Discrete Applied Mathematics 267 (2019) 1–11

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

Generalized Fitch graphs: Edge-labeled graphs that are explained by edge-labeled trees

Marc Hellmuth *

Department of Mathematics and Computer Science, University of Greifswald, Walther-Rathenau-Straße 47, D-17487 Greifswald, Germany
Center for Bioinformatics, Saarland University, Building E 2.1, P.O. Box 151150, D-66041 Saarbrücken, Germany

ARTICLE INFO

Article history:
Received 7 March 2018
Received in revised form 11 June 2019
Accepted 12 June 2019
Available online 27 June 2019

Keywords:
Labeled trees
Forbidden subgraphs
Phylogenetics
Xenology
Fitch graph
Recognition algorithm

ABSTRACT

Fitch graphs $G = (X, E)$ are di-graphs that are explained by $\{\otimes, 1\}$-edge-labeled rooted trees with leaf set X: there is an arc $xy \in E$ if and only if the unique path in T that connects the least common ancestor $\mathrm{lca}(x, y)$ of x and y with y contains at least one edge with label "1". In practice, Fitch graphs represent xenology relations, i.e., pairs of genes x and y for which a horizontal gene transfer happened along the path from $\mathrm{lca}(x, y)$ to y.

In this contribution, we generalize the concept of Fitch graphs and consider complete di-graphs $K_{|X|}$ with vertex set X and a map ε that assigns to each arc xy a unique label $\varepsilon(x, y) \in M \cup \{\otimes\}$, where M denotes an arbitrary set of symbols. A di-graph $(K_{|X|}, \varepsilon)$ is a generalized Fitch graph if there is an $M \cup \{\otimes\}$-edge-labeled tree (T, λ) that can explain $(K_{|X|}, \varepsilon)$.

We provide a simple characterization of generalized Fitch graphs $(K_{|X|}, \varepsilon)$ and give an $O(|X|^2)$-time algorithm for their recognition as well as for the reconstruction of the unique least-resolved phylogenetic tree that explains $(K_{|X|}, \varepsilon)$.

1. Introduction

Edge-labeled graphs that can be explained by *vertex-labeled* trees have been widely studied and range from cographs [8,27] and di-cographs [9] to so-called unp-2-structures [16–18], symbolic ultrametrics [3,30] or three-way symbolic tree-maps [25,33]. Besides their structural attractiveness, those types of graphs play an important role in phylogenomics, i.e., the reconstruction of the evolutionary history of genes and species. By way of example, the concept of *orthologs*, that is, pairs of genes from different species that arose from a speciation event [19], is of fundamental importance in many fields of mathematical and computational biology, including the reconstruction of evolutionary relationships across species [12,32] or functional genomics and gene organization in species [21,45]. The orthology relation Θ is explained by vertex-labeled trees, i.e., a gene pair (x, y) is contained in Θ if and only if the least common ancestor of x and y is labeled as a speciation event. The graph representation of Θ must necessarily be a co-graph [3,27] and provides direct information on the gene history as well as on the history of the species [31,32].

In contrast, *xenology* as defined by Walter M. Fitch [20] is explained by *edge-labeled* rooted phylogenetic trees: a gene y is xenologous with respect to x, if and only if the unique path from the least common ancestor $\mathrm{lca}(x, y)$ to y in the

* Correspondence to: Department of Mathematics and Computer Science, University of Greifswald, Walther-Rathenau-Straße 47, D-17487 Greifswald, Germany.
E-mail address: mhellmuth@mailbox.org.

https://doi.org/10.1016/j.dam.2019.06.015
0166-218X/

gene tree contains a transfer edge. In other words, the xenology relation is explained by an $\{\otimes, 1\}$-edge-labeled rooted tree, where an edge with label "1" is a transfer edge and an edge with label "$\otimes$" is a non-transfer edge. It has been shown by Geiß et al. [22] that the xenology relation forms a *Fitch graph*, that is, an $\{\otimes, 1\}$-edge-labeled di-graph which is characterized by the absence of eight forbidden subgraphs on three vertices. Moreover, for a given Fitch graph $\mathcal{F}$ it is possible to reconstruct the unique minimally resolved phylogenetic tree that explains $\mathcal{F}$ in linear time.

A further example of graphs and relations that are defined in terms of edge-labeled trees are the single-1-relations $\overset{1}{\sim}$ and $\overset{1}{\rightharpoonup}$ [28]. These relations are defined by the existence of a *single* edge with label "1" along the connecting path of two genes and capture the structure of so-called rare genomic changes (RGCs). RGCs have been proven to be phylogenetically informative and helped to resolve many of the phylogenetic questions where sequence data lead to conflicting or equivocal results, see e.g. [4,10,14,15,35,36,38,39,43].

In summary, edge-labeled graphs (or equivalently, binary relations) that can be explained by *edge-labeled* trees provide important information about the evolutionary history of the underlying genes. However, for such type of graphs only few results are available [22,23,28,29].

In this contribution, we extend the notion of xenology and Fitch graphs to *generalized* Fitch graphs, that is, di-graphs that can be derived from $\{\otimes, 1, \ldots, m\}$-edge labeled trees, or equivalently, edge-labeled di-graphs that can be explained by such trees. We show that generalized Fitch graphs are characterized by four simple conditions that are defined in terms of edge-disjoint subgraphs. Moreover, we give an $O(|X|^2)$-time recognition algorithm for generalized Fitch graphs on a set of vertices X and the reconstruction of the *unique* least-resolved phylogenetic tree that explains them. In practice, general Fitch graphs may be used to model different "types" of HGT, e.g. transformation and conjugation as it happens in bacteria [7,13,24,34]; HGTs that are preceded by the different types of gene duplications that are also empirically distinguishable in real data sets [46]; HGTs represented by different degrees of certainty indicating if estimated transfer events between genes are "reliable" or not [37]; or to distinguish pairs of genes between only HGT and HGT with additional RGCs have been taken place.

2. Preliminaries

2.1. Trees, di-graphs and sets

A *rooted tree* $T = (V, E)$ (on X) is an acyclic connected graph with leaf set X, set of *inner* vertices $V^0 = V \setminus X$ and one distinguished inner vertex $\rho_T \in V^0$ that is called the *root of* T. In what follows, we consider always *phylogenetic* trees T, that is, rooted trees such that the root ρ_T has at least degree 2 and every other inner vertex $v \in V^0 \setminus \{\rho_T\}$ has at least degree 3.

We call $u \in V$ an *ancestor* of $v \in V$, $u \succeq_T v$, and v a *descendant* of u, $v \preceq_T u$, if u lies on the unique path from ρ_T to v. We write $v \prec_T u$ ($u \succ_T v$) for $v \preceq_T u$ ($u \succeq_T v$) and $u \neq v$. Edges that are incident to a leaf are called *outer edges*. Conversely, *inner edges* do only contain inner vertices. For a non-empty subset $Y \subseteq X$ of leaves, *the least common ancestor of* Y, denoted as $\mathrm{lca}_T(Y)$, is the unique $\preceq_T$-minimal vertex of T that is an ancestor of every vertex in Y. We will make use of the simplified notation $\mathrm{lca}_T(x, y) := \mathrm{lca}_T(\{x, y\})$ for $Y = \{x, y\}$ and we will omit the explicit reference to T whenever it is clear which tree is considered. For a subset $Y \subseteq X$ of leaves, the tree $T(Y)$ with root $\mathrm{lca}_T(Y)$ has leaf set Y and consists of all paths in T that connect the leaves in Y. The *restriction* $T_{|Y}$ of T to some subset $Y \subseteq X$ is the rooted tree obtained from $T(Y)$ by suppressing all vertices of degree 2 with the exception of the root ρ_T if $\rho_T \in V(T(Y))$.

A *contraction* of an edge $e = xy$ in a tree T refers to the removal of e and identification of x and y. We say that a rooted tree T on L *displays* a rooted tree T' on L', in symbols $T' \preceq T$, if T' can be obtained from $T(L')$ by a sequence of edge contractions. If $T' \preceq T$, then we also say that T *refines* T'.

Rooted triples are binary rooted phylogenetic trees on three leaves. We write $ab|c$ for the rooted triple with leaves a, b and c, if the path from its root to c does not intersect the path from a to b. The definition of "display" implies that a triple $ab|c$ with $a, b, c \in L$ is *displayed* by a rooted tree T if $\mathrm{lca}(a, b) \prec_T \mathrm{lca}(a, b, c)$. The set of all triples that are displayed by T is denoted by $r(T)$. A set of rooted triples R is called *consistent* if there exists a phylogenetic tree T on $L_R := \bigcup_{ab|c \in R}\{a, b, c\}$ that displays R, i.e., $R \subseteq r(T)$. As shown in [1] there is a polynomial-time algorithm, usually referred to as BUILD [42,44], that takes a set R of triples as input and either returns a particular phylogenetic tree $\mathrm{Aho}(R)$ that displays R, or recognizes R as inconsistent.

A set of rooted triples R *identifies* a tree T with leaf set L_R if R is displayed by T and every other tree T' that displays R is a refinement of T. A rooted triple $ab|c \in r(T)$ *distinguishes* an edge uv in T iff a, b, and c are descendants of u; v is an ancestor of a and b but not of c; and there is no descendant v' of v for which a and b are both descendants. In other words, $ab|c \in r(T)$ distinguishes the edge uv if $\mathrm{lca}(a, b) = v$ and $\mathrm{lca}(a, b, c) = u$.

The requirement that a set R of triples is consistent, and thus, that there is a tree displaying all triples, makes it possible to infer new triples from the trees that display R and to define a *closure operation* for R [5,6,26,40]. Let $\langle R\rangle$ be the set of all rooted trees with leaf set L_R that display R. The closure of a consistent set of rooted triples R is defined as

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

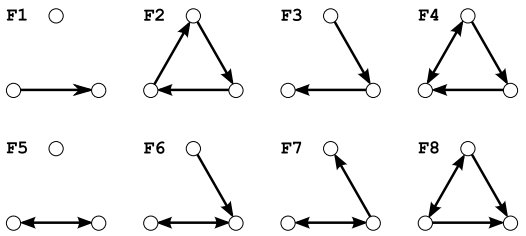

Fig. 1. Shown are the eight forbidden induced subgraphs $F_1, \ldots, F_8$ of Fitch graphs.

Hence, a triple r is contained in the closure $\mathrm{cl}(R)$ if all trees that display R also display r. This operation satisfies the usual three properties of a closure operator [6], namely: (i) expansiveness, $R \subseteq \mathrm{cl}(R)$; (ii) isotony, $R' \subseteq R$ implies that $\mathrm{cl}(R') \subseteq \mathrm{cl}(R)$; and (iii) idempotency, $\mathrm{cl}(\mathrm{cl}(R)) = \mathrm{cl}(R)$. Since $T \in \langle r(T) \rangle$, it is easy to see that $\mathrm{cl}(r(T)) = r(T)$ and thus, $r(T)$ is always closed.

For later reference, we give here an important result from [26] that is closely related to the BUILD algorithm.

Lemma 2.1. *Let T be a phylogenetic tree and let R be a set of rooted triples. Then, R identifies T if and only if $\mathrm{cl}(R) = r(T)$. Moreover, if R identifies T, then $\mathrm{Aho}(R) = T$.*

In this contribution, we will consider phylogenetic trees $T = (V, E)$ together with an edge-labeling map $\lambda : E \to M \cup \{\otimes\}$, where $M = \{1, \ldots |M|\}$ denotes a non-empty set of symbols and we write (T, λ). Edges that have label $m \in M \cup \{\otimes\}$ are called *m-edges*. Furthermore, $M^{\otimes}$ will always denote the set $M \cup \{\otimes\}$.

For a di-graph $G = (V, E)$ and a subset $W \subseteq V$ we denote with $G[W] = (W, F)$ the *induced subgraph* of G, i.e., any arc $xy \in E$ with $x, y \in W$ is also contained in $G[W]$.

In what follows, $[X \times X]_{\mathrm{irr}}$ denotes the set $(X \times X) \setminus \{(x, x) \mid x \in X\}$. To avoid trivial cases, we always assume that $|X| > 1$. The sets $X_1, \ldots, X_k$ form a *quasi-partition* of X, if all sets are pairwisely disjoint, their union is X and at most one X_i is empty.

2.2. Simple Fitch graphs

Let $\lambda : E \to \{1, \otimes\}$ be a map and (T, λ) be an edge-labeled phylogenetic tree on X. We set $(x, y) \in \mathcal{X}_{(T,\lambda)}$ for $x, y \in X$ whenever the uniquely defined path from $\mathrm{lca}_T(x, y)$ to y contains at least one 1-edge. By construction $\mathcal{X}_{(T,\lambda)}$ is irreflexive; hence it can be regarded as a simple directed graph.

An arbitrary di-graph $G = (X, E)$ is *explained* by a phylogenetic tree (T, λ) (on X) and called *simple Fitch graph*, whenever $xy \in E$ if and only if $(x, y) \in \mathcal{X}_{(T,\lambda)}$. Fitch graphs are the topic of Ref. [22,29], which among other results gave a characterization in terms of eight forbidden induced subgraphs. The following theorem summarizes a couple of important results that we need for later reference.

Theorem 2.2 (*[22]*)**.** *A given di-graph $G = (X, E)$ is a simple Fitch graph if and only if it does not contain one of the graphs $F_1, \ldots, F_8$ (shown in Fig. 1) as an induced subgraph.*

Deciding whether G is a simple Fitch graph and, in the positive case, to construct the unique least-resolved tree (T, λ) that explains G can be done in $O(|X| + |E|)$ time.

(T, λ) is a least-resolved tree that explains G, i.e., there is no edge-contracted version T' of T and no labeling λ' such that (T', λ') still explains G, if and only if all its inner edges are 1-edges and for every inner edge uv with $u \succ_T v$ there is an outer $\otimes$-edge vx in (T, λ).

3. Generalized Fitch graphs

To generalize the notion of simple Fitch graphs, we consider complete di-graphs $(K_{|X|}, \varepsilon)$ with vertex X, arc set $[X \times X]_{\mathrm{irr}}$ and a map $\varepsilon : [X \times X]_{\mathrm{irr}} \to M^{\otimes}$ that assigns to each arc xy a unique label $\varepsilon(x, y)$. Clearly, the map ε covers all information provided by $(K_{|X|}, \varepsilon)$. W.l.o.g. we will always assume that for each $m \in M$ there is at least one pair $(x, y) \in [X \times X]_{\mathrm{irr}}$ such that $\varepsilon(x, y) = m$.

Definition 3.1. Let $\varepsilon : [X \times X]_{\mathrm{irr}} \to M^{\otimes}$ be a map. For a given phylogenetic tree (T, λ) with $\lambda : E \to M^{\otimes}$ and two leaves x and y we denote with $\mathbb{P}_{(x,y)}$ the unique path in T from $\mathrm{lca}_T(x, y)$ to y. A pair $(x, y) \in [X \times X]_{\mathrm{irr}}$ is *explained by* a phylogenetic tree (T, λ) on X whenever,

$\varepsilon(x, y) = m \in M$ iff some edge e on the path $\mathbb{P}_{(x,y)}$ has label $\lambda(e) = m$; and

$\varepsilon(x, y) = \otimes$ iff none of the edges e on the path $\mathbb{P}_{(x,y)}$ have label $\lambda(e) \in M$.

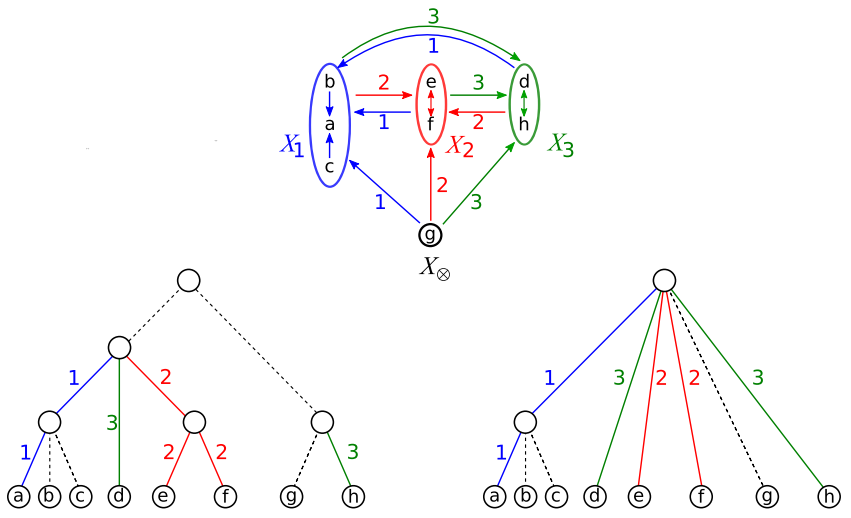

Fig. 2. Lower-left, an edge-labeled tree (T, λ) is shown where $\otimes$-edges are drawn as dashed-lines. The tree (T, λ) explains the graph $(K_{|X|}, \varepsilon)$ in the upper part, where $X = \{a, b, \ldots, h\}$ and $\varepsilon : [X \times X]_{\text{irr}} \to \{1, 2, 3, \otimes\}$. For better readability, $\otimes$-edges are omitted in the drawing of $(K_{|X|}, \varepsilon)$ and only one arc xy with label $\varepsilon(x, y) = m'$ and $\varepsilon(y, x) = m$ for each $x \in X_m$ and $y \in X_{m'}$, $m \neq m'$ is drawn. All arcs between vertices $x, y \in X_m$ have label m. The unique least-resolved tree (T^*, λ^*) constructed with Algorithm 1 is shown at the lower-right part.

The map ε is *tree-like* if each pair $(x, y) \in [X \times X]_{\text{irr}}$ is explained by (T, λ). In this case, we say that (T, λ) *explains* ε and $(K_{|X|}, \varepsilon)$ is a *(generalized) Fitch graph*.

Moreover, a tree (T, λ) is *least-resolved* for a map ε, if (T, λ) explains ε and there is no tree (T', λ') that explains ε, where T' is obtained from T by contracting edges and λ' is an $M \cup \{\otimes\}$-edge-labeling map for T'.

Fig. 2 shows an example of a generalized Fitch graph $(K_{|X|}, \varepsilon)$. We give the following almost trivial result for later reference.

Lemma 3.1. *Let $\varepsilon : [X \times X]_{irr} \to M^\otimes$ be tree-like and (T, λ) be a tree that explains ε. If there is an edge e with $\lambda(e) = m$ on the path P from the root ρ_T to some leaf, then all edges in P are either labeled with m or $\otimes$.*

Proof. Let P be the path from the root ρ_T to the leaf $x \in X$. Let v be the child of ρ_T that is an ancestor of x. Now let $y \in X$ be any leaf that is not a descendant of v and thus $\text{lca}_T(x, y) = \rho_T$. Assume, for contradiction, that there are two edges in P with distinct labels $m, m' \in M$. Since (T, λ) explains ε we would have $\varepsilon(y, x) = m$ and $\varepsilon(y, x) = m'$; a contradiction to ε being a map. □

For each symbol $s \in M^\otimes$ we define the following set

$$X_s := \{x \in X \mid \text{there is a vertex } z \in X \text{ with } \varepsilon(z, x) = s \text{ and for all } z' \in X \setminus \{z, x\} \text{ we have } \varepsilon(z', x) \in \{\otimes, s\}\}$$

that contains for each symbol s those vertices $x \in X$ where at least one incoming arc is labeled s and all other incoming arcs have label s or $\otimes$. Note, by construction for all $x, y \in X_\otimes$ we have $\varepsilon(x, y) = \varepsilon(y, x) = \otimes$ and for all $x, y \in X_m$, $m \in M$ we have $\varepsilon(x, y), \varepsilon(y, x) \in \{m, \otimes\}$.

The intuition behind the sets X_s is sketched in Fig. 3. In this example, let $(K_{|X|}, \varepsilon)$ be the Fitch graph that is explained by the sketched tree and assume that the highlighted m-edge e with $m \neq \otimes$ is the first m-edge that lies on the path from the root to any of the leaves that are located below this edge. Lemma 3.1 implies that all edges on this path that are above e must be $\otimes$-edges and all edges below e must either be $\otimes$- or m-edges. This observation implies that every leaf z located in the subtree T' must "point to" to every leaf $x \in X'_m$ via an m-edge in $(K_{|X|}, \varepsilon)$, i.e., $\varepsilon(z, x) = m$. Moreover, for any two vertices $z', x \in X'_m$ we have $\varepsilon(z', x) \in \{\otimes, m\}$. Thus, the set X'_m in Fig. 3 is a subset of the (possibly larger) set X_m.

Lemma 3.2. *Let $\varepsilon : [X \times X]_{irr} \to M^\otimes$ be a tree-like map and (T, λ) a tree that explains ε. Then, for all $m \in M$ we have $X_m \neq \emptyset$ and $X_m = \{x \in X \mid \exists z \in X \text{ with } \varepsilon(z, x) = m\}$. In particular, the sets $X_1, X_2, \ldots X_{|M|}, X_\otimes$ form a quasi-partition of X.*

Moreover, for all $x \in X_m$ and $y \in X \setminus X_m$ with $m \in M^\otimes$ it holds that $\varepsilon(y, x) = m$.

Proof. To recap, ε is a map such that $\varepsilon^{-1}(m) \neq \emptyset$ for all $m \in M$. Thus, for each $m \in M$ there are two vertices $x, z \in X$ with $\varepsilon(z, x) = m$. Assume for contradiction that there is a vertex $z' \in X$ with $\varepsilon(z', x) = m' \notin \{m, \otimes\}$. Thus, the path from $\text{lca}(z, x)$ to x contains an edge labeled m and the path from $\text{lca}(z', x)$ to x contains an edge labeled m'. However, since both vertices

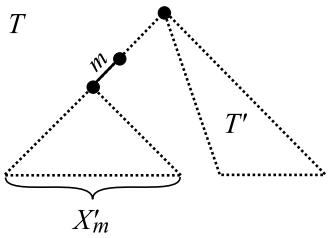

Fig. 3. Shown is a sketched tree T, where the highlighted m-edge e is the first m-edge that is located on the path from the root to any of the leaves below e. Thus, $X'_m \subseteq X_m$, see text for further details.

$\mathrm{lca}(x, z)$ and $\mathrm{lca}(x, z')$ are located on the path from the root of T to x, this path must have two edges, one with label m and one with label m'; a contradiction to Lemma 3.1. Thus, $\varepsilon(z, x) \in \{m, \otimes\}$ for all $z \in X \setminus \{x\}$ and therefore, $x \in X_m$. Thus, $X_m \neq \emptyset$ for all $m \in M$. In particular, the latter arguments imply that whenever there are vertices $x, z \in X$ with $\varepsilon(z, x) = m$, then for all vertices $z' \in X \setminus \{x\}$ we have $\varepsilon(z', x) \in \{m, \otimes\}$ and thus, the sets X_m and $\{x \in X \mid \exists z \in X \text{ with } \varepsilon(z, x) = m\}$ are identical.

We continue to show that $X_1, X_2, \ldots X_{|M|}, X_\otimes$ form a quasi-partition of X. Clearly, for all distinct $m, m' \in M$ the sets $X_m, X_{m'}$ must be disjoint, as otherwise $x \in X_m \cap X_{m'}$ would imply that $\varepsilon(z, x) = m$ for some $z \in X$ and at the same time $\varepsilon(z, x) \in \{m', \otimes\}$; a contradiction to ε being a map. Moreover, for all distinct $m \in M$ the sets $X_m, X_\otimes$ must be disjoint, since $x \in X_\otimes$ if and only if $\varepsilon(z, x) = \otimes$ for all $z \in X \setminus \{x\}$, which is, if and only if $x \notin X_m$ for all $m \in M$.

It remains to show that the union of $X_1, X_2, \ldots X_{|M|}, X_\otimes$ is X and at most one of the sets is empty. Note, for each $m \in M$ there are two vertices $z, x \in X$ with $\varepsilon(z, x) = m$. As argued above, $\varepsilon(z, x) = m \in M$ implies $x \in X_m$. Thus, none of the sets $X_1, X_2, \ldots X_{|M|}$ is empty. In particular, $X_\otimes = \emptyset$ if and only if for every $x \in X$ we have $\varepsilon(z, x) = m$ for some $z \in X$ and $m \in M$. In this case, the union of the sets $X_1, X_2, \ldots X_{|M|}$ is X. Now assume that $X_\otimes \neq \emptyset$ and $x \notin X_m$, $m \in M$. Hence, $\varepsilon(z, x) \neq m$ for all $z \in X \setminus \{x\}$ and all $m \in M$. Thus, $\varepsilon(z, x) = \otimes$ for all $z \in X \setminus \{x\}$, and therefore, $x \in X_\otimes$. Thus, in case $X_\otimes \neq \emptyset$, the union of the sets $X_1, X_2, \ldots X_{|M|}, X_\otimes$ is X.

To prove the last statement, let $x \in X_m$. Clearly, if $m = \otimes$ and thus, $x \in X_\otimes$ then $\varepsilon(y, x) = \otimes$ for all $y \in X \setminus \{x\}$. Now, let $m \in M$ and $m' \in M^\otimes$ with $m \neq m'$. Assume for contradiction that $\varepsilon(y, x) \neq m$ for some $y \in X_{m'}$. Thus, the path from $\mathrm{lca}(x, y)$ to x does not contain an m-edge. By construction of X_m, there is a vertex $z \in X$ with $\varepsilon(z, x) = m$ and thus, the path from $\mathrm{lca}(x, z)$ to x contains an m-edge $e = uv$. Trivially, all ancestors of x are located on the path from the root of T to x and thus, also $\mathrm{lca}(x, z)$ and $\mathrm{lca}(x, y)$. Therefore, the m-edge is located between $\mathrm{lca}(x, z)$ and $\mathrm{lca}(x, y)$ and, in particular, $\mathrm{lca}(x, z) \succeq u \succ v \succeq \mathrm{lca}(x, y)$. Hence, $\mathrm{lca}(x, z) = \mathrm{lca}(y, z)$ and the path from $\mathrm{lca}(y, z)$ to y contains an m-edge; a contradiction to $y \in X_{m'}$. □

For each $m \in M$, we denote with G_m the subgraph of $(K_{|X|}, \varepsilon)$ with vertex set X_m as defined above and arc set

$$E_m = \{xy \mid x, y \in X_m, \varepsilon(x, y) \neq \otimes\}.$$

Note, by definition of X_m, the graph G_m contains only arcs xy with $\varepsilon(x, y) = m$.

Before we can derive the final result, we need one further definition. Let (T, λ) be an edge-labeled phylogenetic tree on X. To recap, the restriction $T_{|X_m}$ of T to X_m is obtained by suppressing all degree-2 vertices of $T(X_m)$. For any edge $uv \in E(T_{|X_m})$, let $S(u, v)$ denote the set of all suppressed vertices on the path from u to v in $T(X_m)$. We define the restriction $\lambda_{|X_m}$ to X_m by putting for all edges $uv \in E(T_{|X_m})$:

$$\lambda_{|X_m}(u, v) = \begin{cases} \lambda(u, v) & \text{, if } S(u, v) = \emptyset \text{ and thus, } uv \in E(T) \\ m & \text{, else if there are } a, b \in S(u, v) \cup \{u, v\} \text{ with } \lambda(a, b) = m \\ \otimes & \text{, else.} \end{cases}$$

Lemma 3.1 implies that the restriction $\lambda_{|X_m}$ of λ is well defined. In particular, $\lambda_{|X_m}(u, v) = m$ if and only if the corresponding unique path between u and v in T contains an m-edge. To characterize tree-like maps ε, we provide Algorithm 1. This algorithm takes as input a tree-like map ε and reconstructs a (least-resolved) tree that explains ε. The steps of this algorithm are explained in more detail in the proof of Theorem 3.3 and are used to show that the conditions *(T1)* to *(T4)* in Theorem 3.3 are sufficient for tree-like maps ε.

We are now in the position to characterize tree-like maps ε.

Theorem 3.3. *A map $\varepsilon : [X \times X]_{irr} \to M^\otimes$ is tree-like (or equivalently $(K_{|X|}, \varepsilon)$ is a generalized Fitch graph) if and only if the following four conditions are satisfied:*

(T1) *The sets $X_1, X_2, \ldots X_{|M|}, X_\otimes$ form a quasi-partition of X.*

(T2) *$G_m = (X_m, E_m)$ is a simple Fitch graph for all $m \in M$.*

Algorithm 1 `Compute Least-Resolved Tree` (T^*, λ^*) `for` ε.

Input: A tree-like map $\varepsilon : [X \times X]_{\text{irr}} \to M^{\otimes}$;
Output: A least-resolved edge-labeled tree (T^*, λ^*) that explains ε;
1: Add a root ρ_{T^*} to T^*;
2: **for** all $m \in M$ **do**
3: Compute the least-resolved tree (T_m, λ_m) the explains $G_m = (X_m, E_m)$;
4: **if** $|X_m| = 1$ *OR* (T_m, λ_m) contains an $\otimes$-edge incident to its root **then**
5: Add a vertex r_m and the edge $\rho_{T^*} r_m$ with label m;
6: Add (T_m, λ_m) by identifying the root of T_m with r_m;
7: Set $\lambda^*(e) = \lambda_m(e)$ for all edges in T_m;
8: **else** Identify the root of T_m with ρ_{T^*} and add (T_m, λ_m);
9: Add an edge $e = \rho_{T^*} x$ with label $\lambda^*(e) = \otimes$ for all $x \in X_{\otimes}$;
10: Return (T^*, λ^*);

(T3) *For all $m \in M$ and $x \in X_m$, $y \in X \setminus X_m$ it holds that $\varepsilon(y, x) = m$.*

(T4) *For all $x \in X_{\otimes}$ and $y \in X \setminus \{x\}$ it holds that $\varepsilon(y, x) = \otimes$.*

In particular, the tree (T^*, λ^*) returned by Algorithm 1 (with input ε) explains ε, whenever ε is tree-like.

Proof. We first establish the 'if' direction. Assume Conditions *(T1)* to *(T4)* are satisfied for ε. Since $G_m = (X_m, E_m)$ is a simple Fitch graph for all $m \in M$, all G_m are explained by a tree (T_m, λ_m) with leaf set X_m.

We show that the tree (T^*, λ^*) constructed with Algorithm 1 explains ε. By construction of (T^*, λ^*) all trees (T_m, λ_m) are exactly the subtrees $T^*(X_m)$ where all edge labels λ_m are kept. Hence, G_m is explained by $(T^*(X_m), \lambda^*_{|X_M})$. Since $\varepsilon(x, y) = m$ (resp. $\varepsilon(x, y) = \otimes$) for any $x, y \in X_m$ if and only if $xy \in E_m$ (resp. $xy \notin E_m$), we can conclude that all pairs $(x, y), (y, x)$ with $x, y \in X_m$ are explained by (T^*, λ^*) for all $m \in M$. Moreover, each $x \in X_{\otimes}$ is linked to the root ρ_{T^*} via an $\otimes$-edge (Line 9 of Algorithm 1). Hence, for each two vertices $x, y \in X_{\otimes}$ we have, by definition of $X_{\otimes}$, $\varepsilon(x, y) = \varepsilon(y, x) = \otimes$, which is trivially explained by (T^*, λ^*). Since the sets $X_1, X_2, \ldots X_{|M|}, X_{\otimes}$ form a quasi-partition of X, it is ensured that there are no overlapping leaf sets when the trees (T_m, λ_m) and the elements $x \in X_{\otimes}$ have been added to (T^*, λ^*) and that the leaf set of T^* is X.

We continue to show that all pairs (x, y) with $x, y \in X$ that satisfy *(T3)* and *(T4)* are explained by (T^*, λ^*). Note first, by construction of (T^*, λ^*) and since (T^*, λ^*) explains G_m for all $m \in M$, all edges along the path from ρ_{T^*} to $x \in X_m$ have label m or $\otimes$. Even more, we show that each path $P_{\rho_{T^*}, x}$ from ρ_{T^*} to each $x \in X_m$, $m \in M$ has always an edge with label m. By construction of (T^*, λ^*) (Algorithm 1, Line 4–6), if $|X_m| = 1$ or there is a leaf $x \in X_m$ adjacent to the root ρ_m of T_m such that $\lambda_m(\rho_m, x) = \otimes$, then the tree (T_m, λ_m) is placed below the particular m-edge $\rho_{T^*} r_m$. Hence, all paths from ρ_{T^*} to $x \in X_m$ contain this m-edge. Since *(T3)* and *(T4)* state that $\varepsilon(y, x) = m$ for all $x \in X_m$, $m \in M$, all pairs (y, x) with $x \in X_m$, $y \in X \setminus X_m$ are explained by (T^*, λ^*), given that (T_m, λ_m) satisfies the Conditions in Algorithm 1 (Line 4). Assume that (T_m, λ_m) does not satisfy the latter conditions. Theorem 2.2 implies that all inner edges of (T_m, λ_m) are m-edges and thus, any $\otimes$-edge in (T_m, λ_m) must be incident to some leaf $x \in X_m$. Since (T_m, λ_m) does not satisfy the if-condition in Line 4 of Algorithm 1, *all* edges that are incident to the root of (T_m, λ_m) have label m. Hence, all paths from ρ_{T^*} to $x \in X_m$ contain an m-edge and, therefore, all pairs (y, x) with $x \in X_m$, $y \in X \setminus X_m$ are explained by (T^*, λ^*). Finally, if $x \in X_{\otimes}$, then *(T4)* claims $\varepsilon(y, x) = \otimes$ for all $y \neq x$ which is trivially explained by (T^*, λ^*), since x is linked to the root ρ_{T^*} via an $\otimes$-edge (Algorithm 1, Line 9). In summary, if the Conditions *(T1)* to *(T4)* are satisfied, then ε is explained by (T^*, λ^*) and therefore, tree-like. This establishes the 'if' direction.

We turn now to the 'only if' direction. Assume that ε is tree-like and let (T, λ) be a tree that explains ε with root ρ_T. Lemma 3.2 implies Conditions *(T1)*, *(T3)* and *(T4)*. We continue to show *(T2)*. To this end, consider the graph $G_m = (X_m, E_m)$, $m \in M$. Since (T, λ) explains ε and therefore, also $(K_{|X|}, \varepsilon)$, it must explain each of its induced subgraphs and thus, any pair (x, y) with $x, y \in X_m$ and $m \in M$ is explained by (T, λ). By construction of the restriction $(T_{|X_m}, \lambda_{|X_m})$ of (T, λ) to X_m we have $\varepsilon(x, y) = m$ if and only if the path in (T, λ) from $\text{lca}_T(x, y)$ to y contains an m-edge which is if and only if there is an m-edge on the path from $\text{lca}_{T_{|X_m}}(x, y)$ to the leaf y in $(T_{|X_m}, \lambda_{|X_m})$. Hence, $(T_{|X_m}, \lambda_{|X_m})$ explains $(K_{|X|}[X_M], \varepsilon)$. By definition of X_m, the graph G_m contains only arcs xy with $\varepsilon(x, y) = m$ and for all $x, y \in X_m$ with $xy \notin E_m$ we have $\varepsilon(x, y) = \otimes$. Thus, G_m is obtained from $(K_{|X|}[X_M], \varepsilon)$ by removing all $\otimes$-edges and is, therefore, explained by $(T_{|X_m}, \lambda_{|X_m})$. Hence, G_m is a simple Fitch graph and *(T2)* is satisfied. This establishes the 'only if' direction.

Thus, Conditions *(T1)* to *(T4)* characterize tree-like maps ε. This together with the proof of the 'if' direction implies the correctness of Algorithm 1. □

Algorithm 2 recognizes tree-like maps ε and, in the affirmative case, returns a least-resolved tree (T^*, λ^*) that explains ε. The steps of this algorithm are explained in more detail in the proof of Theorem 3.4.

Theorem 3.4. *For a given map $\varepsilon : [X \times X]_{irr} \to M^{\otimes}$, Algorithm 2 determines whether ε is tree-like or not, and returns a tree (T^*, λ^*) that explains a tree-like map ε in $O(|X|^2)$-time.*

In particular, if ε is tree-like, then (T^*, λ^*) is a least-resolved tree for ε.

Algorithm 2 `Recognition of tree-like maps` ε.

Input: A map $\varepsilon : [X \times X]_{\text{irr}} \to M^{\otimes}$;
Output: A least-resolved edge-labeled tree (T^*, λ^*) that explains ε or the statement "The map ε is not tree-like";
1: **if** $|M| > 2|X|-2$ **then** Output: "The map ε is not tree-like";
2: **else if** ε satisfies Condition *(T1)* to *(T4)* in Thm. 3.3 **then**
3: Compute (T^*, λ^*) with Alg. 1;
4: **else** Output: "The map ε is not tree-like";

Proof. To establish the correctness of Algorithm 2, note first that for any tree $T = (V, E)$ on X we have $|E| + 1 = |V| \leq 2|X| - 1$ (cf. [32, Lemma 1]). Thus, there is no tree with $|E| > 2|X| - 2$ edges and hence, one can place at most $2|X| - 1$ different symbols on the edges of a tree. Therefore, if $|M| > 2|X|-2$, then ε cannot be tree-like, since we claimed that for any $m \in M$, $\varepsilon^{-1}(m) \neq \emptyset$. This establishes the correctness of Line 1 of Algorithm 2. Now, apply Theorem 3.3 to conclude that Algorithm 2 is correct.

We continue to verify the runtime of Algorithm 2. Clearly, the sets $X_1, \ldots, X_{|M|}, X_{\otimes}$ can be constructed by stepwisely considering each pair $(x, y) \in [X \times X]_{\text{irr}}$ and its label $\varepsilon(x, y)$, which takes $O(|X|^2)$-time. In particular, verifying Condition *(T1)* can be done directly within the construction phase of the sets X_m, $m \in M^{\otimes}$ and, hence stays within the time complexity of $O(|X|^2)$. Theorem 2.2 implies that Condition *(T2)* can be verified in $O(|X| + |E_m|)$ time for each $m \in M$. Due to the 'if-condition' in Line 1 of Algorithm 2, we have $|M| \in O(|X|)$. Furthermore, $\sum_{m \in M} E_m \in O(|X|^2)$. Thus, Condition *(T2)* can be checked in $\sum_{m \in M} O(|X| + |E_m|) = O((|M \parallel X|) + |X|^2) = O(|X|^2)$ time. Finally, for *(T3)* and *(T4)* we need to check if for all $x \in X_m$ and $y \in X \setminus X_m$ it holds that $\varepsilon(y, x) = m$. In other words, we must check for all $x \in X$ which label its $|X| - 1$ incoming arcs zx have. This can be done in $O(|X|^2)$-time. Thus, we end in overall time-complexity of $O(|X|^2)$ for Algorithm 2.

We continue to show that (T^*, λ^*) is a least-resolved tree for ε. By construction of (T^*, λ^*) all trees (T_m, λ_m) are exactly the subtrees $T^*(X_m)$ where all edge labels λ_m are kept. Hence, $(T_m, \lambda_m) = (T^*(X_m), \lambda^*_{|X_M})$. Note that none of the edges can be contracted that are contained in any of the trees (T_m, λ_m) that explains G_m and thus, that explains also any pair (x, y) with $x, y \in X_m$, since (T_m, λ_m) is already the unique least-resolved for the map ε restricted to pairs (x, y) with $x, y \in X_m$ (cf. Theorem 2.2). In particular, Theorem 2.2 implies that the labeling λ_m is unique and can therefore, not be changed. Moreover, no outer-edge of (T^*, λ^*) can be contracted, otherwise we would lose the information of a leaf. Hence, the only remaining edges that might be contracted are the m-edges of the form $\rho_{T^*} r_m$ as constructed in Line 5 of Algorithm 1. However, such an edge $\rho_{T^*} r_m$ was only added if (T_m, λ_m) contains an outer $\otimes$-edge $r_m x$ where $x \in X_m$ and r_m denotes the root of T_m. Thus, contracting the edge $\rho_{T^*} r_m$ would yield $\rho_{T^*} = r_m$. Now, there are two possibilities, either we relabel the resulting edge $\rho_{T^*} x$ or we keep the label $\otimes$. However, relabeling of $\rho_{T^*} x$ is not possible, since λ_m is unique and can therefore, not be changed. Thus, $\rho_{T^*} x$ must remain an $\otimes$-edge. However, due to the definition of X_m there is a pair (z, x) with $\varepsilon(z, x) = m$ which cannot be explained by any tree where x is linked to the root ρ_{T^*} via an $\otimes$-edge; a contradiction. Hence, m-edges of the form $\rho_{T^*} r_m$ cannot be contracted. In summary, there is no tree (T', λ') that explains ε, where T' is obtained from T by contracting an arbitrary edge. Hence, (T^*, λ^*) is least-resolved for ε. □

For maps $\varepsilon : [X \times X]_{\text{irr}} \to M$ that assign to none of the elements (x, y) a label $\otimes$ we obtain the following result.

Corollary 3.5. *A map $\varepsilon : [X \times X]_{\text{irr}} \to M$ is tree-like if and only if Conditions* (T1) *and* (T3) *are satisfied.*

Proof. By Theorem 3.3, *(T1)* and *(T3)* are satisfied if ε is tree-like. Assume that *(T1)* and *(T3)* are satisfied for ε. By construction of X_m, for all $x, y \in X_m$ we have $\varepsilon(x, y) = \varepsilon(y, x) = m$. Therefore, $K_{|X|}[X_m] = G_m$ is a complete di-graph with vertex set X_m. Hence, G_m does not contain any of the forbidden subgraphs $F_1, \ldots, F_8$ (cf. Fig. 1). Therefore, G_m is a simple Fitch graph and *(T2)* is always satisfied. Now, apply Theorem 3.3 to conclude that ε is tree-like. □

3.1. Uniqueness of the least-resolved tree

In general, there may be more than one rooted (phylogenetic) tree that explains a given map ε, see Fig. 2. In particular, if ε is explained by a non-binary tree (T, λ), then there is always a binary tree (T', λ') that refines T and explains the same map ε by setting $\lambda'(e) = \lambda(e)$ for all edges e that are also in T and by choosing the label $\lambda'(e) = \otimes$ for all edges e that are not contained in T. In this section, we will show that whenever a relation ε is explained by an edge-labeled tree (T, λ), then there exists a unique least-resolved tree that explains ε. We mainly follow here the proof strategies as in [22].

To establish the uniqueness of the least-resolved trees, we will consider so-called informative triples as shown in Fig. 4. Due to Lemma 3.1, it is an easy exercise to verify that each edge-labeled graph G_i, $i \in \{1, \ldots, 6\}$ in Fig. 4 is explained by the unique edge-labeled binary tree T_i, i.e., a specific labeled triple

Definition 3.2. An edge-labeled triple $ab|c$ is *informative* if it explains a 3-vertex induced subgraphs of a Fitch graph $(K_{|X|}, \varepsilon)$ isomorphic to one of $G_1, \ldots G_5$ or G_6.

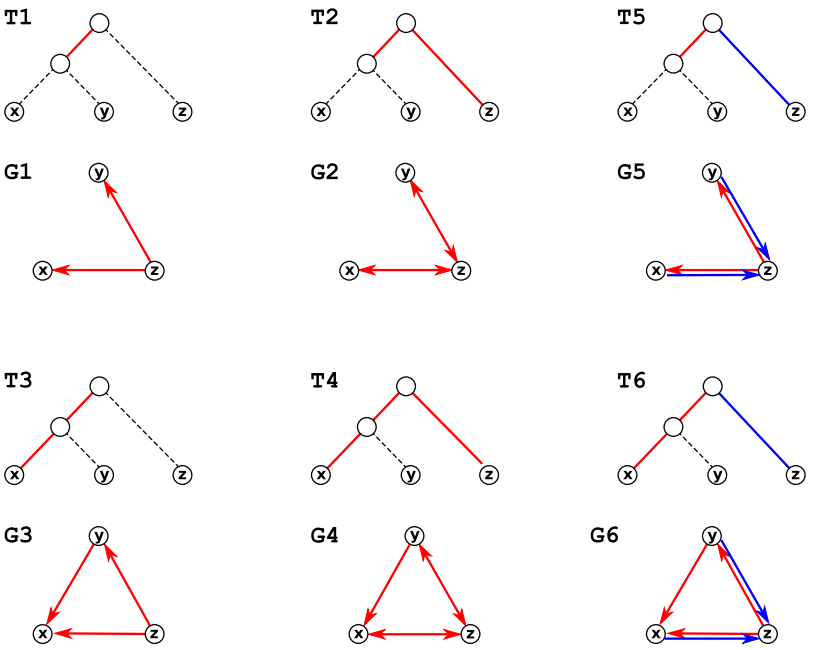

Fig. 4. Shown is the graph representation for six possible 3-vertex induced edge-labeled subgraphs $G_1, \ldots, G_6$ of a generalized Fitch graph$(K_{|X|}, \varepsilon)$ that is explained by a tree (T, λ). The $\otimes$-edges in each graph G_i are omitted. Each subgraph $G_1, \ldots, G_6$ is explained by the unique edge-labeled triple $T_1, \ldots, T_6$, respectively. In each tree T_i, the $\otimes$-edges are drawn as dashed-lines and red-edges and blue-edges correspond to two distinct symbols $m, m' \neq \otimes$. Edges in $T_1, \ldots, T_6$ can be understood as paths in T, whereby red-lined (resp. blue-lined, black-dashed) edges indicate that there is an m-edge (resp. m'-edge, only $\otimes$-edges) on the particular path. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The observation that each graph G_i, $i \in \{1, \ldots, 6\}$ in Fig. 4 is explained by the *unique* edge-labeled binary tree T_i is crucial, as this implies that whenever $(K_{|X|}, \varepsilon)$ contains an induced subgraph of the form $G_1, \ldots G_5$ or G_6, then any tree explaining $(K_{|X|}, \varepsilon)$ must display the corresponding informative triple. Any tree-like relation ε can therefore be associated with a uniquely defined set R_ε of *informative triples* that it displays: $r \in R_\varepsilon$ if and only if r is the unique edge-labeled triple explaining an induced subgraph isomorphic to $G_1, \ldots G_5$ or G_6. For later reference we summarize this fact as

Lemma 3.6. *If (T, λ) explains ε, then all triples in R_ε must be displayed by (T, λ).*

In what follows, we want to show that R_ε identifies the least-resolved tree that explains ε. To this end, we will utilize the following two results.

Lemma 3.7. *If (T^*, λ^*) is a least-resolved tree for the tree-like map $\varepsilon : [X \times X]_{irr} \to M^\otimes$, then (T^*, λ^*) contains no inner $\otimes$-edges and any inner vertex $v \neq \rho_{T^*}$ of (T^*, λ^*) is incident to an outer $\otimes$-edge.*

Proof. First, assume for contradiction that the least-resolved tree (T^*, λ^*) contains an inner $\otimes$-edge $e = uv$. The contraction of the edge e does not change the number of m-edges with $m \neq \otimes$ along the paths connecting any two leaves. It affects the least common ancestor of x and y, if $\mathrm{lca}_T(x, y) = u$ or $\mathrm{lca}_T(x, y) = v$. In either case, however, the number of m-edges between the $\mathrm{lca}_T(x, y)$ and the leaves x and y remains unchanged. Hence, the map ε can still be explained by the tree that is obtained from (T^*, λ^*) after contraction of e. Thus, (T^*, λ^*) is not least-resolved; a contradiction.

We continue to show that any inner vertex v must be incident to some outer $\otimes$-edge. Let $e = uv$ be the edge in T^* with $u \succ_{T^*} v$. Let F be the set of edges that are incident to v and distinct from e. First assume, for contradiction, that all edges in F have a label different from $\otimes$. If there are two edges $f, f' \in F$ with distinct labels, then Lemma 3.1 implies that e must be an $\otimes$-edge. However, (T^*, λ^*) contains no inner $\otimes$-edges and, hence, all edges in F must have the same label $m \neq \otimes$. In this case, Lemma 3.1 implies that the label $\lambda^*(e)$ of e must be m or $\otimes$. In either case, the edge e can be contracted, since every path from u to a leaf contains already an m-edge that is incident to v. Thus, (T^*, λ^*) is not least-resolved; a contradiction. Therefore, v must be incident to at least one $\otimes$-edge $f \in F$. Since (T^*, λ^*) contains no inner $\otimes$-edges, the edge f must be an outer-edge. □

Lemma 3.8. *Each inner edge in a least-resolved tree (T^*, λ^*) for a tree-like map $\varepsilon : [X \times X]_{irr} \to M^\otimes$, is distinguished by at least one informative triple in R_ε.*

Proof. Consider an arbitrary inner edge $e = uv$ of T^* with $u \succ_{T^*} v$. Since (T^*, λ^*) is phylogenetic, there are necessarily leaves x, y, and z such that $\mathrm{lca}(x, y) = v$ and $\mathrm{lca}(x, y, z) = u$. In particular, one can choose y such that vy is an outer $\otimes$-edge, since (T^*, λ^*) is least-resolved and due to Lemma 3.7. Moreover, Lemma 3.7 implies that $\lambda^*(e) = m \neq \otimes$. Lemma 3.1 implies that all edges f that are located in T^* below e must be $\otimes$- or m-edges. Thus, there are two exclusive cases for the path from $\mathrm{lca}(x, y)$ to x: Either the path contains (a) only $\otimes$-edges or (b) at least one m-edge. Moreover, the path $P_{u,z}$ from u to z contains either (A) only $\otimes$-edges or (B) an m-edge or (C) an m'-edge with $m' \neq m, \otimes$. Note, Lemma 3.1 implies that in case (A) (resp. (B)) all edges in $P_{u,z}$ must be m- or $\otimes$-edges (resp. m'- or $\otimes$-edges). Now, the combination of the Cases (a) and (b) with (A), (B) or (C) immediately implies that the tree on $\{x, y, z\}$ displayed by T^* must be one of the trees $T_1, \ldots, T_5$ or T_6 as shown in Fig. 4. Therefore, $xy|z \in R_\varepsilon$. Since $\mathrm{lca}(x, y) = v$ and $\mathrm{lca}(x, y, z) = u$, the edge e is by definition distinguished by the triple $xy|z \in R_\varepsilon$. □

Theorem 3.9. *Let $\varepsilon : [X \times X]_{\mathrm{irr}} \to M^{\otimes}$ be a tree-like map and (T^*, λ^*) be a least-resolved tree that explains ε. Then, the set R_ε identifies (T^*, λ^*) and $\mathrm{Aho}(R_\varepsilon) = T^*$. In particular, (T^*, λ^*) is unique up to isomorphism.*

Proof. We start with showing that R_ε identifies T^*. If $R_\varepsilon = \emptyset$, then (T^*, λ^*) must be a star tree, i.e., an edge-labeled tree that consists of outer edges only. Otherwise, (T^*, λ^*) contains inner edges that are, by Lemma 3.8, distinguished by at least one informative rooted triple in R_ε, contradicting that $R_\varepsilon = \emptyset$. Hence, $r(T^*) = \emptyset$, and therefore, $r(T^*) = \mathrm{cl}(R_\varepsilon)$. Lemma 2.1 implies that R_ε identifies (T^*, λ^*).

In the case $R_\varepsilon \neq \emptyset$, assume for contradiction that $r(T^*) \neq \mathrm{cl}(R_\varepsilon)$. By Lemma 3.6 we have $R_\varepsilon \subseteq r(T^*)$. Isotony of the closure, Theorem 3.1(3) in [5], ensures $\mathrm{cl}(R_\varepsilon) \subseteq \mathrm{cl}(r(T^*)) = r(T^*)$. Our assumption therefore implies $\mathrm{cl}(R_\varepsilon) \subsetneq r(T^*)$, and thus the existence of a triple $ab|c \in r(T^*) \setminus \mathrm{cl}(R_\varepsilon)$. In particular, therefore, $ab|c \notin R_\varepsilon$. Note that neither $ac|b$ nor $bc|a$ can be contained in R_ε, since (T^*, λ^*) explains ε and, by assumption, already displays the triple $ab|c$. Thus, R_ε contains no triples on $\{a, b, c\}$.

Let $u = \mathrm{lca}(a, b, c)$ and $e = uv$ be the edge in T^* with $u \succ_{T^*} v \succeq_{T^*} \mathrm{lca}(a, b)$. By Lemma 3.7, the edge e must be an m-edge with $m \neq \otimes$. Let T_{abc} be the subtree of (T, λ) with leaves a, b, c. Since e is an m-edge, Lemma 3.1 implies that all edges along the paths from v to a and v to b must be m- or $\otimes$-edges. However, since $ab|c \notin R_\varepsilon$, the tree T_{abc} cannot be isomorphic to the subtree $T_1, \ldots, T_6$ and thus, both paths from $\mathrm{lca}(a, b)$ to a and $\mathrm{lca}(a, b)$ to b must contain m-edges.

Moreover, Lemma 3.7 implies that there must be an outer $\otimes$-edge $f = vd$. By the discussion above, $d \neq a, b$. Thus, the subtrees T_{acd} and T_{bcd} of T^* with leaves a, c, d and b, c, d, respectively, correspond to one of the trees T_3, T_4 and T_6. By construction, $ad|c \in R_\varepsilon$ and $bd|c \in R_\varepsilon$. Hence, any tree that explains ε must display $ad|c$ and $bd|c$. As shown in [11], a tree displaying $ad|c$ and $bd|c$ also displays $ab|c$. This implies, however, that $ab|c \in \mathrm{cl}(R_\varepsilon)$, a contradiction to our assumption.

Therefore, $\mathrm{cl}(R_\varepsilon) = r(T)$ and we can apply Lemma 2.1 to conclude that R_ε identifies (T^*, λ^*) and $\mathrm{Aho}(R_\varepsilon) = T^*$.

We continue to show the uniqueness of (T^*, λ^*). Since R_ε identifies (T^*, λ^*), any tree that displays R_ε is by definition a refinement of (T^*, λ^*). In addition, any tree that explains ε must display R_ε (cf. Lemma 3.6). Taking the latter two arguments together, any tree that explains ε must be a refinement of (T^*, λ^*).

To establish uniqueness of (T^*, λ^*) it remains to show that there is no other labeling λ such that (T^*, λ) still explains ε. Let $e = uv$ be an outer edge. Hence, changing the label of e would immediately change the label $\varepsilon(w, v)$ between v and any leaf w located in a subtree rooted at a sibling of v. Since at least one such leaf w exists in a phylogenetic tree, the edge e cannot be re-labeled. Now suppose that $e = uv$ is an inner edge with $u \succ_{T^*} v$. By Lemma 3.7, the edge e must be m-edge and $m \neq \otimes$. and there must be an outer $\otimes$-edge $f = vw$. Let x be a leaf such that $\mathrm{lca}(w, x) = u$. Since T^* is a phylogenetic tree, such a leaf always exists. Then $\varepsilon(x, w) = m$ if and only if $\lambda(e) = m$, i.e., the inner edge e cannot be re-labeled. This establishes the final statement. □

4. Summary and outlook

We have considered maps $\varepsilon : [X \times X]_{\mathrm{irr}} \to M^{\otimes}$ and edge labeled di-graphs $(K_{|X|}, \varepsilon)$ that can be explained by edge-labeled phylogenetic trees. Such graphs generalize the notion of xenology and simple Fitch graphs [22,23]. As a main result, we gave a characterization of Fitch graphs based on four simple conditions *(T1)* to *(T4)* that are defined in terms of underlying edge-disjoint subgraphs. This in turn led to an $O(|X|^2)$-time algorithm to recognize Fitch graphs $(K_{|X|}, \varepsilon)$ and for the reconstruction of the unique least-resolved $M^{\otimes}$-edge-labeled phylogenetic tree that can explain them.

From the combinatorial point of view it might be of interest to consider more general maps $\mathcal{E} : [X \times X]_{\mathrm{irr}} \to \mathcal{P}(M) \cup \{\otimes\}$, where $\mathcal{P}(M)$ denotes the powerset of M. In this case, there are a couple of ways to define when $\mathcal{E}$ is tree-like. The two most obvious ways, which we call "Type-1" and "Type-2" tree-like, are stated here.

The map $\mathcal{E}$ is tree-like

of Type-1, if there is an edge-labeled tree (T, λ) on X such that for at least one $m \in \mathcal{E}(x, y)$ there is an edge on the path from $\mathrm{lca}(x, y)$ to y with label m.

of Type-2, if there is an edge-labeled tree (T, λ) on X such that for all $m \in \mathcal{E}(x, y)$ there is an edge on the path from $\mathrm{lca}(x, y)$ to y with label m.

Note, if $|M| = 1$ or $|\mathcal{E}(x, y)| = 1$ for all $x, y \in X$, then the problem of determining whether $\mathcal{E}$ is Type-1 or Type-2 tree-like reduces to the problem of determining whether $(K_{|X|}, \mathcal{E})$ is a Fitch graph or not. Moreover, if the sets $\mathcal{E}(x, y)$, $x, y \in X$ are pairwise disjoint, we can define a set $N = \{m_{\mathcal{E}(x,y)} \mid x, y \in X\}$ of symbols that identifies each symbol $m_{\mathcal{E}(x,y)}$ with the set $\mathcal{E}(x, y)$. The established results imply the following

Corollary 4.1. *If the map* $\varepsilon : [X \times X]_{irr} \to N \cup \{\otimes\}$ *with* $\varepsilon(x, y) = m_{\mathcal{E}(x,y)}$ *is tree-like, then the map* $\mathcal{E}$ *is tree-like of Type-1.*

It would be of interest to understand such generalized tree-like maps in more detail. To this end, results established in [2,33,41] might shed some light on this question. Moreover, maps that cannot be explained by trees may be explained by phylogenetic networks, an issue that has not been addressed so-far.

References

[1] A.V. Aho, Y. Sagiv, T.G. Szymanski, J.D. Ullman, Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions, SIAM J. Comput. 10 (3) (1981) 405–421.
[2] H.-J. Bandelt, M.A. Steel, Symmetric matrices representable by weighted trees over a Cancellative abelian monoid, SIAM J. Discrete Math. 8 (4) (1995) 517–525.
[3] S. Böcker, A.W.M. Dress, Recovering symbolically dated, rooted trees from symbolic ultrametrics, Adv. Math. 138 (1998) 105–125.
[4] J.L. Boore, The use of genome-level characters for phylogenetic reconstruction, Trends Ecol. Evol. 21 (2006) 439–446.
[5] D. Bryant, Building Trees, Hunting for Trees, and Comparing Trees: Theory and Methods in Phylogenetic Analysis (Ph.D. thesis), University of Canterbury, 1997.
[6] D. Bryant, M. Steel, Extension operations on sets of leaf-labeled trees, Adv. Appl. Math. 16 (4) (1995) 425–453.
[7] I. Chen, D. Dubnau, DNA uptake during bacterial transformation, Nature Rev. Microbiol. 2 (3) (2004) 241.
[8] D.G. Corneil, H. Lerchs, L. Steward Burlingham, Complement reducible graphs, Discrete Appl. Math. 3 (1981) 163–174.
[9] C. Crespelle, C. Paul, Fully dynamic recognition algorithm and certificate for directed cographs, Discrete Appl. Math. 154 (2006) 1722–1741.
[10] E.J. Deeds, H. Hennessey, E.I. Shakhnovich, Prokaryotic phylogenies inferred from protein structural domains, Genome. Res. 15 (2005) 393–402.
[11] M.C.H. Dekker, Reconstruction Methods for Derivation Trees (Ph.D. thesis), Vrije Universiteit, Amsterdam, Netherlands, 1986.
[12] F. Delsuc, H. Brinkmann, H. Philippe, Phylogenomics and the reconstruction of the tree of life, Nature Rev. Genet. 6 (5) (2005) 361–375.
[13] K.M. Derbyshire, T.A. Gray, Distributive conjugal transfer: New insights into horizontal gene transfer and genetic exchange in mycobacteria, Microbiol. Spectr. 2 (1) (2014).
[14] A. Donath, P.F. Stadler, Molecular morphology: Higher order characters derivable from sequence information, in: J.W. Wägele, T. Bartolomaeus (Eds.), Deep Metazoan Phylogeny: The Backbone of the Tree of Life. New Insights from Analyses of Molecules, Morphology, and Theory of Data Analysis, de Gruyter, Berlin, 2014, pp. 549–562.
[15] B.E. Dutilh, B. Snel, T.J. Ettema, M.A. Huynen, Signature genes as a phylogenomic tool, Mol. Biol. Evol. 25 (2008) 1659–1667.
[16] A. Ehrenfeucht, G. Rozenberg, Theory of 2-structures, part I: Clans, basic subclasses, and morphisms, Theoret. Comput. Sci. 70 (1990) 277–303.
[17] A. Ehrenfeucht, G. Rozenberg, Theory of 2-structures, part II: Representation through labeled tree families, Theoret. Comput. Sci. 70 (1990) 305–342.
[18] J. Engelfriet, T. Harju, A. Proskurowski, G. Rozenberg, Characterization and complexity of uniformly nonprimitive labeled 2-structures, Theoret. Comput. Sci. 154 (1996) 247–282.
[19] W.M. Fitch, Distinguishing homologous from analogous proteins, Syst. Biol. 19 (2) (1970) 99–113.
[20] W.M. Fitch, Homology a personal view on some of the problems, Trends Genet. 16 (2000) 227–231, http://dx.doi.org/10.1016/S0168-9525(00)02005-9.
[21] T. Gabaldón, E. Koonin, Functional and evolutionary implications of gene orthology, Nat. Rev. Genet. 14 (5) (2013) 360–366.
[22] M. Geiß, J. Anders, P. Stadler, N. Wieseke, M. Hellmuth, Reconstructing gene trees from Fitch's xenology relation, J. Math. Biol. 77 (5) (2018) 1459–1491.
[23] M. Geiß, M. Hellmuth, Y. Long, P. Stadler, A short note on undirected Fitch graphs, Art Discrete Appl. Math. 1 (1) (2018) #P1.08.
[24] T.A. Gray, J.A. Krywy, J. Harold, M.J. Palumbo, K.M. Derbyshire, Distributive conjugal transfer in mycobacteria generates progeny with meiotic-like genome-wide mosaicism, allowing mapping of a mating identity locus, PLoS Biol. 11 (7) (2013) 1–13.
[25] S. Grünewald, Y. Long, Y. Wu, Reconstructing unrooted phylogenetic trees from symbolic ternary metrics, Bull. Math. Biol. 80 (6) (2018) 1563–1577.
[26] S. Grünewald, M. Steel, M.S. Swenson, Closure operations in phylogenetics, Math. Biosci. 208 (2) (2007) 521–537.
[27] M. Hellmuth, M. Hernandez-Rosales, K.T. Huber, V. Moulton, P.F. Stadler, N. Wieseke, Orthology relations, symbolic ultrametrics, and cographs, J. Math. Biol. 66 (1–2) (2013) 399–420.
[28] M. Hellmuth, M. Hernandez-Rosales, Y. Long, P. Stadler, Inferring phylogenetic trees from the knowledge of rare evolutionary events, J. Math. Biol. 76 (7) (2018) 1623–1653.
[29] M. Hellmuth, C.R. Seemann, Alternative characterizations of Fitch's xenology relation, J. Math. Biol. (2019) http://dx.doi.org/10.1007/s00285-019-01384-x.
[30] M. Hellmuth, P. Stadler, N. Wieseke, The mathematics of xenology: Di-cographs, symbolic ultrametrics, 2-structures and tree-representable systems of binary relations, J. Math. Biol. 75 (1) (2017) 199–237, http://dx.doi.org/10.1007/s00285-016-1084-3.
[31] M. Hellmuth, N. Wieseke, From sequence data incl. orthologs, paralogs, and xenologs to gene and species trees, in: Evolutionary Biology, Springer International Publishing, Cham, 2016, pp. 373–392.
[32] M. Hellmuth, N. Wieseke, M. Lechner, H.-P. Lenhof, M. Middendorf, P. Stadler, Phylogenomics with paralogs, Proc. Natl. Acad. Sci. 112 (7) (2015) 2058–2063, http://dx.doi.org/10.1073/pnas.1412770112.
[33] K. Huber, G. Scholz, V. Moulton, Three-way symbolic tree-maps and ultrametrics, J. Classification (2018) http://dx.doi.org/10.1007/s00357-018-9274-x.
[34] O. Johnsborg, V. Eldholm, L.S. Høavarstein, Natural genetic transformation: prevalence, mechanisms and function, Res. Microbiol. 158 (10) (2007) 767–778, Microbial genomics.
[35] V. Krauss, C. Thümmler, F. Georgi, J. Lehmann, P.F. Stadler, C. Eisenhardt, Near intron positions are reliable phylogenetic markers: An application to Holometabolous Insects, Mol. Biol. Evol. 25 (2008) 821–830.
[36] S.J. Prohaska, C. Fried, C.T. Amemiya, F.H. Ruddle, G.P. Wagner, P.F. Stadler, The Shark HoxN cluster is homologous to the Human HoxD cluster, J. Mol. Evol. (2004) 58, 212-217.
[37] M. Ravenhall, N. Škunca, F. Lassalle, C. Dessimoz, Inferring horizontal gene transfer, PLoS Comput. Biol. 11 (5) (2015) 1–16.
[38] I.B. Rogozin, A.V. Sverdlov, V.N. Babenko, E.V. Koonin, Analysis of evolution of exon-intron structure of eukaryotic genes, Brief. Bioinform 6 (2005) 118–134.
[39] A. Rokas, P.W. Holland, Rare genomic changes as a tool for phylogenetics, Trends Ecol. Evol. 15 (2000) 454–459.
[40] C.R. Seemann, M. Hellmuth, The matroid structure of representative triple sets and triple-closure computation, European J. Combin. 70 (2018) 384–407.
[41] C. Semple, M. Steel, Tree representations of non-symmetric group-valued proximities, Adv. Appl. Math. 23 (3) (1999) 300–321.
[42] C. Semple, M. Steel, Phylogenetics, in: Oxford Lecture Series in Mathematics and its Applications, vol. 24, Oxford University Press, Oxford, 2003.

[43] A.M. Shedlock, N. Okada, SINE insertions: powerful tools for molecular systematics, BioEssays 22 (2000) 148–160.
[44] M. Steel, Phylogeny: Discrete and Random Processes in Evolution, in: CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, 2016.
[45] R.L. Tatusov, M.Y. Galperin, D.A. Natale, E.V. Koonin, The COG database: a tool for genome-scale analysis of protein functions and evolution, Nucleic Acids Res. 28 (1) (2000) 33–36.
[46] J. Zhang, Evolution by gene duplication: an update, Trends Ecol. Evol. 18 (6) (2003) 292–298.