Generalized Fitch Graphs: Edge-labeled Graphs that are explained by Edge-labeled Trees

Marc Hellmuth

Dpt. of Mathematics and Computer Science, University of Greifswald, Walther- Rathenau-Strasse 47, D-17487 Greifswald, Germany
Saarland University, Center for Bioinformatics, Building E 2.1, P.O. Box 151150, D-66041 Saarbrücken, Germany
Email: mhellmuth@mailbox.org

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 $\text{lca}(x, y)$ 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 $\text{lca}(x, y)$ 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 $\varepsilon$ 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 $(T, \lambda)$ 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 $(K_{|X|}, \varepsilon)$.

Keywords: Labeled Gene Trees; Forbidden Subgraphs; Phylogenetics; Xenology; Fitch Graph; Recognition Algorithm

1 Introduction

Edge-labeled graphs that can be explained by vertex-labeled trees have been widely studied and range from cographs [7, 24] and di-cographs [8] to so-called unp-2-structures [14–16], symbolic ultrametrics [3, 26] or three-way symbolic tree-maps [22, 29]. 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 [18], is of fundamental importance in many fields of mathematical and computational biology, including the reconstruction of evolutionary relationships across species [11, 28] or functional genomics and gene organization in species [19, 39]. The orthology relation $\Theta$ is explained by vertex-labeled trees, i.e., a gene pair $(x, y)$ is contained in $\Theta$ if and only if the least common ancestor of $x$ and $y$ is labeled as a speciation event. The graph representation of $\Theta$ must necessarily be a co-graph [3, 24] and provides direct information on the gene history as well as on the history of the species [27, 28].

In contrast, xenology as defined by Walter M. Fitch [17] 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 $\text{lca}(x, y)$ to $y$ in the 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. [20] 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 \(F\) it is possible to reconstruct the unique minimally resolved phylogenetic tree that explains \(F\) in linear time.

A further example of graphs and relations that are defined in terms of edge-labeled trees are the single-1-relations \(\tilde{1}\) and \(\tilde{1}\rightarrow\) [25]. 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, 9, 12, 13, 30–33, 37].

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 [20, 21, 25].

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 these 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.

## 2 Preliminaries

### 2.1 Trees, Di-Graphs and Sets

A rooted tree \(T = (V, E)\) (on \(X\)) is an acyclic connected graphs 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 \(\rho T\) 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 \(v \in V\), \(u \geq_T v\), and \(v\) a descendant of \(u\), \(u \leq_T v\), if \(u\) lies on the unique path from \(\rho T\) to \(v\). We write \(v \triangleleft_T u\) \((u \triangleright_T v)\) for \(v \leq_T u\) \((u \geq_T v)\) 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 \(\text{lca}_T(Y)\), is the unique \(\leq_T\)-minimal vertex of \(T\) that is an ancestor of every vertex in \(Y\). We will make use of the simplified notation \(\text{lca}_T(x, y) := \text{lca}_T(\{x, y\})\) 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 \(\text{lca}_T(Y)\) 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 \(\rho T\) 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 root tree \(T'\) on \(L'\), in symbols \(T' \leq T\), if \(T'\) can be obtained from \(T(L')\) by a sequence of edge contractions. If \(T' \leq T\), 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 \(\text{lca}(a, b) \triangleleft_T \text{lca}(a, b, c)\). 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 [36, 38], that takes a set \(R\) of triples as
input and either returns a particular phylogenetic tree $Aho(R)$ 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 $lca(a, b) = v$ and $lca(a, b, c) = u$.

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, 23, 34]. 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

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

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

For later reference, we give here an important result from [23] that is closely related to the $\text{BUILD}$ 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 $\text{cl}(R) = r(T)$. Moreover, if $R$ identifies $T$, then $Aho(R) = T$.

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 $(T, \lambda)$. 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 digraph $G = (V, E)$ 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, $|X \times X|_{\text{dir}}$ 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 pairwise 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 $(T, \lambda)$ 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 $\text{lca}(T, (x, y))$ 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 digraph $G = (X, E)$ is explained by a phylogenetic tree $(T, \lambda)$ (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. [20], 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** ([20]). A given digraph $G = (X, E)$ is a simple Fitch graph if and only if it does not contain one 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 $(T, \lambda)$ that explains $G$ can be done in $O(|X| + |E|)$ time.

$(T, \lambda)$ is a least-resolved tree that explains $G$, i.e., there is no edge-contracted version $T'$ of $T$ and no labeling $\lambda'$ such that $(T', \lambda')$ still explains $G$, if and only if all its inner edges are 1-edges and for every inner edge $(u, v)$ there is an outer $\otimes$-edge $(v, x)$ in $(T, \lambda)$. 
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 \([X \times X]_{irr}\) and a map \(\varepsilon : [X \times X]_{irr} \to M^\circ\) that assigns to each arc \(xy\) a unique label \(\varepsilon(x, y)\). Clearly, the map \(\varepsilon\) 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 \((x, y) \in [X \times X]_{irr}\) such that \(\varepsilon(x, y) = m\).

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

\[
\varepsilon(x, y) = m \in M \iff \text{some edge } e \text{ on the path } P_{(x,y)} \text{ has label } \lambda(e) = m; \quad \text{and }
\varepsilon(x, y) = \otimes \iff \text{none of the edges } e \text{ on the path } P_{(x,y)} \text{ have label } \lambda(e) \in M.
\]

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

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

Figure 3 shows an example of a generalized Fitch graph \((K_{|X|}, \varepsilon)\). We give the following almost trivial result for later reference.

**Lemma 3.2.** Let \(\varepsilon : [X \times X]_{irr} \to M^\circ\) be tree-like and \((T, \lambda)\) be a tree that explains \(\varepsilon\). If there is an edge \(e\) with \(\lambda(e) = m\) on the path \(P\) from the root \(\rho_T\) to some leaf, then all edges in \(P\) are either labeled with \(m\) or \(\otimes\).

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

For each symbol \(s \in M^\circ\) we define the following set

\[
X_s := \{x \in X \mid \text{there is a vertex } z \in X \text{ with } \varepsilon(z, x) = s \}
\]

and for all \(z' \in X \setminus \{z, x\}\) 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. 2. 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 \( m \neq \emptyset \) is the first \( m \)-edge that lies on the path from the root to any of the leaves below \( e \). Thus, \( X'_m \subseteq X_m \), see text for further details.

Lemma 3.3. Let \( \varepsilon : [X \times X]_{\text{irr}} \to M^\otimes \) be a tree-like map and \((T, \lambda)\) a tree that explains \( \varepsilon \). 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, \( \varepsilon \) 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 \( \varepsilon(z', x) = m' \neq \{ m, \emptyset \} \). Thus, the path from lca\( (z, x) \) to \( x \) contains an edge labeled \( m \) and the path from lca\( (z', x) \) to \( x \) contains an edge labeled \( m' \). However, since both vertices lca\( (x, z) \) and 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 3.2.

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 \( \varepsilon(z, x) \in \{ m', \emptyset \} \); a contradiction to \( \varepsilon \) 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, \( \varepsilon(z, x) = m \) is \( M \)-edges that \( m \neq \emptyset \) 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 \( m = \emptyset \) 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 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 $\lca(x, z)$ 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 $\lca(x, z)$ and $\lca(x, y)$. Therefore, the $m$-edge is located between $\lca(x, z)$ and $\lca(x, y)$ and, in particular, $\lca(x, z) \succeq u \succ v \succeq \lca(x, y)$. Hence, $\lca(x, z) = \lca(y, z)$ and the path from $\lca(y, z)$ to $z$ 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

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

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

Before we can derive the final result, we need one further definition. Let $(T, \lambda)$ 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 $uv \in E(T_{|X_m})$:

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

Lemma 3.2 implies that the restriction $\lambda_{|X_m}$ of $\lambda$ 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.

We are now in the position to characterize tree-like maps $\varepsilon$.

**Theorem 3.4.** The map $\varepsilon : [X \times X]_{\text{irr}} \to M^\otimes$ is tree-like (or equivalently, $(K_{|X|}, \varepsilon)$ is a generalized Fitch graph) if and only if the following four conditions are satisfied:

1. **(T1)** The sets $X_1, X_2, \ldots, X_{|M|}, X_\otimes$ form a quasi-partition of $X$.
2. **(T2)** $G_m = (X_m, E_m)$ is a simple Fitch graph for all $m \in M$.
3. **(T3)** For all $m \in M$ and $x \in X_m$, $y \in X \setminus X_m$, it holds that $\varepsilon(y, x) = m$.
4. **(T4)** For all $x \in X_\otimes$ and $y \in X \setminus \{x\}$, it holds that $\varepsilon(y, x) = \otimes$.

In particular, the tree $(T^*, \lambda^*)$ returned by Algorithm 2 (with input $\varepsilon$) explains $\varepsilon$, whenever $\varepsilon$ is tree-like.

**Proof.** We first establish the ‘if’ direction. Assume Conditions (T1) to (T4) are satisfied for $\varepsilon$. 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 $(T_m, \lambda_m)$ with leaf set $X_m$. 

---

**Figure 3:** Left, an edge-labeled tree $(T, \lambda)$ is shown where $\otimes$-edges are drawn as dashed lines. The tree $(T, \lambda)$ explains the graph $(K_{|X|}, \varepsilon)$ in the middle, 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 least-resolved tree $(T^*, \lambda^*)$ constructed with Alg. 2 is shown at the right-hand side.
Algorithm 1 Recognition of tree-like maps $\varepsilon$.

Input: A map $\varepsilon : [X \times X]_{irr} \rightarrow M^\circ$;
Output: A least-resolved edge-labeled tree $(T^*, \lambda^*)$ that explains $\varepsilon$ or the statement “The map $\varepsilon$ is not tree-like”;
1: if $|M| > 2|X| - 2$ then Output: “The map $\varepsilon$ is not tree-like”;
2: else if $\varepsilon$ satisfies Condition (T1) to (T4) in Thm. 3.4 then
3: Compute $(T^*, \lambda^*)$ with Alg. 2;
4: else Output: “The map $\varepsilon$ is not tree-like”;

We show that the tree $(T^*, \lambda^*)$ constructed with Alg. 2 explains $\varepsilon$. By construction of $(T^*, \lambda^*)$ all trees $(T_m, \lambda_m)$ are exactly the subtrees $T^*(X_m)$ where all edge labels $\lambda_m$ are kept. Hence, $G_m$ is explained by $(T^*(X_m), \lambda^*_m)$. Since $\varepsilon(x, y) = m$ (resp. $\varepsilon(x, y) = \emptyset$) 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 $(T^*, \lambda^*)$ for all $m \in M$. Moreover, each $x \in X_\emptyset$ is linked to the root $p_T$ via an $\emptyset$-edge (Line 9 of Alg. 2). Hence, for each two vertices $x, y \in X_\emptyset$ we have, by definition of $X_\emptyset$, $\varepsilon(x, y) = \varepsilon(y, x) = \emptyset$, which is trivially explained by $(T^*, \lambda^*)$. Since the sets $X_1, X_2, \ldots, X_{|M|}, X_\emptyset$ form a quasi-partition of $X$, it is ensured that there are no overlapping leaf sets when the trees $(T_m, \lambda_m)$ and the elements $x \in X_\emptyset$ have been added to $(T^*, \lambda^*)$ 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 $(T^*, \lambda^*)$. Note first, by construction of $(T^*, \lambda^*)$ and since $(T^*, \lambda^*)$ explains $G_m$ for all $m \in M$, all edges along the path from $p_T$ to $x \in X_m$ have label $m$ or $\emptyset$. Even more, we show that each path $P_{p_T, x}$ from $p_T$ to each $x \in X_m$, $m \in M$ has always an edge with label $m$. By construction of $(T^*, \lambda^*)$ (Alg. 2, Line 4-6), if $|X_m| = 1$ or there is a leaf $x \in X_m$ adjacent to the root $p_m$ of $T_m$ such that $\lambda_m(p_m, x) = \emptyset$, then the tree $(T_m, \lambda_m)$ is placed below the particular $m$-edge $p_T \cdot r_m$. Hence, all paths from $p_T$ 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 $(T^*, \lambda^*)$, given that $(T_m, \lambda_m)$ satisfies the Conditions in Alg. 2 (Line 4). Assume that $(T_m, \lambda_m)$ does not satisfy the latter conditions. Theorem 2.2 implies that all inner edges of $(T_m, \lambda_m)$ are $m$-edges and thus, any $\emptyset$-edge in $(T_m, \lambda_m)$ must be incident to some leaf $x \in X_m$. Since $(T_m, \lambda_m)$ does not satisfy the if-condition in Line 4 of Alg. 2, all edges that are incident to the root of $(T_m, \lambda_m)$ have label $m$. Hence, all paths from $p_T$ 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 $(T^*, \lambda^*)$. Finally, if $x \in X_\emptyset$, then $(T4)$ claims $\varepsilon(y, x) = \emptyset$ for all $y \neq x$ which is trivially explained by $(T^*, \lambda^*)$, since $x$ is linked to the root $p_T$ via an $\emptyset$-edge (Alg. 2, Line 9). In summary, if the Conditions $(T1)$ to $(T4)$ are satisfied, then $\varepsilon$ is explained by $(T^*, \lambda^*)$ and therefore, tree-like. This establishes the ‘if’ direction.

We turn now to the ‘only if’ direction. Assume that $\varepsilon$ is tree-like and let $(T, \lambda)$ be a tree that explains $\varepsilon$ with root $p_T$. Lemma 3.3 implies Condition (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 $(T, \lambda)$ explains $\varepsilon$ 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 $(T, \lambda)$. By construction of the restriction $(T_{|X_m|}, \lambda_{|X_m|})$ of $(T, \lambda)$ to $X_m$ we have $\varepsilon(x, y) = m$ if and only if the path in $(T, \lambda)$ from 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 lca$_T(x, y)$ 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 $\varepsilon(x, y) = \emptyset$. Thus, $G_m$ is obtained from $(K_{|X|}[X_M], \varepsilon)$ by removing all $\emptyset$-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 $\varepsilon$. This together with the proof of the ‘if’ direction implies the correctness of Alg. 2. 

[Ref: 7]
Algorithm 2 Compute Least-Resolved Tree $(T^*, \lambda^*)$ for $\varepsilon$.

**Input:** A tree-like map $\varepsilon : [X \times X]_{irr} \rightarrow M^\circ$;

**Output:** A least-resolved edge-labeled tree $(T^*, \lambda^*)$ that explains $\varepsilon$;

1. Add a root $\rho_{T^*}$ to $T^*$;
2. for all $m \in M$ do
3. Compute the least-resolved tree $(T_m, \lambda_m)$ the explains $G_m = (X_m, E_m)$;
4. if $|X_m| = 1$ OR $(T_m, \lambda_m)$ contains an $\otimes$-edge incident to its root then
5. Add a vertex $r_m$ and the edge $\rho_{T^*} \rightarrow r_m$ with label $m$;
6. Add $(T_m, \lambda_m)$ 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 $\rho_{T^*}$ and add $(T_m, \lambda_m)$;
9. Add an edge $e = \rho_{T^*} \rightarrow x$ with label $\lambda^*(e) = \otimes$ for all $x \in X_{irr}$;
10. Return $(T^*, \lambda^*)$;

**Theorem 3.5.** For a given map $\varepsilon : [X \times X]_{irr} \rightarrow M^\circ$, Algorithm 1 determines whether $\varepsilon$ is tree-like or not, and returns a tree $(T^*, \lambda^*)$ that explains a tree-like map $\varepsilon$ in $O(|X|^2)$-time.

In particular, if $\varepsilon$ is tree-like, then $(T^*, \lambda^*)$ is a least-resolved tree for $\varepsilon$.

**Proof.** To establish the correctness of Alg. 1, note first that for any tree $T = (V,E)$ on $X$ we have $|E|+1 = |V| \leq 2|X|−1$ (cf. [28, 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 $\varepsilon$ 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 Alg. 1. Now, apply Thm. 3.4 to conclude that Alg. 1 is correct.

We continue to verify the runtime of Alg. 1. Clearly, the sets $X_1, \ldots, X_{|M|}, X_{irr}$ can be constructed by stepwisely considering each pair $(x,y) \in [X \times X]_{irr}$, 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 $X_m$, $m \in M^\circ$ and, hence stays within the time complexity of $O(|X|^2)$. Thm. 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 Alg. 1, we have $|M| \in O(|X|)$. Furthermore, $\sum_{m \in M} E_m \in O(|X|^2)$. Thus, Condition (T2) can be checked in $\sum_{m \in M} O(|X| + |E_m|) = O(|M||X| + |X|^2) = O(|X|^2)$ 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 $x$ have. This can be done in $O(|X|^2)$-time. Thus, we end in overall time-complexity of $O(|X|^2)$ for Alg. 1.

We continue to show that $(T^*, \lambda^*)$ is a least-resolved tree for $\varepsilon$. By construction of $(T^*, \lambda^*)$ all trees $(T_m, \lambda_m)$ are exactly the subtrees $T^*(X_m)$ where all edge labels $\lambda_m$ 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 $(T_m, \lambda_m)$ that explains $G_m$ and thus, that explains also any pair $(x,y)$ with $x,y \in X_m$, since $(T_m, \lambda_m)$ is already the unique least-resolved for the map $\varepsilon$ restricted to pairs $(x,y)$ with $x,y \in X_m$ (cf. Thm. 2.2). In particular, Thm. 2.2 implies that the labeling $\lambda_m$ is unique and can therefore, not be changed. Moreover, no outer-edge of $(T^*, \lambda^*)$ can be contracted, otherwise we would loose the information of a leaf. Hence, the only remaining edges that might be contracted are the $m$-edges of the form $\rho_{T^*} \rightarrow r_m$ as constructed in Line 5 of Alg. 2. However, such an edge $\rho_{T^*} \rightarrow r_m$ was only added if $(T_m, \lambda_m)$ contains an outer $\otimes$-edge $r_m \rightarrow x$ where $x \in X_m$ and $r_m$ denotes the root of $T_m$. Thus, contracting the edge $\rho_{T^*} \rightarrow r_m$ would yield $\rho_{T^*} = r_m$. Now, there are two possibilities, either we relabel the resulting edge $\rho_{T^*} \rightarrow x$ or we keep the label $\otimes$. However, relabeling of $\rho_{T^*} \rightarrow x$ is not possible, since $\lambda_m$ is unique and can therefore, not be changed. Thus, $\rho_{T^*} \rightarrow x$ 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 $\rho_{T^*}$ via an $\otimes$-edge; a contradiction. Hence, $m$-edges of the form $\rho_{T^*} \rightarrow r_m$ cannot be contracted. In summary, there is no tree $(T', \lambda')$ that explains $\varepsilon$, where $T'$ is obtained from $T$. 

8
by contracting an arbitrary edge. Hence, \((T^*, \lambda^*)\) is least-resolved for \(\varepsilon\). □

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 3.6.** The map \(\varepsilon : [X \times X]\text{irr} \to M\) is tree-like if and only if Condition \((T1)\) and \((T3)\) are satisfied.

**Proof.** By Thm. 3.4, \((T1)\) and \((T3)\) are satisfied if \(\varepsilon\) is tree-like. Assume that \((T1)\) and \((T3)\) are satisfied for \(\varepsilon\). 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 Thm. 3.4 to conclude that \(\varepsilon\) 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 \(\varepsilon\), see Fig. 3. In particular, if \(\varepsilon\) is explained by a non-binary tree \((T, \lambda)\), then there is always a binary tree \((T', \lambda')\) that refines \(T\) and explains the same map \(\varepsilon\) 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 \(\varepsilon\) is explained by an edge-labeled tree \((T, \lambda)\), then there exists a unique least-resolved tree that explains \(\varepsilon\). We mainly follow here the proof strategies as in [20].

To establish the uniqueness of the least-resolved trees, we will consider so-called informative triples as shown in Fig. 4. Due to Lemma 3.2, 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 3.7.** An edge-labeled triple \(ab\mid c\) 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 \(G_6\).

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 \(\varepsilon\) can therefore be associated with a uniquely defined set \(R_{\varepsilon}\) 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 3.8.** If \((T, \lambda)\) explains \(\varepsilon\), then all triples in \(R_{\varepsilon}\) must be displayed by \((T, \lambda)\).

In what follows, we want to show that \(R_{\varepsilon}\) identifies the least-resolved tree that explains \(\varepsilon\). To this end, we will utilize the following two results.

**Lemma 3.9.** If \((T^*, \lambda^*)\) is a least-resolved tree for the tree-like map \(\varepsilon : [X \times X]\text{irr} \to M\otimes\), then \((T^*, \lambda^*)\) contains no inner \(\otimes\)-edges and any inner vertex \(v \not\in r_{T^*}\) of \((T^*, \lambda^*)\) is incident to an outer \(\otimes\)-edge.

**Proof.** First, assume for contradiction that the least-resolved tree \((T^*, \lambda^*)\) 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 \(\text{lca}_T(x, y) = u\) or \(\text{lca}_T(x, y) = v\). In either case, however, the number of \(m\)-edges between the \(\text{lca}_T(x, y)\) and the leaves \(x\) and \(y\) remains unchanged. Hence, the map \(\varepsilon\) can still be explained by the tree that is obtained from \((T^*, \lambda^*)\) after contraction of \(e\). Thus, \((T^*, \lambda^*)\) 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
Figure 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 $(T, \lambda)$. 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.

Lemma 3.10. Each inner edge in a least-resolved tree $(T^*, \lambda^*)$ for a tree-like map $\varepsilon : [X \times X]_{irr} \rightarrow M^\otimes$, is distinguished by at least one informative triple in $R_e$.

Proof. Consider an arbitrary inner edge $e = uv$ of $T^*$ with $u \nearrow^T v$. Since $(T^*, \lambda^*)$ is phylogenetic, there are necessarily leaves $x, y, z$ such that $\text{lca}(x,y) = v$ and $\text{lca}(x,y,z) = u$. In particular, one can choose $y$ such that $vy$ is an outer $\otimes$-edge, since $(T^*, \lambda^*)$ is least-resolved and due to Lemma 3.9. Moreover, Lemma 3.9 implies that $\lambda^*(e) = m \neq \otimes$. Lemma 3.2 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 $\text{lca}(x,y)$ 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 3.2 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, \( xy|z \in R_e \). Since \( \text{lca}(x,y) = v \) and \( \text{lca}(x,y,z) = u \), the edge \( e \) is by definition distinguished by the triple \( xy|z \in R_e \).

\[ \text{Theorem 3.11.} \quad \text{Let } \varepsilon : [X \times X]_{irr} \to M^\otimes \text{ be a tree-like map and } (T^*, \lambda^*) \text{ be a least-resolved tree that explains } \varepsilon. \text{ Then, the set } R_e \text{ identifies } (T^*, \lambda^*) \text{ and } \text{Aho}(R_e) = T^*. \text{ In particular, } (T^*, \lambda^*) \text{ is unique up to isomorphism.} \]

\[ \text{Proof.} \quad \text{We start with showing that } R_e \text{ identifies } T^*. \text{ If } R_e = \emptyset, \text{ then } (T^*, \lambda^*) \text{ must be a star tree, i.e., an edge-labeled tree that consists of outer edges only. Otherwise, } (T^*, \lambda^*) \text{ contains inner edges that are, by Lemma 3.10, distinguished by at least one informative rooted triple in } R_e, \text{ contradicting that } R_e = \emptyset. \text{ Hence, } r(T^*) = \emptyset, \text{ and therefore, } r(T^*) = \text{cl}(R_e). \text{ Lemma 2.1 implies that } R_e \text{ identifies } (T^*, \lambda^*). \]

In the case \( R_e \neq \emptyset \), assume for contradiction that \( r(T^*) \neq \text{cl}(R_e) \). By Lemma 3.8 we have \( R_e \subseteq r(T^*). \) Isotony of the closure, Theorem 3.1(3) in [5], ensures \( \text{cl}(R_e) \subseteq \text{cl}(r(T^*)) = r(T^*). \) Our assumption therefore implies \( \text{cl}(R_e) \not\subseteq r(T^*), \) and thus the existence of a triple \( ab|c \in r(T^*) \setminus \text{cl}(R_e). \) In particular, therefore, \( ab|c \notin R_e. \) Note that neither \( ac|b \) nor \( bc|a \) can be contained in \( R_e, \) since \( (T^*, \lambda^*) \) explains \( \varepsilon \) and, by assumption, already displays the triple \( ab|c. \)

Thus, \( R_e \) contains no triples on \( \{a, b, c\}. \)

Let \( u = \text{lca}(a, b, c) \) and \( \varepsilon = uv \) be the edge in \( T^* \) with \( u \triangleright_T \ v \triangleright_T \ \text{lca}(a, b) \). By Lemma 3.9, the edge \( \varepsilon \) must be an \( m \)-edge with \( m \neq \emptyset. \) Let \( T_{abc} \) be the subtree of \( (T, \lambda) \) with leaves \( a, b, c. \) Since \( \varepsilon \) is an \( m \)-edge, Lemma 3.2 implies that all edges along the paths from \( v \) to \( a \) and \( v \) to \( b \) must be \( m \)- or \( \otimes \)-edges. However, since \( ab|c \notin R_e, \) the tree \( T_{abc} \) cannot be isomorphic to the subtree \( T_1, \ldots, T_6 \) and thus, both paths from \( \text{lca}(a, b) \) to \( a \) and \( \text{lca}(a, b) \) to \( b \) must contain \( m \)-edges.

Moreover, Lemma 3.9 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 \( a, c \) and \( b, c, d, \) respectively, correspond to one the trees \( T_3, T_4 \) and \( T_6. \) By construction, \( ad|c \in R_e \) and \( bd|c \in R_e. \) Hence, any tree that explains \( \varepsilon \) must display \( ad|c \) and \( bd|c. \) As shown in [10], a tree displaying \( ad|c \) and \( bd|c \) also displays \( ab|c. \) This implies, however, that \( ab|c \in \text{cl}(R_e), \) a contradiction to our assumption.

Therefore, \( \text{cl}(R_e) = r(T) \) and we can apply Lemma 2.1 to conclude that \( R_e \) identifies \( (T^*, \lambda^*) \) and \( \text{Aho}(R_e) = T^*. \)

We continue to show the uniqueness of \( (T^*, \lambda^*) \). Since \( R_e \) identifies \( (T^*, \lambda^*) \), any tree that displays \( R_e \) is by definition a refinement of \( (T^*, \lambda^*) \). In addition, any tree that explains \( \varepsilon \) must display \( R_e \) (cf. Lemma 3.8). Taken the latter two arguments together, any tree that explains \( \varepsilon \) must be a refinement of \( (T^*, \lambda^*) \).

To establish uniqueness of \( (T^*, \lambda^*) \) it remains to show that there is no other labeling \( \lambda \) such that \( (T^*, \lambda) \) still explains \( \varepsilon \). Let \( e = uw \) 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 \( e = uw \) is an inner edge with \( u \triangleright_T v. \) By Lemma 3.9, the edge \( e \) must be an \( m \)-edge and \( m \neq \emptyset. \) and there must be an outer \( \otimes \)-edge \( f = vw. \) Let \( x \) be a leaf such that \( \text{lca}(w, x) = u. \) 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.

\[ \text{■} \]

4 Summary and Outlook

We have considered maps \( \varepsilon : [X \times X]_{irr} \to M^\otimes \) 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 [20, 21]. 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 $\mathcal{E} : [X \times X]_{irr} \rightarrow \mathcal{P}(M) \cup \{\otimes\}$, 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 $(T, \lambda)$ on $X$ such that for at least one $m \in \mathcal{E}(x, y)$ there is an edge on the path from $\lca(x, y)$ to $y$ with label $m$.

of Type-2, if there is an edge-labeled tree $(T, \lambda)$ on $X$ such that for all $m \in \mathcal{E}(x, y)$ there is an edge on the path from $\lca(x, y)$ 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 $\varepsilon : [X \times X]_{irr} \rightarrow N \cup \{\otimes\}$ 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, 29, 35] 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] Alfred V. Aho, Yehoshua Sagiv, Thomas G. Szymanski, and Jeffrey D. Ullman. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal on Computing*, 10(3):405–421, 1981.

[2] Hans-Jürgen Bandelt and Michael Anthony Steel. Symmetric matrices representable by weighted trees over a cancellative abelian monoid. *SIAM Journal on Discrete Mathematics*, 8(4):517–525, 1995.

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

[4] J L Boore. The use of genome-level characters for phylogenetic reconstruction. *Trends Ecol Evol*, 21:439–446, 2006.

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

[6] D. Bryant and M. Steel. Extension Operations on Sets of Leaf-Labeled Trees. *Advances in Applied Mathematics*, 16(4):425–453, December 1995.

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

[8] C. Crespelle and C. Paul. Fully dynamic recognition algorithm and certificate for directed cographs. *Discr. Appl. Math.*, 154:1722–1741, 2006.

[9] Eric J. Deeds, Hooman Hennessey, and Eugene I. Shakhnovich. Prokaryotic phylogenies inferred from protein structural domains. *Genome Res*, 15:393–402, 2005.
[10] M. C. H. Dekker. Reconstruction methods for derivation trees. Master’s thesis, Vrije Universiteit, Amsterdam, Netherlands, 1986.

[11] Frédéric Delsuc, Henner Brinkmann, and Hervé Philippe. Phylogenomics and the reconstruction of the tree of life. Nature Reviews Genetics, 6(5):361–375, 2005.

[12] Alexander Donath and Peter F. Stadler. Molecular morphology: Higher order characters derivable from sequence information. In J. Wolfgang Wägele and Thomas Bartolomaeus, editors, Deep Metazoan Phylogeny: The Backbone of the Tree of Life. New insights from analyses of molecules, morphology, and theory of data analysis, chapter 25, pages 549–562. de Gruyter, Berlin, 2014.

[13] B. E. Dutilh, B. Snel, T. J. Ettema, and M. A. Huynen. Signature genes as a phylogenomic tool. Mol. Biol. Evol., 25:1659–1667, 2008.

[14] A. Ehrenfeucht and G. Rozenberg. Theory of 2-structures, part I: Clans, basic subclasses, and morphisms. Theor. Comp. Sci., 70:277–303, 1990.

[15] A. Ehrenfeucht and G. Rozenberg. Theory of 2-structures, part II: Representation through labeled tree families. Theor. Comp. Sci., 70:305–342, 1990.

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

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

[18] Walter M. Fitch. Distinguishing Homologous from Analogous Proteins. Systematic Biology, 19(2):99–113, June 1970.

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

[20] M. Geiß, J. Anders, P.F. Stadler, N. Wieseke, and M. Hellmuth. Reconstructing gene trees from Fitch’s xenology relation. 2017. arXiv:1711.02152.

[21] M. Geiß, M. Hellmuth, Y. Long, and P.F. Stadler. A short note on undirected fitch graphs. Art Discrete Appl. Math., 1(1):#P1.08, 2018.

[22] S. Grünewald, Y. Long, and Y. Wu. Reconstructing unrooted phylogenetic trees from symbolic ternary metrics. Bulletin of Mathematical Biology, 2018. https://doi.org/10.1007/s11538-018-0413-7.

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

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

[25] M. Hellmuth, M. Hernandez-Rosales, Y. Long, and P.F. Stadler. Inferring phylogenetic trees from the knowledge of rare evolutionary events. J. Math. Biology, 76(7):1623–1653, 2018.

[26] M. Hellmuth, P.F. Stadler, and N. Wieseke. The mathematics of xenology: Di-cographs, symbolic ultrametrics, 2-structures and tree-representable systems of binary relations. J. Math. Biol., 75(1):199–237, 2017.
[27] M. Hellmuth and N. Wieseke. From sequence data incl. orthologs, paralogs, and xenologs to gene and species trees. In *Evolutionary Biology*, pages 373–392, Cham, 2016. Springer International Publishing.

[28] M. Hellmuth, N. Wieseke, M. Lechner, H.-P. Lenhof, M. Middendorf, and P.F. Stadler. Phylogenomics with paralogs. *Proceedings of the National Academy of Sciences*, 112(7):2058–2063, 2015. DOI: 10.1073/pnas.1412770112.

[29] K.T. Huber, G. Scholz, and V. Moulton. Three-way symbolic tree-maps and ultrametrics. *Journal of Classification*, 2018. (in press).

[30] Veiko Krauss, Christian Thümmler, Franziska Georgi, Jörg Lehmann, Peter F. Stadler, and Carina Eisenhardt. Near intron positions are reliable phylogenetic markers: An application to Holometabolous Insects. *Mol. Biol. Evol.*, 25:821–830, 2008.

[31] Sonja J. Prohaska, Claudia Fried, Chris T. Amemiya, Frank H. Ruddle, Günter P. Wagner, and Peter F. Stadler. The shark HoxN cluster is homologous to the human HoxD cluster. *J. Mol. Evol.*, page 58, 2004. 212-217.

[32] I B Rogozin, A V Sverdlov, V N Babenko, and E V Koonin. Analysis of evolution of exon-intron structure of eukaryotic genes. *Brief Bioinform*, 6:118–134, 2005.

[33] A Rokas and P W Holland. Rare genomic changes as a tool for phylogenetics. *Trends Ecol Evol*, 15:454–459, 2000.

[34] Carsten R. Seemann and Marc Hellmuth. The matroid structure of representative triple sets and triple-closure computation. *European Journal of Combinatorics*, 70:384 – 407, 2018.

[35] Charles Semple and Mike Steel. Tree representations of non-symmetric group-valued proximities. *Advances in Applied Mathematics*, 23(3):300 – 321, 1999.

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

[37] A M Shedlock and N Okada. SINE insertions: powerful tools for molecular systematics. *BioEssays*, 22:148–160, 2000.

[38] Mike Steel. *Phylogeny: Discrete and Random Processes in Evolution*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, November 2016.

[39] R L Tatusov, M Y Galperin, D A Natale, and E V Koonin. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, 28(1):33–36, 2000.