BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–5

The minimum evolution problem is hard: A link between tree inference and graph clustering problems

Sarah Bastkowski [1],Vincent Moulton[2], Andreas Spillner[3], Taoyang Wu[2] *

[1]The Genome Analysis Centre, Norwich, United Kingdom. [2] School of Computing Sciences, University of East Anglia, Norwich, United Kingdom. [3]Department of Mathematics and Computer Science, University of Greifswald, Germany.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Distance methods are well suited for constructing massive phylogenetic trees. However, the computational complexity for Rzhetsky and Nei's minimum evolution approach, one of the earliest methods for constructing a phylogenetic tree from a distance matrix, remains open.

Results: We show that Rzhetsky and Nei's minimum evolution problem is NP-complete, and so probably computationally intractable. We do this by linking the minimum evolution problem to a graph clustering problem called the quasi-clique decomposition problem, which has recently also been shown to be NP-complete. We also discuss how this link could potentially open up some useful new connections between phylogenetics and graph clustering.

Contact: taoyang.wu@uea.ac.uk

Supplementary information Supplementary appendix is available at *Bioinformatics* online.

1 INTRODUCTION

One of the earliest distance-based approaches introduced to construct a phylogenetic tree is the minimum evolution (ME) method. It was first suggested by Kidd and Sgaramella-Zonta (1971) and consists of two main steps: First branch lengths are assigned to tree topologies based on a distance matrix, and then a topology is selected which minimizes the sum of the branch lengths. There are several variants of this approach which are reviewed in e.g. Catanzaro (2009); Desper and Gascuel (2005). Although model-based tree construction methods, such as likelihood and Bayesian approaches, are tending to supersede distance-based methods in the literature, ME methods still remain popular. This is in part due to the fact that large-scale sequencing applications such as metagenomics involve constructing massive trees for which distance-based methods are well suited (see e.g. Filipski *et al.*, 2015).

In this paper we are interested in the ME approach introduced by Rzhetsky and Nei (1993). This is based on ordinary least squares (OLS) estimates of branch lengths, served as a motivation for the neighbor-joining method (Saitou and Nei, 1987), and is implemented by Desper and Gascuel (2002) in the popular FastME software. It is commonly believed that, just as the optimization problems arising from the parsimony (Day, 1987) and the likelihood (Addario-Berry *et al.*, 2004) approaches, this version of the ME method also leads to an NP-complete problem and, so, is probably computationally intractable. However, even though this has been stated to be the case in some of the literature (probably because tree construction based solely on OLS for integer branch lengths is NP-complete (Day, 1987)), to our best knowledge this fact has not been formally proven. It should also be noted, however, that the closely related and more recently introduced balanced minimum evolution (BME) problem (Desper and Gascuel, 2002) – in which branch lengths are estimated by a special case of weighted least squares (WLS) (Desper and Gascuel, 2004) – has been shown to be NP-complete (Fiorini and Joret, 2012).

Here, we shall show that the ME problem is NP-complete for trees with integer branch lengths. In particular, to prove our main result, we show that the ME problem is closely related to the so-called *quasi-clique decomposition* problem, a special example of a *graph clustering* problem (see, e.g., Pattillo *et al.*, 2013) which has recently been shown to be NP-complete by Kaya *et al.* (2013). We believe that the link that we describe could open up some interesting and useful new connections between the fields of phylogenetics and graph clustering (Schaeffer, 2007), a burgeoning area with several applications including pedigree construction (Kirkpatrick *et al.*, 2011) and community structure detection (Brunato *et al.*, 2008).

The rest of the paper is organized as follows. In the next section we show that certain OLS weightings for trees relative to a distance matrix are related to clique properties in a graph that can be associated to the distance matrix. In the following section, we then show that a rooted version of the ME problem is NP-complete, and explain how a technique used in Day (1987) can be used to show that the ME problem is NP-complete (we provide the full proof for this in the appendix as it is quite technical in nature). In the last section we discuss a link between phylogenetics and graph clustering which arises from our approach to the ME problem, and some possible future directions.

2 L_2-WEIGHTINGS

In this section we shall show that OLS tree weightings for a certain distance matrix associated to a graph G can be related to a clique property of G.

*to whom correspondence should be addressed

We first recall some definitions concerning trees. For a set X of taxa, a *rooted X-tree* $\mathcal{T} = (V, E)$ is a graph-theoretical tree with (i) leaf set X, (ii) no vertices of degree two and (iii) a specific vertex ρ which is called the root of $\mathcal{T}$ and will not be regarded as part of the leaf set. Given a rooted X-tree $\mathcal{T} = (V, E)$, we let $\leq_{\mathcal{T}}$ be the partial order on V induced by $\mathcal{T}$, that is, $u \leq_{\mathcal{T}} v$, or u is *below* v, if and only if v is contained in the path from the root ρ to u. If in addition we have $u \neq v$, we write $u <_{\mathcal{T}} v$ and say that u is *strictly below* v. The *lowest common ancestor* of two vertices u and v, denoted by $\mathrm{LCA}(u, v)$, is defined as the lowest vertex in $\mathcal{T}$ such that both u and v are below it. Moreover, for each vertex u in $\mathcal{T}$, $\mathcal{C}(u) = \{x \in X : x \leq_{\mathcal{T}} u\}$ denotes the set of leaves below u. Finally, a rooted X-tree with a particularly simple structure is the *star X-tree* $\mathcal{S}_X$ whose vertex set consists of the root ρ and leaf set X.

A *weighting* of a rooted X-tree $\mathcal{T}$ is a map ω that assigns every edge of $\mathcal{T}$ a non-negative real number. Given such a weighting, $D_\omega(u, v)$ denotes the length of the shortest path in $\mathcal{T}$ between any two vertices u and v. Moreover, such a weighting is called an *integer equi-weighting* on $\mathcal{T}$ if $\omega : E \rightarrow \mathbb{Z}_{\geq 0} := \{0, 1, 2, \dots, \}$ and $D_\omega(x, \rho) = D_\omega(y, \rho)$ for all $x, y \in X$. Given a distance matrix D on a set of taxa X and a rooted X-tree $\mathcal{T}$, an *L_2-weighting* ω for $(\mathcal{T}, D)$ is an integer equi-weighting on $\mathcal{T}$ such that

$$\begin{aligned}\Delta(\mathcal{T}; D_\omega, D) := \Delta(D_\omega, D) := ||D_\omega - D||_2^2 :\\ = \sum_{\{x,y\}\subseteq X} |D_\omega(x,y) - D(x,y)|^2\end{aligned}$$

is minimum over all integer equi-weightings on $\mathcal{T}$. In this case, we shall say that $(\mathcal{T}, \omega)$ is an *L_2-representation* of D.

Now, for a graph $G = (X, E)$ with vertex set X, let D_G be the distance matrix on X such that for a pair of distinct elements x and y in X, we have $D_G(x, y) = 2$ if x, y are adjacent in G, and $D_G(x, y) = 4$ otherwise. The edge density of G, denoted by $\gamma(G)$, is defined as $|E|/\binom{X}{2}$ and G is called a *semi-clique* if $\gamma(G) \geq 1/2$. In the following we will also refer to subsets $X' \subseteq X$ as *semi-cliques in G* if the subgraph of G induced by X' is a semi-clique. We now provide a key relationship between the edge density of G and L_2-representations of D_G.

LEMMA 2.1. *Suppose that G is a graph with vertex set X, $|X| \geq 2$, and $\mathcal{S}_X$ is the star X-tree. Let ω_i $(i = 1, 2)$ be the weighting that assigns to each edge of $\mathcal{S}_X$ weight i. Then the following assertions hold:*
(i) If $\gamma(G) > 1/2$, then ω_1 is the unique L_2-weighting for $(\mathcal{S}_X, D_G)$.
(ii) If $\gamma(G) < 1/2$, then ω_2 is the unique L_2-weighting for $(\mathcal{S}_X, D_G)$.
(iii) If $\gamma(G) = 1/2$, then the L_2-weightings for $(\mathcal{S}_X, D_G)$ are ω_1 and ω_2.

PROOF. For simplicity, put $D := D_G$ and let ω_j $(j \in \mathbb{Z}_{\geq 0})$ be a weighting function that assigns weight j to each edge in $\mathcal{S}_X$. Noting that each leaf is incident to the root, we know that an L_2-weighting for $(\mathcal{S}_X, D)$ must equal ω_j for some j in $\mathbb{Z}_{\geq 0}$ because an L_2-weighting is necessarily an integer equi-weighting. Because $D(x, y) \in \{2, 4\}$ for $x \neq y$ in X, a straightforward calculation leads to

$$\min\{\Delta(D_{\omega_1}, D), \Delta(D_{\omega_2}, D)\} < \Delta(D_{\omega_j}, D)$$

for $j \in \mathbb{Z}_{\geq 0} - \{1, 2\}$. In other words, an L_2-weighting for $(\mathcal{S}_X, D)$ is either ω_1 or ω_2.

Let n and m be the number of vertices and edges in G, respectively. Then we have

$$\Delta(D_{\omega_1}, D) - \Delta(D_{\omega_2}, D) = 2[n(n-1) - 4m]. \quad (1)$$

If $\gamma(G) > 1/2$, then we have $2m/(n(n-1)) > 1/2$, and hence $4m > n(n-1)$. Together with Eq. (1), this implies $\Delta(D_{\omega_1}, D) < \Delta(D_{\omega_2}, D)$, and hence ω_1 is the unique L_2-weighting for $(\mathcal{S}_X, D)$. This completes the proof of part (i); parts (ii) and (iii) follow by similar arguments. □

For G as above, we now summarize how the property of being a semi-clique is related to L_2-representations of D_G.

LEMMA 2.2. *Suppose that G is a graph with vertex set X, $|X| \geq 2$, and $\mathcal{T}$ is a rooted X-tree with root ρ. Let ω be an L_2-weighting for $(\mathcal{T}, D_G)$, then $D_\omega(x, y) \geq 2$ for all $x, y \in X$, $x \neq y$. In addition, if $\rho = \mathrm{LCA}(y, z)$ for some $y, z \in X$, then we have $D_\omega(x, \rho) \leq 2$ for all $x \in X$, where equality holds if G is not a semi-clique.*

PROOF. For simplicity, put $D := D_G$ and for a vertex $u \neq \rho$, let $p(u)$ be the parent of u, that is, the vertex on the path from u to ρ in $\mathcal{T}$ that is adjacent to u. Since ω is an L_2-weighting, we know that for every pair of elements $x, y \in X$, we have $D_\omega(u, x) = D_\omega(u, y)$ for every common ancestor u of x and y. In particular, we have $D_\omega(\mathrm{LCA}(x, y), x) = D_\omega(\mathrm{LCA}(x, y), y) = D_\omega(x, y)/2$. Moreover, there exists some integer $k \geq 0$ such that $D_\omega(\rho, x) = k$ for all $x \in X$.

Note that we have $k \geq 1$ because otherwise we have $D_\omega(x, y) = 0$ for all $x, y \in X$, and hence $\Delta(\mathcal{T}; D_{\omega^1}, D) < \Delta(\mathcal{T}; D_\omega, D)$, where ω^1 is the integer equi-weighting on $\mathcal{T}$ that assigns to each pendant edge of $\mathcal{T}$ weight 1, and 0 to all other edges.

First, we shall show that $D_\omega(x, y) \geq 2$ for all $x, y \in X$. If not, then consider a pair $x_1, x_2 \in X$ with $D_\omega(x_1, x_2) < 2$. Let $u = \mathrm{LCA}(x_1, x_2)$. Then by noting that $D_\omega(u, x_1) = D_\omega(x_1, x_2)/2 < 1$ we have $D_\omega(u, x_1) = 0$ and hence $D_\omega(u, \rho) = k$. Let v be the common ancestor of x_1 and x_2 such that $D_\omega(v, \rho) = k$ and $\omega(\{p(v), v\}) > 0$. Let ω' be the weighting function obtained from ω by setting $\omega'(e) = \omega(e) - 1$ for $e = \{p(v), v\}$, $\omega'(e) = 1$ for $e = \{p(x'), x'\}$ with $x' \in \mathcal{C}(v)$, and $\omega'(e) = \omega(e)$ otherwise. Then ω' is an integer equi-weighting with

$$\begin{aligned}&\Delta(\mathcal{T}; D_\omega, D) - \Delta(\mathcal{T}; D_{\omega'}, D)\\ &= \sum_{\{x,y\}\subseteq \mathcal{C}(v), x\neq y} D^2(x,y) - (D(x,y) - 2)^2 > 0,\end{aligned}$$

contradicting that ω is an L_2-weighting for $(\mathcal{T}, D)$.

Now assume that $\rho = \mathrm{LCA}(x_1, x_2)$ for some $x_1, x_2 \in X$. It remains to show that $k \leq 2$, that is, $D_\omega(x, \rho) \leq 2$ for all $x \in X$. If not, then we have $k \geq 3$. Let $\{u_1, \dots, u_t\}$ be the set of vertices in $\mathcal{T}$ such that $D_\omega(\rho, p(u_i)) = 0$ and $D_\omega(\rho, u_i) > 0$ for $1 \leq i \leq t$. Then $\{\mathcal{C}(u_1), \dots, \mathcal{C}(u_t)\}$ is a partition of X. Let ω' be the integer equi-weighting obtained from ω by setting $\omega'(e) = \omega(e) - 1$ for $e = \{p(u_i), u_i\}$ with $1 \leq i \leq t$, and $\omega'(e) = \omega(e)$ otherwise.

Then for $x \in \mathcal{C}(u_i)$ and $x' \in \mathcal{C}(u_j)$ with $i \neq j$, we have

$$D(x,x') \leq 4 \leq 2k-2 = D_\omega(x,x') - 2 = D_{\omega'}(x,x')$$

and hence

$$\begin{aligned}&(D_\omega(x,x') - D(x,x'))^2 - (D_{\omega'}(x,x') - D(x,x'))^2\\ &= (D_{\omega'}(x,x') + 2 - D(x,x'))^2 - (D_{\omega'}(x,x') - D(x,x'))^2\\ &= 4(D_\omega(x,x') - D(x,x') + 1) \geq 4.\end{aligned}$$

Therefore, in view of $t \geq 2$, we have

$$\Delta(\mathcal{T}; D_\omega, D) - \Delta(\mathcal{T}; D_{\omega'}, D) \geq 4 \sum_{1 \leq i < j \leq t} |\mathcal{C}(u_i)| \times |\mathcal{C}(u_j)| > 0,$$

contradicting that ω is an L_2-weighting for $(\mathcal{T}, D)$.

Finally, when G is not a semi-clique, a proof similar to that of Lemma 2.1 shows $k = 2$, and hence completes the proof of the lemma. □

3 MINIMUM EVOLUTION IS NP-COMPLETE

In the last section we saw how L_2-weightings were related to semi-cliques. We now use this information to relate semi-clique decompositions of graphs to the minimum evolution problem, which will also allow us to show that this latter problem is NP-complete.

We begin by presenting a problem that is closely related to the ME-problem. Given a distance matrix D on X, a rooted X-tree $\mathcal{T}$ and an L_2-weighting ω for $(\mathcal{T}, D)$, we let $\omega(\mathcal{T})$ denote the sum of the edge-weights of $\mathcal{T}$.

Problem Ultra-metric Minimum Evolution (UME(D, m))
Instance: A distance matrix D on a finite set X and an integer m.
Question: Does there exist an L_2-representation $(\mathcal{T}, \omega)$ of D such that $\omega(\mathcal{T}) \leq m$?

Now, let G be a graph with vertex set X. We call a partition P of X a *semi-clique decomposition of* G if every set in P is a semi-clique in G. We now relate this concept to the problem of finding a solution to the UME problem.

PROPOSITION 3.1. *Let G be a graph with vertex set X and $k \geq 1$ an integer. Then there exists a semi-clique decomposition of G with size at most k if and only if there exists an L_2-representation $(\mathcal{T}, \omega)$ of D_G with $\omega(\mathcal{T}) \leq |X| + k$.*

PROOF. Put $D := D_G$ and $n = |X|$. In addition, let ω_j ($j \in \mathbb{Z}_{\geq 0}$) be the weighting function that assigns weight j to each edge in a rooted X-tree. To simplify the proof it will be convenient to allow vertices of degree two in a rooted X-tree.

"⇒" Let $\{X_1, X_2, \ldots, X_k\}$ be a semi-clique decomposition of G whose size is minimum over all semi-clique decompositions of G. If $k = 1$, then consider the star tree $\mathcal{S}_X$. Since G is a semi-clique, by Lemma 2.1 we know that ω_1 is an L_2-weighting for $(\mathcal{S}_X, D)$ and, clearly, $\omega_1(\mathcal{S}_X) = n$, as required.

So, assume $k > 1$. Then G is not a semi-clique. For each $1 \leq i \leq k$, let $\mathcal{T}_i := \mathcal{S}^*_{X_i}$ be the X_i-tree obtained from $\mathcal{S}_{X_i}$ by adding a new node adjacent to the root of $\mathcal{S}_{X_i}$, and designating this new node as the root of $\mathcal{S}^*_{X_i}$. Note that, if $|X_i| = 1$, then $\mathcal{S}_{X_i}$ contains one edge while $\mathcal{S}^*_{X_i}$ contains two edges. Considering the X-tree $\mathcal{T}$ obtained by identifying the roots of all $\mathcal{T}_i$ as the root of $\mathcal{T}$, each tree $\mathcal{T}_i$ can be regarded as a subtree of $\mathcal{T}$. Moreover, since $\omega_1(\mathcal{T}) = |X| + k$, it suffices to show that ω_1 is an L_2-weighting for $(\mathcal{T}, D)$. To this end, consider an arbitrary L_2-weighting ω for $(\mathcal{T}, D)$. Since G is not a semi-clique and $k > 1$ implies that ρ is the lowest common ancestor of a pair of elements of X, by Lemma 2.2 we have $D_\omega(\rho, x) = 2$ for all $x \in X$, as well as $D_\omega(x, y) \geq 2$ for $x \neq y$. Therefore, to establish that ω_1 is an L_2-weighting for $(\mathcal{T}, D)$, it remains to show, for all i with $|X_i| \geq 2$, that $\omega(e) = 1$ for all edges e in $\mathcal{T}_i$. Indeed, if this does not hold for some i with $|X_i| \geq 2$, then by $D_\omega(x, \rho) = 2$ and $D_\omega(x, y) \geq 2$ for $x \neq y$ in X_i we must have $\omega(e) = 2$ for all pendant edges e in $\mathcal{T}_i$ and $\omega(e) = 0$ for all other edges. Let ω' be the weighting function on the edges of $\mathcal{T}$ defined as $\omega'(e) = 1$ for edges e in $\mathcal{T}_i$ and $\omega'(e) = \omega(e)$ otherwise. Since X_i is a semi-clique in G, an argument similar to the one used in the proof of Lemma 2.1 either yields $\Delta(D_{\omega'}, D) < \Delta(D_\omega, D)$, contradicting that ω is an L_2-weighting, or $\Delta(D_{\omega'}, D) = \Delta(D_\omega, D)$, as required.

"⇐" Let k be the minimum positive number such that there exists a rooted X-tree $\mathcal{T} = (V, E)$ and an L_2-weighting ω for $(\mathcal{T}, D)$ with $\omega(\mathcal{T}) \leq |X| + k$. Without loss of generality, we may assume that $k < |X|$ (as otherwise the conclusion clearly holds) and that the root ρ of $\mathcal{T}$ is the lowest common ancestor of two elements in X (as the single edge incident to a root of degree one can always be contracted without changing the distance $D_\omega(x, y)$ for any $x, y \in X$). In addition, we may assume that $\omega(e) > 0$ for all edges $e \in E$ (indeed, by Lemma 2.2 we can assume $\omega(e) > 0$ for all pendant edges e of $\mathcal{T}$ and an interior edge with weight 0 can simply be contracted) and may further assume that $\omega = \omega_1$ (as an edge with weight $m > 1$ can be replaced by m edges with weight 1).

Now, if $k = 1$ it follows immediately from the assumptions above that $\mathcal{T} = \mathcal{S}_X$ and, therefore, in view of Lemma 2.1 we can conclude that G is a semi-clique, as required.

So assume $1 < k < |X|$. Then we can further assume that G is not a semi-clique, as otherwise the result clearly holds. Therefore, by Lemma 2.2, we have $D_{\omega_1}(x, \rho) = 2$ for some (and hence all) $x \in X$. This implies that, besides $|X|$ pendant edges, $\mathcal{T}$ contains k edges $\{e_1, \ldots, e_k\}$ that are adjacent to ρ.

For $1 \leq i \leq k$, let X_i be the set of elements x in X such that the path between ρ and x contains e_i and let E_i be the set of pendant edges incident to e_i. It remains to show that, for $1 \leq i \leq k$, X_i is a semi-clique in G. Indeed, if this were not the case for some i, then clearly $|X_i| \geq 2$. Let ω' be the weighting function obtained from ω_1 by setting $\omega'(e) = 0$ for $e = e_i$, $\omega'(e) = 2$ for $e \in E_i$, and $\omega'(e) = \omega_1(e)$ otherwise. Since X_i is not a semi-clique, an argument similar to the proof of Lemma 2.1 leads to the contradiction $\Delta(D_{\omega'}, D) < \Delta(D_{\omega_1}, D)$. □

By the main result of Kaya *et al.* (2013) it follows that the following problem is NP-complete.

Problem Semi-clique decomposition (SCD(G, k))
Instance: A graph G with finite vertex set X and an integer k.
Question: Does there exist a semi-clique decomposition P of G such that $|P| \leq k$?

Using this fact, we immediately obtain the following corollary to Proposition 3.1.

Corollary 3.2. *The problem* UME(D, m) *is NP-complete, even when the non-diagonal entries of the distance matrix* D *are all in* $\{2, 4\}$. □

Now we return to the ME problem mentioned in the Introduction. It refers to unrooted X-trees, that is, we drop the condition of having a distinguished root vertex and, as a consequence, when referring to weightings we also drop the condition that all leaves have the same distance from the root. To avoid any confusion as to whether the latter condition applies or not we will use the term *unrooted* when referring to weightings and L_2-representations for which it does not apply. Formally, the ME problem is stated as below.

Problem Minimum Evolution (ME(D, m))
Instance: A distance matrix D on a finite set X and an integer m.
Question: Does there exist an unrooted L_2-representation (T, ω) of D such that $\omega(T) \leq m$?

Now, using Corollary 3.2 and the following transformation that was presented by Day (1987) we show that the ME problem is NP-complete. Given a distance matrix D on X with $|X| = n$ and two integers m and p, let $Y := \{y_1, \cdots, y_m\}$ be a set disjoint from X and let $\widetilde{D} := f_{m,p}(D)$ be the distance matrix on $X \cup Y$ defined as $\widetilde{D}(x_i, x_j) = D(x_i, x_j)$ for $x_i, x_j \in X$, $\widetilde{D}(x, y) = p$ for $x \in X, y \in Y$, and $\widetilde{D}(y_i, y_j) = 2$ for $y_i \neq y_j$ in Y. The NP-completeness of ME follows from the next result, whose rather technical proof is presented in the appendix.

Proposition 3.3. *Suppose* $|X| = n \geq 4$. *Suppose that* D *is a distance matrix on* X *with* $D(x, x') \in \{2, 4\}$ *for* $x \neq x'$ *in* X. *Let* $p = n^3$, $m = p^3$ *and* $k \geq 1$. *Then* D *has an* L_2*-representation* $(\mathcal{T}, \omega)$ *with* $\omega(\mathcal{T}) \leq n + k$ *if and only if* $f_{m,p}(D)$ *has an unrooted* L_2*-representation* (T, w) *with* $w(T) \leq n + k + m + (p - 3)$.

By Proposition 3.3 and Corollary 3.2 we obtain the main result of this paper.

Theorem 3.4. *The problem* ME(D, m) *is* NP*-complete even when the non-diagonal entries of the distance matrix* D *take on only three values.* □

It would be interesting to see whether the ME problem is hard for the more general case where the edge weights can be set to rational numbers. Note that the hardness of the BME problem mentioned in the introduction includes the case of rational weight (Fiorini and Joret, 2012). On the other hand, Theorem 3.4 does not imply that the rational version of the ME-problem (RME) is hard, and there are many optimization problems which can be solved efficiently once the restriction that the solution must be integral is removed, such as the well-known linear programming problem (cf. Schrijver, 1986). A starting point to explore the complexity of RME could be the observation that the semi-clique decomposition problem is a special case of the γ*-clique* decomposition problem for $\gamma = \frac{1}{2}$, in which the aim is to decompose a graph G into a minimum number of γ-cliques, where γ is a real number with $0 \leq \gamma \leq 1$, and a γ*-clique* in G is a subset C of V having at least $\gamma\binom{|C|}{2}$ edges in G with both endpoints in C (cf. Guo *et al.*, 2011; Pattillo *et al.*, 2013). In addition, Kaya *et al.* (2013) showed that the γ*-clique* decomposition problem is NP-complete. However, to date we have not been able to use this fact to prove that the RME problem is also NP-complete.

4 DISCUSSION

To prove that the ME problem is NP-complete, we first showed that the UME problem is NP-complete by relating it to the semi-clique decomposition problem. Interestingly, this is a special example of a more general link between tree inference and *graph clustering* problems. In particular, we can link the following two types of problem for a set X:

(i) Given a distance matrix D on X and a tree scoring function σ_D on the set $\mathbb{T}_X$ of all rooted X-trees, find a tree that optimizes σ_D.

(ii) Given a graph G with vertex set X and a *cluster scoring* function $\kappa_G : \mathbb{P}_X \to \mathbb{R}$ that assigns to each partition in the set $\mathbb{P}_X$ of all partitions of X a real number κ_G, find a partition of X that optimizes κ_G.

More specifically, this correspondence is obtained by restricting any given tree inference problem to the set of rooted trees in $\mathbb{T}_X$ where every leaf is adjacent to a vertex that is adjacent to the root (the tree in Figure 1(a), for example, has this structure), to edge weightings that assign to every edge weight 1 and to distance matrices that have only off-diagonal entries that are 2 or 4. In this restricted type of rooted tree, every vertex u adjacent to the root induces a cluster of elements in X (namely the leaves that are adjacent to u) and, clearly, for every partition P of X there exists a unique such tree $\mathcal{T}_P$ that induces precisely the clusters in P. Thus, given any graph G with vertex set X and the distance matrix $D = D_G$ on X which is induced by G, we obtain the cluster scoring function κ_G from the scoring function σ_D by putting $\kappa_G(P) = \sigma_D(\mathcal{T}_P)$.

To give another example of this correspondence, consider the L^1-fit problem (see, e.g., Day, 1987; Farach *et al.*, 1995). In this problem, given a distance matrix D the aim is to find a rooted X-tree $\mathcal{T}$ which minimizes the score $\sigma_D(\mathcal{T})$ which is equal to the minimum of $\sum_{x,y \in X} |D(x, y) - D_\omega(x, y)|$ taken over all weightings ω of $\mathcal{T}$. For this example, the corresponding graph clustering problem is known as the *correlation clustering* problem (Bansal *et al.*, 2004) where the cluster scoring function κ_G assigns, for a given graph G with vertex set X, to any partition P in $\mathcal{P}_X$ the number of 2-element subsets (i.e. edges) $e = \{u, v\}$ of X that violate P, that is, either e is an edge of G but u and v do not lie in the same cluster in P or e is not an edge of G but u and v both lie in the same cluster of P. For the partition $P = \{C_1, C_2, C_3\}$ of the graph G in Figure 1(b), for example, this cluster scoring function yields a score of 4. It is not hard to check that the L^1-fit of the tree in Figure 1(a) for the distance matrix D_G derived from the graph G is 4 too. Note that minimizing the cluster scoring function corresponding to the L^1-fit is equivalent to computing the minimum number of edge deletions and insertions that suffice to transform G into a disjoint union of complete graphs. When adopting this latter view, correlation clustering is usually referred to as *cluster editing* (see, e.g., Böcker *et al.*, 2011).

It would be interesting to explore which tree inference problems are related in a similar way to other graph clustering problems, and conversely. This could yield useful new insights into these inference problems, and possibly new algorithms for their solution or approximation.

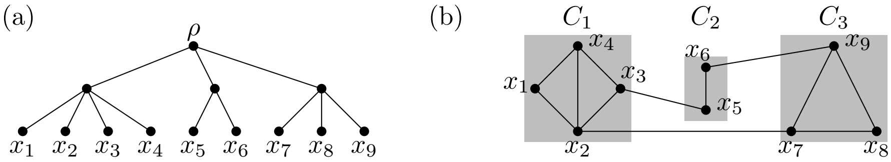

Fig. 1. (a) A rooted graph theoretical tree $\mathcal{T}$ with root ρ and leaf set $X = \{x_1, x_2, \ldots, x_9\}$. For the edge weighting ω that assigns weight 1 to every edge of $\mathcal{T}$, the shortest path distance $D_\omega(x_2, x_5) = 4$. (b) A graph G with vertex set $X = \{x_1, x_2, \ldots, x_9\}$ that is partitioned into the clusters C_1, C_2 and C_3 indicated by the shaded boxes.

ACKNOWLEDGEMENT

We would like to thank three anonymous referees for their helpful and constructive comments on a previous version of this article.

REFERENCES

Addario-Berry, L., Chor, B., Hallett, M., Lagergren, J., Panconesi, A., and Wareham, T. (2004). Ancestral maximum likelihood of evolutionary trees is hard. *Journal of Bioinformatics and Computational Biology*, **2**, 257–271.

Bansal, N., Blum, A., and Chawla, A. (2004). Correlation clustering. *Machine Learning*, **56**, 89–113.

Böcker, S., Briesemeister, S., and Klau, G. (2011). Exact algorithms for cluster editing: evaluation and experiments. *Algorithmica*, **60**, 316–334.

Brunato, M., Hoos, H., and Battit, R. (2008). On effectively finding maximal quasi-cliques in graphs. *Lecture Notes in Computer Science*, **5313**, 41–55.

Catanzaro, D. (2009). The minimum evolution problem: overview and classification. *Networks*, **53**, 112–125.

Day, W. (1987). Computational complexity of inferring phylogenies from dissimilarity matrices. *Bulletin of Mathematical Biology*, **49**, 461–467.

Desper, R. and Gascuel, O. (2002). Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of Computational Biology*, **19**, 687–705.

Desper, R. and Gascuel, O. (2004). Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fittings. *Molecular Biology and Evolution*, **21**, 587–598.

Desper, R. and Gascuel, O. (2005). The minimum-evolution distance-based approach to phylogenetic inference. In O. Gascuel, editor, *Mathematics of evolution and phylogeny*, pages 1–32. Oxford University Press.

Farach, M., Kannan, S., and Warnow, T. (1995). A robust model for finding optimal evolutionary trees. *Algorithmica*, **13**, 155–179.

Filipski, A., Tamura, K., Billing-Ross, P., Murillo, O., and Kumar, S. (2015). Phylogenetic placement of metagenomic reads using the minimum evolution principle. *BMC Genomics*, **16**, S13.

Fiorini, S. and Joret, G. (2012). Approximating the balanced minimum evolution problem. *Operation Research Letters*, **40**, 31–35.

Guo, J., Kanj, I., Komusiewicz, C., and Uhlmann, J. (2011). Editing graphs into disjoint unions of dense clusters. *Algorithmica*, **61**, 949–970.

Kaya, O., Kayaaslan, E., and Uçar, B. (2013). On the minimum edge cover and vertex partition by quasi-cliques problems. *HAL Research Report*, **RR-8255**.

Kidd, K. and Sgaramella-Zonta, L. (1971). Phylogenetic analysis: Concepts and methods. *American Journal of Genetics*, **23**, 235–252.

Kirkpatrick, B., Li, S., Karp, R., and Halperin, E. (2011). Pedigree reconstruction using identity by descent. *Journal of Computational Biology*, **18**, 1481–1493.

Pattillo, J., Veremyev, A., Butenko, S., and Boginski, V. (2013). On the maximum quasi-clique problem. *Discrete Applied Mathematics*, **161**, 244–257.

Rzhetsky, A. and Nei, M. (1993). Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular Biology and Evolution*, **10**, 1073–1095.

Saitou, N. and Nei, M. (1987). The Neighbor-Joining method: a new method for reconstructing phylogenetictrees. *Molecular Biology and Evolution*, **4**, 406–425.

Schaeffer, S. (2007). Survey: graph clustering. *Computer Science Review*, **1**, 27–64.

Schrijver, A. (1986). *Theory of linear and integer programming*, Wiley.