD` as it is based on finding (nearly) maximal cliques in graphs, for which many different algorithms have been developed in the literature.

Suppose that $\delta : X \times X \to M^{\odot}$ is a symbolic ultrametric, and that $(T;t)$ is some symbolic representation of δ. For every maximal clique $C \in \mathfrak{C}(\delta)$ let $x_C \in C$ denote an arbitrary but fixed element in C. Then it is easy to check that the map

$$d'_{(T,t)} : \mathfrak{C}(\delta) \times \mathfrak{C}(\delta) \to M^{\odot},\ d'_{(T;t)}(C,C') = d_{(T;t)}(x_C, x_{C'}), \tag{10}$$

$C, C' \in \mathfrak{C}(\delta)$, is well-defined. A key observation that we shall use in the `BOTTOM-UP` algorithm is that the map $d'_{(T,t)}$ is in fact a symbolic ultrametric on $\mathfrak{C}(\delta)$.

In order to prove this last statement, we shall associate a phylogenetic tree T' on $\mathfrak{C}(\delta)$ plus a symbolic dating map $t' : \mathfrak{C}(\delta) \to M^{\odot}$ for T' as follows. Note that by Proposition 4, every element in $\mathfrak{C}(\delta)$ is a pseudo-cherry of T_δ; we put $v_C = \mathrm{lca}_{T_\delta}(C)$, for all $C \in \mathfrak{C}(\delta)$, and fix some leaf $x_C \in L(T_\delta)$ contained in C. Next, we remove all leaves in $C \setminus \{x_C\}$ from T together with all edges in $\{\{v_C, y\} \in E(T) : y \in C \setminus \{x_C\}\}$. If $v_C \neq \rho_T$ and this process has rendered v_C a degree two vertex then suppress v_C, and if $v_C = \rho_T$ and this process has rendered v_C an outdegree one vertex then identify v_C with its unique leaf. Let $T' = (V', E')$ denote the resulting tree. Then the restriction $t'|_{V'}$ of t to V' is clearly a discriminating symbolic dating map for T'. Moreover, since

$$d'_{(T;t)}(C,C') = d_{(T;t)}(x_C, x_{C'}) = t(\mathrm{lca}_T(x_C, x_{C'})) = t'(\mathrm{lca}_{T'}(C,C')) = d_{(T';t')}(C,C')$$

holds for all $C, C' \in \mathfrak{C}(\delta)$, it follows that $(T';t')$ is the (necessarily unique) discriminating symbolic representation of $d'_{(T;t)}$. Thus, by Theorem 2 we have:

Proposition 5 *Let T be a phylogenetic tree on X and $t : V(T) \to M^{\odot}$ be a symbolic dating map. Then the map $d'_{(T;t)} : \mathfrak{C}(d_{(T;t)}) \times \mathfrak{C}(d_{(T;t)}) \to M^{\odot}$ defined in Equ. 10 is a symbolic ultrametric on $\mathfrak{C}(d_{(T;t)})$.*

We now establish a second result which will be central to the `BOTTOM-UP` algorithm. Given a map $\delta : X \times X \to M^{\odot}$ satisfying (U0)–(U2) we denote the set of connected components of $G(\delta)$ by $\pi(\delta)$ and, for future reference, we let $\pi_2(\delta)$ denote those elements in $\pi(\delta)$ with size at least two.

Lemma 3 *Suppose that $\delta : X \times X \to M^{\odot}$ is a map that satisfies Properties (U0)–(U2), and $K \in \pi_2(\delta)$. Then the following hold:*

(i) If $\{x,y,z\} \in \binom{K}{3}$ is such that x,y,z is a path in K of length two, then $\delta(x,y) = \delta(y,z)$.

(ii) If $\{x,y,z\} \in \binom{K}{3}$ is such that $\{x,y\}$ and $\{y,z\}$ are edges in K, then $\{z,x\}$ must also be an edge in K.
(iii) K is a clique and $\delta(x,y) = \delta(u,v)$ holds for all $\{x,y\},\{u,v\} \in \binom{K}{2}$.

Proof (i) Suppose for contradiction that there exists $\{x,y,z\} \in \binom{K}{3}$ such that x,y,z is a path of length two but $m_1 := \delta(x,y) \neq \delta(y,z) =: m_2$. Then Property (U2) implies $\delta(x,z) \in \{m_1,m_2\}$. Without loss of generality we may assume that $\delta(x,z) = m_1$. Then $z \in N_{m_1}(x)$ and, since $\delta(x,z) \neq m_1$, we also have $z \notin N_{m_1}(y)$. Hence, $z \in N_{m_1}(x) - N_{m_1}(y) \subseteq N_{m_1}(x)\Delta N_{m_1}(y) = \{x,y\}$ since $\{x,y\}$ is an edge in K, a contradiction.

(ii) Suppose for contradiction that there exists $\{x,y,z\} \in \binom{K}{3}$ such that $\{x,y\}$ and $\{y,z\}$ are edges in K but $\{x,z\}$ is not an edge in K. Then Assertion (i) implies that $\delta(x,y) = \delta(y,z) =: m$. We distinguish the cases $\delta(x,z) = m$ and $\delta(x,z) \neq m$.

First, suppose $\delta(x,z) = m$. Then there must exist some $u \in K - \{x,z\}$ such that $u \in N_m(x)\Delta N_m(z)$ as otherwise $\{x,z\}$ would be an edge in K. Without loss of generality, we may assume that $u \in N_m(x) - N_m(z)$. Note that since both $\{x,y\}$ and $\{y,z\}$ are edges in K it follows that $y \in N_m(x) \cap N_m(z)$ and so $u \neq y$. Moreover, since $\{x,y\}$ is an edge in K, $\{x,y\} = N_m(x)\Delta N_m(y)$ must hold, and so $u \in N_m(y)$. Similarly, since $\{y,z\}$ is an edge in K, $u \in N_m(z)$ which is impossible. Thus $\{x,z\}$ must be an edge of K.

Now suppose $\delta(x,z) \neq m$. Then $z \notin N_m(x)$. Since $\{y,z\}$ is an edge in K we have $z \in N_m(y)$ and so $z \in N_m(y) - N_m(x) \subseteq N_m(y)\Delta N_m(x) = \{x,y\}$ as $\{x,y\}$ is an edge of K, a contradiction. Thus $\{x,z\}$ must be also an edge of K in this case.

(iii) This is an immediate consequence of Assertions (i) and (ii). □

We now present the BOTTOM-UP algorithm. The pseudo-code for this algorithm is given in Fig. 5. BOTTOM-UP works in a similar way to the UPGMA algorithm (Sneath and Sokal, 1973) for constructing phylogenetic trees from distance matrices. Essentially BOTTOM-UP works by iteratively looking for pseudo-cherries and, if it finds them, defining a new map on the set of these pseudo-cherries along the lines of Proposition 5.

We now prove a result that is analogous to Proposition 2.

Theorem 5 *Suppose $\delta : X \times X \to M^{\odot}$ is a map. Then the algorithm BOTTOM-UP is a polynomial-time algorithm that either:*

(i) outputs a symbolic discriminating representation for δ if δ is a symbolic ultrametric, or
(ii) the statement "δ is not a symbolic ultrametric"

Proof We first remark that if the input map $\delta : X \times X \to M^{\odot}$ satisfies Properties (U0)–(U2) then, at each execution step of the while loop at Line 3, if Line 5 is not executed then the map δ' defined in Line 12 must also satisfy (U0)–(U2). Moreover, the map t_C defined in Line 8 is well-defined since, in view of Lemma 3, $\delta(C_1,C_2) = \delta(C_3,C_4)$ holds for all $\{C_1,C_2\},\{C_3,C_4\} \in \binom{C}{2}$. In addition, since the set of connected components of a graph can be found in polynomial time and the size of the set F defined in Line 11 decreases by at least one in each execution of the while loop in Line 3 (in case Line 5 is not executed), it follows that the run time for BOTTOM-UP is polynomial in $|X|$.

```
BOTTOM-UP
Input:   Non-empty finite sets $X$ and $M$ with $|X| \geq 3$ and a map $\delta : X \times X \rightarrow M^{\odot}$.
Output:  Discriminating symbolic representation of $\delta$ or the statement "$\delta$ is not a symbolic ultrametric
         on $X$".

1.  If $\delta$ does not satisfy Property (U0), (U1), or (U2) then return the statement "$\delta$ is not a symbolic
    ultrametric on $X$" and stop.
2.  Let $F = \{(T_{\{x\}}, t_{\{x\}}) : x \in X\}$, where, for $x \in X$, $T_{\{x\}}$ is the tree consisting of one vertex $x$ and
    $t_{\{x\}}$ is the map on $V(T_{\{x\}})$ given by putting $t_{\{x\}}(x) = \odot$.
3.  While $|F| \geq 2$ do
4.      Compute the sets $\pi(\delta)$ and $\pi_2(\delta)$.
5.      If $\pi_2(\delta) = \emptyset$, then return the statement "$\delta$ is not a symbolic ultrametric on $X$" and stop.
6.      For all $C \in \pi_2(\delta)$ do
7.          Let $T_C$ be the phylogenetic tree obtained by adding a new vertex $w$ and edges
            $\{w, \rho_{C'}\}$ from $w$ to the root $\rho_{C'}$ of $T_{C'}$, for all $C' \in C$.
8.          Define $t_C : V(T_C) \rightarrow M^{\odot}$ by putting $t_C(w)$ equal to $\delta(C_1, C_2)$ for any $C_1 \neq C_2 \in C$,
            and $t_{C'}(v)$ for any $v \in V(T_{C'})$, $C' \in C$.
9.          Collapse edges of $T_C$ as necessary to ensure that the restriction of $t_C$ to the vertex
            set of the resulting tree is discriminating. Denote the resulting pair also by $(T_C; t_C)$.
10.     end do Line 6.
11.     Let $F = \{(T_C; t_C) : C \in \pi(\delta)\}$, where we identify each singleton set in $\pi(\delta)$ with
        its unique element.
12.     For all $C \in \pi(\delta)$ choose some $x_C \in C$ and define $\delta' : \pi(\delta) \times \pi(\delta) \rightarrow M^{\odot}$ to be the map
        given by setting $\delta'(C_1, C_2) := \delta(x_{C_1}, x_{C_2})$ for all $C_1 \neq C_2 \in \pi(\delta)$.
13.     Let $\delta = \delta'$.
14. end do Line 3.
15. Return the unique element in $F$.
```

Fig. 5 The BOTTOM-UP algorithm.

Now, to complete the proof, given a map $\delta : X \times X \rightarrow M^{\odot}$ we will prove the following claims: (i) if δ is a symbolic ultrametric, then BOTTOM-UP will output a phylogenetic tree T on X and a discriminating symbolic dating map for T, and (ii) if BOTTOM-UP returns a phylogenetic tree T on X and a discriminating symbolic dating map t on T, then $(T;t)$ is a discriminating symbolic representation for δ. This will complete the proof of the theorem in view of Theorem 2.

Proof of (i): Assume δ is a symbolic ultrametric so that, in particular, δ satisfies Properties (U0)–(U2). We first remark that, since $\pi_2(\delta) \neq \emptyset$ (in view of Proposition 4(iii)), Line 6 is not executed at the first execution of the while loop on Line 5. Moreover, as in each execution step of that loop the map δ' defined in Line 12 is a symbolic ultrametric, in view of Proposition 5 we must also have $\pi_2(\delta') \neq \emptyset$.

We now show that BOTTOM-UP returns a pair $(T^{\delta}; t^{\delta})$ where T^{δ} is a phylogenetic tree on X and t^{δ} is a discriminating symbolic dating map for T^{δ}. Note that it suffices to show that at the end of each execution of the while loop in Line 3, every element $(T_C; t_C)$ in the set F defined at Line 11 consists of a phylogenetic tree T_C and a discriminating symbolic dating map t_C for T_C.

To this end, assume that $k \geq 1$ executions of that loop have been carried out, and denote the map computed in Line 12 at execution l by δ_l, for $l = k-1, k$, where we set $\delta_0 := \delta$. Let $C \in \pi(\delta_{k-1})$. If $C \notin \pi_2(\delta_{k-1})$ then, by assumption, T_C and t_C are of the required form, where we identify C with its unique element. So assume that $C \in \pi_2(\delta_{k-1})$. Then, by construction, the tree T_C generated in Line 7 is a phylogenetic tree on $\bigcup_{A\in C} L(T_A)$. Since δ satisfies properties (U0)–(U2), the map t_C defined in Line 8 is well-defined in view of the remark at the beginning of the proof. Now, note that there can be at most one $C' \in C$ such that $t_C(w) = t_C(\rho_{C'})$. If there exists no such element, then t_C is a discriminating symbolic dating map for T_C. Moreover, if such an element C' exists, then the map obtained by restricting t_C to the vertex set of the phylogenetic tree obtained from T_C by collapsing the edge $\{w, \rho_{C'}\}$ is a discriminating symbolic dating map for that tree. Thus the pair $(T_C;t_C)$ in Line 9 is of the required form and so (i) follows.

Proof of (ii): Suppose that δ is an arbitrary map, and that BOTTOM-UP returns a pair $(T^{\delta};t^{\delta})$ with T^{δ} a phylogenetic tree on X and t^{δ} a discriminating symbolic dating map for T^{δ}. Note that in this case, δ must satisfy Properties (U0)–(U2). To show that (ii) holds, it suffices to show that in each execution of the while loop in Line 3 every element $(T_C;t_C)$ in the set F defined in Line 11 is a discriminating symbolic representation of δ restricted to $L(T_C)$.

To this end, assume that $k \geq 1$ executions of the while loop have been carried out and, as before, denote the map defined in Line 12 at execution l by δ_l, $l = k-1, k$ where $\delta_0 := \delta$. Let $C \in \pi(\delta_{k-1})$. If $C \notin \pi_2(\delta_{k-1})$ then, by assumption, $(T_C;t_C)$ is a discriminating symbolic representation for δ restricted to $L(T_C)$, where we identify C with its unique element.

So, assume that $C \in \pi_2(\delta_{k-1})$. Suppose $x, y \in L(T_C)$. Since, by assumption, $(T_{C'}, t_{C'})$ is a discriminating symbolic representation of δ restricted to $L(T_{C'})$, for all $C' \in C$, we may assume without loss of generality that there exist distinct $C_1, C_2 \in C$ such that $x \in L(T_{C_1})$ and $y \in L(T_{C_2})$. Note that the definition of the tree T_C and the map t_C imply that $w = \mathrm{lca}_{T_C}(c_1, c_2)$ holds for all $c_1 \in L(T_{C_1})$ and all $c_2 \in L(T_{C_2})$, and so $\delta(c_1, c_2) = t_C(w)$ for all such c_1 and c_2. But then $d_{(T_C;t_C)}(x,y) = t_C(\mathrm{lca}_{T_C}(x,y)) = t_C(w) = \delta(x,y)$. Thus, again, $(T_C;t_C)$ is a discriminating symbolic representation of δ restricted to $L(T_C)$. This completes the proof of (ii). □

7 Discussion

The case of most immediate practical relevance for the results presented in this paper is the case $|M| = 2$, where the events in M are simply speciation and duplication. Here, we assume that we are given an arbitrary orthology relation $\delta : X \times X \to \{0,1\}^{\odot}$ on a set X of genes (i.e., a map that satisfies (U0) and (U1) and that assigns the value 1 to pairs of genes that are (co-)orthologs and 0 to pairs that are paralogs), a relation that can be reliably estimated from X using various bioinformatics techniques; cf. e.g. (Lechner et al, 2011) and the reference therein. We then aim to obtain a symbolic representation $(T;t)$ of δ, such that $x, y \in X$ are orthologs if $\mathrm{lca}_T(x,y)$ corresponds to a speciation event and paralogs if $\mathrm{lca}_T(x,y)$ corresponds to a duplication event within a single lineage (i.e. $t(\mathrm{lca}_T(x,y))$ equals 1 or 0, respectively).

The above results immediately provide the following characterizations of orthology relations for which a symbolic representation exists:

Corollary 4 *Suppose that* $\delta : X \times X \rightarrow \{0,1\}^{\odot}$ *is an orthology relation. Then the following are equivalent:*

(i) δ *has a symbolic representation.*
(ii) δ *is a symbolic ultrametric.*
(iii) $G_1(\delta) = \overline{G_0(\delta)}$ *is a cograph.*

Somewhat surprisingly, this simple characterization of "ideal" orthology relations does not seem to appear in the literature, even though Falls et al (2008) describes clusters of orthologous genes as Turán graphs, a subclass of cographs. Related methods, which use clustering algorithms to help identify orthologs, have been developed e.g. by Tatusov et al (2000), Li et al (2003), Berglund et al (2008), Wheeler et al (2008) or Lechner et al (2011).

We suspect that Corollary 4 could have far-reaching consequences for the area of orthology detection. In particular, instead of employing clustering techniques, given an arbitrary orthology relation δ, it suggests looking for either a symbolic ultrametric or a cograph that is 'close' to δ, from which a (partially resolved) gene tree could then be constructed. Clearly this is not a trivial endeavor since in practical applications any estimate of δ will be plagued by noise and hence will be neither a symbolic ultrametric nor a cograph.

More specifically, for finding symbolic representations, it could be of interest to try modifying the `BUILD` or `BOTTOM-UP` algorithms to enable them to handle arbitrary orthology relations. For example, ideas behind the `MIN-CUT supertree` algorithm (Semple and Steel, 2000), an algorithm extending `BUILD` which outputs a tree given *any* set of rooted triples, could be explored, as well as related approaches for finding compatible sets of triples that have as many triples as possible in common with a given set of triples, such as those in e.g. Byrka et al (2010). Alternatively, Proposition 4 suggests that heuristics for finding maximum cliques (or subsets that are close to being maximum cliques) in graphs might be useful for modifying the `BOTTOM-UP` algorithm.

For finding cographs there is a large literature that could be useful for analyzing orthology relations. For example, in the cograph editing problem, given a graph $G = (V,E)$ one aims to convert G into a cograph $G^* = (V,E^*)$ such that the number $|E \vartriangle E^*|$ of inserted or deleted edges is minimized. Recently it has been proven that this optimization problem is NP-complete (Liu et al, 2011) which, in view of the above results, implies the following:

Corollary 5 *Let* $\delta : X \times X \rightarrow \{0,1\}^{\odot}$ *be an orthology relation map, and K be a positive integer. Then the problem of deciding if there is a map* $\delta^* : X \times X \rightarrow \{0,1\}^{\odot}$ *such that* $G_1(\delta^*)$ *is a cograph (or, equivalently,* δ^* *a symbolic ultrametric) with* $|E_1(\delta) \vartriangle E_1(\delta^*)| \leq K$ *is NP-complete.*

Even so, it should be noted that the cograph editing problem is fixed parameter tractable (Protti et al, 2009), and so there may be off-the-shelf solutions to help get around this difficulty. Alternatively, efficient ILP approaches might be worth investigating.

Before concluding it is worth mentioning that the general theory developed above for $|M| > 2$ is potentially useful for more refined applications. More specifically, gene duplications have several different mechanistic causes that are also empirically distinguishable in real data sets. Thus it could be of interest, for example, to consider sets M which, as well as representing speciation and duplication events could also take into account events such as local segmental duplications, duplications by retrotransposition, or whole-genome duplications (Zhang, 2003). Moreover, in addition to such events, it might be of interest to consider lineage sorting and horizontal gene transfer, both of which play an important role in genome evolution (Maddison, 1997; Page and Charleston, 1998). From the point of view of gene trees, these behave in a similar manner to speciations, although they introduce incongruencies between the gene and species trees. Hence it might be of interest to investigate whether some of the theory developed in this paper could be extended to phylogenetic networks, graph theoretical structures generalizing phylogenetic trees which are commonly used for modeling horizontal gene transfer (see e.g. (Huson et al, 2010)).

Acknowledgements We thank Andreas W. M. Dress, Hans-Peter Lenhof, and Sonja J. Prohaska for stimulating discussions. This work is supported in part by the *Volkswagen Stiftung* (proj. no. I/82719) and the *Deutsche Forschungsgemeinschaft* (SPP-1174 “Deep Metazoan Phylogeny”, proj. nos. STA 850/2 and STA 850/3).

References

Aho AV, Sagiv Y, Szymanski TG, Ullman JD (1981) Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. SIAM J Comput 10:405–421

Altenhoff AM, Dessimoz C (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput Biol 5:e1000,262

Berglund AC, Sjölund E, Ostlund G, Sonnhammer EL (2008) `InParanoid 6`: eukaryotic ortholog clusters with inparalogs. Nucleic Acids Res 36:D263–D266

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. SIAM Monographs on Discrete Mathematics and Applications, Soc. Ind. Appl. Math., Philadephia

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

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

Datta RS, Meacham C, Samad B, Neyer C, Sjölander K (2009) Berkeley PHOG: Phylofacts orthology group prediction web server. Nucl Acids Res 37:W84–W89

Falls C, Powell B, Snœyink J (2008) Computing high-stringency COGs using Turán-type graphs. Tech. rep., http://www.cs.unc.edu/∼snoeyink/comp145/cogs.pdf

Goodstadt L, Ponting CP (2006) Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. PLoS Comput Biol 2:e133

Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson

N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E (2007) Ensembl 2007. Nucleic Acids Res 35:D610–D617

Huson D, Rupp R, Scornavacca C (2010) Phylogenetic Networks. Cambridge University Press

Kristensen DM, Wolf YI, Mushegian AR, Koonin EV (2011) Computational methods for gene orthology inference. Briefings Bioinf Doi:10.1093/bib/bbr030

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

Li H, Coghlan A, Ruan J, Coin LJ, Hériché JK, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, Wong GK, Zheng W, Dehal P, Wang J, Durbin R (2006) `TreeFam`: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res 34:D572–D580

Li L, Stoeckert CJ Jr, Roos DS (2003) Orthomcl: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189

Liu Y, Wang J, Guo J, Chen J (2011) Cographs editing: Complexity and parametrized algorithms. In: Fu B, Du DZ (eds) COCOON 2011, Springer-Verlag, Berlin, Heidelberg, Lect. Notes Comp. Sci., vol 6842, pp 110–121

Maddison WP (1997) Gene trees in species trees. Syst Biol 46:523–536

Page RDM, Charleston MA (1998) Trees within trees: phylogeny and historical associations. Trends Ecol Evol 13:356–359

Protti F, Dantas da Silva M, Szwarcfiter JL (2009) Applying modular decomposition to parameterized cluster editing problems. Th Computing Syst 44:91–104

Pryszcz LP, Huerta-Cepas J, Gabaldón T (2011) `MetaPhOrs`: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. Nucleic Acids Res 39:e32

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

Semple C, Steel M (2000) A supertree method for rooted trees. Discrete Applied Mathematics 105:147–158

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

Sneath P, Sokal R (1973) Numerical Taxonomy. W.H. Freeman and Company, San Francisco, pp. 230–234

Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28:33–36

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Ostell J, Pruitt KD, Schuler GD, Shumway M, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E (2008) Database resources of the national center for biotechnology information. Nucleic Acids Res 36:D13–D21

Zhang J (2003) Evolution by gene duplication: an update. Trends Ecol Evol 18:292–298