


Parsimonious Migration History Problem: Complexity and Algorithms

Mohammed El-Kebir

Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL
melkebir@illinois.edu

 <https://orcid.org/0000-0002-1468-2407>

Abstract

In many evolutionary processes we observe extant taxa in different geographical or anatomical locations. To reconstruct the migration history from a given phylogenetic tree T , one can model locations using an additional character and apply parsimony criteria to assign a location to each internal vertex of T . The *migration criterion* assumes that migrations are independent events. This assumption does not hold for evolutionary processes where distinct taxa from different lineages comigrate from one location to another in a single event, as is the case in metastasis and in certain infectious diseases. To account for such cases, the *comigration criterion* was recently introduced, and used as an optimization criterion in the PARSIMONIOUS MIGRATION HISTORY (PMH) problem. In this work, we show that PMH is NP-hard. In addition, we show that a variant of PMH is fixed parameter tractable (FPT) in the number of locations. On simulated instances of practical size, we demonstrate that our FPT algorithm outperforms a previous integer linear program in terms of running time.

2012 ACM Subject Classification Applied computing → Molecular evolution, Mathematics of computing → Trees, Mathematics of computing → Combinatorial optimization

Keywords and phrases Reconciliation, maximum parsimony, metastasis, infection, phylogenetics, phylogeography, fixed parameter tractability

Digital Object Identifier 10.4230/LIPIcs.WABI.2018.24

1 Introduction

A phylogenetic tree, or phylogeny for short, models an evolutionary process such as those that arise in the development of cancer, species, pathogens and languages. In a character-based phylogeny, taxa are described by the same set of traits, where each trait is modeled by a single character with discrete states. Mathematically, a *character-based phylogeny* is a tree T whose vertices are taxa labeled by a vector of character states, assigning a single state to each character in each taxon. While the leaves of T correspond to extant taxa with observed character states, the internal vertices and edges of T are typically inferred algorithmically using different criteria such as maximum parsimony [4] or maximum likelihood [3].

In some evolutionary processes, extant taxa occur in different geographical or anatomical locations and one wishes to reconstruct the locations of ancestral taxa. Slatkin and Maddison [10] noted that locations can be modeled by an additional character and introduced the *migration criterion*. That is, given a phylogeny T with a location $\ell(u)$ assigned to each leaf u , the authors proposed to assign a location $\ell(v)$ to each internal vertex v of T such that the number $\mu(T, \ell)$ of migrations, i.e. edges (v, w) where $\ell(v) \neq \ell(w)$, is minimized (Fig. 1). Slatkin and Maddison [10] noted that this is an instance of the small phylogeny problem and can be solved in polynomial time [4, 9]. Later, McPherson et al. [7] used the migration criterion to study the migration history in metastatic ovarian cancers, where locations are distinct anatomical locations with metastasis that occur within the same patient.



© Mohammed El-Kebir;

licensed under Creative Commons License CC-BY

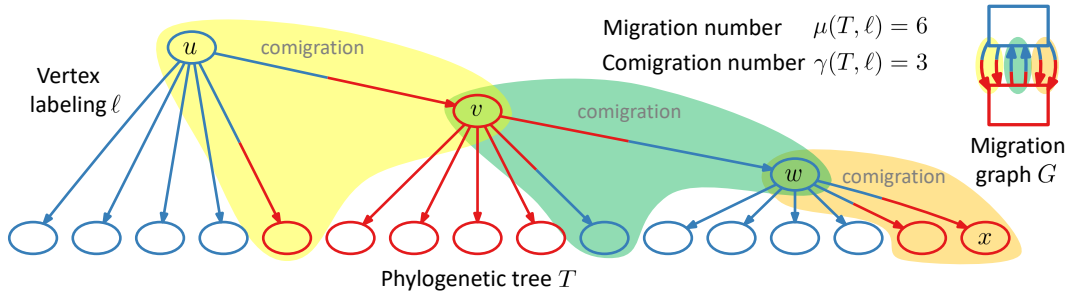
18th International Workshop on Algorithms in Bioinformatics (WABI 2018).

Editors: Laxmi Parida and Esko Ukkonen; Article No. 24; pp. 24:1–24:14

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



■ **Figure 1** Migration histories are labelings of the vertices of a phylogenetic tree by locations. Vertex labeling ℓ assigns to each vertex u a location $\ell(u)$. A migration is an edge (u, v) where $\ell(u) \neq \ell(v)$. Here, vertex labeling ℓ of T incurs 6 migrations (bichromatic edges) and thus has migration number $\mu(T, \ell) = 6$. Although migrations (u, v) and (w, x) have identical source and target locations, they could not have happened simultaneously as taxon w is a descendant of taxon v . A comigration is a set of migrations between the same pair of locations that occur on distinct branches of the phylogenetic tree T . The comigration number is the smallest partition of migrations into comigrations. Here, vertex labeling ℓ of T has comigration number $\gamma(T, \ell) = 3$. The set of migrations determines the migration graph G , a multi-graph whose vertices are locations.

Underlying the migration criterion is the assumption that migrations are independent events. While this assumption is justified for evolutionary processes where shared location of taxa is uninformative, in certain evolutionary processes migrations of distinct taxa between the same locations are not independent events. For instance, in cancer, metastases may be seeded by groups of tumor cells from different clones (taxa) that migrate together in the bloodstream or lymphatics. Moreover, in certain infectious diseases, such as those caused by Hepatitis-B and C, HIV and Ebola, different strains of the pathogen can infect a person through a single transmission event. As such, the migration criterion does not always apply.

Recently, El-Kebir et al. [2] introduced the *comigration criterion*, where multiple migrations between the same pair of locations are counted as a single event (Fig. 1). The authors showed that the problem of assigning locations to ancestral taxa under the migration and comigration criteria is a multi-objective optimization problem, with tradeoffs between both criteria and the topology of the *migration graph*, a directed multi-graph capturing all migrations between locations (Fig. 2). This problem was called the PARSIMONIOUS MIGRATION HISTORY (PMH) and was solved using an integer linear program (ILP). The hardness of PMH was an open problem.

In this work, we show that PMH is NP-hard. On the positive side, we show that when the migration graph is restricted to a tree, PMH is fixed parameter tractable (FPT) in the number of locations. Using simulated data of metastatic cancers, we show that for practical PMH instances with a small number of locations, the FPT algorithm outperforms the ILP in terms of running time.

2 Preliminaries

Let T be a tree rooted at vertex $r(T)$. We denote the edge set by $E(T)$, the vertex set by $V(T)$ and the leaf set by $L(T)$. We denote by T_v the subtree of T rooted at vertex v . We denote the parent vertex of a non-root vertex $v \neq r(T)$ by $\pi(v)$. We write $u \preceq_T v$ if and only if vertex u occurs on the unique path from $r(T)$ to v . Note that \preceq_T is reflexive, i.e. $v \preceq_T v$ for all vertices $v \in V(T)$. We write $u \prec_T v$ if and only if $u \preceq_T v$ and $u \neq v$. We denote the lowest common ancestor of a vertex subset $U \subseteq V(T)$ by $\text{LCA}_T(U)$. We say that two vertices

u, v are *incomparable* or occur on *distinct branches* if and only if $u \not\leq_T v$ and $v \not\leq_T u$. Note that $u \not\leq_T v$ and $v \not\leq_T u$ if and only if $\text{LCA}_T(\{u, v\}) \notin \{u, v\}$.

A *phylogenetic tree* T is a rooted tree, whose leaves correspond to taxa that are labeled by different locations. We denote the set of locations by Σ . Since the leaf set $L(T)$ is composed of taxa observed at the present time, we know their locations and are thus given the function $\hat{\ell} : L(T) \rightarrow \Sigma$, assigning location $\hat{\ell}(v)$ to each leaf $v \in L(T)$. We use $\hat{\ell}(L(T_u))$ where $u \in V(T)$ as a shorthand for $\{\hat{\ell}(v) \mid v \in L(T_u)\}$. Typically, the edges of a phylogenetic tree are labeled by mutations. As the problem that we consider does not use mutations, they are omitted.

3 Problem Statement

Given a phylogenetic tree T with leaf labeling $\hat{\ell} : L(T) \rightarrow \Sigma$, the task is to reconstruct the migration history by inferring the location of origin of each ancestral taxon. In other words, we wish to extend the given leaf labeling $\hat{\ell}$ to a *vertex labeling* $\ell : V(T) \rightarrow \Sigma$ such that $\ell(v) = \hat{\ell}(v)$ for each leaf $v \in L(T)$. To distinguish different vertex labelings ℓ of a phylogenetic tree T , we use two different optimization criteria.

The first criterion was introduced by Slatkin and Maddison [10] and considers migrations, which are defined as follows.

► **Definition 1** ([2, 10]). A *migration* is an edge $(u, v) \in E(T)$ that connects two vertices with different locations, i.e. $\ell(u) \neq \ell(v)$.

The *migration number* $\mu(T, \ell)$ is the number of migrations incurred by ℓ . Slatkin and Maddison [10] noted that the problem of finding a vertex labeling with minimum migration number is an instance of the small phylogeny maximum parsimony problem with a single multi-state character (with states Σ). As such, this problem can be solved in polynomial time using either the Fitch [4] or the Sankoff algorithm [9].

The second criterion, which was recently introduced in [2], allows for the simultaneous migration, or *comigration*, of individuals from different (ancestral) taxa between the same locations. This criterion is applicable to evolutionary processes where locations are seeded by individuals from distinct taxa through a single event. Two migrations $(u, v) \neq (w, x)$, where $v \leq_T w$, between the same pair of locations, i.e. $\ell(u) = \ell(w)$ and $\ell(v) = \ell(x)$, could never have occurred simultaneously. This is because taxon w is a descendant of taxon v , and thus migration (u, v) must have occurred prior to migration (w, x) (Fig. 1). To account for such scenarios, we define comigrations as follows.

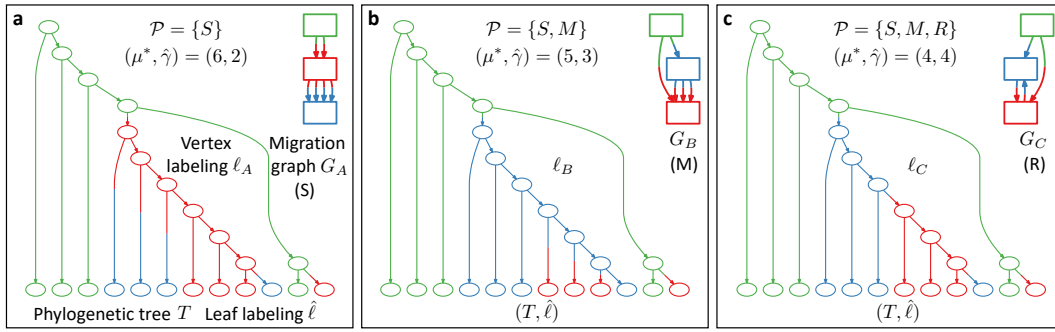
► **Definition 2** ([2]). A *comigration* is a subset X of pairwise incomparable migrations between the same locations. That is, for all distinct pairs $(u, v), (u', v') \in X$ of migrations it holds that (i) $\ell(u) \neq \ell(v)$, (ii) $\ell(u) = \ell(u')$, (iii) $\ell(v) = \ell(v')$, (iv) $v \not\leq_T v'$ and (v) $v' \not\leq_T v$.

Employing the principle of parsimony, we define the *comigration number* $\gamma(T, \ell)$ as the smallest partition of migrations into comigrations [2]. As shown in [2], we have that

$$\gamma(T, \ell) = \sum_{s, t \in \Sigma : s \neq t} \gamma(T, \ell, s, t) \quad (1)$$

where $\gamma(T, \ell, s, t)$ is the maximum number of edges $(u, v) \in E(T)$ with $\ell(u) = s$ and $\ell(v) = t$ that are on the same path of T starting from the root $r(T)$.

The set of migrations incurred by a vertex labeling ℓ of T determines the *migration graph* G . More formally, the vertices of G are locations, and there is a directed edge (s, t) for each edge (u, v) in T where (i) $s \neq t$, (ii) $\ell(u) = s$ and (iii) $\ell(v) = t$. The migration



■ **Figure 2** The Parsimonious Migration History (PMH) problem is a constrained multi-objective optimization problem, with tradeoffs between the migration number, comigration number and the migration pattern. Given a phylogenetic tree T whose leaves are labeled by locations via $\hat{\ell}$ and a set \mathcal{P} of allowed migration patterns, the task is to extend $\hat{\ell}$ to a vertex labeling ℓ such that: (i) the resulting migration graph G adheres to \mathcal{P} , (ii) ℓ has the minimum migration number $\mu^*(T)$ and (iii) subsequently smallest comigration number $\hat{\gamma}(T)$. (a) In the most restrictive case, $\mathcal{P} = \{S\}$, the migration graph is required to be a tree, yielding vertex labeling ℓ_A with $\mu(T, \ell_A) = 6$ and $\gamma(T, \ell_A) = 2$. The resulting migration graph G_A has a single-source seeding (S) pattern. (b) In the case $\mathcal{P} = \{S, M\}$, the migration graph is required to not contain a directed cycle, yielding vertex labeling ℓ_B with $\mu(T, \ell_B) = 5$, $\gamma(T, \ell_B) = 3$, and migration graph G_B with a multi-source seeding (M) pattern. (c) In the case $\mathcal{P} = \{S, M, R\}$, the migration graph is left unrestricted, yielding vertex labeling ℓ_C with minimum migration number $\mu(T, \ell_C) = 4$, smallest comigration number $\gamma(T, \ell_C) = 4$, and migration graph G_C with a reseeded (R) pattern.

graph G is a multi-graph because there may exist multiple directed edges between the same pair of locations. Different topologies of G correspond to different *migration patterns*. We distinguish three migration patterns: (i) in *single-source seeding* (S), the migration graph G is a tree (Fig. 2a); (ii) in *multi-source seeding* (M), some locations are the target of migrations from distinct source locations but G itself does not have a directed cycle (Fig. 2b); and (iii) in *reseeding* (R), G has a directed cycle (Fig. 2c).

There are tradeoffs between the migration number, the comigration number and the migration pattern, as shown in [2] and briefly reviewed in the following. Let T be a phylogenetic tree whose leaves are labeled by $m = |\Sigma|$ locations. Any vertex labeling ℓ of T has migration number $\mu(T, \ell) \geq m - 1$ and comigration number $\gamma(T, \ell) \geq m - 1$. While there may not exist a vertex labeling ℓ with migration number $\mu(T, \ell) = m - 1$, there always exists a vertex labeling ℓ with comigration number $\gamma(T, \ell) = m - 1$ for any leaf-labeled tree T . For instance, a labeling ℓ that assigns the same location to all internal vertices has comigration number $\gamma(T, \ell) = m - 1$. Moreover, a vertex labeling ℓ incurs a single-source seeding (S) pattern if and only if $\gamma(T, \ell) = m - 1$. Let $\mu^*(T) = \min_{\ell} \mu(T, \ell)$ be the minimum migration number of T . In general, there may not exist a vertex labeling ℓ of T that simultaneously achieves the minimum migration number $\mu(T, \ell) = \mu^*(T)$ and minimum comigration number $\gamma(T, \ell) = m - 1$ (Fig. 2). To examine these tradeoffs, El-Kebir et al. [2] introduced the following constrained multi-objective optimization problem that considered three different sets \mathcal{P} of allowed migration patterns: (i) $\mathcal{P} = \{S\}$, requiring the migration graph G to be an S pattern (Fig. 2a); (ii) $\mathcal{P} = \{S, M\}$, requiring the migration graph to be either an S or M pattern (Fig. 2b); (iii) $\mathcal{P} = \{S, M, R\}$ meaning that G is unrestricted (Fig. 2c).

► **Problem 1** (PARSIMONIOUS MIGRATION HISTORY (PMH) [2]). *Given a phylogenetic tree T with leaf labeling $\hat{\ell} : L(T) \rightarrow \Sigma$ and a set \mathcal{P} of allowed migration patterns, find a vertex labeling ℓ such that: (i) $\ell(v) = \hat{\ell}(v)$ for each leaf $v \in L(T)$; (ii) ℓ has the minimum migration*

number $\mu(T, \ell) = \mu^*(T) = \min_{\ell'} \mu(T, \ell')$ and subsequently the smallest comigration number $\gamma(T, \ell) = \hat{\gamma}(T) = \min_{\ell': \mu(T, \ell') = \mu^*(T)} \gamma(T, \ell')$; and (iii) the resulting migration graph G is a tree if $\mathcal{P} = \{S\}$, a directed acyclic graph if $\mathcal{P} = \{S, M\}$ or unrestricted if $\mathcal{P} = \{S, M, R\}$.

4 Results

We have the following two results for the case where the migration graph G is restricted to a tree (i.e. $\mathcal{P} = \{S\}$).

► **Theorem 3.** PMH is NP-hard when $\mathcal{P} = \{S\}$.

► **Theorem 4.** PMH is fixed parameter tractable in $|\Sigma|$ when $\mathcal{P} = \{S\}$.

Both theorems rely on the following important proposition that we prove first.

► **Proposition 5.** Let T be a phylogenetic tree, and let ℓ be a vertex labeling of T such that the migration graph G is a tree. Then, $\ell(u) \preceq_G \text{LCA}_G(\hat{\ell}(L(T_u)))$ for any vertex $u \in V(T)$.

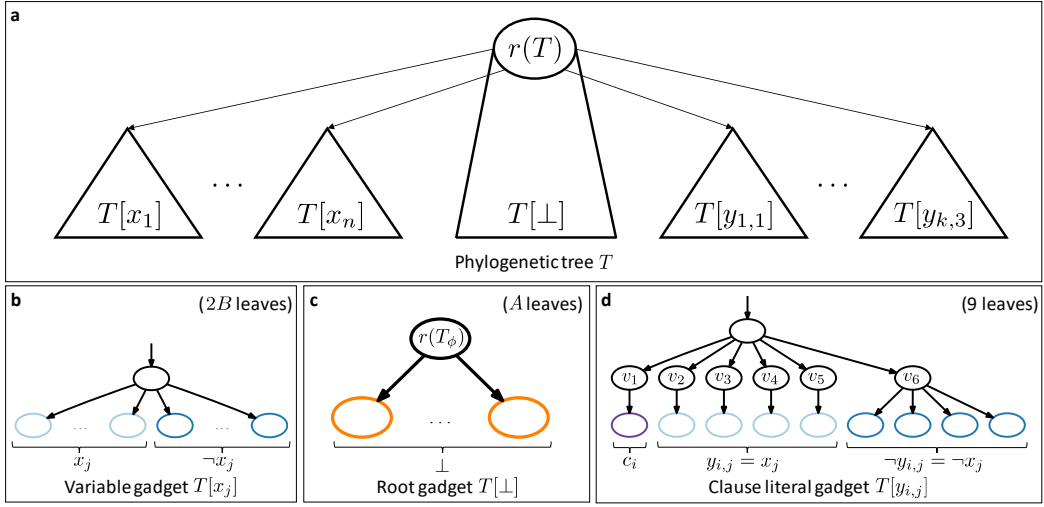
Proof. Suppose for a contradiction that there exists a vertex labeling ℓ of T such that $\ell(u) \not\preceq_G \text{LCA}_G(\hat{\ell}(L(T_u)))$ for a vertex u of T . For brevity, we define $s = \text{LCA}_G(\hat{\ell}(L(T_u)))$. By our premise, there exists a leaf $v \in L(T_u)$ such that $\hat{\ell}(v) \notin V(G_{\ell(u)})$. By definition of LCA_G , there exists a unique path from s to $\hat{\ell}(v)$ in G . Moreover, since v is reachable from u in T , there exists a path from $\ell(u)$ to $\hat{\ell}(v)$ in G . We distinguish two cases. First, $\hat{\ell}(v) \preceq_G \pi(\ell(u))$. This case contradicts that G is a tree, as there would be a directed cycle between $\ell(u)$ and $\hat{\ell}(v)$ in G . Second, $\hat{\ell}(v)$ and $\ell(u)$ are incomparable in G . Thus, the path from s to $\hat{\ell}(v)$ does not contain $\ell(u)$. This, however, contradicts that there exists a path from $\ell(u)$ to $\hat{\ell}(v)$. ◀

4.1 NP-hardness

We show NP-hardness of the PMH problem in the case where $\mathcal{P} = \{S\}$ by reduction from 3-SATISFIABILITY (3-SAT). In 3-SAT, we are given a Boolean formula $\phi = \bigwedge_{i=1}^k (y_{i,1} \vee y_{i,2} \vee y_{i,3})$ in 3-conjunctive normal form (3-CNF) with n variables and k clauses, and wish to decide whether there exists a truth assignment $\theta : [n] \rightarrow \{0, 1\}$ that satisfies all the clauses of ϕ . We define $\psi(y_{i,j}) = 1$ if literal $y_{i,j}$ is of the form x , and define $\psi(y_{i,j}) = 0$ if literal $y_{i,j}$ is of the form $\neg x$. Truth assignment θ satisfies clause $(y_{i,1} \vee y_{i,2} \vee y_{i,3})$ provided there exists a $j \in \{1, 2, 3\}$ such that $\theta(x) = \psi(y_{i,j})$ where x is the variable corresponding to literal $y_{i,j}$. Without loss of generality, we may assume that each clause of ϕ consists of three distinct variables. We denote the n variables of ϕ by x_1, \dots, x_n and the k clauses of ϕ by c_1, \dots, c_k . 3-SAT is among the 21 problems proven to be NP-hard by Karp [6].

We construct a tree T with location set $\Sigma = \{\perp, x_1, \dots, x_n, \neg x_1, \dots, \neg x_n, c_1, \dots, c_k\}$ such that the migration graph G^* of any optimal vertex labeling ℓ^* models a truth assignment θ . More specifically, we want to ensure that: (i) ℓ^* labels the root of T by \perp ; (ii) for each variable x of ϕ , either $\{(\perp, x), (x, \neg x)\} \subseteq E(G^*)$ if $\theta(x) = 1$, or $\{(\perp, \neg x), (\neg x, x)\} \subseteq E(G^*)$ if $\theta(x) = 0$; and (iii) θ is satisfiable if and only if ℓ^* encodes a truth assignment that satisfies all clauses of θ . We accomplish these three requirements using three types of gadgets that form subtrees of T (Fig. 3a): (i) n variable gadgets $T[x_1], \dots, T[x_n]$, each with $2B$ leaves (Fig. 3b); (ii) a single root gadget $T[\perp]$ with A leaves (Fig. 3c); and (iii) $3k$ clause literal gadgets $T[y_{1,1}], \dots, T[y_{k,3}]$, each with 9 leaves (Fig. 3d). Fig. 4 shows an example reduction.

Let ℓ^* be an optimal vertex labeling of T under the restriction $\mathcal{P} = \{S\}$, and let G^* be the resulting migration graph. By setting $B > 10k + 1$ and $A > 2Bn + 27k$, we accomplish the first two requirements, as shown by the following two lemmas.



■ **Figure 3** Given a Boolean formula $\phi = \bigwedge_{i=1}^k (y_{i,1} \vee y_{i,2} \vee y_{i,3})$ in 3-conjunctive normal form, we construct the phylogenetic tree T composed of three types of gadgets. (a) Tree T has n variable gadgets (panel b), a single root gadget (panel c) and $3k$ clause literal gadgets (panel d). The location set Σ contains a location corresponding to a positive and negative literal of each variable, a location corresponding to each clause and a special location \perp . That is, $\Sigma = \{x_1, \dots, x_n, \neg x_1, \dots, \neg x_n, c_1, \dots, c_k, \perp\}$. (b) For each variable x_j , the corresponding variable gadget $T[x_j]$ is composed of $2B$ leaves where $B = 10k + 2$. This gadget enforces that either $(x_j, \neg x_j) \in E(G^*)$ or $(\neg x_j, x_j) \in E(G^*)$. (c) The root gadget $T[\perp]$ is composed of $A = 2Bn + 27k + 1$ leaves and enforces that $\ell^*(r(T)) = \perp$. (d) For each literal $y_{i,j}$, the corresponding clause literal gadget $T[y_{i,j}]$ has 9 leaves, and links variables to clauses. We refer to Fig. 4 for an example.

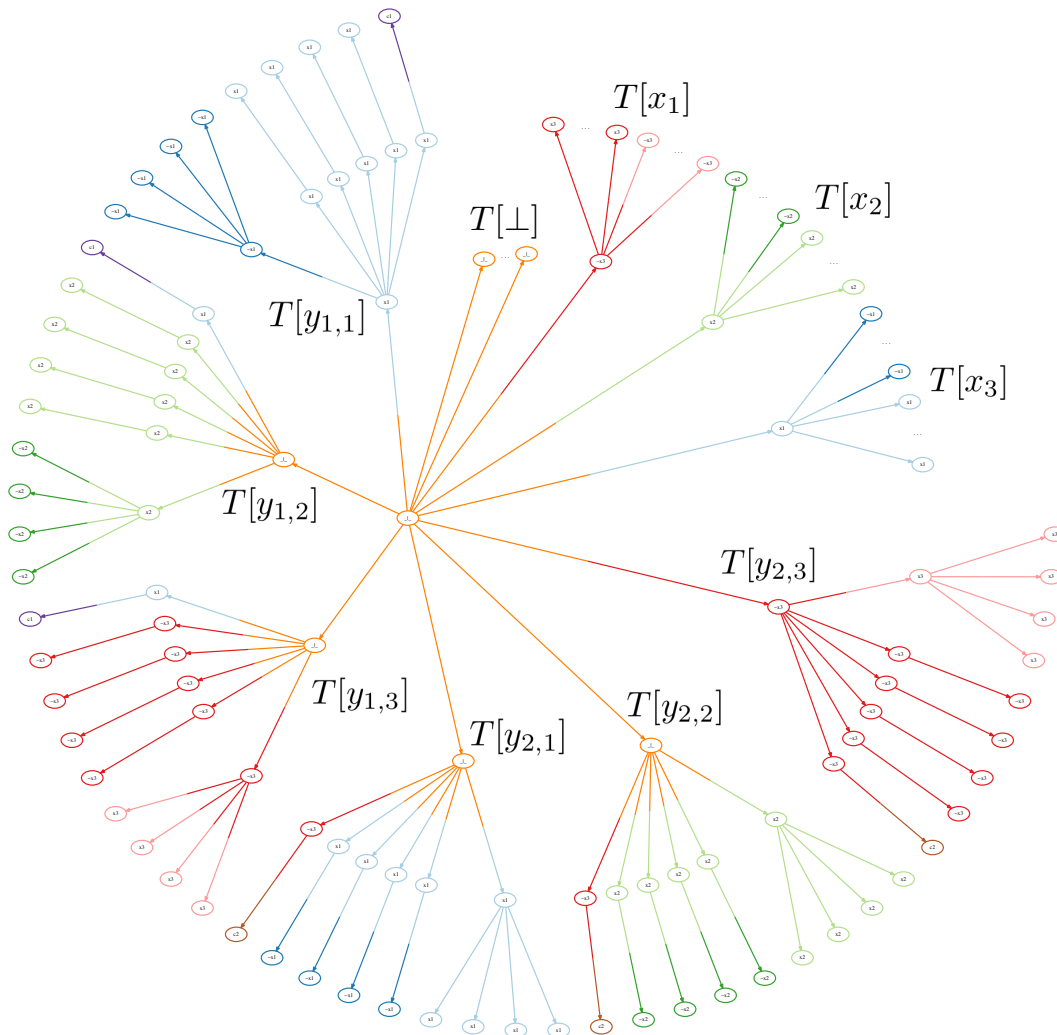
► **Lemma 6.** *The root vertex $r(T)$ has location $\ell^*(r(T)) = \perp$.*

Proof. Suppose for a contradiction that $\ell^*(r(T)) \neq \perp$. Thus, we have that each of the incoming edges to the leaves of the subtree $T[\perp]$ is a migration. Hence, $\mu(T, \ell^*) \geq A > 2Bn + 27k$. Consider the vertex labeling ℓ of T where $\ell(v) = \perp$ for each vertex v . Observe that the resulting migration graph of ℓ has an S pattern. Only the leaves of subtree $T[\perp]$ are labeled by \perp ; the remaining $2Bn + 27k$ have leaf labels that differ from \perp . We thus have $\mu(T, \ell) = 2Bn + 27k$. Therefore, $\mu(T, \ell) < \mu(T, \ell^*)$, which contradicts the premise that ℓ^* is an optimal vertex labeling. Hence, $\ell^*(r(T)) = \perp$. ◀

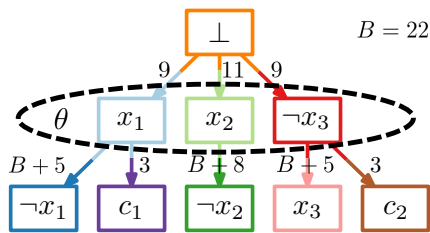
► **Lemma 7.** *For all variables x , $\{(\perp, x), (x, \neg x)\} \subseteq E(G^*)$ or $\{(\perp, \neg x), (\neg x, x)\} \subseteq E(G^*)$.*

Proof. Suppose for a contradiction that there exists a variable x of ϕ such that neither $\{(\perp, x), (x, \neg x)\} \subseteq E(G^*)$ nor $\{(\perp, \neg x), (\neg x, x)\} \subseteq E(G^*)$. Let $T[x]$ be the subtree of T corresponding to the variable gadget of x . Let r_x be the root vertex of $T[x]$. By Lemma 6 and the premise, we have that $\ell^*(r_x) \notin \{x, \neg x\}$. Therefore, the $2B$ edges incoming to the leaves of $T[x]$ are migrations in ℓ^* .

We construct a vertex labeling ℓ with fewer migrations than ℓ^* . Intuitively, vertex labeling ℓ corresponds to a truth assignment θ where $\theta(x) = 1$ for all variables x . Initially, we set $\ell = \ell^*$. Next, we set $\ell(r_x) = x$. By definition of 3-SAT, each clause c_i of ϕ contains at most one literal $y_{i,j}$ of the form x or $\neg x$. Let c_i be such a clause with literal $y_{i,j}$ of the form x or $\neg x$. Let $T[y_{i,j}]$ be the subtree of T corresponding to the clause literal gadget of $y_{i,j}$, and let $r_{i,j}$ be the root vertex of $T[y_{i,j}]$. We set $\ell(r_{i,j}) = \perp$, $\ell(v_1) = \perp$, $\ell(v_2) = \ell(v_3) = \ell(v_4) = \ell(v_5) = x$ and $\ell(v_6) = x$ (where v_1, \dots, v_6 are defined in Fig. 3d). We distinguish two cases. First,

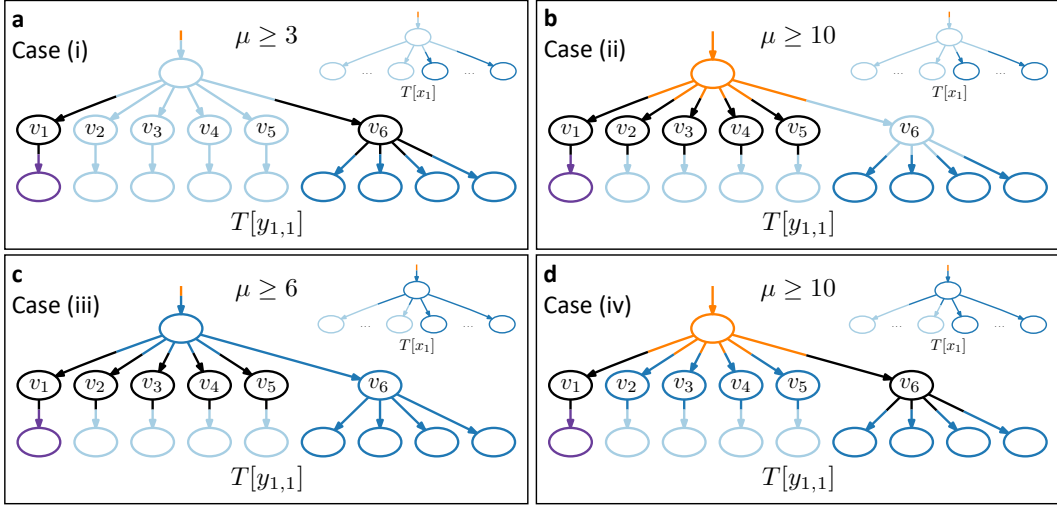


(a) Phylogenetic tree T and optimal vertex labeling ℓ^* .



(b) Migration graph G^* .

■ **Figure 4 Example reduction.** Consider the Boolean formula $\phi = (y_{1,1} \vee y_{1,2} \vee y_{1,3}) \wedge (y_{2,1} \vee y_{2,2} \vee y_{2,3}) = (x_1 \vee x_2 \vee \neg x_3) \wedge (\neg x_1 \vee \neg x_2 \vee \neg x_3)$ with $k = 2$ clauses and $n = 3$ variables. Truth assignment $\theta(x_1) = 1, \theta(x_2) = 1, \theta(x_3) = 0$ satisfies ϕ . (a) The corresponding tree T has location set $\Sigma = \{x_1, x_2, x_3, \neg x_1, \neg x_2, \neg x_3, c_1, c_2, \perp\}$. The variable gadgets $T[x_1], T[x_2], T[x_3]$ each have $2B = 2(10k + 2) = 44$ leaves. The root gadget $T[\perp]$ has $A = 2Bn + 27k + 1 = 187$ leaves. The variable literal gadgets $T[y_{1,1}], \dots, T[y_{2,3}]$ each have 9 leaves. The shown vertex labeling ℓ^* corresponds to truth assignment θ and has migration number $\mu(T, \ell^*) = (B + 1)n + 25k = (22 + 1) \cdot 3 + 25 \cdot 2 = 119$. Thus, by Lemma 10, the truth assignment θ encoded by ℓ^* satisfies ϕ . (b) The corresponding migration graph G^* , where each edge is labeled by its multiplicity incurred by ℓ^* .



■ **Figure 5** A clause literal gadget $T[y_{i,j}]$ imposes four different sets of constraints on optimal vertex labelings ℓ^* , as shown in Lemma 8. The figure reproduces clause literal gadget $T[y_{1,1}]$, where $y_{1,1} = x_1$, depicted in Fig. 4a. (a) If $\ell^*(r(T[x_1])) = x_1$ and $\ell^*(r(T[y_{1,1}])) = x_1$, vertices v_2, \dots, v_5 must be labeled by x_1 and the migration number is at least 3. (b) If $\ell^*(r(T[x_1])) = x_1$ and $\ell^*(r(T[y_{1,1}])) = \perp$, vertex v_6 must be labeled by x_1 and the migration number is at least 10. (c) If $\ell^*(r(T[x_1])) = \neg x_1$ and $\ell^*(r(T[y_{1,1}])) = \neg x_1$, vertex v_6 must be labeled by $\neg x_1$ and the migration number is at least 6. (d) If $\ell^*(r(T[x_1])) = \neg x_1$ and $\ell^*(r(T[y_{1,1}])) = \perp$, vertices v_2, \dots, v_5 must be labeled by $\neg x_1$ and the migration number is at least 10.

$y_{i,j} = x$. In this case, the outgoing edge of v_1 is a migration. The incoming edges of v_2, v_3, v_4, v_5, v_6 are migrations. The four outgoing edges of v_6 are migrations. Thus, the number of migrations incurred by ℓ in $T[y_{i,j}]$ is 10. Second, $y_{i,j} = \neg x$. In this case, the outgoing edge of v_1 is a migration. The incoming edges of v_2, v_3, v_4, v_5, v_6 are migrations. The outgoing edges of v_2, v_3, v_4, v_5 are migrations. Thus, the number of migrations incurred by ℓ in $T[y_{i,j}]$ is 10.

In both cases, we have that $\mu(T[x], \ell^*) > 2B$. On the other hand, $\mu(T[x], \ell) = B + 1$, and $\mu(T[y_{i,j}], \ell) = 10$ for each literal $y_{i,j}$ of the form $\{x, \neg x\}$. As there are at most k literals of the form $\{x, \neg x\}$, we have that there are at most $B + 1 + 10k$ migrations incurred by ℓ in subtree $T[x]$ and the corresponding clause literal subtrees $T[y_{i,j}]$. Since $B > 1 + 10k$, we have that $2B > B + 1 + 10k$. Thus,

$$\mu(T[x], \ell^*) > 2B > B + 1 + 10k > \mu(T[x], \ell) + \sum_{y_{i,j} \text{ of form } \{x, \neg x\}} \mu(T[y_{i,j}], \ell).$$

Compared to ℓ^* , the new vertex labeling ℓ differs in subtree $T[x]$ and in at most k subtrees $T[y_{i,j}]$. Hence, $\mu(T, \ell^*) > \mu(T, \ell)$, yielding a contradiction. The lemma now follows. ◀

Thus, our reduction enables us to encode a truth assignment. We now focus on the third requirement that links literals to clauses. We start by showing that the clause literal gadget, when combined with the variable and root gadgets, imposes very specific constraints on ℓ^* .

► **Lemma 8.** *Let x be a variable of ϕ and let $y_{i,j}$ be a literal corresponding to x . Then, one of the following four cases must hold, where $C = \{c_i \mid i \in [k]\}$ and $X = \{\ell^*(r(T[x_i])) \mid i \in [n]\}$.*

case	$\ell^*(r(T[x]))$	$\ell^*(r(T[y_{i,j}]))$	$\ell^*(v_1)$	$\ell^*(v_2), \dots, \ell^*(v_5)$	$\ell^*(v_6)$
(i)	$y_{i,j}$	$y_{i,j}$	$\{y_{i,j}, \neg y_{i,j}\} \cup C$	$y_{i,j}$	$\{y_{i,j}, \neg y_{i,j}\}$
(ii)	$y_{i,j}$	\perp	$\{\perp\} \cup C \cup X$	$\{\perp, y_{i,j}\}$	$y_{i,j}$
(iii)	$\neg y_{i,j}$	$\neg y_{i,j}$	$\{y_{i,j}, \neg y_{i,j}\} \cup C$	$\{y_{i,j}, \neg y_{i,j}\}$	$\neg y_{i,j}$
(iv)	$\neg y_{i,j}$	\perp	$\{\perp\} \cup C \cup X$	$\neg y_{i,j}$	$\{\perp, \neg y_{i,j}\}$

Proof. The lemma follows by case analysis, with four cases that each correspond to a unique case in the above table. First, it holds that $\ell^*(r(T[x])) \in \{y_{i,j}, \neg y_{i,j}\}$ by Lemma 7. Thus, we distinguish two cases.

1. $\ell^*(r(T[x])) = y_{i,j}$: By construction, $T[y_{i,j}]$ has leaves labeled by $y_{i,j}$ and $\neg y_{i,j}$. Hence, by Proposition 5, Lemma 6 and the fact that $\ell^*(r(T[x])) = y_{i,j}$, we have that $\ell^*(r(T[y_{i,j}])) \in \{y_{i,j}, \perp\}$. Thus, we distinguish two additional subcases.
 - a. $\ell^*(r(T[y_{i,j}])) = y_{i,j}$: Vertices v_2, v_3, v_4, v_5 must be labeled by $y_{i,j}$ as their parent $r(T[y_{i,j}])$ is labeled by $y_{i,j}$ and they themselves are the parents of leaves labeled by $y_{i,j}$. The four children of v_6 are leaves labeled by $\neg y_{i,j}$, and the parent $r(T[y_{i,j}])$ of v_6 is labeled $y_{i,j}$. Thus, v_6 must be labeled by either $y_{i,j}$ or $\neg y_{i,j}$. Vertex v_1 is the child of $r(T[y_{i,j}])$ labeled by $y_{i,j}$, and thus the only literals that can label v_1 are $y_{i,j}$ and $\neg y_{i,j}$. In addition, v_1 can be labeled by clauses c_1, \dots, c_k . This subcase corresponds to case (i) of the above table, and is depicted in Fig. 5a.
 - b. $\ell^*(r(T[y_{i,j}])) = \perp$: Vertex v_6 must be labeled by $y_{i,j}$ as its parent $r(T[y_{i,j}])$ is labeled by \perp and its four children are leaves labeled by $\neg y_{i,j}$. Vertices v_2, v_3, v_4, v_5 must each be labeled by either $y_{i,j}$ or \perp as their parent $r(T[y_{i,j}])$ is labeled by \perp and their children are leaves labeled by $y_{i,j}$. As vertex v_1 is the child of $r(T[y_{i,j}])$ and its only child a leaf labeled by c_i , we have that v_1 can be labeled by \perp , all clauses $C = \{c_1, \dots, c_k\}$ and all active literals X . This subcase corresponds to case (ii) of the above table, and is depicted in Fig. 5b.
2. $\ell^*(r(T[x])) = \neg y_{i,j}$: By construction, $T[y_{i,j}]$ has leaves labeled by $y_{i,j}$ and $\neg y_{i,j}$. Hence, by Proposition 5, Lemma 6 and the fact that $\ell^*(r(T[x])) = \neg y_{i,j}$, we have that $\ell^*(r(T[y_{i,j}])) \in \{\neg y_{i,j}, \perp\}$. Thus, we distinguish two additional subcases.
 - a. $\ell^*(r(T[y_{i,j}])) = \neg y_{i,j}$: Vertex v_6 must be labeled by $\neg y_{i,j}$ as its parent $r(T[y_{i,j}])$ is labeled by $\neg y_{i,j}$ and its four children are leaves labeled by $\neg y_{i,j}$. Vertices v_2, v_3, v_4, v_5 must each be labeled by either $\neg y_{i,j}$ or $y_{i,j}$ as their parent $r(T[y_{i,j}])$ is labeled by $\neg y_{i,j}$ and their children are leaves labeled by $\neg y_{i,j}$. Vertex v_1 is the child of $r(T[y_{i,j}])$ labeled by $y_{i,j}$, and thus the only literals that can label v_1 are $y_{i,j}$ and $\neg y_{i,j}$. In addition, v_1 can be labeled by clauses c_1, \dots, c_k . This subcase corresponds to case (iii) of the above table, and is depicted in Fig. 5c.
 - b. $\ell^*(r(T[y_{i,j}])) = \perp$: Vertices v_2, v_3, v_4, v_5 must each be labeled by $\neg y_{i,j}$ as their parent $r(T[y_{i,j}])$ is labeled by \perp and their children are leaves labeled by $y_{i,j}$. Vertex v_6 must be labeled by either \perp or $\neg y_{i,j}$, as its parent $r(T[y_{i,j}])$ is labeled by \perp and its four children are leaves labeled by $\neg y_{i,j}$. As vertex v_1 is the child of $r(T[y_{i,j}])$ and its only child a leaf labeled by c_i , we have that v_1 can be labeled by \perp , all clauses $C = \{c_1, \dots, c_k\}$ and all active literals X . This subcase corresponds to case (iv) of the above table, and is depicted in Fig. 5d. ◀

Next, we prove a lower bound on the minimum migration number of T given $\mathcal{P} = \{S\}$.

► **Lemma 9.** *It holds that $\mu(T, \ell^*) \geq (B + 1)n + 25k$.*

Proof. Consider clause i composed of literals $y_{i,1}$, $y_{i,2}$ and $y_{i,3}$. By Lemma 8, the vertex labeling ℓ^* of each of the subtrees $T[y_{i,1}]$, $T[y_{i,2}]$ and $T[y_{i,3}]$ must adhere to one of four cases. It is easy to verify that case (i) has migration number $\mu(T[y_{i,j}]) \geq 3$ (Fig. 5a), case (ii) has

migration number $\mu(T[y_{i,j}]) \geq 10$ (Fig. 5b), case (iii) has migration number $\mu(T[y_{i,j}]) \geq 6$ (Fig. 5c) and case (iv) has migration number $\mu(T[y_{i,j}]) \geq 10$ (Fig. 5d). At most one of $T[y_{i,1}]$, $T[y_{i,2}]$ and $T[y_{i,3}]$ can be labeled by ℓ^* according to cases (i) or (iii), due to the restriction that G^* must be a tree (i.e. $\mathcal{P} = \{S\}$). Without loss of generality, we assume that only $T[y_{i,1}]$ adheres to case (i) or (iii). We distinguish three cases.

1. $T[y_{i,1}]$ is of case (i), and $T[y_{i,2}]$ and $T[y_{i,3}]$ are of cases (ii) or (iv): In $T[y_{i,1}]$, the parent of vertex v_1 is labeled by $y_{i,j}$. Moreover, in both $T[y_{i,2}]$ and $T[y_{i,3}]$, the parent of vertex v_1 is labeled by \perp . As such, these two vertices in $T[y_{i,2}]$ and $T[y_{i,3}]$ must be assigned label $y_{i,j}$. Hence, the minimum number of migrations induced by ℓ^* on the subtree of T induced by vertices $r(T)$, $V(T[y_{i,1}])$, $V(T[y_{i,2}])$ and $V(T[y_{i,3}])$ is $3 + 11 + 11 = 25$.
2. $T[y_{i,1}]$ is of case (iii), and $T[y_{i,2}]$ and $T[y_{i,3}]$ are of cases (ii) or (iv): In $T[y_{i,1}]$, the parent of vertex v_1 is labeled by $\neg y_{i,j}$. Moreover, in both $T[y_{i,2}]$ and $T[y_{i,3}]$, the parent of vertex v_1 is labeled by \perp . As such, these two vertices in $T[y_{i,2}]$ and $T[y_{i,3}]$ must be assigned label $\neg y_{i,j}$. Hence, the minimum number of migrations induced by ℓ^* on the subtree of T induced by vertices $r(T)$, $V(T[y_{i,1}])$, $V(T[y_{i,2}])$ and $V(T[y_{i,3}])$ is $6 + 11 + 11 = 28$.
3. $T[y_{i,1}]$, $T[y_{i,2}]$ and $T[y_{i,3}]$ are of cases (ii) or (iv): The minimum number of migrations induced by ℓ^* on the subtree of T induced by vertices $r(T)$, $V(T[y_{i,1}])$, $V(T[y_{i,2}])$ and $V(T[y_{i,3}])$ is $10 + 10 + 10 = 30$.

Thus, for each clause i , the minimum number of migrations induced by ℓ^* on $r(T)$, $V(T[y_{i,1}])$, $V(T[y_{i,2}])$ and $V(T[y_{i,3}])$ is 25. Thus, all k clauses lead to a least $25k$ migrations. By Lemmas 6 and Lemma 7, we have that $\ell^*(r(T)) = \perp$ and $\ell^*(r(T[x])) \in \{x, \neg x\}$ for each variable x . Thus, all variable subtrees $T[x]$ induce $(B+1)n$ migrations. Hence, $\mu(T, \ell^*) \geq (B+1)n + 25k$. \blacktriangleleft

Finally, we show that the above lower bound is tight if and only if ϕ is satisfiable.

► **Lemma 10.** *Boolean formula ϕ is satisfiable if and only if $\mu(T, \ell^*) = (B+1)n + 25k$.*

Proof. (\Rightarrow) We construct a vertex labeling ℓ with migration number $\mu(T, \ell) = (B+1)n + 25k$ such that the resulting migration graph G has an S pattern.

First, we set $\ell(r(T)) = \perp$. For each variable x , we set $\ell(r(T[x])) = x$ if $\theta(x) = 1$ and set $\ell(r(T[x])) = \neg x$ if $\theta(x) = 0$. Next, we consider each clause i . By the premise, there must exist a literal in clause i that satisfies the clause. Without loss of generality, we assume that $\theta(y_{i,1}) = 1$. We assign vertex labels to the vertices of $T[y_{i,1}]$ by setting $\ell(r(T[y_{i,1}])) = y_{i,1}$, $\ell(v_1) = \dots = \ell(v_5) = y_{i,1}$ and $\ell(v_6) = \neg y_{i,1}$, according to case (i) of Lemma 8 (Fig. 5a). For the other two literals $y_{i,j}$ (where $j \in \{2, 3\}$), we set $\ell(r(T[y_{i,j}])) = \perp$. Let x' be the variable corresponding to $y_{i,j}$. If $\theta(x') = \psi(y_{i,j})$, we set $\ell(v_2) = \dots = \ell(v_5) = y_{i,j}$ and $\ell(v_6) = y_{i,2}$, according to case (ii) of Lemma 8 (Fig. 5b). Otherwise, we set $\ell(v_2) = \dots = \ell(v_5) = \neg y_{i,j}$ and $\ell(v_6) = \neg y_{i,2}$, according to case (iv) of Lemma 8 (Fig. 5d).

It is easy to verify that $\mu(T, \ell) = (B+1)n + 25k$. By construction, each literal label $y_{i,j}$, as well as each clause label c_i , has a unique parent in G . Thus, G adheres to an S pattern. By Lemma 9, it holds that $\mu(T, \ell^*) \geq (B+1)n + 25k$. Hence, ℓ is an optimal vertex labeling and $\mu(T, \ell^*) = (B+1)n + 25k$.

(\Leftarrow) We construct a truth assignment θ from ℓ^* . By Lemma 7, we have that $\ell^*(r(T[x])) \in \{x, \neg x\}$ for each variable x . We set $\theta(x) = \psi(\ell^*(r(T[x])))$. We claim that θ satisfies at least one literal for each clause i of ϕ .

By Lemmas 6, 7 and 9, we have that ℓ^* induces, for each clause i , 25 migrations in the subtree induced by vertices $V(T[y_{i,1}]) \cup V(T[y_{i,2}]) \cup V(T[y_{i,3}]) \cup \{r(T)\}$. Thus, one subtree among $T[y_{i,1}]$, $T[y_{i,2}]$ and $T[y_{i,3}]$ must have 3 migrations, corresponding to case (i) of Lemma 8 (Fig. 5a). Let $T[y_{i,j}]$ be the subtree with 3 migrations and let x be the variable

corresponding to $y_{i,j}$. As $\ell^*(r(T[y_{i,j}])) = y_{i,j}$, we have that $\ell^*(r(T[x])) = y_{i,j}$ and thus $\theta(x) = \psi(\ell^*(r(T[x]))) = \psi(y_{i,j})$. Thus, θ satisfies literal $y_{i,j}$ and thereby clause i . The same argument applies to each clause of ϕ . Hence, θ is a truth assignment that satisfies ϕ . ◀

Phylogenetic tree T has $1 + A + (2B + 1)n + 16kn$ vertices can be constructed in polynomial time from ϕ . For instance, setting $B = 10k + 2$ and $A = 2Bn + 27k + 1$ (respecting the requirement that $B > 10k + 1$ and $A > 2Bn + 27k$), yields a total of $56kn + 27k + 9n + 2$ vertices in T , which is polynomial in the number k of clauses and the number n of variables. Thus, by Lemma 10, we have a polynomial time reduction from 3-SAT to PMH, proving Theorem 3.

4.2 Fixed parameter tractability

A PMH instance $(T, \hat{\ell})$ has $m = |\Sigma|$ locations that label $n = |V(T)|$ vertices. Typically, $m \ll n$ in practical problem instances. Here, we show that PMH is fixed parameter tractable in m and give an algorithm that runs in time $O(nm^m)$.

Our algorithm relies on the notion of a *migration tree* \hat{G} , a rooted tree whose vertex set is Σ . Unlike a migration graph, which is a directed multi-graph, \hat{G} is a simple directed graph with at most one edge between every pair of vertices. We say that a vertex labeling ℓ is *consistent* with migration tree \hat{G} provided that for any distinct pair (s, t) of locations there exists an edge $(u, v) \in E(T)$ such that $\ell(u) = s$ and $\ell(v) = t$ if and only if $(s, t) \in E(\hat{G})$.

For a given PMH instance $(T, \hat{\ell})$ and migration tree \hat{G} there may not exist a vertex labeling consistent with \hat{G} (Fig. 6). Let $\text{LCA}_{\hat{G}}(u) = \text{LCA}_{\hat{G}}(\hat{\ell}(L(T_u)))$ for vertex $u \in V(T)$. For $u \preceq_T v$, let $d_T(u, v)$ be the number of edges on the unique path from u to v . We show that the condition

$$d_T(u, v) \geq d_{\hat{G}}(\text{LCA}_{\hat{G}}(u), \hat{\ell}(v)) \quad \forall u, v \in V(T) \text{ such that } u \preceq_T v, \quad (2)$$

is both necessary and sufficient for the existence of a vertex labeling of T consistent with \hat{G} .

► **Lemma 11.** *If there exists a vertex labeling ℓ for T consistent with \hat{G} then (2) holds.*

Proof. Let ℓ be a vertex labeling for T consistent with \hat{G} . Let u and v be vertices of T such that $u \preceq_T v$. By Proposition 5, we have that $\ell(u) \preceq_{\hat{G}} \text{LCA}_{\hat{G}}(u)$. Thus, the number of migrations on the path from u to v must be at least $d_{\hat{G}}(\text{LCA}_{\hat{G}}(u), \hat{\ell}(v))$. To accommodate these migrations, the path from u to v must have at least $d_{\hat{G}}(\text{LCA}_{\hat{G}}(u), \hat{\ell}(v))$ edges. Hence, $d_T(u, v) \geq d_{\hat{G}}(\text{LCA}_{\hat{G}}(u), \hat{\ell}(v))$. ◀

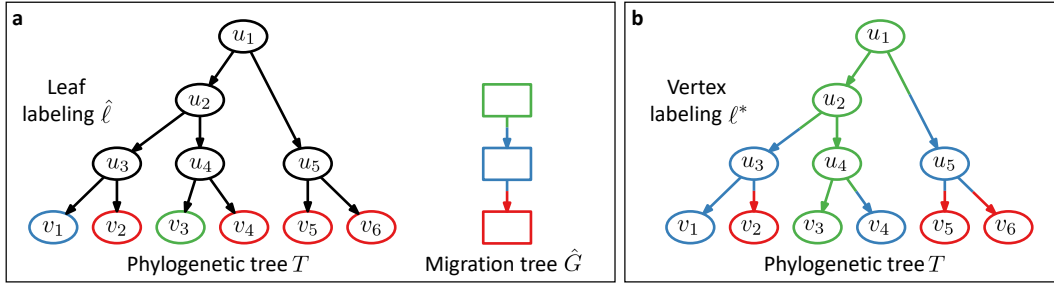
We prove sufficiency constructively, using the vertex labeling ℓ^* of T defined as

$$\ell^*(v) = \begin{cases} \text{LCA}_{\hat{G}}(r(T)), & \text{if } v = r(T), \\ \sigma(\ell^*(\pi(v)), \text{LCA}_{\hat{G}}(v)), & \text{if } v \neq r(T), \end{cases} \quad (3)$$

where $\sigma(s, t) = s$ if $s = t$ and otherwise $\sigma(s, t)$ is the unique child of s that lies on the path from s to t in \hat{G} . In the case where $\sigma(\ell^*(\pi(v)), \text{LCA}_{\hat{G}}(v)) = \text{LCA}_{\hat{G}}(v)$ for all vertices v , vertex labeling ℓ^* is identical to the LCA mapping [5] used in the context of gene-tree species-tree reconciliation to minimize the duplication number.

► **Lemma 12.** *If (2) holds then vertex labeling ℓ^* for T is consistent with \hat{G} .*

Proof. For ℓ^* to be consistent with \hat{G} it must hold that: (i) for any distinct pair (s, t) of locations there exists an edge $(u, v) \in E(T)$ such that $\ell(u) = s$ and $\ell(v) = t$ if and only if $(s, t) \in E(\hat{G})$, and (ii) $\ell^*(v) = \hat{\ell}(v)$ for each leaf v . Condition (i) holds by definition of ℓ^* .



■ **Figure 6** Given a phylogenetic tree T and migration tree \hat{G} , there may not exist a vertex labeling of T consistent with \hat{G} . (a) Here, any vertex labeling ℓ consistent with \hat{G} must label u_1, u_2, u_4 by green. This, however, introduces a migration from green to red, violating \hat{G} . By Lemma 11, there does not a vertex labeling of T consistent with \hat{G} . (b) Equivalently, labeling ℓ^* defined in (3) is not a vertex labeling of T , as it changes the labeling of leaf v_4 , i.e. $\ell^*(v_4) \neq \hat{\ell}(v_4)$.

That is, each vertex u has either the same label as its parent $\pi(u)$ or is labeled by a child of $\ell^*(\pi(u))$ in \hat{G} . As for condition (ii), assume for a contradiction that there exists a leaf v such that $\ell^*(v) \neq \hat{\ell}(v)$. Consider the unique path from $r(T)$ to v . Let u_1 be the vertex closest to v such that either $u_1 = r(T)$ or $\ell^*(u_1) = \ell^*(\pi(u_1))$. Thus, each edge on the path u_1, \dots, u_t, v is a migration. By definition of ℓ^* , we have that $\ell^*(u_1), \dots, \ell^*(u_t), \ell^*(v)$ forms a path of \hat{G} .

We distinguish two cases. First, $u_1 = r(T)$. We have that $\ell^*(r(T)) = \text{LCA}_{\hat{G}}(r(T))$. As $\ell^*(v) \neq \hat{\ell}(v)$, we have that $d(r(T), v) < d_{\hat{G}}(\text{LCA}_{\hat{G}}(r(T)), \hat{\ell}(v))$. This contradicts the premise. Second, $u_1 \neq r(T)$. Let $u_0 = \pi(u_1)$. As $\ell^*(u_0) = \ell^*(u_1)$, we have that $\ell^*(u_1) = \text{LCA}_{\hat{G}}(u_1)$. Since $\ell^*(v) \neq \hat{\ell}(v)$, we have that $d(u_1, v) < d_{\hat{G}}(\text{LCA}_{\hat{G}}(u_1), \hat{\ell}(v))$. This contradicts the premise. Hence, ℓ^* is a vertex labeling of T that is consistent with \hat{G} . ◀

Vertex labeling ℓ^* has minimum migration number $\mu(T, \ell^*)$. To show this, we prove the following lemma that states that ℓ^* assigns each vertex a location nearest to the leaf set $L(\hat{G})$.

► **Lemma 13.** *Let ℓ be a vertex labeling ℓ of T that is consistent with \hat{G} . Then, $\ell(u) \preceq_{\hat{G}} \ell^*(u)$ for each vertex $u \in V(T)$.*

Proof. Assume for a contradiction that vertex u is the nearest vertex to $r(T)$ such that $\ell(u) \not\preceq_{\hat{G}} \ell^*(u)$ (i.e. $d_T(r(T), u)$ is minimum). By Proposition 5, we have that $\ell^*(u) \preceq_{\hat{G}} \text{LCA}_{\hat{G}}(u)$ and $\ell(u) \preceq_{\hat{G}} \text{LCA}_{\hat{G}}(u)$. Thus, $\ell(u) \not\preceq_{\hat{G}} \ell^*(u)$ implies that $\ell^*(u) \prec_{\hat{G}} \ell(u)$. Moreover, we have that $u \neq r(T)$, as $\ell^*(r(T)) = \ell(r(T)) = \text{LCA}_{\hat{G}}(r(T))$ by the same proposition. Since u is the nearest vertex to $r(T)$ such that $\ell(u) \not\preceq_{\hat{G}} \ell^*(u)$, we have that $\ell^*(\pi(u)) = \ell(\pi(u))$.

We distinguish two cases. First, $\ell(u) = \ell(\pi(u))$. Thus, $\ell(u) = \ell^*(\pi(u))$. Therefore, $\ell^*(u) \prec_{\hat{G}} \ell(u)$ implies that $\ell^*(u) \prec_{\hat{G}} \ell^*(\pi(u))$. This contradicts the definition of ℓ^* . Second, $\ell(u) \neq \ell(\pi(u))$. As $\ell^*(\pi(u)) = \ell(\pi(u))$, we have that $(\pi(u), u)$ is a migration from $\ell^*(\pi(u)) = \ell(\pi(u))$ to $\ell(u)$ in ℓ and to $\ell^*(u)$ in ℓ^* . By Lemma 12, we have that ℓ^* is consistent with \hat{G} . As $\ell^*(u) \prec_{\hat{G}} \ell(u)$, the labeling by ℓ of edge $(\pi(u), u)$ results in an edge $(\ell(\pi(u)), \ell(u))$ that is not in \hat{G} . Hence, ℓ is not consistent with \hat{G} , yielding a contradiction. Both cases yield a contradiction. Hence, the lemma follows. ◀

Finally, we prove that ℓ^* achieves the minimum migration number among all vertex labelings that are consistent with a given migration tree \hat{G} .

► **Lemma 14.** *Let T be a phylogenetic tree with leaf labeling $\hat{\ell}$ satisfying (2) for a given migration graph \hat{G} . Then, ℓ^* is a minimum migration labeling of T that is consistent with \hat{G} .*

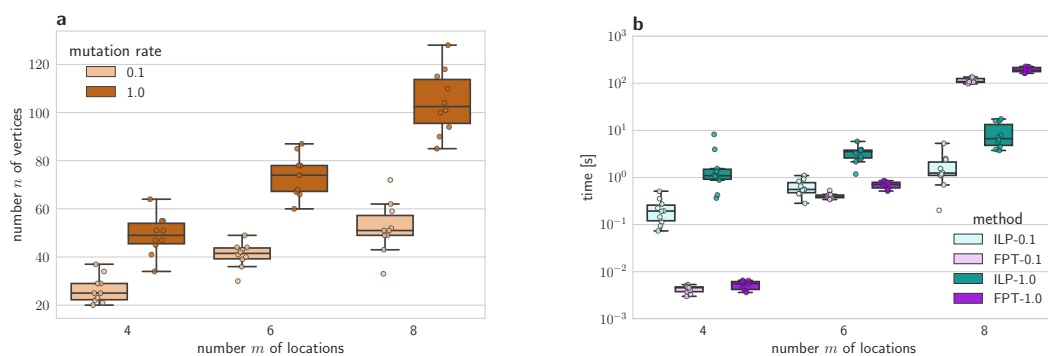


Figure 7 The fixed parameter tractable (FPT) algorithm is faster than the previously published [2] integer liner program (ILP) for small number m of locations. Using a tool for simulating metastatic cancers, we generated 60 PMH instances with varying number m of locations and number n of vertices. (a) The number of vertices increased with increasing mutation rate. (b) Running time in seconds (logarithmic scale) for FPT algorithm and ILP for instances with varying number m of locations and mutation rate.

Proof. Assume for a contradiction that ℓ is a vertex labeling of T consistent with \hat{G} such that $\mu(T, \ell) < \mu(T, \ell^*)$. Consider a location t such that the number of migrations with target t is greater in ℓ^* than ℓ . Observe that $t \neq r(\hat{G})$. Let s be the parent of t in \hat{G} . Let $Y = \{v \in V(T) \mid \ell(\pi(v)) = s, \ell(v) = t\}$ and $Y^* = \{v \in V(T) \mid \ell^*(\pi(v)) = s, \ell^*(v) = t\}$. By the premise we have that $|Y^*| > |Y|$.

For any vertex labeling ℓ' consistent with \hat{G} , distinct edges $(u, v), (u', v') \in E(T)$ labeled by $\ell'(u) = \ell'(u') = s$ and $\ell'(v) = \ell'(v') = t$ are incomparable. Thus, the vertices in Y (Y^*) are pairwise incomparable in T . Let $L(t)$ be the set of leaves u of T labeled by $\ell(u) = t$ where $t \preceq_{\hat{G}} t'$. Let $L(T_Y)$ ($L(T_{Y^*})$) be the combined set of leaves in the subtrees rooted at each vertex $v \in Y$ ($v \in Y^*$). As both ℓ and ℓ^* are consistent with \hat{G} , we have that $L(t) = L(T_Y) = L(T_{Y^*})$. Since $|Y^*| > |Y|$ and $L(T_Y) = L(T_{Y^*})$, there must exist two distinct vertices $v, w \in Y^*$ and vertex $u \in Y$ such that $u \prec_T v$ and $u \prec_T w$. Since $u \preceq_T \pi(v)$, $u \preceq_T \pi(w)$ and $\ell^*(\pi(v)) = \ell^*(\pi(w)) = s$, we have that $\ell^*(u) \preceq_{\hat{G}} s$. Given that $\ell(u) = t$ and $s \prec_{\hat{G}} t$, we have that $\ell^*(u) \preceq_{\hat{G}} s \prec_{\hat{G}} t \preceq_{\hat{G}} \ell(u)$, which contradicts Lemma 13. Hence, ℓ^* is a minimum migration labeling of T that is consistent with \hat{G} . \blacktriangleleft

We note that ℓ^* can be computed in $O(nm)$ time—i.e., each $\text{LCA}_{\hat{G}}(\cdot)$ query can be resolved in $O(m)$ time using a simple post-order tree traversal of \hat{G} . The number of unrooted trees on m labeled vertices is m^{m-2} , which is known as Cayley's formula [1]. Since every unrooted tree can be rooted at each of its m vertices, the number of migration trees is m^{m-1} . Thus, a brute-force algorithm that computes ℓ^* for each migration tree \hat{G} requires $O(nm^m)$ time, proving Theorem 4.

5 Experimental Evaluation

We implemented the FPT algorithm in C++ and used Prüfer sequences [8] to enumerate migration trees. The code is available at <https://github.com/elkebir-group/PMH-S>. Using simulated data of metastatic cancers, we compared the FPT algorithm and the previously published integer linear program (ILP) [2]. More specifically, for each combination of $m \in \{4, 6, 8\}$ locations and mutation rates $N \in \{0.1, 1.0\}$, we simulated 10 metastatic cancer where each of the $m - 1$ metastases was seeded by clones from a single location. The simulation algorithm is described in more detail in [2]. In total, we simulated 60 PMH problem instances where $\mathcal{P} = \{S\}$.

Increasing the mutation rate N resulted in larger phylogenetic trees (Fig. 7a). We ran both the FPT algorithm and the ILP in single-threaded mode on a computer with two Intel Xeon CPUs at 2.6 GHz (32 cores) and 512 GB of RAM. As a sanity check, we found that both the ILP and the FPT algorithm resulted in the same minimum migration number for each instance (data not shown). We found that the running time of the FPT algorithm is only moderately affected by the size of the phylogenetic tree in contrast to the ILP (Fig. 7b). In addition, we found that the FPT algorithm outperformed the ILP for $m \in \{4, 6\}$ locations in terms of running time, but was slower for $m = 8$ locations (Fig. 7b). These findings are in line with the asymptotic running time of $O(nm^m)$ for the FPT algorithm. For modest number m of locations the FPT algorithm is the method of choice due to its simplicity.

6 Conclusion

In this work, we studied the complexity of the PARSIMONIOUS MIGRATION HISTORY (PMH) problem, a constrained multi-objective optimization problem for reconstructing the migration history of a given phylogenetic tree T whose leaves are labeled by locations. For the case where the resulting migration graph G is restricted to a tree (i.e. $\mathcal{P} = \{S\}$), we showed that PMH is NP-hard and fixed parameter tractable in the number m of locations. We demonstrated that our FPT algorithm runs in time $O(nm^m)$ and outperforms the previously integer linear program for small m on simulated instances of practical size. While we did not consider the polytomy resolution variant of PMH, as introduced in [2], our hardness proof easily extends to this case by appropriately binarizing the gadgets used in the reduction. The hardness of PMH for the cases where G is restricted to a DAG (i.e. $\mathcal{P} = \{S, M\}$) or left unrestricted (i.e. $\mathcal{P} = \{S, M, R\}$) remains open.

References

- 1 A Cayley. A theorem on trees. *The Quart. J. of Pure and Appl. Math.*, 23:376–8, 1889.
- 2 Mohammed El-Kebir, Gryte Satas, and Benjamin J Raphael. Inferring parsimonious migration histories for metastatic cancers. *Nature Genetics*, 50(5):718–726, 2018.
- 3 Joseph Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, nov 1981.
- 4 Walter M Fitch. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Zoology*, 20(4):406, 1971.
- 5 M Goodman et al. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Biology*, 28(2):132–163, 1979.
- 6 Richard M. Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, pages 85–103. Springer, 1972.
- 7 A W McPherson et al. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature Genetics*, may 2016.
- 8 H Prüfer. Neuer Beweis eines Satzes über Permutationen. *Arch Math Phys*, 27:742–4, 1918.
- 9 David Sankoff. Minimal Mutation Trees of Sequences. *SIAM Journal on Applied Mathematics*, 28(1):35–42, jan 1975.
- 10 M Slatkin and W P Maddison. A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics*, 123(3):603–613, 1989.