On the extension of a partial metric to a tree metric

Alain Guénoche[1], Bruno Leclerc[2], Vladimir Makarenkov[3]

[1]Institut de Mathématiques de Luminy,
163 avenue de Luminy, F-13009 MARSEILLE, FRANCE,
guenoche@iml.univ-mrs.fr

[2]Centre d'Analyse et de Mathématique Sociales,
École des Hautes Études en Sciences Sociales,
54 bd Raspail, F-75270 PARIS CEDEX 06, FRANCE,
leclerc@ehess.fr

[3]Département de Sciences Biologiques,
Université de Montréal,
C.P. 6128, succ. Centre-ville, Montréal, Québec H3C 3J7, CANADA,
and Institute of Control Sciences,
65 Profsoyuznaya, Moscow 117806, RUSSIA,
makarenv@magellan.umontreal.ca

September 2000

Abstract. Farach, Kannan and Warnow (1995) have defined Problem **MCA** (matrix completion to additive) and proved it to be NP-complete: given a partial dissimilarity d on a finite set X, does there exist a tree metric extending d to all pairs of elements of X. We use a previously described simple method of phylogenetic reconstruction, and its extension to partial dissimilarities, to characterize some classes of polynomial instances of **MCA** and of a related problem. We point out that these problems admit many other polynomial instances. Our main tool consists of two classes of generalized cycles, together with the corresponding maximal acyclic graphs (2-trees and 2d-trees).

Résumé. Farach, Kannan et Warnow (1995) ont posé le problème **MCA** (matrix completion to additive) suivant et ont démontré sa NP-complétude : étant donné une dissimilarité d partielle sur un ensemble fini X, est-il possible de l'étendre en une distance d'arbre définie sur toutes les paires d'éléments de X. Nous utilisons une méthode simple de reconstruction phylogénétique, précédemment décrite, et son extension aux dissimilarités partielles pour caractériser des classes d'instances polynomiales de **MCA** et d'un problème voisin. Nous montrons qu'en fait beaucoup d'autres instances sont aussi polynomiales. L'outil principal est constitué par deux classes de cycles généralisés, avec les graphes acycliques maximaux (2-arbres et 2d-arbres) correspondants.

Keywords: tree, 2-tree, partial distance

1. Introduction

We consider a partial metric d on a fixed finite set X. Precisely, the value of d is known on a subset E of undirected pairs of elements of X. The following decision problem **MCA** (Matrix Completion to Additive) arises in several application domains, e.g. phylogenetic tree reconstruction: does there exist a valued X-tree T, such as the tree metric d_T associated with it satisfies the following condition: for any $xy \in E$, $d(x,y) = d_T(x,y)$.

In other terms, is it possible to complete d into a tree metric? We also consider a non-metric version **WMCA** (WeakMatrix Completion to Additive) of **MCA**, where negative values on some edges of the X-tree T are allowed and, as a consequence, d_T does not necessarily satisfy the metric triangle inequality.

Farach, Kannan and Warnow (1995) proved that Problem **MCA** is NP-complete. Here we characterize some polynomial instances of both Problems **WMCA** (Section 4) and **MCA** (Section 5). Our approach is based on a simple phylogenetic reconstruction method recalled in Section 3. This method was previously described in Leclerc and Makarenkov (1998) and recently extended to an approximation method to fit a tree metric to a partial metric with (Guénoche and Leclerc 2000). Two types of generalized aciclicities will be extensively used in this paper. One of them is defined, whereas the other is recalled, in Section 2. In Section 6, we point out that Problems **MCA** and **WMCA** are, in fact, polynomial in a wide class of instances.

2. Notations and definitions

2.1. *Graphs and XLL-trees.* We consider here only undirected simple graphs without loops or multiple edges. In such a graph $G = (V,E)$, a vertex v is a *leaf* if its degree $\partial(v)$ is equal to 1. In a path $(vv_1, v_1v_2, \ldots, v_{k-1}v')$ of G between two vertices v and v', all the vertices are distinct except, possibly, when $v = v'$ and the path P is a *cycle* of G. The graph G is a *tree* if it is connected and has no cycles. The unique path between two distinct vertices v and v' of a tree T is denoted as $T(v,v')$. The graph G is a k-clique if $|V| = k$ and $uv \in E$ for all $u, v \in V$. A *triangle* of G is a subset of V inducing a 3-clique; such a subset is denoted xyz instead of $\{x,y,z\}$.

A *valued graph* is an ordered pair (G,ℓ), where G is a graph and ℓ is a real length function on the edge set E of G. When the graph G is connected and has no circuits of negative length, we set, for any two distinct vertices v and v' of G,
$d_G(v,v') = \mathrm{Min}_{P \text{ path of } G \text{ between } v \text{ and } v'} \sum_{e \in P} \ell(e)$.
In the case of a tree T, $d_T(v,v') = \sum_{e \in T(vv')} \ell(e)$.

An *XLL–tree* (leaf labelled according to X tree) is a tree T satisfying two properties: (i) the leaf set of T is X; (ii) for any $v \in V(T)$-X, $\partial(v) \geq 3$. In an XLL-tree, the vertices in $V(T)$-X are called *latent vertices*. The maximum number of latent vertices of T is n-2, where $n = |X|$; when it is reached, all the latent vertices have degree 3 and the tree T is said to be *resolved*. For more definitions and properties of such trees, see the book of Barthélemy and Guénoche (1991).

2.2. *Dissimilarities and metrics.* A *dissimilarity* on X is a real function d on $X \times X$ satisfying $d(x,y) = d(y,x)$ and $d(x,y) \geq d(x,x) = 0$ for all $x, y \in X$:

A dissimilarity d is a *metric* (or a *metric*) if it satisfies the classical metric triangle inequality: for all $x, y, z \in X$, $d(x,z) \leq d(x,y) + d(y,z)$. It is well known that this

property is satisfied by the minimum path length function of any positively valued connected and undirected graph. So, a metric d^m is associated in this way to the complete graph on X valued by a dissimilarity d.

A dissimilarity d is a *tree metric* if it satisfies the following four-point condition: for all $x, y, z, w \in X$, the inequality (F) holds:

$$d(x,y) + d(z,w) \leq \max\{\ d(x,z) + d(y,w),\ d(x,w) + d(y,z)\ \}. \qquad \text{(F)}$$

It is now well-known that a tree metric is uniquely representable by the lengths of the paths between the leaves of a non-negatively valued XLL-tree T_d, called its *tree representation* (Buneman 1971).

An extension of the previous result (Leclerc 1995) consists of considering the *weak four-point condition*, where the inequality (F) is met only for all *distinct* $x, y, z, w \in X$. A dissimilarity d satisfying this condition is not necessarily a metric. Such a dissimilarity is called a *tree dissimilarity*. A real function d on $X \times X$ satisfying the weak four-point condition is called a *tree function*. A tree function (resp. dissimilarity) is easily transformable into a tree dissimilarity (resp. metric) by addition of a convenient positive constant $2C$ to each of its values. Conversely, reducing by C the lengths of all terminal edges in a positively valued XLL-tree T is equivalent to reducing by $2C$ the path lengths between leaves of T. As a consequence, a tree function has again a unique XLL-tree representation, possibly with negative lengths on the external edges (incident to the leaves).

Sometimes, the dissimilarity d is *partial*, in that sense that it is defined only on a set E of unordered pairs of elements of X. Thus, we have a *support graph* $G = (X,E)$, valued by d. We say that a dissimilarity d' *extends* d, or d *completes into* d' if $xy \in E$ implies $d'(xy) = d(xy)$. Without loss of generality, it will be assumed in the sequel that G is connected. We say that d is a *partial metric* if, for any $xy \in E$, $d(xy) = d^m(xy)$. For the complete graph as G, d is a partial metric if and only if it is a metric. The following property is well-known and easy to obtain. Clearly, a partial metric d may be always completed into its associated minimum path length metric d^m.

Proposition 2.1. *A partial dissimilarity d completes into a metric on X if and only if it is a partial metric.*

2.3. *Two problems.* Assume that a partial dissimilarity on X with a support graph $G = (X,E)$ is given. The following "Matrix Completion to Additive" (**MCA**) problem has been shown to be *NP*-hard by Farach et al. (1995):

Problem MCA: given a partial metric d on X, does it complete into a tree metric?

According to Proposition 2.1 above, when the given partial dissimilarity is not a metric, the answer to Problem **MCA** is negative. The following "Weak Matrix Completion to Additive" (**WMCA**) problem remains of interest since such a completion still provides a tree structure (but negative lengths do not fit most of evolutionary models). In such an extension, it is not important to distinguish tree dissimilarities from tree functions in the completion ouput. Here we do not address the complexity status of **WMCA**, and just exhibit some polynomial classes of instances.

3. 2d-trees and 2-trees

We recall and complete the description of two classes of graphs which constitute major tools for this study.

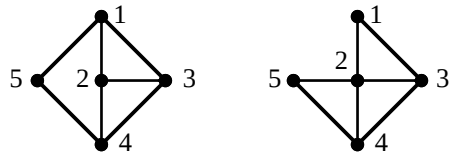
3.1. 2*d-acyclic graphs.* Let $G = (X,E)$ be a finite undirected simple graph, and $A \subseteq E$ a set of edges of G. Then, X_A denotes the set of all vertices incident to one edge of A at less, and G_A the subgraph (X_A,A) of G. A set $C \subseteq E$ is said to be a *kd-cycle* (d for degree) of G if all the vertices of X_C have degree at least k+1 in G_C and C is minimal for inclusion with this property. Clearly, a 1d-cycle is a cycle. Here we are concerned with the case $k = 2$.

Examples. If G_C is isomorphic to the complete graph K_{k+2} or to the complete bipartite graph $K_{k+1,k+1}$, then C is a kd-cycle. If G_C is a wheel, then C is a 2d-cycle.

A graph with no kd-cycles is said *kd-acyclic*. The maximal *kd-acyclic* graphs are called here *kd-trees*. They have been characterized in a recursive way by Todd (1989):

- the complete graph K_k with k vertices is a kd-tree;
- if $G = (X,E)$ is a kd-tree, then, for any subset $Y \subseteq X$ of cardinality k and new vertex $x \notin X$, the graph $G' = (X \cup \{x\}, E \cup \{xy: y \in Y\})$ is a kd-tree.

Then, a graph $G = (X,E)$ is a 2d-tree if there exists an ordering $(x_1,x_2,\ldots,x_n)$ of X such that $x_1x_2 \in E$ and, for i = 3,…, n, the vertex x_i has degree 2 in the subgraph G^i induced by the vertex set $\{x_1,x_2,\ldots,x_i\}$ (such an ordering is a *reversed elimination order*, abbreviated as *RE order*). A 2d-tree with n vertices is 2-connected and has $2n$-3 edges. It has at least one vertex of degree 2. Both graphs G and G' of Figure 1 are 2d-trees.



G *G'*

Figure 1

Given a set of edges A, Todd proposes a procedure for deciding whether it includes a kd-cycle. This procedure determines a subset Peel(A) of A, called the *kd-peeling* of A, as follows: search a vertex of degree at most k in G_A; if no such vertex exists, then Peel(A) = A; otherwise, delete the vertex found with its incident edges, and repeat the operation until no vertices of degree 1 remain. The set of remaining edges is Peel(A). The set A is a kd-acyclic if and only if Peel(A) = $\varnothing$. Such an algorithm clearly runs in $O(n)$ time.

A connected 2d-acyclic graph completes in many ways into a 2d-tree; we give here a procedure that will be useful in Section 4:

- If there exists a vertex x of degree 1, add a new edge between x and an arbitrary other vertex y, not already adjacent to x. Repeat the operation until no vertex of degree less than 2 remains.
- List all the pairs not included in E in an arbitrary order and check them according to the list order. For each such pair xy, use the 2d-peeling algorithm above to determine whether the graph $(X, E\cup\{xy\})$ is acyclic; add xy to E if the answer is positive, and reject it otherwise. Stop when $|E| = 2n-3$.

Algorithm 3.1. *Completion of 2d-acyclic graph into a 2d-tree.*

While there exists $x \in X$ such that $\partial(x) = 1$
  Select $y \in E$ such that $y \neq x$ and $xy \notin E$
  $E := E \cup \{xy\}$
End While
If $|E| = 2n-3$ then **stop** Algorithm
Else
  Make a list L of all pair xy not included in E
  While $|E| < 2n-3$
    Select any pair xy from L
    Apply *2d-peeling algorithm* to check whether the graph
      $G = (X, E\cup\{xy\})$ is acyclic
    $L := L \setminus \{xy\}$
    If G is acyclic
      $E := E \cup \{xy\}$
  End While
End Else

Proposition 3.2. *If G is a 2d-acyclic graph, the above algorithm extends it into a 2d-tree in $O(n^3)$ time.*

Proof. Clearly, adding an edge to a vertex of degree 0 or 1 cannot create a 2d-cycle. This justifies the first part of the algorithm. In the second part, the final graph is a maximal 2d-acyclic graph, that is a 2d-tree; otherwise, further pairs would be retained during the scanning of of the list.

As far as the algorithmic complexity is concerned, the first part is in $O(n)$. In the second one, we have to check $O(n^2)$ pairs, the peeling procedure being in $O(n)$ each time. ❑

The notion of chain generalizes to 2d-trees. Let $G = (X,E)$ be a 2d-tree on X and a pair $xy \notin E$. The graph $G' = (X,E \cup \{xy\})$ is no longer 2d-acyclic. It has a unique 2d-cycle $C_{xy} = \mathrm{Peel}(E \cup \{xy\})$. At each step of the peeling algorithm, a vertex is eliminated together with two edges. So, setting $Y = X_{\mathrm{Peel}(E \cup \{xy\})}$ and $n' = |Y|$, the equality $|\mathrm{Peel}(E \cup \{xy\})| = 2n'\text{-}2$ holds. Then, the graph $H = (Y, C_{xy} - \{xy\})$ is 2d-acyclic with $2n'\text{-}3$ edges and, so, is a 2d-tree. This 2d-tree has one or two vertices of degree two, taken in the set $\{x,y\}$; it is called the *2d-chain* $G[xy]$ *of* G *between* x *and* y. If x' and y' are two vertices of $G[xy]$, then the 2d-chain $G[x'y']$ is a subgraph of $G[xy]$. The *length* of a 2d-chain, comprised between 2 and n-2, is the number of its vertices minus 2. We then have the following property:

Proposition 3.3. *For any RE order on* X, *the last vertex of* $G[xy]$ *is* x *or* y *and has degree* 2 *in* $G[xy]$.

Proof. Assume that there exists a RE order L on X and a vertex z of $G[xy]$ such that both x and y are predecessors of z in L. Then, the construction above would lead to a 2d-chain between x and y without z as a vertex. ❑

3.2. 2-*acyclic graphs.* Another generalization of cycles and trees is more classical than the previous one, and has prompted important litterature. Recall that, given a graph G, a reduced graph is obtained from G by successive contractions of edges incident to a vertex of degree 2 until no such possible operation remains. For instance, a cycle reduces to a 3-clique. A graph G is *homeomorphic* to a graph H without vertices of degree 2 if its reduced graph is isomorphic to H.

For $k \geq 2$, a set $C \subseteq E$ is said to be a *k-cycle* of G if the graph G_C is homeomorphic to K_{k+2}. Especially, a 1-cycle is a cycle. A graph with no k-cycles is said *k-acyclic*. The maximal k-acyclic graphs are the classical k-trees, which are recursively characterized as follows:

- the complete graph K_k is a k-tree ;
- if $G = (X,E)$ is a k-tree, then, for any k-clique $Y \subseteq X$ of G and new vertex $z \notin X$, the graph $G' = (X \cup \{z\}, E \cup \{zy : y \in Y\})$ is a k-tree.

So, k-trees are particular cases of kd-trees. The subgraphs of 2-trees are exactly the graphs with no subgraph homeomorphic to K_4; they are called partial 2-trees or series-parallel graphs in the literature (Wald and Colbourn 1983). k-trees are also the maximal triangulated (or chordal, or rigid circuit) graphs with no $(k+2)$-clique (Rose 1974), that is the maximal graphs of treewidth k (see e.g. Bodlaender 1997). Such properties make them to constitute an interesting class in algorithmic graph theory.

We are again interested in the case where $k = 2$. A graph $G = (X,E)$ is a 2-tree if there exists a RE order $(x_1,x_2,\ldots,x_n)$ of X such that $x_1x_2 \in E$ and, for i = 3,…, n, the vertex x_i has degree 2 and belongs to a unique triangle in the subgraph G^i induced by the vertex set $\{x_1,x_2,\ldots,x_i\}$. A 2-tree with n vertices has at least two vertices of degree 2. 2-trees are the maximal triangulated graphs without 4-clique. The graph G' of Figure 1 is a 2-tree while G is not (note that it is a 2-cycle). An $O(\max(m,n))$ algorithm to decide whether a given graph G is 2-acyclic was devised by Liu and Geldmacher (1980). The analogous problem is *NP*-hard for $k \geq 3$ (Arnborg *et al.* 1987).

As a consequence of the above recursive characterizations of 2d-trees and 2-trees, one obtains the following property which, for $k = 2$, is a variant of a well-known result of Dirac (1952; see, e.g., Welsh 1976, p. 238, or Aigner 1979, p. 387):

Proposition 3.4. *A kd-cycle includes at least one k-cycle.*

Proof. Otherwise, let C be a kd-cycle such that G_C has no k-cycle. So, G_C is a subgraph of a maximal graph G with no k-cycle, that is a k-tree. But G is also a kd-tree, a contradiction with the hypothesis that C is a kd-cycle. ❑

4 The triangle method

4.1. *A solution of **WMCA** in* 2*d-trees.* We first assume that the support graph G of the given partial dissimilarity d is a 2d-tree. In that case, an extension of d to a tree function always exists and is obtained by the *triangle method*. This procedure, introduced in Leclerc (1995), was then formalized and studied in Makarenkov (1997), Leclerc and Makarenkov (1998), and Guénoche and Leclerc (2000).

In fact, the triangle method builds a valued XLL-tree T such that $d_T(x,y) = d(x,y)$ for any $xy \in E$. The basic observation is that a triangle $\{x,y,z\}$, weighted according to d, defines a valued $\{x,y,z\}$LL-tree T of the 3-star type, that is, with a unique latent vertex u. The values $d(x,y)$, $d(x,z)$ and $d(y,z)$ are uniquely obtained as path lengths in T after resolving the following system of linear equations $2d_T(xu) = d(x,y) + d(x,z) - d(y,z)$, $2d_T(yu) = d(y,x) + d(y,z) - d(x,z)$ and $2d_T(zu) = d(z,x) + d(z,y) - d(x,y)$.

The order of vertices in G is an arbitrary RE order $x_1, x_2,\ldots, x_n$. So, for every x_i, there exist exactly two elements $y, z \in \{x_1,\ldots, x_{i-1}\}$ such that both xy and xz belong to E. The triangle $\{x_i,y,z\}$ will be changed into an $\{x_i,y,z\}$LL-tree of the 3-star type, and the obtained 3-stars will be successively glued to finally obtain an XLL-tree.

- First, the triangle $\{x_1,x_2,x_3\}$ is represented as a 3-star T_3. Then, the same operation is made on the triangle $\{x_4,y,z\}$, where $y, z \in \{x_1,x_2,x_3\}$.

- A second 3-star T_4 is obtained with the path $T_4(yz)$ common with $T_3(yz)$, with the same length $d(y,z)$. The trees T_3 and T_4 are glued on this path to obtain an $\{x_1,x_2,x_3,x_4\}$LL-tree.
- A new triangle with the vertices x_i, y, z such that $y, z \in \{x_1,x_2,\ldots,x_{i-1}\}$ is considered at each step; the existence of such a triangle is guaranteed by the properties of RE orders. If $yz \notin E$, then its value is fixed as $d_{T^{i-1}}(y,z)$, the length of the path between y and z in the current $\{x_1,x_2,\ldots,x_{i-1}\}$LL-tree T^{i-1}. So, the 3-star corresponding to the triangle $\{x_i,y,z\}$ provides a grafting of the new vertex x_i onto this tree.
- Finally, an XLL-tree $T = T^n$ is obtained, preserving all the dissimilarity values in d. Applying the triangle method is polynomial with $O(n^2)$ complexity. Thus:

Proposition 4.1. *If $G = (X, E)$ is a 2d-tree, then there exists a valued XLL-tree T such that the tree function d_T extends d.*

Theorem 4.2. *If $G = (X, E)$ is 2d-acyclic, then there exists a valued XLL-tree T such that the tree function d_T extends d.*

Proof. Assume G is 2d-acyclic and use Algorithm 3.1 to obtain a 2d-tree $G' = (X, E')$ with $E \subseteq E'$. Give arbitrary positive lengths to all the pairs in E'-E. Now d extends to a partial dissimilarity with a 2d-acyclic support graph and the result follows from Proposition 4.1. ❑

Since Algorithm 3.1 runs in $O(n^3)$ and the triangle method runs in $O(n^2)$, we are able to conclude:

Corollary 4.3. *Partial dissimilarities with 2d-acyclic support graphs constitute a polynomial class of Problem* **WMCA**.

In the particular case where G is a 2-tree, Leclerc and Makarenkov (1998) showed that the final X-tree does not depend on the order on triangles. Their arguments extend to 2d-trees.

Theorem 4.4. *The triangle method uniquely extends a partial dissimilarity d with a 2d-tree support graph $G = (X,E)$ to a tree function d_T, independently of the used RE order.*

Proof. Let $x, y \in X$ such that $xy \notin E$. We proceed by induction on the length k of the 2d-chain $G[xy]$. The result is obvious for $k \leq 4$. Assume that it is true for all 2d-trees of length at most k-1 and, without lost of generality, that x has degree 2 in $G[xy]$. Let z and z' be the vertices adjacent to x in this graph. As observed in Section 3.1, either the pair zz' belongs to E, or $G[zz'] \subset G[xy]$. In both cases, $d_T(z,z')$ is given or, by the induction hypothesis, uniquely determined by the triangle method applied to $G[zz']$; the same for

both pairs yz and yz'. Finally, one has $d_T(x,y) = \max\{\ d(x,z) + d_T(y,z')\ ,\ d_T(x,z') + d_T(y,z\ \}- d_T(z,z')$. ❑

Example 4.5. Consider Figure 2, which shows a 2d-tree endowed with a partial metric d. Figure 3 shows the 3-stars associated to its triangles and their successive incorporation, until the final X-tree is obtained. In all our examples, the alphabetic order on the vertices will be an RE order.

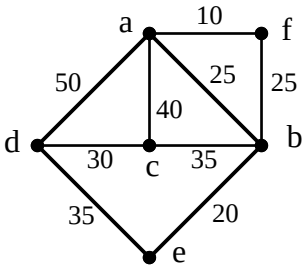


Figure 2: a 2d-tree endowed with a partial metric d

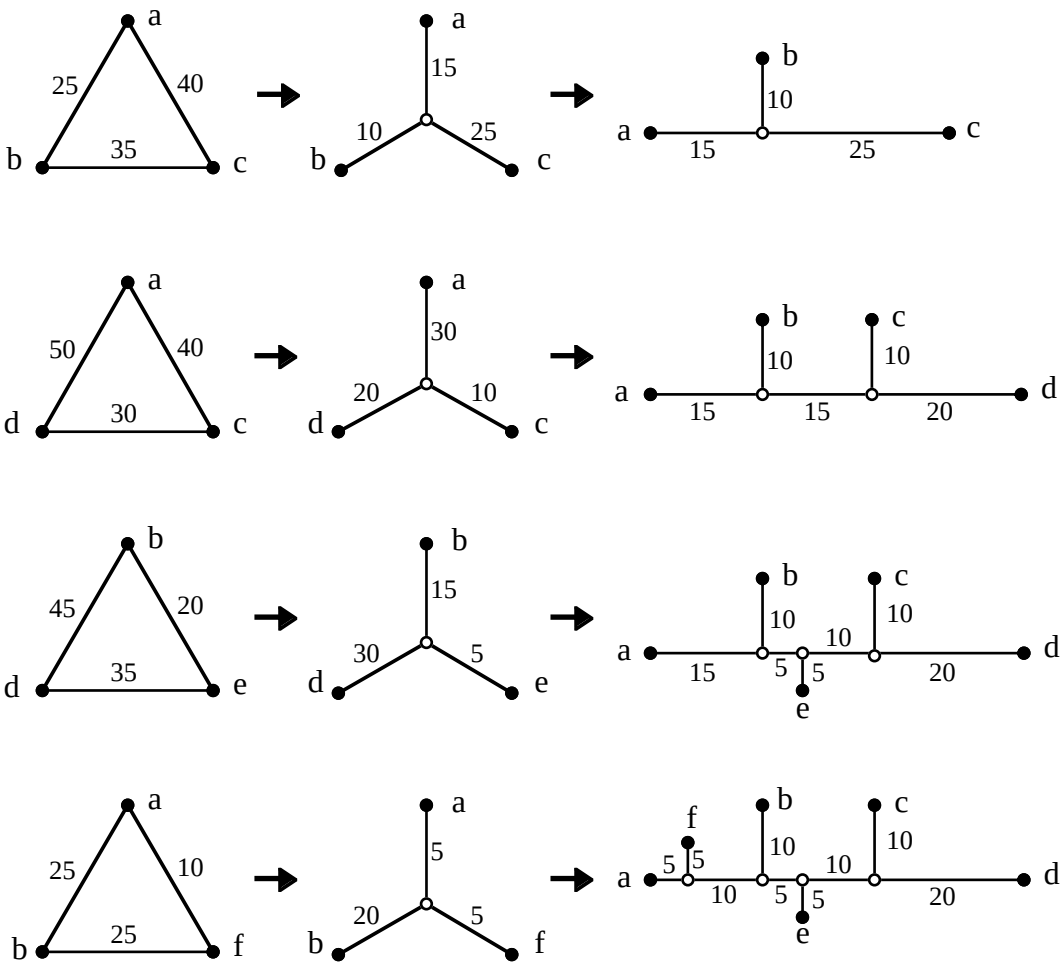


Figure 3: construction of an X-tree from the valued 2d-tree of Figure 2

4.2. *Resolved* 2*d-trees and the* ***MCA*** *problem.* The triangle method provides a unique tree function extension of any partial dissimilarity defined on a 2d-tree. But this extension is not always the unique possible one.

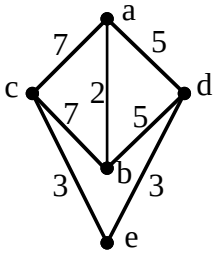


Figure 4

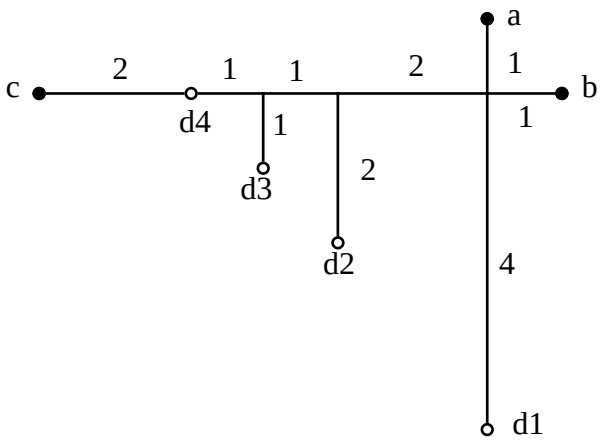


Figure 5

Example 4.6. Consider the valued 2d-tree of Figure 4, with the RE order (a, b, c, d, e). The tree of Figure 5 shows four possible graftings of d on the initial { a, b, c }-tree, among an infinity; here, d1 is the grafting provided by the triangle method. The reason of such an ambiguity is the equality of the sums d(a,c)+d(b,d) and d(a,d)+d(b,c). So, the third sum d_T(a,b)+d(c,d) can take any value inferior to 12 (corresponding to the triangle method), and superior to 4 if a metric is required. Then, the grafting of e provides a tree metric only for $2 \leq d_T(c,d) \leq 6$: d2 and d4 are the extreme placements of d compatible with this condition.

With a slight change on the given partial dissimilarity, say d(b,c) = 7.1, the triangle method extension becomes the unique possible one since, according to the four-point condition, one then obtains d_T(a,d)+d(b,c) = 12.1 and, so, d_T(c,d) = 10.1. Although a and d have no longer the same grafting point on the path T(yz), the obtained XLL-tree is very close to the previous one with the d1 placement for d . In that sense, the triangle method extension gives a particularly stable tree. Note also that, although the data constitute a partial metric, the unique possible extension is not a tree metric.

Definition 4.7. A valued 2d-tree G on X is said to be *resolved* if it leads, by applying the triangle method, to a resolved XLL-tree.

Theorem 4.8. *The tree function extension of a partial dissimilarity with a* 2*d-tree support graph G is unique if and only if G is resolved.*

Proof. If G is resolved, each of the n-2 latent vertices is placed in turn at an interior point of an edge of the current tree. There is no choice for this placement and the proof is easily

obtained by induction on n. For the converse, assume that, at some step of the triangle method, a new vertex $x_i = x$ is grafted on the path $T(yz)$ at a point u which is already a latent vertex. So, two edges xu and uv, both not belonging to the path $T(yz)$, are obtained, with respective lengths $\ell(xu')$ and $\ell(uu'')$. Determine a new tree T' by replacing xu and uv with three edges uu', $u'x$ and $u'v$, and give them lengths respectively equal to $\ell'(uu') = \varepsilon$, $\ell'(u'x) = \ell(u'x)-\varepsilon$, $\ell'(uu') = \ell(uu')-\varepsilon$, where ε is a small enough strictly positive constant. The metric in the obtained valued tree T' is an alternative extension of the values of d between all the pairs of predecessors of x in the considered RE order. Starting from T' to continue the triangle method process leads to an extension of d that differs from the triangle method one.❑

Example 4.9. Set X= { a, b, c, d, e} and consider the valued 2d-tree G (here, a 2-tree) in Figure 6. The triangle method gives the XLL-tree of Figure 7, with the vertex u of degree 4. With $\varepsilon = 1$, the operation described in the above proof provides the alternative tree T' of Figure 8, where all the lengths of edges of G are still preserved as path lengths. On the contrary, the valued 2d-tree of Example 4.5 is resolved. Consequently, it leads to a unique XLL-tree, which is depicted in Figure 3.

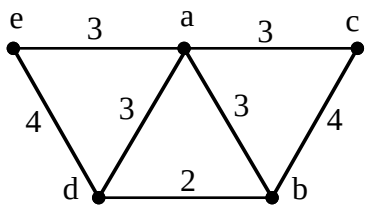


Figure 6

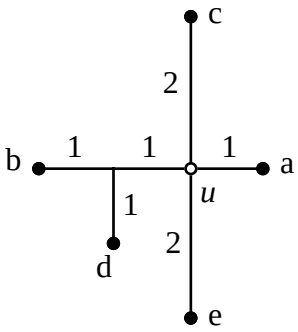


Figure 7

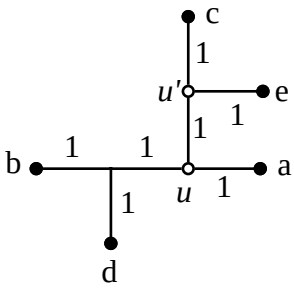


Figure 8

Given a partial dissimilarity d with a 2d-tree support graph G, one may use the triangle method (in $O(n^2)$) to determine the corresponding XLL-tree T. If T is resolved, then the extension of d is unique. This extended measure will meet the metric condition if and only if none of external edges in T has a strongly negative length. So:

Corollary 4.10. *Partial dissimilarities with resolved* 2*d-tree support graphs constitute a polynomial class of Problem* **MCA**.

Theorem 4.8 will be useful in Section 6 to find other polynomial classes. We end this section with three remarks:

Remark 4.11. As Example 4.6 shows, Problem **MCA** remains difficult in an unresolved 2d-tree. In that case, many tree function extensions are possible, and the

problem is to determine whether some of them are metric. Example 4.9 shows an easy case, where the triangle method extension is already metric.

Remark 4.12. For similar reasons, Corollary 4.10 cannot be extended to 2d-acyclic graphs. In this case again, many tree function extensions exist. Although it is always possible to extend a partial metric from a 2d-acyclic to a 2d-tree support graph (fix the length of a new edge as the minimum path length between its extremities whenever this quantity is defined), such an extension is not guaranteed to be a partial tree metric.

Remark 4.13. The case of resolved 2d-trees may be considered as the general one, since unresolved ones correspond to additional linear dependencies between the values of d. For instance, in Example 4.9, we have $2\ell(a,u) = d(a,b)+d(a,c)-d(b,c) = d(a,d)+d(a,e)-d(d,e)$.

5 Partial 2-trees and Problem MCA

Compared to 2d-trees, 2-trees provide, as support graphs, additional information about the corresponding valued *X*LL-tree (Leclerc and Makarenkov 1998). Especially, we have the following result:

Proposition 5.1. *Let d be a partial dissimilarity with a 2-tree support graph G, and d_T the tree function extending d obtained by the triangle method. Then, d_T is a tree metric if and only if d is a partial metric.*

Since 2-trees are chordal graphs, it is easy to verify that d, with a 2-tree support graph $G = (X, E)$, is a partial metric if and only if all the triangles of G are metric. Assume that G is just 2-acyclic. We then can add new pairs to E until a 2-tree is obtained. To obtain a tree metric extension by the triangle method, we have to give to each new edge xy a length preserving the property of being a partial metric. For that purpose, a simple solution consists of taking the minimum length $d^m(x,y)$ of a path of G between x and y as $d(x,y)$.

As recalled in Section 3.2, fast algorithms exist to recognize partial 2-trees, that are graphs without subgraphs homeomorphic to K_4 (that is, 2-cycles). Here we give a simple algorithm that combines this recognition with a 2-tree extension of a given partial metric d. The algorithm is based on the construction of a RE order, together with marking some new edges. Let x be a vertex with minimum degree $\partial(x)$ in the current graph:

- if $\partial(x) > 2$, the algorithm stops; G is not a partial 2-tree;
- if $\partial(x) = 1$, let y be the vertex adjacent to x and z a vertex, different from x, adjacent to y; a convenient length is assigned to the pair yz, which is marked, and x is eliminated together with the edge xy;

- if $\partial(x) = 2$, let y and z be the vertices adjacent to x; if the pair xy is not already an edge of the current graph, then this edge is marked and added, a convenient length is assigned to it, and x is eliminated together with the edges xy and xz.

Algorithm 5.2. *Completion-elimination procedure.*
While $|X| \geq 3$
  While there exists $x \in X$ such that $\partial(x) = 1$
    Select y such that $xy \in E$ and $z \neq x$ such that $yz \in E$
    Mark xz
    $d(x,z) := d^m(x,z)$
    $X := X \setminus \{x\}$ and $E := E \setminus \{xy\}$
  End While
  While there exists $x \in X$ such that $\partial(x) = 2$
    Select y and z such that $xy \in E$ and $xz \in E$
    If $yz \notin E$
      Mark yz
      $E := E \cup \{yz\}$
      $d(y,z) := d^m(y,z)$
    $X := X \setminus \{x\}$
    $E := E \setminus \{xy,xz\}$
  End While
End While

The *irreducible part* of G, notée Irr(G) is the graph obtained at the end of this procedure; either Irr(G) = K_3 or all its vertices have degree at least 3. Let E' be the set E augmented with all the marked pairs.

Theorem 5.3. *The above procedure extends a given partial metric with the support graph G to a partial metric with a 2-tree support graph if and only if* Irr(G) = K_3.

Proof. The completion-elimination procedure determines a RE order on X. We first show that the elimination of x cannot change the eventual 2-acyclicity of G. This is obvious when $\partial(x) = 1$. For $\partial(x) = 2$, when deleting x, a 2-cycle C of G including the edges xy and xz could be suppressed. The existence of such a 2-cycle implies another one C' obtained by substituting yz to these two edges; if $yz \in E$, then the 2-cycle C' exists in G and is not affected by the deletion of x; if $yz \notin E$, then C' is substituted to C before deleting x. In all cases, the new graph is 2-acyclic if and only if G is. So, if Irr(G) = K_3, all the successively considered graphs are 2-acyclic. Otherwise, Irr(G) has a 2d-cycle, as defined in Section 3.1 and, by Proposition 3.2, is not 2-acyclic.
Assume Irr(G) = K_3 and consider the graph $G' = (X,E')$: on this graph, the above valuation and elimination procedure consists of successive eliminations of a vertex belonging to a unique triangle, which is metric. So, G' is a metric 2-tree and the result follows. ❑

Corollary 5.4. *Partial metrics with* 2*-acyclic support graphs constitute a polynomial class of Problem* **MCA**.

Example 5.5. Consider the partial metric d of Figure 9 with a cycle support graph. While eliminating vertices a and c, edges be and bd are added with respective lengths 10 and 7. The resulting 2-tree corresponds to the *X*LL-tree of Figure 10, with null lengths for the edges adjacent to a and c, which gives a tree metric extension of d.

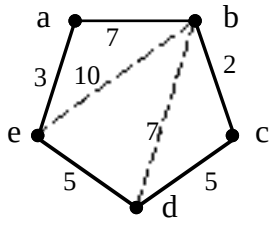


Figure 9

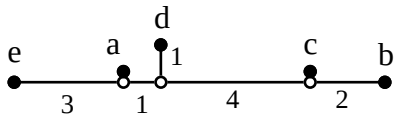


Figure 10

Remark 5.6. When the completion-elimination procedure leads to a complete graph Irr(G) with a vertex set Y of cardinality at least 4, it could be expected that we just have to determine whether $d_{|Y}$ is a tree metric. In fact, this is only the case if all of the edges of Irr(G) have not received their lengths during the procedure. Otherwise, many possible lengths were convenient, provided they are compatible with the triangle metric condition. So, it is not possible to give a general conclusion.

Example 5.7. Applying the completion-elimination procedure to the valued graph of Figure 11 leads to a K_4. In that example, no possible system of lengths on the new edges can give a tree metric.

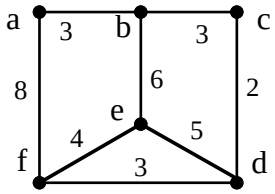


Figure 11

The elimination of vertices a and c implies addition of edges bf and bd. These operations lead to the complete graph on Y = { b, d, e, f}. One may have $5 \leq d(b,e) \leq 10$ and $1 \leq d(b,d) \leq 5$; so, $d(b,e) + d(d,f) = 9$, $10 \leq d(b,f) + d(d,e) \leq 15$ and $5 \leq d(b,d) + d(e,f) \leq 9$. The four-point condition cannot be satisfied.

In the example above, the answer to Problem **MCA** was obtained with a polynomial number of elementary operations. In a more general case, the graph Irr(G) is endowed with edge lengths, some of them comprised into intervals. One has then an instance of the "sandwich problem", also proved *NP*-complete by Farach et al. (1995).

6 Further polynomial cases

6.1. *Skew C_4's.* A C_4 of G is a cycle of length 4 (the usual term 4-cycle has another meaning in this paper). A C_4 $xyzw$ is said *skew* if $d(x,y) + d(z,w) \neq d(x,w) + d(y,z)$. Then, if, say, $xz \in E$, we have $d(y,w) = \max\{ d(x,y) + d(z,w) , d(x,w) + d(y,z) \} - d(x,y)$. If $xyzw$ does not admit a chord, we have the linear equation $d(y,w) + d(x,y) = \max\{ d(x,y) + d(z,w) , d(x,w) + d(y,z) \}$ with two variables. A graph G with enough skew 4-cycles leads in this way to a system of linear equations, whose resolution may give an answer to problems **MCA** and **WMCA** in a polynomial time.

Example 6.1. The support graph of Figure 12 has 9 edges; so, it remains 6 undetermined values for an extension of *d*.

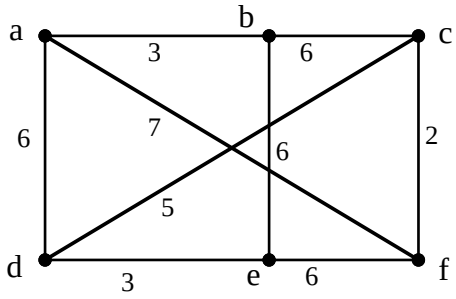


Figure 12

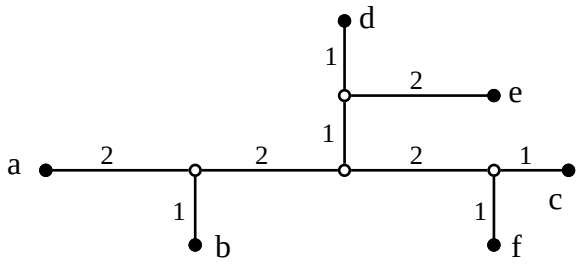


Figure 13

G has 9 skew C_4's leading to the following system of equations:

abcd	:	d(a,c)+d(b,d) = 12	abcf	:	d(a,c)+d(b,f) = 13
abed	:	d(a,e)+d(b,d) = 12	abef	:	d(a,e)+d(b,f) = 13
adef	:	d(a,e)+d(d,f) = 12	adcf	:	d(a,c)+d(d,f) = 12
bcde	:	d(b,d)+d(c,e) = 11	bcfe	:	d(b,f)+d(c,e) = 12
cdef	:	d(c,e)+d(d,f) = 11			

This system has the solution: d(a,c) = d(a,e) = 7, d(b,d) =d(d,f) = 5, d(b,f) = d(c,e) = 6. The corresponding tree metric is represented by the *X*LL-tree of Figure 13.

6.2. *G includes a* 2*d-tree*. 2d-acyclic graphs are sparse, because they have at most 2n-3 edges. When the number of edges increases, it may be expected that the support graph G admits a 2d-tree $H = (Y,F)$ as a subgraph (we call H a 2*d-subtree* of G). The triangle method then provides an *Y*LL-tree T and, if H is resolved, a unique tree function d_T extending the restriction $d_{|F}$. One may then, in a first step, compare the obtained values of d_T to the values of d on pairs of E not in F, but with extremities in Y. If the values are not identical, Problem **WMCA** (and so **MCA**) has a negative answer for the data. Otherwise, the positive answer to **WMCA** obtained for a part of the data can be also, according to Section 4.2, a negative one for **MCA**. In both problems, one may, with a positive answer on Y, try to extend the obtained solution to the remaining pairs of E, or to seek another 2d-tree in G.

This approach allows us to solve Problems **WMCA** and **MCA** in many cases. We present here a more formalized procedure that works as soon as G admits a subgraph H which is a resolved 2d-tree on X.

Definition 6.2. A *diamond* D of G is a quadruple $xyzw$ of elements of X such that $\{xy, yz, zw, wx, xz\} \subseteq E$ and $yw \notin E$ (so, D is a C_4 with a unique chord). The diamond D is *resolved* if the C_4 $xyzw$ is skew.

It was observed in Section 6.1 above that Condition (F) uniquely determines the value of the second chord of a resolved diamond. A step of the procedure is as follows:

- scan all the quadruples of elements of X. When a quadruple is a 4-clique of G, check whether it satisfies the four-point condition (F); if not, stop: d is not a tree function. When a quadruple $xyzw$ is a diamond, check whether it is resolved; if it is the case, set $d(y,w) = \max\{ d(x,y) + d(z,w) , d(x,w) + d(y,z) \}-d(x,y)$ and add the pair yw to E.

This step is iterated until either all the pairs have received a tree function value, or a 4-clique not satisfying Condition (F) is found, or no new pairs can be valued; in the last case, the problems remain undetermined.

Assume that G includes a resolved 2d-tree H and consider the elements of X in an RE order. The first four vertices of a resolved 2d-tree constitute a resolved diamond in H; at the next step, the fith vertex constitutes a resolved diamond with some triples in the previous vertices, and so on. According to Theorem 4.8, the extension is unique or leads to contradict Condition (F); in both cases, problems **WMCA** and **MCA** are solved. The scanning of quadruples is in $O(n^4)$-* and the number of iterations is bounded by the number of pairs not in E, that is $O(n^2)$. Finally, the procedure needs at worst $O(n^6)$ time.

Theorem 6.3. *Partial dissimilarities with support graphs including a resolved 2d-tree constitute polynomial instances of Problems **WMCA** and **MCA**.*

There are generally few missing values in phylogenetic applications. Therefore, an algorithm based on the above procedure generally provides an answer. In Guénoche and Leclerc (2000), the triangle method was applied, in a tree metric approximation purpose, to more than 500 partial metric tables corresponding to vertebrate homologous genes issued from the HOVERGEN database (Duret et al. 1994). Among them, the answer to **MCA** remained undetermined for only a dozen tables, with unconnected support graphs.

Example 6.4. Consider the valued graph in Figure 14; it has $n = 6$ vertices and $m = 10$ edges. The deletion of the edge ab provides a resolved 2d-tree (while the deletion of af gives an unresolved one). In the above procedure, a first scanning of quadruples gives the diamonds bacd (unresolved), cbde (d(b,e) = 7) and dcef (d(c,f) = 6). The edges be and cf are added to E and a second scanning is performed. From the diamonds cafd , cafe and abcf , we find, respectively, d(a,d) = 6, d(a,e) = 6 and d(b,f) = 10. then d does

not satisfy Condition (F) on the 4-clique bcdf . The conclusion is that this partial metric cannot be completed as a tree function.

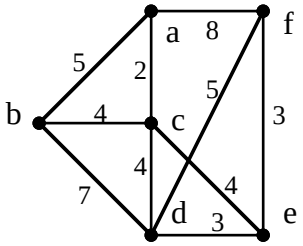


Figure 14

7. Conclusion

We described and established new properties of two classes of acyclic-like graphs (Section 3). Problems **WMCA** and **MCA** were proved (Sections 4 and 5) to be polynomial for, respectively, 2d-acyclic and 2-acyclic graphs. We also observed (Section 6) that, although **MCA** is NP-complete, most of the practical instances of this problem are polynomially resolvable.

Here is our last example, presenting a seemingly difficult instance. The Petersen graph of Figure 15 is a 2d-cycle and has no 2d-subtrees. It is endowed with a partial tree metric, except the value of one edge (in bold), increased by 1. It may be expected that such a partial metric no longer extends to a tree metric. How to prove (or disprove) that?

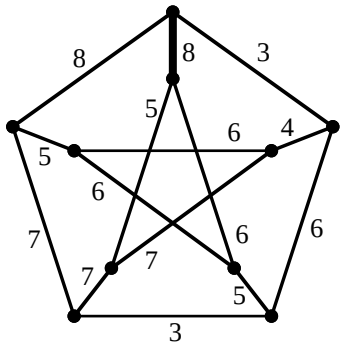


Figure 15

Références

M. Aigner (1979), *Combinatorial Theory*, Springer-Verlag, Berlin.

S. Arnborg, D.G. Corneil, A. Proskurowski (1987), Complexity of finding embeddings in a *k*-tree, *SIAM J. Alg. Disc. Meth.* 8, 277-284.

J.P. Barthélemy, A. Guénoche (1991), *Trees and Proximity Representations*, London, J. Wiley.

H.L. Bodlaender (1997), Treewidth: algorithmic techniques and results, in *Proceedings 22nd International Symposium on Mathematical Foundations of Computer Sciences, MFCS'97*, I. Privara, P. Ruzicka (eds), *Lecture Notes in Computer Sciences* 1295, Berlin, Springer-Verlag, pp. 19-36.

P. Buneman (1971), The recovery of trees from measures of dissimilarity *in Mathematics in Archaeological and Historical Sciences*, F.H. Hodson, D.G. Kendall, P. Tautu (Eds.), Edimburg, Edimburg University Press, pp. 387-395.

G.A. Dirac (1952), A property of 4-chromatic graphs and some remarks on critical graphs, *J. London Math. Soc.* 27, 85-92.

L. Duret, D. Mouchiroud, M. Gouy (1994), HOVERGEN: a database of homologous vertebrate genes, *Nucleic Acids Res.* 22, 2360-2365.

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

A. Guénoche, S. Grandcolas (1999), Approximation par arbre d'une distance partielle, *Mathématiques, Informatique et Sciences humaines* 146, 51-64.

A. Guénoche, B. Leclerc (2000), The triangle method to build phylogenetic trees from incomplete distance matrices, *RAIRO Operations Research*, to appear.

B. Leclerc (1995), Minimum spanning trees for tree metrics: abridgements and adjustments, *J. of Classification* 12 (1995) 207-241.

B. Leclerc, V. Makarenkov (1998), On some relations between 2-trees and tree metrics, *Discrete Math.* 192, 223-249.

P.C. Liu, R.C. Geldmacher (1980), An $O(\max(m,n))$ algorithm for finding a subgraph homeomorphic to K_4, in *Proceedings 11th Southeastern Conference on Combinatorics, Graph Theory and Computing*, pp. 597-609.

V. Makarenkov (1997), *Propriétés combinatoires des distances d'arbre : Algorithmes et applications*, Thèse de l'EHESS, Paris.

R.E. Pippert, L.W. Beineke (1969), Characterisation of 2-dimentional trees, in *The Many Facets of Graph Theory*, G. Chatrand, SF. Kapoor (Eds.), *Lecture Notes in Mathematics* 110, Berlin, Springer-Verlag, pp. 263-270.

A. Proskurowski (1984), Separating subgraphs in *k*-trees: cables and caterpillars, *Discrete Math.* 49, 275-295.

D.J. Rose (1974), On simple characterizations of *k*-trees, *Discrete Math.* 7, 317-322.

P. Todd (1989), A *k*-tree generalization that characterizes consistency of dimensioned engineering drawings, *SIAM J. Disc. Math.* 2 (2), 255-261.

A. Wald, C.J. Colbourn (1983), Steiner trees, partial 2-trees and minimum IFI networks, *Networks* 13, 159-167.

D.J.A. Welsh (1976) *Matroid Theory*, London, Academic Press.