Orthology relation and Gene Tree correction: Complexity results

Manuel Lafond [1] and Nadia El-Mabrouk [1]

[1]Department of Computer Science, Université de Montréal, Montréal (QC), Canada

Abstract. Tree-oriented methods for inferring orthology and paralogy relations between genes are based on reconciling a gene tree with a species tree. On the other hand, many tree-free methods, mainly based on sequence similarity, are also available. The link between orthology relations and gene trees has been formally considered recently from the angle of reconstructing phylogenies from orthology relations. Here, we rather consider this link from a correction point of view. While a gene tree induces a set of relations, the converse is not always true, as a set of relations is not necessarily in agreement with any gene tree. How can we minimally correct an infeasible set of relations? On the other hand, given a gene tree and a set of relations, how to minimally correct a gene tree in order to fit the set of relations? In this paper, various objective functions are considered for the minimality criterion, among them the Robinson-Foulds distance between the initial and corrected gene tree. All considered problem variants are shown to be NP-complete.

1 Introduction

Genes are the molecular units of heredity, holding the information to build and maintain cells. In the course of evolution, they are duplicated, lost, and passed to organisms through speciation. Genes originating from the same ancestral copy are called *homologs*. They are usually inferred from sequence similarity and grouped into *Gene Families*. Two homologous genes are *orthologous* if their parental origin is a speciation, and *paralogous* if it is a duplication. From the orthology conjecture, orthologs tend to be more similar in function than paralogs [29]. This is a major motivation for inferring gene evolution, as it is a prerequisite for functional prediction purposes.

The tree-based method requires to build, classically from a DNA or protein sequence alignment, a phylogenetic tree for the considered gene family. Reconciliation [12] with the species tree then allows to label internal nodes as duplications and speciations, inducing a full orthology and paralogy set of relations between gene pairs. On the other hand, tree-free orthology detection methods are also available. They are based

on gene clustering according to sequence similarity, (cf. e.g. the COG database [34], OrthoMCL [24], InParanoid [3], Proteinortho [22]), synteny [20,21] or functional annotation of genes [7]. Only partial sets of relations are usually inferred from these methods.

Recent papers have been dedicated to the formal study of the link between trees and orthology/paralogy relations (we just say "relations" in the following) [15,16]. Given a gene family Γ and a set $\mathcal{C}$ of pairwise relations, can we reconstruct a labeled gene tree for Γ inducing $\mathcal{C}$? The question can be subdivided into two parts: 1. Is $\mathcal{C}$ *satisfiable*, i.e. is there an event-labeled gene tree G in agreement with $\mathcal{C}$? However satisfiability is not sufficient to ensure the possibility for the relation set to reflect a true history, as nodes of G labeled as speciations can be contradictory. This raises the second question; 2. Is there an event-labeled gene tree G which is *S-consistent*, i.e. obtained from reconciliation, with a species tree S? A simple characterization of satisfiability is given in [15] in the case of $\mathcal{C}$ being a full set of relations (i.e. each pair of genes of Γ is in $\mathcal{C}$). Moreover, a polynomial-time algorithm can be devised to check for S-consistency [1,17]. In [19], we generalized these results to partial relations.

In this paper we explore the link between relations and trees for the purpose of relation and tree correction. Several gene tree databases from whole genomes are available, including for instance Ensembl Compara [36], Hogenom [30], Phog [8], MetaPHOrs [31], PhylomeDB [18], Panther [26]. However, due to various limitations such as alignment errors, systematic artifacts of inference methods or unsufficient differentiation between sequences, trees are known to contain errors and uncertainties. Consequently, a great deal of effort has been put towards tools for gene tree editing [5,6,13,14,9,33,35]. Most of them are based on selecting, in a neighborhood of an input tree, one best fitting the species tree.

Recently, we developed the first algorithm for gene tree correction using orthology relations [20]. Here we address, from a complexity point of view, the more general problem of correcting a gene tree according to a set of orthology and paralogy relations. Two objective functions are considered: the number of unchanged relations and the number of unchanged clades (the Robinson-Foulds distance [32]). Conversely, we also address the problem of correcting a set of relations so that it represents a valid history in terms of an S-consistent gene tree. Two criteria are considered: maximize the number of unchanged relations, and minimize the number of genes that should be removed for the relation set to be S-consistent. These problems are all shown to be NP-complete.

We introduce the notations and known results in Section 2, and show the NP-completeness of two relation correction problems in Section 3, namely the Minimum Edge-Removal Consistency and Minimum Node-Removal Consistency problems. In Section 4, we then provide analogous complexity results for two gene tree correction problem: the Maximum Homology Correction and the Maximum Clade Correction problems. Algorithmic avenues are discussed in Section 5. Due to space constraints, some of the proof have been relegated to Supplementary materials, which can be accessed at `http://www-ens.iro.umontreal.ca/~lafonman/publications.php`.

2 Trees and orthology relations

All trees considered in this paper are assumed to be rooted. We also assume that trees have no nodes of degree 2, except possibly the root. Given a set X, a *tree T for X* is a tree whose leafset $\mathcal{L}(T)$ is in bijection with X. We denote by $V(T)$ the set of nodes and by $r(T)$ the root of T. 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, and we then call v a descendant of u. If u and v are connected by an edge of T, then v is a *direct descendant* of u. We denote by $ch(u)$ the set of direct descendants (children) of u. 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 say that u and v are *separated* iff $lca_T(u, v) \notin \{u, v\}$ (i.e. none is an ancestor of the other). We define $lca_T(U)$ analogously for a set U of nodes. Let L' be a subset of $\mathcal{L}(T)$. The restriction $T|_{L'}$ of T to L' is the tree with leaf set L' obtained from the subtree of T rooted as $lca_T(L')$ by removing all leaves that are not in L', and all internal nodes of degree 2, except the root. Let T' be a tree such that $\mathcal{L}(T') = L' \subseteq \mathcal{L}(T)$. We say that T displays T' iff $T|_{L'}$ is label-isomorphic to T'.

2.1 Evolution of a Gene Family

Species evolve through *speciation*, which is the separation of one species into distinct ones. 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. Inside the species' genomes, genes undergo speciation when the species to which they belong do, but also duplications, and losses (other events such as transfers can happen, but we ignore

them here). A *gene family* is a set of genes Γ accompanied by a *mapping function* $s : \Gamma \to \Sigma$ 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_G), where G is a tree with $\mathcal{L}(G) = \Gamma$, and $ev_G : V(G) \setminus \mathcal{L}(G) \to \{Dup, Spec\}$ is a function labeling each internal node of G as a duplication or a speciation node (we drop the G subscript from ev_G when it is clear from the context). Given a species tree S, the *LCA-mapping* function s_G maps each gene, ancestral or extant, to a species as follows: if $g \in \mathcal{L}(G)$, then $s_G(g) = s(g)$; otherwise, $s_G(g) = lca_S(\{s(g') : g' \in \mathcal{L}(G_g)\})$. An example is given in Figure 1, where the label of each node of G represents its LCA-mapping with respect to S.

According to the Fitch [11] 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 $ev(lca_G(x, y)) = Dup$. We denote by $\mathcal{O}(G)$, respectively $\mathcal{P}(G)$, the set of all gene pairs that are orthologous, respectively paralogous in G. By $xy \in \mathcal{O}(G)$ we mean $\{x, y\} \in \mathcal{O}(G)$ (the same applies for $\mathcal{P}(G)$). In Figure 1, $a_1c_1 \in \mathcal{O}(G)$ while $a_1b_1 \in \mathcal{P}(G)$. We say that a_1c_1 (respec. a_1b_1) is an orthology (respec. paralogy) relation *induced* by G.

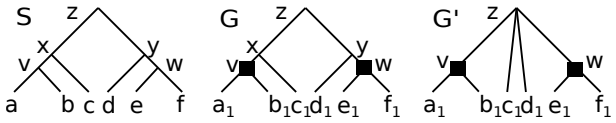

Fig. 1: A species tree S, a binary DS-tree G and a non-binary DS-tree G'. In DS-trees, *Dup* nodes are indicated by squares, and each leaf α_i denotes a gene belonging to the genome α. G is a refinement of G' such that $\mathcal{O}(G) = \mathcal{O}(G')$ and $\mathcal{P}(G) = \mathcal{P}(G')$.

While a history for Γ can be represented as a DS-tree, the converse is not always true, as 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 [19]. Formally G should be *S-consistent*, as defined below.

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 is*

S-consistent *iff for any* $v_1, v_2 \in ch(v)$, $s_G(v_1)$ *and* $s_G(v_2)$ *are separated in* S.

We say that G *is* S-consistent *iff every speciation node of* G *is* S-*consistent.*

Notice that G and S are not required to be binary. In particular, the definition of S-consistency for a speciation node v of G does not require v to be binary, even if S is binary. In this case, one can "refine" v into a set of binary S-consistent speciation nodes based on the topology of S. This operation does not affect the orthology and paralogy relations of by G (see Figure 1). Duplication nodes can be refined as well. Lemma 1 formalizes this intuition - we leave the proof to the Supplementary materials.

Lemma 1. *Let* G *be an* S*-consistent* DS*-tree for some binary species tree* S*. Then there is a binary* DS*-tree* G' *such that* G' *is* S*-consistent,* $\mathcal{O}(G) = \mathcal{O}(G')$ *and* $\mathcal{P}(G) = \mathcal{P}(G')$.

We can verify that both DS-trees in Figure 1 are S-consistent. For example, the speciation node in G' has children from species v, c, d and w, which are pairwise separated in S. Notice that, from Definition 1, if G is a DS-tree, then the lca of two leaves of G belonging to the same species must be a duplication node. The converse is not true. For example, in the S-consistent gene tree G of Figure 1, the parental node of e_1 and f_1 is a duplication node even though e_1 and f_1 belong to two different species.

2.2 Relation graph

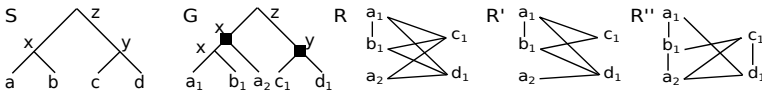

Fig. 2: A species tree S and a DS-tree G. The full orthology set induced by G is represented by the relation graph R. The following graph R' is an example of a not satisfiable graph, as $\{c_1, b_1, d_1, a_2\}$ induces a P_4, while R'' is an example of a satisfiable (it has no induced P_4), but not S-consistent graph.

A set of orthology/paralogy relations on Γ (or simply a relation set) is a pair $C = (C_O, C_P)$ of subsets $C_O, C_P \subseteq \binom{\Gamma}{2}$ such that $C_O \cap C_P = \emptyset$ and if $s(x) = s(y)$, then $\{x, y\} \in C_P$. The relation set is said *full* if $C_O \cup C_P = \binom{\Gamma}{2}$. A DS-tree G induces a full set $(\mathcal{O}(G), \mathcal{P}(G))$ of relations.

We adopt the graph representation considered in [19] for full relation sets. A *relation graph* R on a gene family Γ is a graph with vertex set $V(R) = \Gamma$, in which we interpret each edge uv of the edge set $E(R)$ of R as an orthology relation between u and v, and each missing edge (non-edge) $uv \notin E(R)$ as a paralogy relation [1]. Note that if $s(u) = s(v)$, then $uv \notin E(R)$. The relation graph of a DS-tree G, denoted by $R(G)$, is the graph with vertex set $\mathcal{L}(G)$ and edge set $\mathcal{O}(G)$ (for example, see the relation graph R in Figure 2).

A DS-tree for a gene family Γ leads to a relation graph, but the converse is not always true. A relation graph R is *satisfiable* if there exists a DS-tree G such that $R(G) = R$. The problem of relation graph satisfiability has been addressed in [15]. The following theorem is a reformulation of one of the main results of this paper.

Theorem 1 ([15]). *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 4.*

For example, in Figure 2, the relation graphs R and R'' are satisfiable, while the graph R' is not. As a DS-tree does not necessarily represent a true history for Γ (see previous section and Definition 1), satisfiability of a relation graph does not ensure a possible translation in terms of a history for Γ. For this to hold, R should 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.*

For example the graph R in Figure 2 is S-consistent. Note that S-consistency implies satisfiability. Results from [19] complete the characterization of S-consistent graphs through Theorem 2. A triplet is a binary tree with leafset L of size three. For $L = \{x, y, z\}$, we denote by $xy|z$ the unique triplet T on L for which $lca_T(x, y) \neq r(T)$ holds. Now $P_3(R)$ is the subset of triplets of species induced by paths of length 3 in $R = (V, E)$:

$$P_3(R) = \{s(x)s(y)|s(z) : \ zx, zy \in E \text{ and } xy \notin E \text{ and } s(x) \neq s(y)\}$$

Theorem 2. *Let $R = (V, E)$ be a satisfiable relation graph. Then R is S-consistent if and only if S displays all the triplets of $P_3(R)$.*

[1] It has been pointed out to us that the term 'relation graph' is also used in phylogenetics in the form of a generalization of a median network to a set of partitions. To make it clear, relation graphs in this paper have nothing to do with this notion

Theorem 2 is an immediate consequence of Theorem 5 in [19]. For the sake of completeness, we include the full proof in the Supplementary materials. As an example, the graph R'' in Figure 2 is satisfiable but not S-consistent as the path of length 3 containing $\{a_1, b_1, c_1\}$ induces the triplet $ac|b$, while the triplet displayed by S is $ab|c$.

We end this section with additional notations that will be of use later. A *subgraph* H' of H is a graph with $V(H') \subseteq V(H)$ and $E(H') \subseteq E(H)$. For a graph H and some $V' \subseteq V(H)$, the *subgraph of H induced by V'*, denoted $H[V']$, is the subgraph of H with vertex-set V' having the maximum number of edges. We say that H' is an *induced subgraph of H* if there is a subset $V' \subseteq V(H)$ such that $H' = H[V']$. 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. Finally, for some edge set $E' \subseteq E(H)$, $H - E'$ is the subgraph H' with $V(H') = V(H)$ and $E(H') = E(H) \setminus E'$.

3 Relation Correction Problems

We raise the issue of leaving out a minimum of information from a relation graph R in order to reach satisfiability or S-consistency. The problem limited to satisfiability reduces to modifying, i.e. adding or removing, a minimum number of edges of R in order to make it P_4-free, which is known to be NP-Hard [25]. In [16], an integer linear programming formulation is used to correct relation graphs of reasonable size.

We first extend the above problem to S-consistency: given a relation graph R and a species tree S, what is the minimum number of edges that need to be modified in order to reach S-consistency? Then, we study the problem of removing as few genes as possible from the gene family in order for the set of relations to be consistent.

3.1 The Minimum Edge-Removal Consistency Problem

Based on the same construction used in paper [10], we show that adding the information on the species tree S does not make the problem of removing the minimum number of edges leading to a P_4-free graph simpler. Although a similar reduction is likely to hold in the general case of edge-modification (removal or insertion) [25], here we focus on edge removal, as this formulation is needed in subsequent developments (Section 4). We show the NP-Completeness of this problem, even when every gene from the family Γ comes from a distinct species.

Minimum Edge-Removal Consistency Problem:
Input: A relation graph R for a gene family Γ, a species tree S and an integer k;
Output: An S-consistent subgraph R' of R with $V(R) = V(R')$ such that $|E(R) \setminus E(R')| \leq k$.

Theorem 3. *The* Minimum Edge-Removal Consistency Problem *is NP-Complete, even if for any distinct* $g_1, g_2 \in \Gamma$, $s(g_1) \neq s(g_2)$.

Proof. Given R', Theorem 2 easily translates into a polynomial-time algorithm to verify that R' is S-consistent. It is also clear that verifying if $|E(R) \setminus E(R')| \leq k$ can be done quickly. The problem is therefore in NP. As for the NP-Hardness, the reduction is from the exact 3-cover problem, a classic NP-Hard problem [27]: given a set $W = \{w_1, \ldots, w_{3t}\}$ and a collection $Z = \{Z_1, \ldots, Z_r\}$ of 3-elements of W, does there exists $Z' \subseteq Z$ such that $|Z'| = t$ and Z' is a partition of W ?

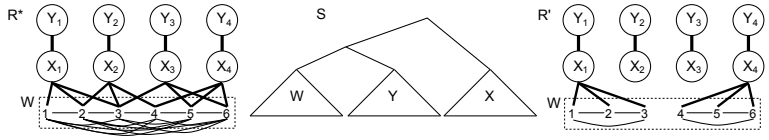

Fig. 3: S represents the species tree and R^* the relation graph constructed from the sets W, Z, X and Y. The illustration is given for $W = \{1, 2, 3, 4, 5, 6\}$ and $Z = \{\{1, 2, 3\}, \{2, 3, 4\}, \{3, 5, 6\}, \{4, 5, 6\}\}$. $Z' = \{\{1, 2, 3\}, \{4, 5, 6\}\}$ is a subset of Z which is a partition of W. R' is the "corrected" relation graph corresponding to Z'.

Given arbitrary W and Z, we construct R and S by first defining the species set Σ. Let $\alpha = \binom{3t}{2}$ and let $X = \{X_1, \ldots, X_r\}$ and $Y = \{Y_1, \ldots, Y_r\}$ be two collections of all disjoint sets (i.e. for any distinct set $A, B \in X \cup Y$, $A \cap B = \emptyset$), with $|X_i| = \alpha$ and $|Y_i| = r^2\alpha$, for all $1 \leq i \leq r$. Let $X_\Sigma = \bigcup_{1 \leq i \leq r} X_i$ and $Y_\Sigma = \bigcup_{1 \leq i \leq r} Y_i$ be the species in X and Y. Then the species set is $\Sigma = W \cup X_\Sigma \cup Y_\Sigma$. Let S_W, S_X and S_Y be three trees such that $\mathcal{L}(S_W) = W, \mathcal{L}(S_X) = X_\Sigma$ and $\mathcal{L}(S_Y) = Y_\Sigma$. Then S is obtained by first connecting $r(S_Y)$ with $r(S_W)$ to obtain a new tree S_{WY}, then connecting $r(S_{WY})$ with $r(S_X)$ (see Fig. 3). Therefore S has exactly $|\Sigma| = 3t + r(\alpha + r^2\alpha)$ leaves. The gene family Γ is then constructed so that it contains exactly one gene per species, as mentioned in the Theorem statement. In other words the mapping $s : \Gamma \to \Sigma$ is one-to-one. Since s is a bijection, we make

no distinction between a gene g and its species $s(g)$. We then define R with $V(R) = \Sigma$ such that each of the sets $W, X_1, \ldots, X_r, Y_1, \ldots, Y_r$ forms an individual clique. Finally we add two edge-sets E_1 and E_2 to R, where $E_1 = \{g_1g_2 : g_1 \in X_i, g_2 \in Z_i, \text{ for a given } 1 \leq i \leq r\}$ and $E_2 = \{g_1g_2 : g_1 \in X_i, g_2 \in Y_i, \text{ for a given } 1 \leq i \leq r\}$. Then R has $2r + 1$ cliques, namely $W, X_1, \ldots, X_r, Y_1, \ldots, Y_r$. Also, for $1 \leq i \leq r$, all edges between X_i and Y_i are present, as well as all edges between X_i and Z_i. Figure 3 gives an example with $t = 2$ and $W = \{1, 2, 3, 4, 5, 6\}$.

We show that W and Z admit an exact 3-cover if and only if R admits an S-consistent DS-tree after the deletion of at most $3\alpha(r-t)+(\alpha-3t)$ edges. Notice that the construction of R described above can clearly be done in polynomial time.

($\Rightarrow$) : let $Z' \subseteq Z$ be a partition of W, $|Z'| = t$. Let R' be the subgraph of R in which all edges between Z_i and X_i are removed iff $Z_i \notin Z'$ (which removes $3\alpha(r-t)$ edges), and the only edges not removed from the W-clique are those belonging to a Z_i triangle with $Z_i \in Z'$ (which removes $\alpha - 3t$ edges). An example of R' is given in Figure 3. Thus there are exactly $3\alpha(r-t)+(\alpha-3t)$ edges of R missing from R', as desired. Clearly, R' is P_4-free and thus satisfiable. To see that R' is S-consistent, we use Theorem 2. Notice that any path of length 3 in R' has the form wx_iy_i with $w \in W, x_i \in X_i$ and $y_i \in Y_i$ for some i, inducing the $wy_i|x_i$ speciation triplet, which is in agreement with S. Therefore there exists an S-consistent gene tree G' satisfying R'.

($\Leftarrow$) : The construction of R is exactly the same as in Theorem 3 in [10], and the proof is directly applicable to our case. Still, we have included a complete proof in the Supplementary materials. □

3.2 The Minimum Node-Removal Consistency Problem

Minimum Node-Removal Consistency Problem:
Input: A relation graph R for a gene family Γ, a species tree S and an integer k;
Output: An S-consistent induced subgraph R' of R with $|V(R')| \geq k$.

We use a reduction similar to that in [23], where it was remarkably shown that finding a maximum induced subgraph of some graph H having some property Π is NP-Hard whenever Π is a hereditary property, i.e. applies to any induced subgraph of H. Though it can be shown that S-consistency is indeed hereditary, the reduction assumes H is unlabeled and unconstrained, which is not the case of R .

Theorem 4. *The* Minimum Node-Removal Consistency Problem *is NP-Complete.*

Proof. Again by Theorem 2, verifying that R' is indeed a solution can be done in polynomial time and the problem is thus in NP. The reduction is from the maximum independent set problem. That is, given a graph H, is there an induced subgraph H' of H having at least k nodes such that H' has no edge. Let $n = |V(H)|$. We construct R and S from H as follows: R starts as a copy of H, and for each node x of R, we add a single neighbor x^* (i.e. xx^* is an edge of R and x^* is of degree one). Denote by X the nodes of R originally from H, and by X^* the newly added nodes. Each gene in R is assigned to a distinct species. To construct S, first let S_X be a tree with leafset $s(X)$, and S_{X^*} be a tree with leafset $s(X^*)$. Then S is obtained by connecting $r(S_X)$ and $r(S_{X^*})$ under a common parent. We show that H has an independent set of size at least k if and only if R admits an induced subgraph of size at least $n + k$ that is S-consistent.

Let H' be a solution to the independent set problem with $|V(H')| \geq k$, and let X' be the nodes of X corresponding to $V(H')$. Let $R' = R[X' \cup X^*]$. Now, no two nodes of X' share an edge, and thus the only edges left in R' are of the form xx^*. Therefore, R' is P_3-free and thus, by Theorem 2, is S-consistent. Moreover, $|V(R')| = |X' \cup X^*| \geq k + n$.

Conversely, let R' be an S-consistent induced subgraph of R with $|V(R')| \geq n + k$. Let $W = \{x \in X : x \in V(R') \text{ and } x^* \in V(R')\}$. We first claim that no two nodes $x, y \in W$ share an edge in R'. For otherwise, x^*xy induce a P_3 with x in the center, inducing the $s(x^*)s(y)|s(x)$ speciation triplet. This contradicts the triplet $s(x)s(y)|s(x^*)$ found in S, and R' is not S-consistent. Therefore, by letting W' denote the nodes of H corresponding to W, we get that $H[W']$ is an independent set. Our final claim is that $|W| \geq k$. Indeed if $|W| < k$, then there are strictly more than $n-k$ node pairs $\{x, x^*\}$ from which at least one of x or x^* is missing in R'. This implies that $|V(R')| < 2n-(n-k) = n+k$, contradicting our initial assumption. □

4 Gene Tree Correction Problems

In this section, we consider we are given a gene family Γ, a species tree S, an S-consistent DS-tree G for Γ, and a set $C = (O, P)$ of orthology/paralogy constraints (not necessarily full). We focus on the problem of correcting G according to C in a minimal way. The goal is thus to find a DS-tree G' inducing C such that the difference between G and G' is minimum. We consider two ways of measuring the difference (or

symetrically the similarity) between gene trees, one based on conserved orthology/paralogy relations induced by the two trees, and one based on the number of conserved clades between the two trees, which is the Robinson-Foulds in the case that G, G' and S are all binary trees.

4.1 The Maximum Homology Correction Problem

Maximum Homology Correction Problem :
Input: A species tree S, an S-consistent DS-tree G for a gene family Γ, an integer k, a set O of orthology and a set P of paralogy relations;
Output: An S-consistent DS-tree G' for Γ with $O \subseteq \mathcal{O}(G')$, $P \subseteq \mathcal{P}(G')$ such that $|\mathcal{O}(G) \cap \mathcal{O}(G')| + |\mathcal{P}(G) \cap \mathcal{P}(G')| \geq k$.

Theorem 5. *The* Maximum Homology Correction Problem *is NP-Complete, even if S, G and G' are required to be binary.*

Proof. The problem is clearly in NP, as verifying S-consistency can be done in polynomial time, as well as counting the common orthologs/paralogs relations (the set of relations is quadratic in size). For our reduction, we use the Minimum Edge-Removal Consistency Problem for the case of a gene family with at most one gene per genome, which is NP-Hard by Theorem 3. Given a species tree S, a relation graph R with $V(R)$ in bijection with $\mathcal{L}(S)$ and an integer k, we construct an instance of the Maximum Homology Correction Problem, i.e. a species tree S', a DS-tree G, an orthologous set O and paralogous set P. We show that that there is an S-consistent subgraph R' of R obtained by removing at most k edges iff there is an S'-consistent DS-tree G' satisfying O and P with at most $|P| + k$ relations that are not induced by G.

Let $S' = S$ and construct G by mimicking S - that is by first copying S and its leaf labels, then replacing each leaf ℓ of G by the gene $s^{-1}(\ell)$. Note that if S is binary, then so is G. All internal nodes of G are labeled as speciations, so all genes of Γ are pairwise orthologous. Thus $R(G)$ is a clique. Finally, let $O = \emptyset$ and $P = \{g_1g_2 : g_1g_2 \notin E(R)\}$. Notice that $R = R(G) - P$.

$\Rightarrow$: Let R' be a solution to the Minimum Edge-Removal Consistency Problem for R and S. Then there exists a S-consistent DS-tree G' satisfying R', which is obtained by deleting at most k edges from R. By Lemma 1, we may assume that if S is binary, then so is G'. Now, since R' has at most $|P| + k$ non-edges, G' has at most $k + |P|$ paralogs and is therefore a solution to the constructed instance of the Maximum Homology Correction Problem that breaks at most $k + |P|$ orthologies.

$\Leftarrow$: Let G' be a solution, binary or not, to the constructed Maximum Homology Correction Problem instance and let $R' = R(G')$. Since G' satisfies P and breaks at most $|P| + k$ orthologies, R' must have P as non-edges, plus at most k other non-edges. Thus R' can be obtained by removing at most k edges from $R(G) - P = R$, as desired. □

4.2 The Maximum Clade Correction Problem

Maximum Clade Correction Problem:
Input: A gene tree G, a species tree S, a set O of orthology and a set P of paralogy relations and an integer k;
Output: An S-consistent DS-tree G' satisfying O and P such that G and G' have at least k clades in common.

Notice that if S, G and G' are required to be binary, the effective measure between G and G' is the Robinson-Foulds distance. This special case is handled as part of the general proof.

Theorem 6. *The* Maximum Clade Correction Problem *is NP-Complete, even if S, G and G' are required to be binary.*

The proof of Theorem 6 is a bit involving, and due to space constraints we only provide the construction and intuition of the NP-Hardness reduction. The complete proof can be accessed in the supplementary materials.

We use the Minimum Node-Removal Consistency Problem for our reduction. Let R be the input relation graph with $V(R) = \{v_1, \ldots, v_n\}$, S be the species tree and k be an integer. Let $\alpha = n(n-1-k) + 2k$. The constructed instance of the Maximum Clade Correction Problem uses the same species tree S. Construct G as follows: first consider G as a binary tree with n leaves $l_1, \ldots, l_n$, where each leaf l_i is mapped to v_i. Then, replace each leaf l_i by a subtree T_i constructed as follows: T_i is a caterpillar tree with $n - 1 + \alpha$ leaves, and each leaf ℓ of T_i is such that $s(\ell) = s(v_i)$ (a caterpillar tree is a path to which we add a leaf child to each internal node). Let L_i denote the set of the $n-1$ deepest leaves of T_i (the depth of a leaf ℓ being the number of nodes on the path between ℓ and the root). Each leaf of L_i is mapped to a distinct node of $V(R)\backslash\{v_i\}$. Denote by $\ell_{i,j}$ the leaf of T_i mapped to v_j. Then G has exactly $n(n-1+\alpha)$ leaves and $n(n-1+\alpha)-1$ clades (since it is binary). Finally define $O = \{\{\ell_{i,j}, \ell_{j,i}\} : v_iv_j \in E(R)\}$ the set of orthology relations and $P = \{\{\ell_{i,j}, \ell_{j,i}\} : v_iv_j \notin E(R)\}$ the set of paralogy relations. Note that each $\ell_{i,j}$ is present in exactly one relation.

It can be shown that R admits an S-consistent induced subgraph R' with at least k nodes if and only if G, O and P admit an S-consistent DS-tree G' satisfying O and P such that G and G' share at least $k(\alpha+n-2)$ clades. The idea is that given R', we can construct an S-consistent gene tree H satisfying R. To each leaf v_i of H corresponds a subtree T_i in G. We obtain G^* by replacing each such leaf v_i by its corresponding T_i, which guarantees that the required number of clades were preserved (as there were k such leaves in H). Noting that G^* does not include every gene of G, the difficulty of the proof consists in including every such missing gene whilst satisfying the relations of O and P.

In the other direction, i.e. if we are given a solution G' that preserves enough clades, it can be shown that G' must preserve at least k of the T_i subtrees intact, and restricting G' to these k subtrees, then replacing each such T_i by its corresponding vertex v_i in R, we obtain a gene tree G^* whose relation graph R' is the solution we are looking for.

5 Algorithmic avenues

As the problems presented in this work are NP-complete, non-polynomial exact algorithms or approximation algorithms avenues should be explored. Let us generalize the Minimum Edge-Removal Consistency Problem to the minimum *editing* problem (i.e. minimzing edge removals and insertions). It is not hard to imagine a branch-and-bound algorithm that solves the problem. Call an induced subgraph H of a relation graph R *bad* if it is either a P_4, or a P_3 in contradiction with S. Each P_4 can be solved by 6 possible edge editings, and each contradictory P_3 can be solved by 3 possible editings. Therefore, in a branch-and-bound process, one would verify if a given graph R' contains a bad subgraph and if so, proceed recursively on each graph obtained by an editing that removes it. If no bad subgraph exists, then R' is a possible solution and its number of editings is retained. If, at any point, R' has had more editings than the best solution encountered so far, the algorithm can stop the recursion. Notice however that an edge should not be edited more than once in order to avoid infinite loops. The idea of this branch-and-bound algorithm can also be applied to the Minimum Node-Removal Consistency problem. It is known that a P_4, if one exists, can be found in linear time [4]. It remains to see if we can find a contradictory P_3 in time better than $O(n^3)$.

As for approximations, an algorithm proposed in [28] can be directly applied to the Minimum Edge-Removal Consistency Problem and guarantees that we do not remove more than $4\Delta(R)$ times more edges than the

optimal solution, where $\Delta(R)$ is the maximum degree of R. The idea is simple : as long as R has a bad subgraph H, remove every edge incident to a vertex of H and continue. Even though this is the best known approximation algorithm so far, it has the undesirable effect of isolating many vertices, motivating the exploration of alternative algorithms. One direction would be to consider existing ideas on the problem of satisfiability, i.e. what is the minimum number of editings required to make a graph P_4-free, and adapt them to the consistency problem - for instance the Min-Cut algorithm proposed in [2].

For gene tree correction, we have developed in [19] a polynomial-time algorithm which, given a species tree S and a partial set of relations O and P, verifies if there exists an S-consistent gene tree G' satisfying O and P and if so, constructs one among the set of all possible solutions. It would be interesting to explore the possibility of providing an input gene tree G to the algorithm in order to pick a solution that is close to G (either in terms of common homology relations or clades).

It is also worth mentioning that relations are not always fully known, and instead of a yes or no orthology assignment between two genes, existing methods for orthology prediction can rather motivate a way of assigning a probabilistic score to a given relation [19]. A natural extension to the edge removal/editing problems is therefore to add a weight to each edge and non-edge, so that each insertion/removal has its own weight. The objective then becomes to minimize the total weight of a set of edited edges. Notice that the branch-and-bound algorithm given above can easily be adapted to support weights on editings. This generalization actually encompasses the Maximum Homology Correction problem. Indeed, given a gene tree G and relations O and P to satisfy, one can create a weighted relation graph R in this way: each relation in O (resp. P) is an edge (resp. non-edge) with infinite weight, and each relation in $\mathcal{O}(G) \setminus O$ (resp. $\mathcal{P}(G) \setminus P$) is an edge (resp. non-edge) with a weight of one. Therefore a minimum S-consistent edge-editing of R corresponds to a gene tree G' that satisfies O and P and has a maximum number of common homologies with G.

6 Conclusion

A gene tree induces a set of orthology and paralogy relations between members of a gene family, but the converse is not always true. In this paper we show that attempting to modify a set of relations as least as possible in order to ensure consistency with a species tree leads to the

formulation of NP-Complete problems. Moreover, even assuming that the given relations are error-free, it remains computationally difficult to correct a gene tree in order to fit the given set of relations. As various model-free methods are available to infer orthology and paralogy, these correction problems are of practical biological interest. A future direction would be to explore fast approximation algorithms for the relation graph and gene tree editing.

References

1. A.V. Aho, Y. Sagiv, T.G. Szymanski, and J.D. Ullman. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM J. Comp.*, 10:405- 421, 1981.
2. A. M. Altenhoff, M. Gil, G. H. Gonnet, and C. Dessimoz. Inferring hierarchical orthologous groups from orthologous gene pairs. *PLoS One*, 8(1):e53786, 2013.
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. A. Bretscher, D. Corneil, M. Habib, and C. Paul. A simple linear time lexbfs cograph recognition algorithm. In *Graph-Theoretic Concepts in Computer Science*, pages 119–130. Springer, 2003.
5. R. Chaudhary, J.G. Burleigh, and O. Eulenstein. Efficient error correction algorithms for gene tree reconciliation based on duplication, duplication and loss, and deep coalescence. *BMC Bioinformatics*, 13(Supp.10):S11, 2011.
6. K. Chen, D. Durand, and M. Farach-Colton. Notung: Dating gene duplications using gene family trees. *J. Comp. Biol.*, 7:429–447, 2000.
7. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nat. Genet.*, 25(1):25 - 29, 2000.
8. R.S. Datta, C. Meacham, B. Samad, C. Neyer, and K. Sjölander. Berkeley PHOG: PhyloFacts orthology group prediction web server. *Nucleic Acids Res.*, 37:W84-W89, 2009.
9. A. Doroftei and N. El-Mabrouk. Removing noise from gene trees. In *WABI 2011*, volume 6833 of *Lecture Notes in Bioinformatics*, pages 76-91, 2011.
10. Ehab S El-Mallah and Charles J Colbourn. The complexity of some edge deletion problems. *Circuits and Systems, IEEE Transactions on*, 35(3):354–362, 1988.
11. W. M. Fitch. Homology. a personal view on some of the problems. *TIG*, 16(5):227-231, 2000.
12. 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.
13. P. Gorecki and O. Eulenstein. Algorithms: simultaneous error-correction and rooting for gene tree reconciliation and the gene duplication problem. *BMC Bioinformatics*, 13(Supp 10):S14, 2011.
14. P. Gorecki and O. Eulenstein. A linear-time algorithm for error-corrected reconciliation of unrooted gene trees. In *ISBRA 2011*, volume 6674 of *Lecture Notes in Bioinformatics*, pages 148-159, 2011.
15. 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.

16. Marc Hellmuth, Nicolas Wieseke, Markus Lechner, Hans-Peter Lenhof, Martin Middendorf, and Peter F Stadler. Phylogenomics with paralogs. *PNAS*, 2014.
17. M. Hernandez-Rosales, M. Hellmuth, N. Wieseke, K.T. Huber, V. Moulton, and P. Stadler. From event-labeled gene trees to species trees. *BMC Bioinformatics*, 13 (Suppl. 19):56, 2012.
18. J. Huerta-Cepas J, S. Capella-Gutierrez S, L.P. Pryszcz LP, I. Denisov, D. Kormes, M. Marcet-Houben, and T. Gabald'on. Phylomedb v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res.*, 39:D556-D560, 2011.
19. M. Lafond and N. El-Mabrouk. Orthology and paralogy constraints: satisfiability and consistency. *BMC Genomics*, 15(Suppl 6):S12, 2014.
20. 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.
21. M. Lafond, K. Swenson, and N. El-Mabrouk. *Models and algorithms for genome evolution*, chapter Error detection and correction of gene trees. Springer, 2013.
22. 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.
23. J. M Lewis and M. Yannakakis. The node-deletion problem for hereditary properties is np-complete. *J. of Computer and System Sciences*, 20(2):219–230, 1980.
24. L. Li, C.J. Jr. Stoeckert, and D.S. Roos. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13:2178- 2189, 2003.
25. Y. Liu, J. Wang, J. Guo, and J. Chen. Cograph editing: Complexity and parameterized algorithms. In *Computing and Combinatorics*. Springer, 2011.
26. H. Mi, A. Muruganujan, and P.D. Thomas. Panther in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucl. Acids Res.*, 41:D377-D386, 2012.
27. R Garey Michael and S Johnson David. Computers and intractability: a guide to the theory of np-completeness. *WH Freeman & Co., San Francisco*, 1979.
28. A. Natanzon, R. Shamir, and R. Sharan. Complexity classification of some edge modification problems. *Discrete Applied Mathematics*, 113(1):109–128, 2001.
29. S. Ohno. *Evolution by gene duplication*. Springer, Berlin, 1970.
30. S. Penel, A.M. Arigon, J.F. Dufayard, A.S. Sertier, V. Daubin, L. Duret, M. Gouy, and G. Perrière. Databases of homologous gene families for comparative genomics. *BMC Bioinf.*, 10 S6:S3, 2009.
31. L.P. Pryszcz, J. Huerta-Cepas, and T. Gabaldón. MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucl. Acids Res.*, 39:e32, 2011.
32. D. Robinson and L. Foulds. Comparison of phylogenetic trees. *Math. Biosc.*, 53:131–147, 1981.
33. K. M. Swenson, A. Doroftei, and N. El-Mabrouk. Gene tree correction for reconciliation and species tree inference. *Algorithms Mol. Biol.*, 7(31), 2012.
34. 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.
35. T.H. Nguyen TH, V. Ranwez, S. Pointet S, A.M. Chifolleau, J.P. Doyon, and V. Berry. Reconciliation and local gene tree rearrangement can be of mutual profit. *Alg Mol. Biol.*, 8(8):12, 2013.
36. A.J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, and E. Birney. EnsemblCompara gene trees: Complete, duplication-aware phylogenetic trees in vertebrates. *Gen. Res.*, 19:327-335, 2009.