Correction of Weighted Orthology and Paralogy Relations - Complexity and Algorithmic Results

Riccardo Dondi,[1], Nadia El-Mabrouk [2] and Manuel Lafond [2]

[1] Dipartimento di Lettere, Filosofia e Comunicazione, Università degli Studi di Bergamo, Bergamo - Italy,
[2] Department of Computer Science, Université de Montréal, Montréal (QC), Canada

Abstract. As inferring orthology relations between genes is a major concern in genomics, several methods have been developed for this purpose. A relation graph for a gene family is a graph with vertices representing the genes, edges connecting pairs of orthologous genes and "missing" edges representing paralogs. While a gene tree directly leads to a set of orthology and paralogy relations, the converse is not always true. Indeed a relation graph cannot necessarily be inferred from any tree, and even if it is "satisfiable" by a tree, this tree is not necessarily "consistent", i.e. does not necessarily reflect a valid history for the genes, in agreement with a species tree. Here, we consider the problems of minimally correcting a relation graph for satisfiability and consistency, from a new perspective. In fact, as different orthology-inferrence methods may lead to conflicting results, a degree of confidence can be assigned to each orthology or paralogy relation, leading to a weighted relation graph. We provide complexity and algorithmic results for minimizing corrections on a weighted graph, and also for the maximization variant of the problems for unweighted graphs.

1 Introduction

As genes are the basic molecular units of heredity, key for understanding genetic diversity, a first step of most genomic studies is to group genes into families. Gene families are usually inferred from sequence similarity, the underlying idea being that similar sequences reflect "homologous" genes that have diverged from a common ancestral sequence.

Given a gene family, it is important to discriminate between two types of homologs: *orthologs* being gene copies originating from a speciation event, and thus likely to be more similar in function, and *paralogs* originating from a duplication, and thus representing two different copies of an ancestral gene. For this purpose, tree-based methods consist in first constructing a phylogenetic tree for the gene family, and then, given a species tree, applying a reconciliation approach for inferring speciation and duplication nodes [8]. Results are strongly dependent upon the inferred gene tree, as few errors may lead to a completely different history. On the other hand, tree-free methods are based on gene clustering according to sequence similarity (cf. e.g. the COG database [21], OrthoMCL [17], InParanoid [3], Proteinortho [16]), synteny [13,14] or functional annotation of genes [5]. Results of these methods are pairwise orthology relations, or groups of orthologs, that can be represented as relation graphs, where vertices are genes and edges are orthology relations. Assuming a full inference of pairwise orthology relations, "missing" edges of the relation graph represent paralogy relations. In addition, as different inference methods may lead to different predictions, instead of a yes or no orthology assignment, existing methods can rather motivate a way of assigning a probabilistic score to a given relation [12], leading to a weighted relation graph. Surprisingly, as far as we know, weighted orthology/paralogy relation graphs have not been formally considered in the literature.

While a gene tree induces a set of relations betwen genes, the converse is not true, as a set of relations may or may not represent a valid history for the gene family. Two underlying questions are: (1) is the set of relations "satisfiable" i.e. is there a tree, with internal nodes labeled as duplication or speciation, containing them all? (2) is the set of relations "S-consistent" with the known species tree S, i.e. is there a tree containing the relations that is a "valid" gene tree "in agreement" with S? Polynomial-time algorithms are known to exist for deciding satisfiability and S-consistency for full [9,10] or partial [12] pairwise gene relations.

In this paper, we address the problem of correcting a full relation graph R, and more specifically a full weighted relation graph, so that it represents a satisfiable and S-consistent set of relations. The related minimization problems consist in editing, i.e. adding or removing, edges of minimum total weight. In the unweighted case, the satisfiability correction problem reduces to editing a minimum number of edges of R in order to make it P_4-free, which is known to be NP-hard [18]. In [10], an integer linear programming formulation is used to correct relation graphs of small size, which is also applicable to weighted graphs. In [19], the authors propose an approximation algorithm of factor 4Δ, where Δ is the maximum degree of the input graph. The algorithm, however, offers no guarantees in the case of weighted graphs, as there are weighted instances on which it is arbitrarily far from optimal. It is shown in [1] that the minimum edge editing problem cannot be approximated within an "additive" factor of $n^{2-\epsilon}$, for any $\epsilon > 0$. Yet, the authors give a class of polynomial time algorithms that are approximable within an additive factor of ϵn^2, for any $\epsilon > 0$. This implies a constant factor algorithm for graphs with an edit distance of $\Omega(n^2)$, but offers no guarantee in the other cases. Moreover, this algorithm only applies to unweighted graphs, and does not consider that two genes from the same species must remain paralogs. Finally in [18], parameterized versions of the algorithm are explored. As for the S-consistency correction problem, we proved in a previous paper [15] that it is NP-hard, which is the only result so far.

Here, we consider the satisfiability and S-consistency problems for weighted relation graphs. We show in Section 3 that the weighted problems are not approximable within a constant factor, assuming the Unique Games Conjecture. In Section 4, we then show that they can be approximated within a factor of n (the number of vertices of the relation graph), and provide n-approximation algorithms for both the Satisfiability and S-consistency problems. We end this paper by giving, in Section 5, few results on the maximization variants of the problems for the unweighted case, which consists in maximizing the number of preserved relations.

2 Trees and orthology relations

In this section, we introduce the notations and definitions required for the rest of this paper, and state our optimization problems.

A graph H is denoted $H = (V_H, E_H)$, where V_H is its set of vertices and E_H its set of edges. If H is a tree, we may call members of V_H *nodes*, and degree one nodes are *leaves*.

2.1 Trees

All considered trees are rooted and binary. Given a set X, a *tree T for X* is a tree whose leafset $\mathcal{L}(T)$ is in bijection with X. Given an internal node u of T, the subtree rooted at u is denoted T_u and we call the leafset $\mathcal{L}(T_u)$ the *clade of* u. A node u is an *ancestor* of v if u is on the (inclusive) path between v and the root. The *lowest common ancestor* (lca) of u and v, denoted $lca_T(u, v)$, is the ancestor common to both nodes that is the most distant from the root. We define $lca_T(U)$ analogously for a set $U \subseteq V(T)$.

A *species tree S* for a species set Σ represents an ordered set of speciation events that have led to Σ: an internal node is an ancestral species at the moment of a speciation event, and its children are the new descendant species. For simplicity, we will assume that species trees are binary.

A *gene family Γ* is a set of genes accompanied with a function $s : \Gamma \rightarrow \Sigma$ mapping each gene to its corresponding species. The evolutionary history of Γ can be represented as a *node-labeled gene tree* for Γ, where each internal node refers to an ancestral gene at the moment of an event (either speciation or duplication), and is labeled as a speciation (*Spec*) or duplication (*Dup*) accordingly. Formally, we call a *DS-tree* for Γ a pair (G, ev), where G is a tree with $\mathcal{L}(G) = \Gamma$, and $ev : V_G \setminus \mathcal{L}(G) \rightarrow \{Dup, Spec\}$ is a function labeling each internal node of G as a duplication or a speciation. For example, in Figure 1, G_1 and G_2 are two DS-trees.

According to the Fitch [7] terminology, we say that two genes x, y of Γ are *orthologous in* G if $ev(lca_G(x, y)) = Spec$, and *paralogous in* G if $ev(lca_G(x, y)) = Dup$.

A DS-tree G for Γ does not necessarily represent a valid history. For this to hold, any speciation node of G should reflect a clustering of species "in agreement" with S [12]. Formally G should be S*-consistent*, as defined below, where s_G is the *LCA-mapping* function, mapping each gene, ancestral or extant, to a species as follows: if $g \in \mathcal{L}(G)$, then $s_G(g) = s(g)$; otherwise, $s_G(g) = lca_S(\{s(g') : g' \in \mathcal{L}(G_g)\})$.

Definition 1. *Let S be a species tree and G be a DS-tree. Let v be an internal node of G such that $ev(v) = Spec$. Then the speciation node v, with children v_1 and v_2, is S-*consistent *iff none of $s_G(v_1)$ and $s_G(v_2)$ is an ancestor of the other. We say that G is S-*consistent *iff every speciation node of G is S-consistent.*

For example, in Figure 1, G_1 is not S-consistent as the root of G_1 is not S-consistent.

2.2 Relation graphs

Let first introduce some preliminary notations. For a graph $H = (V_H, E_H)$, we denote the complementary set of E_H by $\overline{E_H} = \{\{u, v\} : u, v \in V_H, \{u, v\} \notin E_H\}$. Let V' be a subset of V_H. The *subgraph of H induced by V'*, denoted $H[V']$, is the subgraph of H with vertex-set V' having every edge $\{u, v\}$ of H such that $u, v \in V'$. If I is another graph, we say H is I*-free* if there is no $V' \subseteq V_H$ such that $H[V']$ is isomorphic to I.

A *relation graph* R on a gene family Γ is a graph with vertex set $V_R = \Gamma$, in which we interpret each edge $\{u, v\}$ of E_R as an orthology relation between u and v, and each "missing" edge $\{u, v\} \in \overline{E_R}$, also called *non-edge*, as a paralogy relation. Notice that if $s(u) = s(v)$, then $\{u, v\}$ must be a non-edge. We denote $n = |V_R|$.

A DS-tree G leads to a relation graph, denoted $R(G)$, with vertex set $\mathcal{L}(G)$ and edge set corresponding to all gene pairs that are orthologous in G. Conversely, a relation graph R does not necessarily lead to a DS-tree.

If this is the case, i.e. if there exists a DS-tree G such that $R(G) = R$, then R is said *satisfiable*. As shown in [9], a relation graph R is satisfiable if and only if R is P_4-free, meaning that no four vertices of R induce a path of length 3 (number of edges). The P_4-free graphs are sometimes called *cographs*. See Figure 1 for an example.

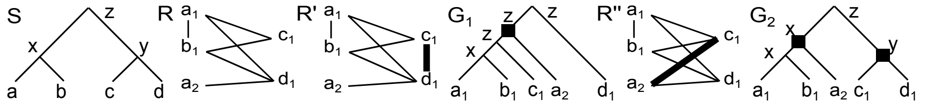

Fig. 1: S is the species tree for $\Sigma = \{a, b, c, d\}$. A gene name corresponds to the species it belongs to (e.g. $s(a_1) = a$). R is a relation graph. It is not satisfiable as the set of vertices $\{c_1, b_1, d_1, a_2\}$ induces a P_4. R' is a satisfiable relation graph obtained from R by inserting the edge $\{c_1, d_1\}$, and G_1 is a DS-tree tree displaying every relation of R' (each internal node v is labeled by $s_{G_1}(v)$). However, G_1 is not consistent with the species tree S. R'' is another correction of R that is S-consistent, as the tree G_2 displays the relations in R'' and is S-consistent. *Dup* nodes in DS-trees are marked by a square; all other nodes are speciation nodes.

As a DS-tree does not necessarily represent a true history for Γ, satisfiability of a relation graph does not ensure a possible translation in terms of a history for Γ. For this to hold, R should also be *consistent* with the species tree, according to the following definition.

Definition 2. *A relation graph R for Γ is S-consistent if and only if R is satisfiable by a DS-tree G which is itself S-consistent.*

2.3 Problem statements

We call a *weight* for a relation graph $R = (V_R, E_R)$ a function $w : \binom{V_R}{2} \rightarrow \mathbb{R}^+$ on its vertex pairs. Notice that w assigns a weight to both edges (orthologies) and non-edges (paralogies). We shall assume that if $s(u) = s(v)$ for two genes u and v, then $\{u, v\} \in \overline{E_R}$ and $w(\{u, v\}) = \infty$. The weight function w is extended to any $I_R \subseteq \binom{V_R}{2}$ by defining $w(I_R) = \sum_{\{x,y\} \in I_R} w(\{x, y\})$.

Given a relation graph $R = (V_R, E_R)$, an *edge-editing* of R is a pair $E^*_R = (E^+_R, E^-_R)$ with $E^+_R \subseteq \overline{E_R}$ and $E^-_R \subseteq E_R$. We denote by $R(E^*_R)$ the graph $R(E^*_R) = (V_R, (E_R \cup E^+_R) \setminus E^-_R, w)$. In other words, E^+_R (respectively E^-_R) denote inserted (respec. removed) edges. Given a relation graph $R' = (V_{R'}, E_{R'})$ computed from R by edge insertion and removal, the set of removed edges is $E_R^- = E_R \setminus E_{R'}$, and the set of inserted edges is $E_R^+ = E_{R'} \setminus E_R$. For example, for the graph R' of Figure 1, $E_R^+ = \{\{c_1, d_1\}\}$ and $E_R^- = \emptyset$. An *edge-editing* E_R^* *is said* P_4*-free* if $R(E_R^*)$ is itself P_4-free. The problems considered in Sections 3 and 4 are the following (corresponding maximization problems are introduced in Section 5).

Minimum Weighted Editing for Satisfiability (MinWES):
Input: A weighted relation graph $R = (V_R, E_R, w)$;
Output: A satisfiable relation graph $R' = (V_R, E_{R'})$, obtained from R by an edge-editing $E_R^* = (E_R^+, E_R^-)$ that minimizes $w(E_R^+) + w(E_R^-)$.

Minimum Weighted Editing for Consistency (MinWEC):
Input: A weighted relation graph $R = (V_R, E_R, w)$ for a gene family with genes belonging to genomes in Σ, a species tree S for Σ;
Output: An S-consistent relation graph $R' = (V_R, E_{R'})$, obtained from R by an edge-editing $E_R^* = (E_R^+, E_R^-)$ that minimizes $w(E_R^+) + w(E_R^-)$.

3 Hardness of Approximation of Minimum Weighted Editing for Satisfiability and Consistency

We show that MinWES is unlikely to be approximable within a constant factor, by presenting a gap-preserving reduction from Minimum Multi-Cut. First, we consider the variant of MinWES, called Minimum Weighted Removal for Satisfiability (MinWRS), where only edge removal is allowed, then we easily extend the result to MinWES.

Given a graph $H = (V_H, E_H)$, and a set $X \subseteq \binom{V_H}{2}$ (i.e. a set of pairs), Minimum Multi-Cut asks for a set E'_H of minimum cardinality such that each pair $\{v_i, v_j\} \in X$ is disconnected in $H' = (V_H, E_H \setminus E'_H)$.

Given an instance $H = (V_H, E_H, X)$ of Minimum Multi-Cut, we construct an instance $R = (V_R, E_R, w)$ of MinWRS as follows. The vertex set V_R includes, for each $v_i \in V_H$, two vertices $v_{i,R}$ and $v'_{i,R}$. For any distinct $x, y \in V_R$, we set $s(x) \neq s(y)$, and hence there are no "forced" paralogs. As for E_R, it is defined as follows, where $q = |V_H|^5 + 1$.

- For each $v \in V_H$, define an edge $\{v_{i,R}, v'_{i,R}\}$ in E_R of weight $q' = q|E_H| + 2\left(\binom{|V_H|}{2} - |E_H|\right)$;
- For each $\{v_i, v_j\} \in X$, define an edge $\{v_{i,R}, v_{j,R}\}$ in E_R with weight q if $\{v_i, v_j\} \in E_H$, and with weight 1 if $\{v_i, v_j\} \notin E_H$;

– For each $\{v_i, v_j\} \notin X$, define the edges $\{v_{i,R}, v'_{j,R}\}$ and $\{v'_{i,R}, v_{j,R}\}$ in E_R, each with weight $q/2$ if $\{v_i, v_j\} \in E_H$, and with weight 1 if $\{v_i, v_j\} \notin E_H$.

For each $\{u_R, v_R\} \in \overline{E_R}$, $\{u_R, v_R\}$ has weight q'. Notice however, that, since edge insertion is not allowed in MinWRS, the weight of $\{u_R, v_R\}$ never contributes to the cost of a solution of MinWRS.

We first show (in the Appendix) that there is a correspondance between solutions to the two problems on our constructed instances.

Lemma 1. *Let $H = (V_H, E_H, X)$ be an instance of Minimum Multi-Cut and let $R = (V_R, E_R, w)$ be the corresponding instance of MinWRS. Given a solution E'_H of Minimum Multi-Cut, we can compute in polynomial time a solution of MinWRS of weight at most $q|E'_H| + 2\left(\binom{|V_H|}{2} - |E_H|\right)$.*

Lemma 2. *Let $H = (V_H, E_H, X)$ be an instance of Minimum Multi-Cut and let $R = (V_R, E_R, w)$ be the corresponding instance of MinWRS. Given a solution R' of MinWRS of weight at most $qW + 2\left(\binom{|V_H|}{2} - |E_H|\right)$, we can compute in polynomial time a multicut E'_H of H of size at most W.*

Assuming the Unique Games Conjecture, the inapproximability of MinWRS is deduced from the inapproximability of Minimum Multi-Cut [4].

Theorem 1. *MinWRS is not approximable within a constant factor assuming the Unique Games Conjecture.*

The result of Theorem 1 can be easily extended to MinWES.

Corollary 1. *MinWES is not approximable within a constant factor assuming the Unique Games Conjecture.*

Proof. The result follows by a gap-preserving reduction similar to that for MinWRS. Recall that for each pair $\{u_R, v_R\} \in \overline{E_R}$, a weight of q' is associated with $\{u_R, v_R\}$. Consider a solution R' of MinWES on instance R that has cost not greater than $qW + \left(\binom{|V_H|}{2} - |E_H|\right) + \binom{|V_H|}{2}$. It is easy to see that R' is obtained without any edge insertion. □

The inapproximability result for MinWES is easily extended to MinWEC. This is achieved by defining a species tree S on V_R such that the root of S is connected to two subtrees, one with leafset $\{v_{i,R} : v_i \in V_H\}$, one with leafset $\{v'_{i,R} : v_i \in V_H\}$, and showing that any solution to our instance of MinWRS must agree with this species tree.

Corollary 2. *MinWEC is not approximable within a constant factor assuming the Unique Games Conjecture.*

4 A Bounded Approximation Algorithm for Minimum Weighted Editing for Satisfiability and Consistency

While MinWES and MinWEC are not approximable within a constant factor, we show here that they can be approximated within factor of n, where n is the number of vertices of the graph, and we give the corresponding algorithms. Despite being a large approximation factor, this is the best known bound so far and shows that at least the problems have polynomially bounded approximability, unlike some other weighted graph problems. We first describe the approximation algorithm for MinWES.

Denote by $\overline{R} = (V_R, \overline{E_R})$ the *complement* of the graph $R = (V_R, E_R)$. A well-known property of cographs is given by the following lemma.

Lemma 3. *[6] A graph R is P_4-free if and only if for any $X \subseteq V_R$, one of $R[X]$ or $\overline{R[X]}$ is disconnected.*

This motivates a greedy min-cut approach for MinWES, performing an edge-editing of minimum weight disconnecting the graph or its complement, and iterating recursively on the resulting components. This is the main idea of Algorithm MinCut-Cograph-Editing below. Note that assuming forced paralogs have infinite weight, this algorithm will never make two genes from the same species orthologs.

More formally, let $R = (V_R, E_R, w)$ be a weighted relation graph. Define a *cut* $C = \{X, Y\}$ as a partition of V_R with X and Y being non-empty sets, and denote $E_R(C) = \{\{x, y\} \in E_R : x \in X, y \in Y\}$. The weight of C is $w(C) = w(E_R(C))$. The cut C is a *minimum cut* or *MinCut* if no other cut has a smaller weight $w(C)$. *Applying a cut C* to R consists in removing all edges of $E_R(C)$ from R.

```
Algorithm MinCut-Cograph-Editing($R$):
    If $R$ has at most 2 vertices Then Return;
    Find a MinCut $C = \{X, Y\}$ for $R$;
    Find a MinCut $\overline{C} = \{\overline{X}, \overline{Y}\}$ of $\overline{R}$;
    If $w(C) < w(\overline{C})$ Then
        Remove all edges between $X$ and $Y$ in $R$;
        MinCut-Cograph-Editing($R[X]$);
        MinCut-Cograph-Editing($R[Y]$);
    Else
        Add all possible edges between $\overline{X}$ and $\overline{Y}$ in $R$;
        MinCut-Cograph-Editing($R[\overline{X}]$);
        MinCut-Cograph-Editing($R[\overline{Y}]$);
    End If
End Algorithm
```

Complexity: A MinCut of a given graph of n vertices and m edges can be found in time $O(nm + n^2 \log n)$ using the Stoer-Wagner algorithm [20]. In the MinCut-Cograph-Editing algorithm, MinCut is applied to both R and $\overline{R}$. As at least one of these two graphs has $\Omega(n^2)$ edges, the required time for MinCut is therefore $O(n^3)$. This step is repeated at most n times, hence the overall time complexity of MinCut-Cograph-Editing is $O(n^4)$.

The remaining of this section is dedicated to proving Theorem 2, which states that MinCut-Cograph-Editing is an n-approximation algorithm. We denote by σ_R the minimum weight of a P_4-free edge-editing of R. If $X \subseteq V_R$, we denote $\sigma_{R[X]}$ by σ_X.

Lemma 4. *Let C be a minimum cut of R, and let $\overline{C}$ be a minimum cut of $\overline{R}$. Then $\sigma_R \geq \min\{w(C), w(\overline{C})\}$.*

Proof. Let E^*_R be a P_4-free edge-editing of R. By Lemma 3, either $R(E^*_R)$ or its complement is disconnected, implying that E^*_R must apply some cut on either R or $\overline{R}$. This cut is at best a minimum cut. □

Lemma 5. *Let $\{X, Y\}$ be a partition of V. Then, $\sigma_R \geq \sigma_X + \sigma_Y$.*

Proof. Let E^*_R be a P_4-free edge-editing of weight σ_R, and let $R' = R(E^*_R)$. Assume that E^*_R has a weight stricly smaller than $\sigma_X + \sigma_Y$. Then, since $R'[X]$ and $R'[Y]$ are P_4-free, there must either be an edge-editing of $R[X]$ of weight smaller than σ_X, or an edge-editing of $R[Y]$ of weight smaller than σ_Y, contradicting the definition of σ_X and σ_Y. □

Theorem 2. MinCut-Cograph-Editing *is an n factor approximation algorithm for* MinWES*.*

Proof. Denote by $\beta(R)$ the weight of the edge-editing found by the algorithm on R. We proceed by induction on $n = |V_R|$ to show that $\beta(R) \leq n\sigma_R$. The statement is trivial for $n \leq 3$ (as there is nothing to correct), so assume that the algorithm finds a solution of weight $\beta(R) \leq k\sigma_R$ for any graph of size at most $k < n$. The algorithm applies a minimum cut $C = \{X, Y\}$ on R or $\overline{R}$, and proceeds recursively on X and Y, with $|X|, |Y| \leq n - 1$. By the induction hypothesis, we have

$$\begin{aligned}\beta(R) &\leq |X|\sigma_X + |Y|\sigma_Y + w(C) \leq (n-1)(\sigma_X + \sigma_Y) + w(C) \\ &\leq (n-1)\sigma_R + \sigma_R = n\sigma_R\end{aligned}$$

where the last inequality holds due to Lemma 4 and Lemma 5. □

It is possible to show that the approximation factor of MinCut-Cograph-Editing is tight.

By modifying MinCut-Cograph-Editing, it is possible to design an n factor approximation algorithm for MinWEC. The main difference with respect to MinCut-Cograph-Editing, is that the algorithm considers a minimum cut on a subset of R and a cut on a subset of $\overline{R}$ induced by the species tree S. The detailed algorithm, along with the proof of the following Theorem, are given in the Appendix. It also requires time $O(n^4)$.

Theorem 3. MinCut-Cograph-Editing-Cons *is an n factor approximation algorithm for MinWEC.*

5 Polynomial Time Approximation Schemes for the Maximization Variant of Graph Correction

Here, we consider the complementary maximization problem, which consists in maximizing conservation between the original and corrected graphs. Although sharing the same objectives, the minimization and maximization variants are not equivalent from an approximation point of view.

Below is a formal statement of the corresponding maximization version of MinWES (see Section 2) for unweighted graphs. Remember that edges represent orthologies, while non-edges are paralogies. Maximizing conservation therefore requires accounting for both edges and non-edges.

Maximum Editing for Satisfiability (MaxES):
Input: A relation graph $R = (V_R, E_R)$;
Output: A satisfiable relation graph $R' = (V_R, E_{R'})$ obtained from R by an edge-editing, such that its *value* $|E_R \cap E_{R'}| + |(\overline{E_R} \cap \overline{E_{R'}})|$ is maximized.

Given a relation graph R, the value of a solution R' for MaxES over instance R is called the *agreement* value of R'.

Lemma 6. *Given a relation graph R, an optimal solution of MaxES over instance R has an agreement value of at least $\frac{n^2}{8}$.*

Proof sketch: Consider the two 'extreme' solutions: either make all genes from two distinct species orthologs, or all genes paralogs. In R, either at least half the genes are orthologs, or at least half the genes are paralogs. Thus one extreme solution preserves at least half the total number of relations, which is $\binom{n}{2}/2 > \frac{n^2}{8}$. The detailed proof is in the Appendix. □

Note that the above gives, almost trivially, a factor $1/2$ approximation (i.e. preserving at least half as many relations as the optimal). Using

Lemma 6 and results from [1], one can devise a PTAS for MaxES in the case that every gene belongs to a distinct species. Let $OPT(R)$ be the value of an optimal solution on R, and let c be such that $OPT(R) = cn^2$. The additive εn^2 approximation algorithm for cograph editing [1] yields a solution of value $(c - \varepsilon)n^2$. As $c \geq 1/8$ by Lemma 6, ε can be adjusted so that, for any $0 < \varepsilon' < 1$, $(c - \varepsilon)n^2 \geq (1 - \varepsilon')cn^2$, hence yielding a PTAS. In the more general case, this algorithm does not ensure that genes from the same species remain paralogs. However, the authors of [1] claim that their approximation algorithm applies to any hereditary graph property (i.e. preserved after vertex-deletion), which holds for satisfiability.

Finally, we end this paper with few additional results on the maximization version of graph correction for consistency, that we call MaxEC. Notice that the lower bound $\frac{n^2}{8}$ of lemma 6 also holds for an optimal solution of MaxEC. However, the PTAS for MaxES does not guarantee that the returned relation graph R' is S-consistent with the given species tree S. We can show however that a PTAS for MaxEC can be obtained, based on smooth-polynomial integer programming [2], a technique that has been applied to problems like Maximum Quartet Consistency [11]. Proofs are quite involved, and require several technical arguments, that will be included in a journal version of this extended abstract.

6 Conclusion

This paper explores a new direction in the field of orthology and paralogy prediction. Taking advantage of the many existing prediction tools, a set of relation is better represented as a weighted relation graph, where the weight of a relation represents its degree of confidence. In case of non-satisfiability or unconsistency, the goal is to minimally correct the corresponding relation graph. While the problem has been largely explored in the case of unweighted graphs, the weighted version of the problem remains largely unexplored. Here, we provide complexity results and polynomial approximation algorihms for this problem.

For real application to biological datasets, the challenge remains to assign appropriate weights to relations. This can be done in many different ways, depending on the considered prediction tools and the degree of confidence given to each of them. A full bioinformatics study on simulated and real datasets remains to be undertaken for this purpose.

References

1. Noga Alon and Uri Stav. Hardness of edge-modification problems. *Theor. Comput. Sci.*, 410(47-49):4920–4927, 2009.

2. Sanjeev Arora, Alan M. Frieze, and Haim Kaplan. A new rounding procedure for the assignment problem with applications to dense graph arrangement problems. *Math. Program.*, 92(1):1–36, 2002.
3. A.C. Berglund, E. Sjolund, G. Ostlund, and E.L. Sonnhammer. InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucl. Acids Res.*, 36, 2008.
4. Shuchi Chawla, Robert Krauthgamer, Ravi Kumar, Yuval Rabani, and D. Sivakumar. On the hardness of approximating multicut and sparsest-cut. *Computational Complexity*, 15(2):94–114, 2006.
5. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nat. Genet.*, 25(1):25 - 29, 2000.
6. D. G. Corneil, Y. Perl, and L. K Stewart. A linear recognition algorithm for cographs. *SIAM J. Comput.*, 14(4):926-934, 1985.
7. W. M. Fitch. Homology. a personal view on some of the problems. *TIG*, 16(5):227-231, 2000.
8. M. Goodman, J. Czelusniak, G.W. Moore, A.E. Romero-Herrera, and G. Matsuda. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zoology*, 28:132–163, 1979.
9. M. Hellmuth, M. Hernandez-Rosales, K. Huber, V. Moulton, P. Stadler, and N. Wieseke. Orthology relations, symbolic ultrametrics, and cographs. *J. Math. Biol.*, 66(1–2):399–420, 2013.
10. Marc Hellmuth, Nicolas Wieseke, Markus Lechner, Hans-Peter Lenhof, Martin Middendorf, and Peter F Stadler. Phylogenomics with paralogs. *PNAS*, 2014.
11. Tao Jiang, Paul E. Kearney, and Ming Li. A polynomial time approximation scheme for inferring evolutionary trees from quartet topologies and its application. *SIAM J. Comput.*, 30(6):1942–1961, 2000.
12. M. Lafond and N. El-Mabrouk. Orthology and paralogy constraints: satisfiability and consistency. *BMC Genomics*, 15(Suppl 6):S12, 2014.
13. M. Lafond, M. Semeria, K.M. Swenson, E. Tannier, and N. El-Mabrouk. Gene tree correction guided by orthology. *BMC Bioinformatics*, 14 (supp 15)(S5), 2013.
14. M. Lafond, K. Swenson, and N. El-Mabrouk. *Models and algorithms for genome evolution*, chapter Error detection and correction of gene trees. Springer, 2013.
15. Manuel Lafond, Riccardo Dondi, and Nadia El-Mabrouk. The link between orthology relations and gene trees: a correction perspective. *Algorithms for Molecular Biology*, 11(1):1, 2016.
16. M. Lechner, S.Sven Findeib, L. Steiner, M. Marz1, P.F. Stadler, and S.J. Prohaska. Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics*, 12:124, 2011.
17. L. Li, C.J. Jr. Stoeckert, and D.S. Roos. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13:2178- 2189, 2003.
18. Yunlong Liu, Jianxin Wang, Jiong Guo, and Jianer Chen. Complexity and parameterized algorithms for cograph editing. *Theor. Comput. Sci.*, 461:45–54, 2012.
19. A. Natanzon, R. Shamir, and R. Sharan. Complexity classification of some edge modification problems. *Discrete Applied Mathematics*, 113(1):109–128, 2001.
20. Mechthild Stoer and Frank Wagner. A simple min-cut algorithm. *Journal of the ACM (JACM)*, 44(4):585–591, 1997.
21. 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. *Nucl. Acids Res.*, 28:33- 36, 2000.

7 Appendix

A Proof of Lemma 1

We first bound the number of edges of weight 1 in R.

Claim. Let $H = (V_H, E_H, X)$ be an instance of Minimum Multi-Cut and let $R = (V_R, E_R, w)$ be the corresponding instance of MinWRS. Then, R contains at most $2\left(\binom{|V|}{2} - |E_H|\right)$ edges of weight 1.

Proof. Consider the edges connecting vertices $v_{i,R}$ and $v_{j,R}$; $v_{i,R}$ and $v_{j,R}$ are connected by an edge of weight 1 if and only if $\{v_i, v_j\} \notin E_H$ and $\{v_i, v_j\} \in X$.

Consider the edges connecting vertices $v_{i,R}$ and $v'_{j,R}$, $v'_{i,R}$ and $v_{j,R}$. $v_{i,R}$, $v'_{j,R}$ (and $v'_{i,R}$,$v_{j,R}$) are connected by an edge of weight 1 if $\{v_i, v_j\} \notin E_H$ and $\{v_i, v_j\} \notin X$.

Any other edge has weight greater than 1, hence the lemma follows. □

We are now ready to prove Lemma 1.

Proof. Given a set E' that defines a multicut in H, let $V_{H,1}, \ldots, V_{H,p}$ be the sets of vertices of the connected components in the graph $V'_H = (V'_H, E_H \setminus E'_H)$.

We define a solution of MinWRS over instance R as follows. We construct the partition $V_{R,1}, \ldots, V_{R,p}$ of the vertices of R such that $v_{j,R}$ and $v'_{j,R}$ belong to set $V_{R,i}$ if and only if $v_j \in V_{H,i}$. All edges having their endpoints in two distinct $V_{R,i}, V_{R,j}$ are removed.

We claim that the computed graph R' induced by the partition is P_4-free. By construction, for each $v_{j,R}$, $v'_{j,R}$, $v_{h,R}$, $v'_{h,R}$ that belong to $V_{R,i}$, the edges $\{v_{j,R}, v'_{h,R}\}$ and $\{v'_{j,R}, v_{h,R}\}$ belong to E_R (because $\{v_j, v_h\} \notin X$). Moreover, there is no edge between $v_{j,R}$ and $v_{h,R}$, nor between $v'_{j,R}$ and $v'_{h,R}$. Thus any path on four vertices in the graph on vertex set $V_{i,R}$ must be either of the form $v_{j,R}v'_{h,R}v_{k,R}v'_{\ell,R}$, or of the form $v'_{j,R}v_{h,R}v'_{k,R}v_{\ell,R}$. In both cases, the endpoints of the path share an edge, and thus cannot induce a P_4.

Now, consider the edges $\{v_i, v_j\} \in E'_H$. If $\{v_i, v_j\} \in X$, the corresponding solution of MinWRS removes an edge of weight q, namely $\{v_{i,R}, v_{j,R}\}$. If $\{v_i, v_j\} \notin X$, the corresponding solution of MinWRS removes two edges of weight $q/2$, namely $\{v_{i,R}, v'_{j,R}\}$ and $\{v'_{i,R}, v_{j,R}\}$. Hence those edges have a total weight $q|E'_H|$. Since at most $2\left(\binom{|V_H|}{2} - |E_H|\right)$ edges of weight 1 are removed (see Claim A), we can conclude that the lemma holds. □

B Proof of Lemma 2

Proof. Consider a solution $R' = (V_R, E'_R, w)$ of MinWRS over instance $R = (V_R, E_R, w)$ of weight at most $qW + 2\left(\binom{|V_H|}{2} - |E_H|\right)$, with $W \leq |E_H|$. First, notice that no edge $\{v_{i,R}, v'_{i,R}\}$, with $1 \leq i \leq |V|$, is removed to obtain R', since the weight of such an edge is greater than $qW + 2\left(\binom{|V_H|}{2} - |E_H|\right)$.

Consider now two vertices $v'_{i,R}$, $v'_{j,R}$, such that, given the corresponding vertices v_i, v_j in H, we have $\{v_i, v_j\} \in X$. By construction there is a P_4 in R, namely $v'_{i,R}, v_{i,R}, v_{R,j}, v'_{j,R}$. It follows that the edge $\{v_{i,R}, v_{j,R}\}$ must be removed in R'. Moreover, we claim that in R', the vertices $v'_{i,R}$, $v'_{j,R}$ must be disconnected. Assume by contradiction that this does not hold, and that $v'_{i,R}$, $v'_{j,R}$ belong to the same connected component of R'. Consider the shortest path P that connects vertices $v_{i,R}$ and $v_{j,R}$ in R'. Then P has length at least 2. Note that as P is a shortest path, it has no chord, i.e. non-consecutive vertices of P cannot share an edge.

Suppose that P does not include the vertex $v'_{i,R}$. Then we can assume that $v_{i,R}$ is adjacent in P to a vertex $v'_{t,R}$, since if it is adjacent to a vertex $v_{q,R}$, then the vertices $v_{i,R}$, $v'_{i,R}$, $v_{q,R}$, and $v'_{q,R}$ would induce a P_4. Now, if $v'_{t,R}$ is adjacent to $v_{j,R}$, then $v'_{i,R}$, $v_{i,R}$, $v'_{t,R}$ and $v_{j,R}$ induce a P_4. If there is no such $v'_{t,R}$, then P has length at least 3 and it must therefore contain an induced P_4.

So suppose instead that P includes the vertex $v'_{i,R}$. Since by construction $v'_{i,R}$ is not adjacent to $v_{j,R}$ and it is not adjacent to any $v'_{t,R}$, with $t \neq i$, while it is adjacent to $v_{i,R}$, P has length at least 3, and again must have an induced P_4.

We can conclude that when $\{v_i, v_j\} \in X$, the corresponding vertices $v'_{i,R}$, $v'_{j,R}$ belong to disconnected connected components of R'. Hence we can compute a multi-cut of H as follows:

$$E'_H = \{\{v_i, v_j\} : \{v_{i,R}, v_{j,R}\}, \text{ of weight } q, \text{ or } \{v_{i,R}, v'_{j,R}\}, \{v'_{i,R}, v_{j,R}\}, \text{ of weight } \frac{q}{2}, \text{ are removed in } R' .\}$$

E'_H is a multi-cut, since each $\{v_i, v_j\} \in X$ is disconnected. Now, recall that R' is obtained by removing edges of overall weight at most $qW + 2\left(\binom{|V_H|}{2} - |E_H|\right)$. Since edge edge in E'_H corresponds to edges of overall weight q in R (an edge $\{v_{i,R}, v_{j,R}\}$ of weight q if $\{v_i, v_j\} \in X$, or two edges of weight $q/2$, namely $\{v_{i,R}, v'_{j,R}\}$ and $\{v'_{i,R}, v_{j,R}\}$ if $\{v_i, v_j\} \notin X$), we must have $|E'_H| \leq W$. □

C Proof of Theorem 1

Proof. Given a graph H instance of Minimum Multi-Cut and the corresponding instance R of MinWRS, denote by OPT_M (AP_M, respectively) the value of an optimal solution (of an approximation solution, respectively) of Minimum Multi-Cut on instance H, and denote by OPT_C (AP_C, respectively) the value of an optimal solution (of an approximation solution, respectively) of MinWRS on instance R. Define $z = 2\left(\binom{|V_H|}{2} - |E_H|\right)$. By Lemma 1, we assume that $AP_C(R) \leq AP_M(H)/q$, as there exists an algorithm that always outputs at most such a value, and thus any approximation algorithm can be adapted to output at most this value. Also, by Lemma 2, we have $OPT_C(R) \leq OPT_M(H)q + z$. We have that

$$\frac{AP_C(R)}{OPT_C(R)} \geq \frac{AP_M(H)q}{OPT_M(H)q + z} = \frac{AP_M(H)q + AP_M(H)z - AP_M(H)z}{OPT_M(H)q + z} =$$

$$= \frac{AP_M(H)q + AP_M(H)z}{OPT_M(H)q + z} - \frac{AP_M(H)z}{OPT_M(H)q + z}$$

$$\geq \frac{AP_M(H)q + AP_M(H)z}{OPT_M(H)q + OPT_M(H)z} - \frac{AP_M(H)z}{OPT_M(H)q + z}$$

$$= \frac{AP_M(H)(q + z)}{OPT_M(H)(q + z)} - \frac{AP_M(H)z}{OPT_M(H)q + z}$$

$$= \frac{AP_M(H)}{OPT_M(H)} - \frac{AP_M(H)z}{OPT_M(H)q + z}$$

where we assume $OPT_M(H) \geq 1$ for the second inequality (the case $OPT_M(H) = 0$ can be checked in polynomial time). Since Minimum Multi-Cut is not approximable within a constant factor assuming the Unique Games Conjecture [4], even on unweighted graphs, it follows that

$$\frac{AP_M(H)}{OPT_M(H)} \geq \alpha$$

on an infinity of instances of H for any constant $\alpha \geq 1$. As a consequence, for any constant $\alpha \geq 1$, an infinity of instances of R yield:

$$\frac{AP_C(R)}{OPT_C(R)} \geq \alpha - \frac{AP_M(H)z}{OPT_M(H)q + z}$$

Since $q = n^5+1$, $AP_M(H) \leq n^2$ and $z \leq n^2$, it follows that $\frac{AP_M(H)z}{OPT_M(H)q+z} \leq 1/n$. Combining the last two inequalities, we have that

$$\frac{AP_C(R)}{OPT_C(R)} \geq \alpha - 1/n \geq \beta$$

for any constant $\beta \geq 1$, which concludes the proof. □

D Proof of Corollary 2

Proof. The result follows by a gap-preserving reduction similar to that for MinWRS and MinWES. Define a species tree S on V_R such that the root of S is connected to two subtrees, one with leafset $\{v_{i,R} : v_i \in V_H\}$, one with leafset $\{v'_{i,R} : v_i \in V_H\}$.

Consider the partition $V_{R,1}, \ldots, V_{R,p}$ of the vertices of a solution R' of MinWRS and MinWES. Each connected component $V_{R,t}$ that contains vertices $v_{i,R}$, $v'_{i,R}$, $v_{j,R}$, $v'_{j,R}$, contains only edges $\{v_{i,R}, v'_{i,R}\}$, $\{v_{j,R}, v'_{j,R}\}$, $\{v_{i,R}, v'_{j,R}\}$, $\{v_{j,R}, v'_{i,R}\}$.

For each set $V_{R,i}$, we construct a tree $G_{R,i}$ by defining two subtrees $G^1_{R,i}$ and $G^2_{R,i}$ such that $G^1_{R,i}$ has leafset $\{v_{j,R} : v_{j,R} \in V_{R,i}\}$ and $G^2_{R,i}$ has leafset $\{v'_{j,R} : v'_{j,R} \in V_{R,i}\}$. Each node of $G^1_{R,i}$ and $G^2_{R,i}$ is associated with a duplication. $G_{R,i}$ is obtained by joining $G^1_{R,i}$ and $G^2_{R,i}$ in a root, associated with a speciation. Finally, the subtrees $G_{R,1}, \ldots, G_{R,p}$ are joined in a gene tree G by duplication nodes (with any topology). By construction, G is S-consistent, thus the hardness result can be extended to MinWEC. □

E Proof of Theorem 3

We first provide the detailed MinCut-Cograph-Editing-Cons algorithm, and show that it also is a n-factor approximation.

Given a species tree S and a set $Z \subseteq V_R$, let $\Sigma(Z) = \{s(x) : x \in Z\}$. Let $S|\Sigma(Z)$ be the subtree of S restricted to $\Sigma(Z)$ and let X_S, Y_S be the clades of the left and right child, respectively, of the root of $S|\Sigma(Z)$. Consider the sets $X = \{x : s(x) \in X_S\}$ and $Y = \{y : s(y) \in Y_S\}$, the cut $C_S(Z)$ on $\overline{R}[Z]$ is defined as $C_S(Z) = \{X_R, Y_R\}$. Observe that $C_S(Z)$ is the only possible cut on $\overline{R}$ that maintains S-consistency, as this cut corresponds to a speciation in a DS-tree, and speciations must separate genes according to S. Therefore, it suffices to modify MinCut-Cograph-Editing by forcing the cut $\overline{C}$ to be $C_S(Z)$. Call this modified algorithm MinCut-Cograph-Editing-Cons.

```
Algorithm MinCut-Cograph-Editing-Cons($R$):
    If $R$ has at most 2 vertices Then Return;
    Find a MinCut $C = \{X, Y\}$ for $R$;
    Let $C_S(V_R) = \{\overline{X}, \overline{Y}\}$;
    If $w(C) < w(C_S(V_R))$ Then
        Remove all edges between $X$ and $Y$ in $R$;
        MinCut-Cograph-Editing-Cons($R[X]$);
        MinCut-Cograph-Editing-Cons($R[Y]$);
    Else
        Add all possible edges between $\overline{X}$ and $\overline{Y}$ in $R$;
        MinCut-Cograph-Editing-Cons($R[\overline{X}]$);
        MinCut-Cograph-Editing-Cons($R[\overline{Y}]$);
    End If
End Algorithm
```

Proof. Denote by $\beta(R)$ the weight of the edge-editing found by the algorithm on R. We proceed by induction on $n = |V_R|$ to show that $\beta(R) \leq n\sigma_R$. The statement is trivial for $n \leq 2$ (as there is nothing to correct), so assume that the algorithm finds a solution of weight $\beta(R) \leq k\sigma_R$ for any graph of size at most $k < n$.

The algorithm applies a cut $C = \{X, Y\}$ which is either a minimum cut on R or it is the cut $C_S(V_R)$, and proceeds recursively on X and Y, with $|X|, |Y| \leq n - 1$. By the induction hypothesis, we have

$$\beta(R) \leq |X|\sigma_X + |Y|\sigma_Y + w(C) \leq (n-1)(\sigma_X + \sigma_Y) + w(C)$$

Now, similarly to Lemma 4, we have that $w(C) \leq \sigma_R$. First, let G' be the gene tree associated with a solution of MinWEC over instance R. If C is a minimum cut on R, it holds due to the proof Lemma 4. If C is $C_S(V_R)$, then notice that, in order to guarantee the consistency with S, the root of G' must be exactly $C_S(V_R)$.

Lemma 5 holds also for MinWEC, hence

$$\begin{aligned}\beta(R) &\leq |X|\sigma_X + |Y|\sigma_Y + w(C) \leq (n-1)(\sigma_X + \sigma_Y) + w(C) \\ &\leq (n-1)\sigma_R + \sigma_R = n\sigma_R\end{aligned}$$

hence the theorem holds. □

F Proof of Lemma 6

Given a relation graph R, the value of a solution R' for MaxES over instance R is called the *agreement* value of R' and it is denoted by $A(R', R)$.

Moreover, given a gene tree G, we denote by $A(G, R)$ the agreement between the relation graph associated with G and R.

Proof. Let $X = \{\{u, v\} : u, v \in V_R \text{ and } s(u) = s(v)\}$ be the set of 'must-be' paralogs. Consider the relation graphs $R' = (V_R, \emptyset)$ and $R'' = (V_R, \binom{V_R}{2} \setminus X)$, where $\binom{V_R}{2}$ is the set of all unordered pairs of V_R. It is not hard to see that R' and R'' are both feasible solutions of MaxES and of MaxEC. For each $\{u, v\} \in \binom{V_R}{2} \setminus X$, the u, v relation in R agrees with exactly one of R' or R'', and for each $\{u, v\} \in X$, the u, v relation agrees with both R' and R''. It follows that

$$A(R, R') + A(R, R'') \geq \binom{n}{2}$$

But then, for this inequality to hold, at least one of R', R'' must have an agreement value of at least $\frac{1}{2}\binom{n}{2}$, hence an optimal solution of MaxES and MaxEC has an agreement value of at least $\frac{1}{2}\binom{n}{2} \geq \frac{n^2}{8}$. □