Article

Arc-Completion of 2-Colored Best Match Graphs to Binary-Explainable Best Match Graphs

David Schaller [1,2,*] **, Manuela Geiß** [3] **, Marc Hellmuth** [4] **and Peter F. Stadler** [1,2,5,6,7,8]

1   Max Planck Institute for Mathematics in the Sciences, D-04103 Leipzig, Germany; studla@bioinf.uni-leipzig.de
2   Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, Universität Leipzig, D-04107 Leipzig, Germany
3   Software Competence Center Hagenberg GmbH (SCCH), A-4232 Hagenberg, Austria; manuela.geiss@scch.at
4   Department of Mathematics, Faculty of Science, Stockholm University, SE-10691 Stockholm, Sweden; marc.hellmuth@math.su.se
5   German Centre for Integrative Biodiversity, Research (iDiv) Halle-Jena-Leipzig, Competence Center for Scalable Data Services and Solutions, Leipzig Research Center for Civilization Diseases, and Leipzig Research Center for Civilization Diseases (LIFE), Leipzig University, D-04103 Leipzig, Germany
6   Institute for Theoretical Chemistry, University of Vienna, A-1090 Wien, Austria
7   Facultad de Ciencias, Universidad National de Colombia, CO-111321 Bogotá, Colombia
8   Santa Fe Institute, Santa Fe, NM 87501, USA
*   Correspondence: sdavid@bioinf.uni-leipzig.de

Abstract: Best match graphs (BMGs) are vertex-colored digraphs that naturally arise in mathematical phylogenetics to formalize the notion of evolutionary closest genes w.r.t. an a priori unknown phylogenetic tree. BMGs are explained by unique least resolved trees. We prove that the property of a rooted, leaf-colored tree to be least resolved for *some* BMG is preserved by the contraction of inner edges. For the special case of two-colored BMGs, this leads to a characterization of the least resolved trees (LRTs) of binary-explainable trees and a simple, polynomial-time algorithm for the minimum cardinality completion of the arc set of a BMG to reach a BMG that can be explained by a binary tree.

Keywords: best matches; least resolved trees; graph completion; polynomial-time algorithm

1. Introduction

Best match graphs (BMGs) are vertex-colored digraphs that appear in mathematical phylogenetics as a representation of a gene's evolutionary closest relatives in another species [1,2]. That is, given a rooted tree T, a vertex (gene) x in the BMG $G(T, \sigma)$ is colored by the species $\sigma(x)$ in which it resides, and there is an arc (x, y) if there is no other gene y' in species $\sigma(y') = \sigma(y) \neq \sigma(x)$ with a later last common ancestor than the last common ancestor $\text{lca}_T(x, y)$ of x and y in T. Although rooted trees are crucial for the definition of BMGs, they are, however, unknown in practice, and we are often only left with estimates of their BMGs. In general, there are multiple trees that "explain" the same BMG. There is, however, a unique least resolved tree (LRT) for each BMG, which can be obtained from T by contracting certain edges [1]. The LRTs will play a central role in this contribution. The subgraph of a BMG induced by the vertices of some subset of colors is again a BMG. Every BMG therefore can be viewed as the disjoint union of (the arc sets of) 2-colored BMGs (2-BMGs). These 2-BMGs [1,3,4] are bipartite and form a common subclass of the *sink-free* digraphs [5,6] and the *bi-transitive* digraphs [7].

Estimates of graphs from real-life data tend to be affected by noise and thus typically will violate the defining properties of the desired graph class. The solution of a corresponding graph modification problem [8] therefore can by employed as a means of noise reduction, see, e.g., in [9]. The arc modification problems (deletion, completion, and editing) for BMGs are NP-complete, in general [10], and remain hard even for the special case of 2 colors.

Phylogenetic trees are often considered to be binary in theory. Most polytomies are therefore considered a limitation of the available data or method of tree reconstruction [11,12] rather than a biological reality [13,14]. In the setting of BMGs, this distinction is important because not all BMGs can be derived from binary gene trees. Instead, *binary-explainable* BMGs (beBMGs) form a proper subclass [15] that is distinguished by a single forbidden induced subgraph, the *hourglass*, from other BMGs [16]. As "true phylogenies" are often assumed to be binary, BMGs that are not binary-explainable will be considered as a poor model. It is of interest, therefore, to consider the problem of modifying a BMG to a beBMG. Conceptually, this is similar to fully resolving phylogenetic trees starting from a tree with multifurcations. The arc modification problems for beBMGs are, as the more general case, NP-complete [10,15]. Thus, heuristic algorithms must be employed in practice [17]. It is useful to know, therefore, whether an approximation of a graph by a BMG that is not binary-explainable may be helpful towards finding a beBMG. That is, whether a solution to the more general problem makes is easier to find a solution of the constrained problem. To this end, we naturally ask whether the problem of modifying a BMG to a beBMG is as difficult as the general case. It is, in fact, not unusual that graph modification problems that are hard in general become easy when the input is confined to a—usually restrictive—class of graphs, see, e.g., in [18,19]. Here, we show that the problem of completing a 2-colored BMG to a beBMG can indeed be solved in polynomial time.

To prove this result, we make use of the fact that every BMG is associated with a unique *least resolved tree* (LRT). Theorem 1 shows that the property of being the LRT for some BMG is preserved under contraction of inner edges. This observation leads to the explicit construction of a "collapsed tree" from the LRT of the input BMG (G, σ), which not only is the LRT of a 2-colored beBMG, but also minimizes the number of arcs that need to be inserted to obtain a beBMG from (G, σ). The construction does not generalize to more than 2 colors.

2. Notation

We consider simple directed graphs (digraphs) $G = (V, E)$ with vertex set V and *arc* set $E \subseteq V \times V \setminus \{(v, v) \mid v \in V\}$ and rooted (undirected) trees T with root ρ. Correspondingly, we write (x, y) for directed arcs from x to y, and xy for undirected tree edges. We write $G[W]$ for the subgraph of G induced by a set of vertices $W \subseteq V$. Given a tree T, we write $V(T)$ and $E(T)$ for its set of vertices and edges, respectively; $L(T)$ for the set of leaves; and $V^0(T) = V(T) \setminus L(T)$ for the set of inner vertices.

A vertex coloring of a graph is a map $\sigma : V \to S$, where S is a non-empty set of colors. A vertex coloring of G is *proper* if $\sigma(x) \neq \sigma(y)$ for all $(x, y) \in E(G)$. We will also consider *leaf-colorings* $\sigma \colon L(T) \to S$ for trees T and denote by (G, σ) and (T, σ) vertex-colored (di)graphs and leaf-colored trees, respectively.

Given a rooted tree, we write $x \preceq_T y$ if y is an *ancestor* of x, i.e., if y lies along the unique path from ρ to x in T. We write $x \prec_T y$ if $x \preceq_T y$ and $x \neq y$. The relation $\preceq_T$ is a partial order on T. If $xy \in E(T)$ and $x \prec_T y$, then y is the unique *parent* of x, denoted by $\mathrm{par}_T(x)$, and x a *child* of y. The set of children of a vertex $u \in V(T)$ is denoted by $\mathrm{child}_T(u)$. A rooted tree T is phylogenetic if every inner vertex $x \in V^0(T)$ has at least two children. All trees in this contribution are assumed to be phylogenetic. Furthermore, we write $T(u)$ for the subtree rooted in u, i.e., $V(T(u)) = \{y \in V(T) \mid y \preceq_T u\}$. The *last common ancestor* of a non-empty subset $A \subseteq V(T)$ is the unique $\preceq_T$-minimal vertex of T, that is, an ancestor of every $u \in A$. For convenience, we write $\mathrm{lca}_T(x, y, \dots)$ instead of $\mathrm{lca}_T(\{x, y, \dots\})$.

A triple $xy|z$ is a rooted tree with the three leaves x, y, and z such that $\mathrm{lca}_T(x, y) \prec \mathrm{lca}_T(x, y, z)$. If $e \in E(T)$, we denote by T_e the tree obtained by contracting the edge e. We will only be interested in contractions of inner edges, i.e., those that preserve the leaf set. We say that T *displays* a tree T', in symbols $T' \leq T$, if T' can be obtained from T as the minimal subtree of T that connects all elements in $L(T')$ with root $\mathrm{lca}_T(L(T'))$ and by suppressing all inner vertices that only have one child left, which can, e.g., be achieved by a stepwise contraction of one of their two incident edges until no such vertices remain.

3. Best Match Graphs, Least Resolved Trees, and Binary-Explainable BMGs

In this section, we first summarize some properties of best match graphs and their least resolved trees. We then show that the contraction of inner edges in least resolved trees always leads to least resolved trees. Furthermore, we recall some properties of binary-explainable best match graphs that will be needed later.

Definition 1. *Let (T, σ) be a leaf-colored tree. A leaf $y \in L(T)$ is a* best match *of the leaf $x \in L(T)$ if $\sigma(x) \neq \sigma(y)$ and $\operatorname{lca}_T(x, y) \preceq_T \operatorname{lca}_T(x, y')$ holds for all leaves y' of color $\sigma(y') = \sigma(y)$.*

Given (T, σ), the digraph $G(T, \sigma) = (V, E)$ with vertex set $V = L(T)$, vertex-coloring σ, and with arcs $(x, y) \in E$ if and only if y is a best match of x w.r.t. (T, σ) is called the *best match graph* (BMG) of (T, σ) [1]:

Definition 2. *An arbitrary vertex-colored digraph (G, σ) is a* best match graph (BMG) *if there exists a leaf-colored tree (T, σ) such that $(G, \sigma) = G(T, \sigma)$. In this case, we say that (T, σ) explains (G, σ).*

Proposition 1 ([16], Lemma 8). *If T_A is obtained from a tree T by contracting all edges in a subset A of inner edges in T, then $G(T, \sigma) \subseteq G(T_A, \sigma)$.*

An edge e of a leaf-colored tree is *redundant* (w.r.t. (G, σ)) if it can be contracted without affecting the BMG, i.e., if $G(T, \sigma) = G(T_e, \sigma)$.

Definition 3. *A leaf-colored tree (T, σ) is* least resolved *if there is no non-empty subset $A \subseteq E(T)$ such that $G(T, \sigma) = G(T_A, \sigma)$.*

We define the notion of being least resolved here as a property of the tree (T, σ) alone. Of course, every least resolved tree is also *least resolved w.r.t. some BMG*, namely, the (uniquely defined) digraph $G(T, \sigma)$.

It is shown in [1] that (T, σ) is least resolved if and only if it does not contain a redundant edge.

Proposition 2 ([1], Thm. 8). *Every BMG (G, σ) is explained by a unique least resolved tree (LRT), which is obtained from an arbitrary tree (T, σ) that explains (G, σ) by contraction of all redundant edges of (T, σ).*

In particular, therefore, there is a bijection between BMGs and LRTs. Surprisingly, the property of being least resolved for some BMG is preserved under contraction of inner edges of T.

Theorem 1. *Suppose (T, σ) is least resolved and let A be a set of inner edges of T, and denote by T_A the tree obtained from a tree T by contracting all edges in A. Then, (T_A, σ) is again least resolved.*

Proof. Assume that (T, σ) is least resolved, i.e., it does not contain any redundant edges, and set $(G, \sigma) := G(T, \sigma)$. Lemma 7 in [16] states that an inner edge $e = uv$ with $v \prec_T u$ in (T, σ) is non-redundant if and only if there is an arc $(a, b) \in E(G)$ such that $\operatorname{lca}_T(a, b) = v$ and $\sigma(b) \in \sigma(L(T(u)) \setminus L(T(v)))$. The statement trivially holds if (T, σ) has at most one inner edge. Therefore, we assume that (T, σ) has at least two distinct inner edges $e = uv$ and e'. We show that every non-redundant edge e in T remains non-redundant in $T_{e'}$. Thus, let e be a non-redundant edge in T. Therefore, there is an arc $(a, b) \in E(G)$ such that $\operatorname{lca}_T(a, b) = v$ and $\sigma(b) \in \sigma(L(T(u)) \setminus L(T(v)))$. Now consider the tree $T_{e'}$ obtained from T by contraction of the inner edge $e' \neq e$. Clearly, we also have $\operatorname{lca}_{T_{e'}}(a, b) = v$ and $\sigma(b) \in \sigma(L(T_{e'}(u)) \setminus L(T_{e'}(v)))$. Proposition 1 implies $G(T, \sigma) \subseteq G(T_{e'}, \sigma)$, and thus $(a, b) \in E(G(T_{e'}, \sigma))$. Making use of the characterization of redundant edges in ([16], Lemma 7) again, we conclude that e is non-redundant in $(T_{e'}, \sigma)$.

As both e and e' were chosen arbitrarily, we observe that the contraction of a single inner edge does not produce new redundant edges. We can therefore apply this argument for each step in the consecutive contraction of all edges in A (in an arbitrary order) to conclude that (T_A, σ) does not contain redundant edges. Therefore, Proposition 2 implies that (T_A, σ) is least resolved.　□

Corollary 1. *If (T, σ) is least resolved and A is a non-empty set of inner edges of T, then $G(T, \sigma) \subsetneq G(T_A, \sigma)$.*

Proof. By Proposition 1, we have $G(T, \sigma) \subseteq G(T_A, \sigma)$. By Theorem 1, (T_A, σ) is least resolved. As the LRT of a BMG is unique (cf. Proposition 2), we have $G(T, \sigma) \neq G(T_A, \sigma)$.　□

As another immediate consequence of Theorem 1 and uniqueness of the LRT of a BMG (Proposition 2), we obtain the following.

Corollary 2. *If e and e' are two distinct inner edges of a least resolved tree (T, σ), then $G(T_e, \sigma) \neq G(T_{e'}, \sigma)$.*

Let us now turn to the subclass of BMGs that can be explained by a binary tree.

Definition 4. *A* binary-explainable BMG *(beBMG) is a BMG (G, σ) such that there is a binary leaf-colored tree (T, σ) that explains (G, σ).*

As shown in [16], beBMGs can be characterized among BMGs by means of a simple forbidden colored induced subgraph:

Definition 5. *An* hourglass *in a properly vertex-colored digraph (G, σ), denoted by $[xy \Join x'y']$, is a subgraph $(G[Q], \sigma_{|Q})$ induced by a set of four pairwise distinct vertices $Q = \{x, x', y, y'\} \subseteq V(G)$ such that (i) $\sigma(x) = \sigma(x') \neq \sigma(y) = \sigma(y')$, (ii) $(x, y), (y, x)$ and $(x'y'), (y', x')$ are bidirectional arcs in G, (iii) $(x, y'), (y, x') \in E(G)$, and (iv) $(y', x), (x', y) \notin E(G)$.*

An hourglass together with a (non-binary) tree explaining it is illustrated in Figure 1A. A properly vertex-colored digraph that does not contain an hourglass as an induced subgraph is called *hourglass-free*.
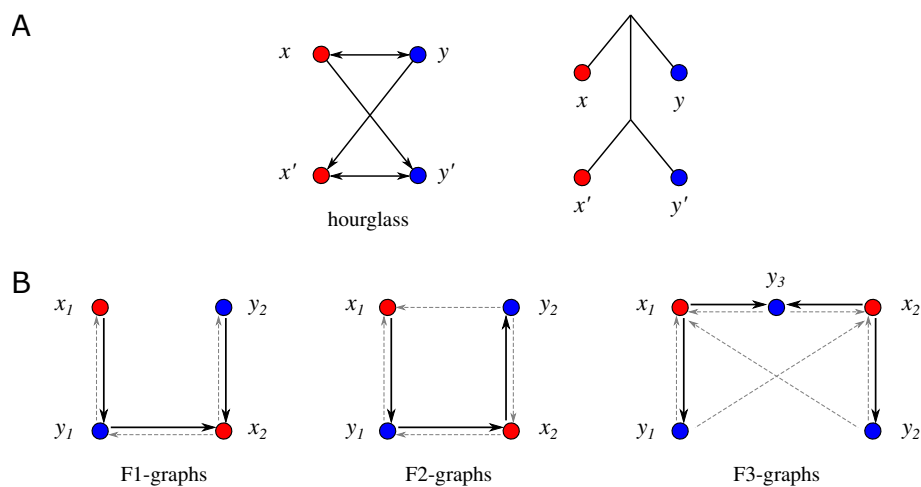

Figure 1. (**A**) An hourglass as the characteristic forbidden induced subgraph of beBMGs and its non-binary explaining tree. (**B**) The three classes of forbidden induced subgraphs of 2-colored BMGs (see Definition 6 below). The gray-dashed arcs may or may not exist.

Proposition 3 ([16], Lemma 31 and Prop. 8). *For every BMG (G, σ) explained by a tree (T, σ), the following three statements are equivalent:*

1. (G, σ) *is binary-explainable.*
2. (G, σ) *is hourglass-free.*
3. *There is no vertex $u \in V^0(T)$ with three distinct children—v_1, v_2, and v_3—and two distinct colors—r and s—satisfying*
   - (a) $r \in \sigma(L(T(v_1)))$, $r, s \in \sigma(L(T(v_2)))$, *and* $s \in \sigma(L(T(v_3)))$, *and*
   - (b) $s \notin \sigma(L(T(v_1)))$, *and* $r \notin \sigma(L(T(v_3)))$.

Note that the LRTs of beBMGs are usually not binary. In fact, it is shown in [15] that, for a beBMG (G, σ), there exists a unique *binary refinable tree* (BTR) $B(G, \sigma)$ with the property that every binary tree (T, σ) that displays $B(G, \sigma)$ explains (G, σ). The BRT is in general much better resolved than the LRT of (G, σ).

4. Two-Colored BMGs

Let us now briefly focus on 2-colored BMGs (2-BMGs). As arcs in BMG can only connect vertices with different colors, every 2-BMG is bipartite. Furthermore, every leaf x in a tree with two leaf colors has at least one best match y. Every 2-BMG is therefore *sink-free*, i.e., every vertex has at least one out-neighbor. Furthermore, Schaller et al. [10] showed that the following graphs (see also Figure 1B) are forbidden induced subgraphs for 2-BMGs.

Definition 6 (F1-, F2-, and F3-graphs)**.**

(F1) *A properly 2-colored digraph on four distinct vertices $V = \{x_1, x_2, y_1, y_2\}$ with coloring $\sigma(x_1) = \sigma(x_2) \neq \sigma(y_1) = \sigma(y_2)$ is an* F1-graph *if $(x_1, y_1), (y_2, x_2), (y_1, x_2) \in E$ and $(x_1, y_2), (y_2, x_1) \notin E$.*

(F2) *A properly 2-colored digraph on four distinct vertices $V = \{x_1, x_2, y_1, y_2\}$ with coloring $\sigma(x_1) = \sigma(x_2) \neq \sigma(y_1) = \sigma(y_2)$ is an* F2-graph *if $(x_1, y_1), (y_1, x_2), (x_2, y_2) \in E$ and $(x_1, y_2) \notin E$.*

(F3) *A properly 2-colored digraph on five distinct vertices $V = \{x_1, x_2, y_1, y_2, y_3\}$ with coloring $\sigma(x_1) = \sigma(x_2) \neq \sigma(y_1) = \sigma(y_2) = \sigma(y_3)$ is an* F3-graph *if $(x_1, y_1), (x_2, y_2), (x_1, y_3), (x_2, y_3) \in E$ and $(x_1, y_2), (x_2, y_1) \notin E$.*

Proposition 4 ([10], Thm. 4.4)**.** *A properly 2-colored digraph is a BMG if and only if it is sink-free and does not contain an induced F1-, F2-, or F3-graph.*

A peculiar property of 2-BMGs is that their LRTs can be constructed efficiently by recursively decomposing an input 2-BMG into non-trivial induced subgraphs and individual vertices [15]. Although we will not need this construction here, one of its corner stones plays an important role:

Definition 7 (Support Leaves)**.** *For a given tree T, the set $S_u := \text{child}_T(u) \cap L(T)$ is the set of all support leaves of vertex $u \in V(T)$.*

We note in passing that every inner vertex u of the LRT of a 2-BMG (G, σ), with the possible exception of the root ρ, has a non-empty set of support leaves S_u, and $S_\rho \neq \emptyset$ if and only if (G, σ) is connected [20]. In the following, we will make use of a connection between a 2-BMG and its LRT.

Lemma 1. *Let (G, σ) be a 2-BMG, (T, σ) its LRT, and $x, y \in L(T) = V(G)$. Then, $(x, y) \in E(G)$ if and only if $\sigma(x) \neq \sigma(y)$ and $y \in L(T(\text{par}_T(x)))$.*

Proof. First note that as (G, σ) is 2-colored, (T, σ) has at least two leaves and $u := \text{par}_T(x)$ is always defined. First, assume $\sigma(x) \neq \sigma(y)$, and thus $x \neq y$, and let $y \in L(T(u))$. As x is a child of u, we have $\text{lca}_T(x, y) = u$. Moreover, as u is the parent of x, there is no vertex y'

of color $\sigma(y)$ such that $\mathrm{lca}_T(x, y') \prec_T \mathrm{lca}_T(x, y) = u$. Therefore, y is a best match of x, i.e., $(x, y) \in E(G)$.

Now suppose, for contraposition, that $\sigma(x) = \sigma(y)$ or $y \notin L(T(u))$. If $\sigma(x) = \sigma(y)$, then, by definition, $(x, y) \notin E(G)$. If $y \notin L(T(u))$, then $u \prec_T \rho_T$. Therefore, we can apply Corollary 1 in [20] to the inner vertex u to conclude that $|\sigma(L(T(u)))| > 1$, i.e., the subtree $L(T(u))$ contains both colors. Thus, we can find a vertex y' of color $\sigma(y)$ such that $\mathrm{lca}_T(x, y') \preceq_T u \prec_T \mathrm{lca}_T(x, y)$ which implies that $(x, y) \notin E(G)$. $\square$

As an immediate consequence, we find the following.

Corollary 3. *Let (G, σ) be a 2-BMG, (T, σ) its LRT and $x, y \in V(G) = L(T)$. Then, $(x, y), (y, x) \in E(G)$ if and only if $\sigma(x) \neq \sigma(y)$ and $\mathrm{par}_T(x) = \mathrm{par}_T(y)$.*

5. Completion of a 2-BMG to a 2-beBMG

Writing $G + F := (G, E \cup F)$ for a digraph $G = (V, E)$ and arc set $F \subseteq V \times V \setminus \{(v, v) \mid v \in V\}$, consider the following graph completion problem:

Problem 1 (2-BMG COMPLETION RESTRICTED TO BINARY-EXPLAINABLE GRAPHS (2-BMG CBEG))**.**

Input:	*A properly 2-colored digraph $(G = (V, E), \sigma)$ and an integer k.*
Question:	*Is there a subset $F \subseteq V \times V \setminus (\{(v, v) \mid v \in V\} \cup E)$ such that $|F| \leq k$ and $(G + F, \sigma)$ is a binary-explainable 2-BMG?*

In the general case, 2-BMG CBEG is NP-complete ([10], Cor. 5.11). Here, we are interested in the restriction of the 2-BMG CBEG problem with BMGs as input.

The following result holds for BMGs and their completions to beBMGs with an arbitrary number of colors.

Lemma 2. *Let (G', σ) be a completion of a BMG (G, σ) to a beBMG, and let $[xy \bowtie x'y']$ be an induced hourglass in (G, σ). Then, (G', σ) contains both arcs (x', y) and (y', x).*

Proof. It is shown in ([1], Obs. 1) that the subgraphs of a BMG induced by all vertices with any two given colors is a 2-BMG. As (G', σ) is a (binary-explainable) BMG, all of its 2-colored induced subgraphs are therefore 2-BMGs. By assumption, (G, σ) is not binary-explainable as it contains the hourglass $[xy \bowtie x'y']$ as an induced subgraph (cf. Proposition 3). The hourglass contains all possible arcs between vertices of different colors except (x', y) and (y', x). As (G', σ) contains no hourglass and G' is a completion of G, i.e., $E(G) \subseteq E(G')$, we conclude that (G', σ) contains at least one of the arcs (x', y) and (y', x).

Assume for contradiction that, w.l.o.g., (G', σ) only contains (x', y). We have (y', x'), $(y, x) \in E(G')$ and $\sigma(y') = \sigma(y) \neq \sigma(x') = \sigma(x)$ by the definition of hourglasses, and by assumption $(x', y) \in E(G')$ and $(y', x) \notin E(G')$. Therefore, the four vertices x, x', y, y' induce an F2-graph in (G', σ). By Proposition 4, the 2-colored subgraph of (G', σ) induced by the two colors $\sigma(x)$ and $\sigma(y)$ is not a BMG. Consequently, (G', σ) is not a BMG either, it is a contradiction. Therefore, (G', σ) contains both arcs (x', y) and (y', x). $\square$

Definition 8. *Let (T, σ) be a tree with a 2-colored leaf set, i.e., $|\sigma(L(T))| = 2$. Denote by (T^*, σ) the* collapsed tree *obtained from (T, σ) by contraction of all inner edges in $T(u)$ for all $u \in V^0(T)$ that have support leaves of both colors.*

In other words, (T^*, σ) is obtained from (T, σ) by collapsing every subtree $T(u)$ to a star if u has support leaves of both colors.

Lemma 3. *The collapsed tree (T^*, σ) of (T, σ) is uniquely defined and can be computed from (T, σ) in $O(|V(T)|)$-time.*

Proof. The collapsed tree (T^*, σ) is well defined because whenever $v \prec_T u$, collapsing the subtree $T(v)$ to a star does not change the set of support leaves S_u. Similarly, collapsing $T(v)$ if v is not $\prec_T$-comparable with u does not change S_u. Thus, (T^*, σ) is uniquely defined. To see that (T^*, σ) can be computed in $O(|V(T)|)$ operations, we observe that it suffices to collapse all subtrees $T(u)$ such that $u \in V^0(T)$ has support leaves of both colors and there is no $u' \prec_T u$ with this property, i.e., u is $\preceq_T$-maximal in that sense. These vertices u for which $T(u)$ is replaced by a star are found by a top-down traversal of T and evaluating $|\sigma(S_u)|$, all of which can be computed in linear total time. $\square$

As an immediate consequence of the uniqueness of T^* and the construction in the second part of the proof of Lemma 3, we obtain the following.

Corollary 4. *The collapsed tree (T^{**}, σ) of a collapsed tree (T^*, σ) satisfies $T^{**} = T^*$.*

Lemma 4. *If (T^*, σ) is the collapsed tree of an LRT (T, σ) with 2-colored leaf set, then $G(T^*, \sigma)$ is binary-explainable.*

Proof. As the collapsed tree (T^*, σ) is obtained from the LRT (T, σ) by contraction of edges, Theorem 1 implies that (T^*, σ) is also least resolved. Now suppose, for contradiction, that $G(T^*, \sigma)$ is not binary-explainable. By, Proposition 3(3), (T^*, σ) has a vertex $u \in V^0(T^*)$ with three distinct children—v_1, v_2, and v_3—and two distinct colors—r and s— satisfying (i) $r \in \sigma(L(T^*(v_1)))$, $r, s \in \sigma(L(T^*(v_2)))$, and $s \in \sigma(L(T^*(v_3)))$, and (ii) $s \notin \sigma(L(T^*(v_1)))$ and $r \notin \sigma(L(T^*(v_3)))$. As (G, σ) is only 2-colored, the latter arguments imply that $|\sigma(L(T^*(v_1)))| = |\sigma(L(T^*(v_3)))| = 1$ and $|\sigma(L(T^*(v_2)))| = 2$. As moreover (T^*, σ) is least resolved and none of the vertices v_1, v_2, and v_3 is the root of T^*, we can apply Corollary 1 in [20] to conclude that v_1 and v_2 are leaves, and that v_3 is an inner vertex, respectively. In particular, $\sigma(v_1) = r \neq s = \sigma(v_3)$. Therefore, $T^*(u)$ is not a star tree and u has support leaves of both colors in T^*; a contradiction to its construction. Therefore, we conclude that $G(T^*, \sigma)$ is binary-explainable. $\square$

Theorem 2. *The optimization version of 2-BMG CBEG with a 2-BMG (G, σ) as input has the unique solution $F := E(G(T^*, \sigma)) \setminus E(G)$, where (T^*, σ) is the collapsed tree of the LRT (T, σ) of (G, σ).*

Proof. First note that the optimization version of 2-BMG CBEG always has a solution. To see this, consider the complete bipartite and properly 2-colored digraph (G', σ) with vertex set $V(G)$. This digraph is explained by the star tree with leaf set $V(G)$. Moreover, (G', σ) is clearly hourglass-free as hourglasses require non-arcs (between vertices of distinct colors). By Proposition 3, the BMG (G', σ) is binary-explainable.

Now, consider the collapsed tree (T^*, σ) of (T, σ). As T^* is obtained from T by contraction of inner edges, Proposition 1 implies $(G, \sigma) = G(T, \sigma) \subseteq G(T^*, \sigma) =: (G^*, \sigma)$. Furthermore, (G^*, σ) is binary-explainable by Lemma 4. Therefore, (G^*, σ) is a valid completion of (G, σ) to a beBMG.

We continue by showing the existence of certain arcs in every (not necessarily optimal) completion (G', σ) of (G, σ) to a beBMG. To this end, consider a $\preceq_T$-maximal vertex u such that the subtree $T(u)$ is not a star tree and u has support leaves S_u of both colors in T. We will make frequent use of the fact that $E(G) \subseteq E(G')$. We consider the following cases in order to show that all arcs between vertices $x, y \in L(T(u))$ with $\sigma(x) \neq \sigma(y)$ exist in (G', σ):

(i) $\quad x, y \in S_u$,
(ii) $\quad x \in L(T(u)) \setminus S_u$ and $y \in S_u$, and
(iii) $\quad x, y \in L(T(u)) \setminus S_u$.

In Case (i), the leaves x and y are both children of u. Together with Corollary 3, this implies $(x, y), (y, x) \in E(G) \subseteq E(G')$.

In Case (ii), we can find a vertex $x' \in S_u$ of color $\sigma(x)$ as S_u contains vertices of both colors. As in Case (i), we have $(x', y), (y, x') \in E(G) \subseteq E(G')$. As $x \in L(T(u)) \setminus S_u$, we can conclude that $v := \mathsf{par}_T(x) \prec_T u$ by the definition of support leaves. Therefore, the inner vertex v is not the root of T and we can apply Cor. 1 in [20] to conclude that the subtree $T(v)$ of the inner vertex v contains both colors. The latter together with Lemma 10 in [21] implies that there are arcs $(x'', y''), (y'', x'') \in E(G) \subseteq E(G')$ with $x'', y'' \in L(T(v))$ and $\sigma(x) = \sigma(x'') \neq \sigma(y) = \sigma(y'')$. Note that $x = x''$ is possible. As x, x'', y'' in $L(T(v)) \subset L(T(u))$, $x', y \in L(T(u)) \setminus L(T(v))$ and $v \prec_T u$, we can apply Lemma 1 to conclude that $(x', y''), (y, x), (y, x'') \in E(G) \subseteq E(G')$ and $(y'', x'), (x, y), (x'', y) \notin E(G) \subseteq E(G')$. Together with $(x', y), (y, x'), (x'', y''), (y'', x'') \in E(G)$ and the coloring, this implies that x', y, x'', y'' induce an hourglass $[x'y \bowtie x''y'']$ in (G, σ). By Lemma 2, we have arcs $(x'', y), (y'', x') \in E(G')$. If $x = x''$, we immediately obtain $(x, y), (y, x) \in E(G')$. Now, suppose $x \neq x''$, i.e., it remains to show that $(x, y) \in E(G')$. Thus assume, for contradiction, that $(x, y) \notin E(G')$. Lemma 1 together with $\sigma(x) \neq \sigma(y'')$ and $y'' \in L(T(\mathsf{par}_T(x) = v))$ implies that $(x, y'') \in E(G) \subseteq E(G')$. Therefore, we have the arcs $(x, y''), (y'', x'), (x', y) \in E(G')$ but $(x, y) \notin E(G')$, i.e., x, x', y, y'' induce a forbidden F2-graph. Together with Proposition 4, this is a contradiction to (G', σ) being a 2-BMG. Therefore, we conclude that $(x, y) \in E(G')$.

In Case (iii), we have $x, y \in L(T(u)) \setminus S_u$. We can find two vertices $x', y' \in S_u$, which are distinct from x and y, and satisfy $\sigma(x) = \sigma(x') \neq \sigma(y) = \sigma(y')$. From Cases (i) and (ii), we obtain $(x', y'), (y', x') \in E(G')$ and $(x', y), (y, x'), (x, y'), (y', x) \in E(G')$, respectively. Now assume for contradiction that $(x, y) \notin E(G')$. Thus, we have $(x, y'), (y', x'), (x', y) \in E(G')$ and $(x, y) \notin E(G')$, i.e., x, x', y, y' induce a forbidden F2-graph in (G', σ); a contradiction to (G', σ) being a 2-BMG. Therefore, we conclude that $(x, y) \in E(G')$. The existence of the arc $(y, x) \in E(G')$ can be shown by analogous arguments.

We will now show that $E(G^*) \subseteq E(G')$ for every (not necessarily optimal) completion (G', σ) of the 2-BMG (G, σ) to a beBMG. To this end, consider an arbitrary arc $(x, y) \in E(G^*)$. If $(x, y) \in E(G)$, then $(x, y) \in E(G')$ follows immediately. Now, assume that $(x, y) \in F = E(G^*) \setminus E(G)$. As (G, σ) is a 2-BMG and thus properly-colored and sink-free (cf. Proposition 4), there must be a vertex y' of color $\sigma(y)$ such that $(x, y') \in E(G)$. As $(x, y) \notin E(G)$, we have $\mathsf{lca}_T(x, y') \prec_T \mathsf{lca}_T(x, y)$ and thus the LRT (T, σ) displays the triple $xy'|y$. However, $(x, y), (x, y') \in E(G^*)$ implies that (T^*, σ) does not display the triple $xy'|y$, i.e., all edges on the path from $\mathsf{lca}_T(x, y')$ to $\mathsf{lca}_T(x, y)$ have been contracted. Therefore, there is a $\preceq_T$-maximal inner vertex $u \in V^0(T)$ such that $x, y \in L(T(u))$, $T(u)$ is not a star tree and u has support leaves of both colors in T. By the arguments above, we can conclude that $(x, y) \in E(G')$.

In summary, F is a solution for 2-BMG CBEG with the 2-BMG (G, σ) (and some integer $k \geq |F|$) as input, and $F \subseteq F'$ for every other solution $F' = E(G') \setminus E(G)$. Therefore, we conclude that F is the unique optimal solution. □

As a direct consequence of Theorem 2, the fact that LRTs can be constructed in $O(|V| + |E| \log^2 |V|)$ (cf. [20]) and Lemma 3, we have

Corollary 5. 2-BMG CBEG *with a 2-BMG as input can be solved in* $O(|V| + |E| \log^2 |V|)$ *time.*

We also immediately obtain a characterization of the LRTs of 2-beBMGs.

Corollary 6. *A 2-colored least resolved tree* (T, σ) *is the LRT of 2-beBMG if and only if it is a collapsed tree.*

6. Concluding Remarks

Starting from the observation that the property of being least resolved is preserved under contraction of inner edges, we have obtained a characterization of the LRTs that explains 2-colored beBMGs. The construction of these "collapsed trees" corresponds to the

completion of BMGs to beBMGs, resulting in a simple, polynomial-time algorithm for this problem. This result is primarily of theoretical interest.

In contrast to the 2-colored case, ℓ-BMG CBEG with a BMG as input and $\ell \geq 3$ in general does not have a unique optimal solution. In the example in Figure 2, the missing arcs (a_2, b_1) and (b_2, a_1) in the induced hourglass $[a_1 b_1 \bowtie a_2 b_2]$ must be inserted. The resulting digraph is not a BMG. To obtain a beBMG, it suffices to insert in addition either the arc (c, a_1) or the arc (c, b_1) (cf. Proposition 3). We suspect, therefore, that ℓ-BMG CBEG does not admit an efficient solution in general. The solutions of 2-BMG CBEG problems for all 2-colored induced subgraphs that are not binary-explainable are nevertheless an appealing starting point for constructing heuristics for ℓ-BMG CBEG. We refer to the work in [17] for a detailed analysis of a class of approximation algorithms for BMG and beBMG modification.
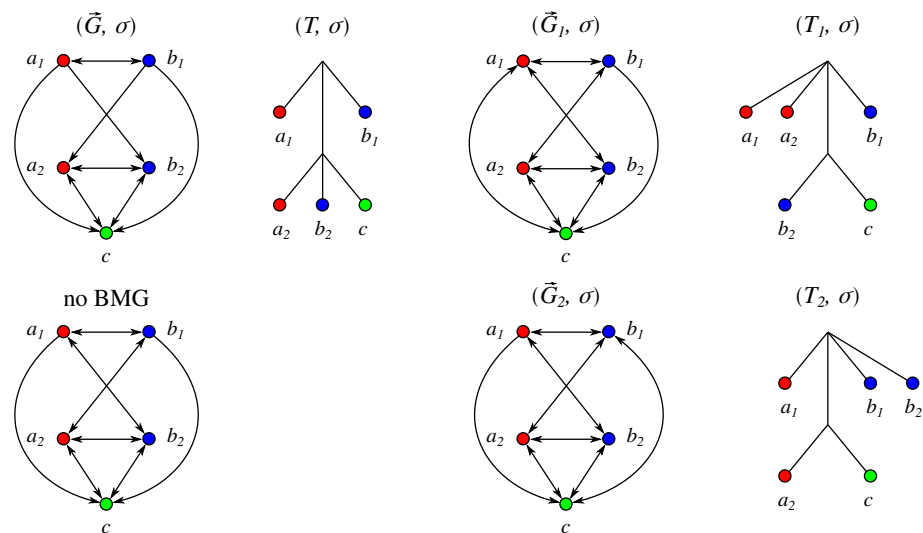

Figure 2. Example for 3-BMG CBEG with the 3-BMG (G, σ) (explained by the LRT (T, σ)) as input that has no unique optimal solution. Insertion of the missing arcs (a_2, b_1) and (b_2, a_1) produces a digraph that is not a BMG. At least one of the arcs (c, a_1) or (c, b_1) has to be inserted additionally to obtain the beBMGs (G_1, σ) and (G_2, σ) (shown with their LRTs (T_1, σ) and (T_2, σ)), respectively.

The simple solution of 2-BMG CBEG begs the question whether other arc modification problems for beBMGs, in particular the corresponding deletion and editing problems, have a similar structure. This does not seem to be case however. Neither 2-BMG EBEG nor 2-BMG DBEG with a 2-BMG as input has a unique optimal solution. To see this, consider the 2-BMG consisting of the hourglass $[xy \bowtie x'y']$ which is explained by the unique non-binary tree $(x, y, (x', y'))$ (in Newick format, see also Figure 1A). Deletion of the arcs (x, y) or (y, x) results in a digraph that is explained by the binary trees $(y, (x, (x', y')))$ or $(x, (y, (x', y')))$, respectively. We suspect that a BMG as input does not make these problems easier than the general case—the complexity of which remains an open question however.

Abbreviations

The following abbreviations are used in this manuscript:

BMG   Best Match Graph
beBMG   Binary-explainable Best Match Graph
LRT   Least resolved tree

References

1. Geiß, M.; Chávez, E.; González Laffitte, M.; López Sánchez, A.; Stadler, B.M.R.; Valdivia, D.I.; Hellmuth, M.; Hernández Rosales, M.; Stadler, P.F. Best Match Graphs. *J. Math. Biol.* **2019**, *78*, 2015–2057. [CrossRef]
2. Schaller, D.; Geiß, M.; Chávez, E.; González Laffitte, M.; López Sánchez, A.; Stadler, B.M.R.; Valdivia, D.I.; Hellmuth, M.; Hernández Rosales, M.; Stadler, P.F. Corrigendum to "Best Match Graphs". *J. Math. Biol.* **2021**, in press.
3. Korchmaros, A. The Structure of 2-Colored Best Match Graphs. *arXiv* **2020**, arXiv:2009.00447v3.
4. Korchmaros, A. Circles and Paths in 2-Colored Best Match Graphs. *arXiv* **2020**, arXiv:2006.04100v1.
5. Cohn, H.; Pemantle, R.; Propp, J.G. Generating a random sink-free orientation in quadratic time. *Electr. J. Comb.* **2002**, *9*, R10. [CrossRef]
6. Abrams, G.; Sklar, J.K. The Graph Menagerie: Abstract Algebra and the Mad Veterinarian. *Math. Mag.* **2010**, *83*, 168–179. [CrossRef]
7. Das, S.; Ghosh, P.; Ghosh, S.; Sen, S. Oriented Bipartite Graphs and the Goldbach Graph. *arXiv* **2020**, arXiv:1611.10259v6.
8. Natanzon, A.; Shamir, R.; Sharan, R. Complexity Classification of Some Edge Modification Problems. *Discr. Appl. Math.* **2001**, *113*, 109–128. [CrossRef]
9. Hellmuth, M.; Wieseke, N.; Lechner, M.; Lenhof, H.P.; Middendorf, M.; Stadler, P.F. Phylogenetics from Paralogs. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 2058–2063. [CrossRef] [PubMed]
10. Schaller, D.; Stadler, P.F.; Hellmuth, M. Complexity of Modification Problems for Best Match Graphs. *Theor. Comp. Sci.* **2021**, *865*, 63–84. [CrossRef]
11. Maddison, W. Reconstructing character evolution on polytomous cladograms. *Cladistics* **1989**, *5*, 365–377. [CrossRef]
12. DeSalle, R.; Absher, R.; Amato, G. Speciation and phylogenetic resolution. *Trends Ecol. Evol.* **1994**, *9*, 297–298. [CrossRef]
13. Hoelzer, G.A.; Meinick, D.J. Patterns of speciation and limits to phylogenetic resolution. *Trends Ecol. Evol.* **1994**, *9*, 104–107. [CrossRef]
14. Slowinski, J.B. Molecular Polytomies. *Mol. Phylog. Evol.* **2001**, *19*, 114–120. [CrossRef] [PubMed]
15. Schaller, D.; Geiß, M.; Hellmuth, M.; Stadler, P.F. Best Match Graphs with Binary Trees. In *Proceedings of the 8th International Conference on Algorithms for Computational Biology; Missoula, MO, USA, 8–11 November 2021*; Lecture Notes in Computer Science; Springer Nature: Cham, Switzerland, 2021; Volume 12715, in press, arXiv:2011.00511.
16. Schaller, D.; Geiß, M.; Stadler, P.F.; Hellmuth, M. Complete Characterization of Incorrect Orthology Assignments in Best Match Graphs. *J. Math. Biol.* **2021**, *82*, 20. [CrossRef] [PubMed]
17. Schaller, D.; Geiß, M.; Hellmuth, M.; Stadler, P.F. Heuristic Algorithms for Best Match Graph Editing. *arXiv* **2021**, arXiv:2103.07280.
18. Liu, Y.; Wang, J.; Guo, J.; Chen, J. Cograph Editing: Complexity and Parametrized Algorithms. In *COCOON 2011*; Fu, B., Du, D.Z., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2011; Volume 6842, pp. 110–121. [CrossRef]
19. Gao, Y.; Hare, D.R.; Nastos, J. The cluster deletion problem for cographs. *Discret. Math.* **2013**, *313*, 2763–2771. [CrossRef]
20. Schaller, D.; Geiß, M.; Hellmuth, M.; Stadler, P.F. Least resolved trees for two-colored best match graphs. *arXiv* **2021**, arxiv:2101.07000.
21. Geiß, M.; Stadler, P.F.; Hellmuth, M. Reciprocal Best Match Graphs. *J. Math. Biol.* **2020**, *80*, 865–953. [CrossRef] [PubMed]