Soft Comput (2005)
DOI 10.1007/s00500-005-0507-7

FOCUS

Anne Berry · Alain Sigayret · Christine Sinoquet

Maximal sub-triangulation in pre-processing phylogenetic data

Published online: 7 June 2005

Abstract In order to help infer an evolutionary tree (phylogeny) from experimental data, we propose a new method for pre-processing the corresponding dissimilarity matrix, which is related to the property that the distance matrix of a phylogeny (called an *additive matrix*) describes a sandwich family of *chordal graphs*. As experimental data often yield distance values which are known to be under-estimated, we address the issue of correcting the data by *increasing* the distances which are incorrect. This is done by computing, for each graph of the sandwich family, a maximal chordal subgraph.

1 Introduction

Inferring *evolutionary trees* (also called *phylogenies*) from dissimilarity data remains one of the major challenges in the field of computational biology. Problems like multiple sequence alignment, gene function prediction and protein structure prediction involve phylogenetic reconstruction.

Dissimilarity matrices are obtained experimentally by considering a set of 'objects', which in a phylogenetic context are *taxa*, (but which can, for other biological problems, be genes or proteins for example), and by measuring, using some feasible criterion, the distance between the elements of each pair of taxa. Data are thus described as positive-valued symmetric matrices; when this matrix does indeed correspond to a phylogeny (it is then called an *additive matrix*), reconstructing the tree is easy, and can be done in polynomial time, yielding a unique tree topology [2, 13, 23].

A. Berry (✉) · A. Sigayret
LIMOS (Laboratoire d'Informatique de Modélisation et d'Optimisation des Systèmes), CNRS UMR 6158, Université Clermont-Ferrand II, Ensemble scientifique des Cézeaux, 63177 Aubière Cedex, France
E-mail: berry@isima.fr, sigayret@isima.fr

C. Sinoquet
LINA (Laboratoire d'Informatique de Nantes-Atlantique), FRE CNRS 2729, BP 92208, 44322 Nantes Cedex 3, France
E-mail: christine.sinoquet@univ-mantes.fr

Experimental results, however, are not exactly additive matrices, and the phylogeny has to be inferred from real data. Many methods have been proposed for this (see e.g. [9, 13, 24, 26]), but they remain costly and inaccurate, so that new approaches are still being sought.

One of the recent trends in this field of research is to examine the *ordinal properties* of the matrix. Dissimilarity matrices describe a succession of thresholds, and it turns out to be interesting to examine the *structure* of these thresholds, rather than only the values themselves (see [21, 24, 26]), partly because they seem to be less sensitive to small data variations (see [11]), but also because biologists are more concerned with reconstructing the structure (topology) of the phylogenies than by finding the exact valuations of the edges of the phylogeny, so that it seems promising to look for structural properties.

Our approach here is to consider the family of undirected graphs defined by the dissimilarity matrix, each graph of the family corresponding to one of the thresholds of the matrix; we call this the *threshold family of graphs defined by a dissimilarity matrix*.

Huson, Nettles and Warnow in [24] use the property that if the matrix is additive, all the graphs of the threshold family are *chordal* (or *triangulated*), and give experimental evidence that the graphs obtained in real-world data are "almost triangulated".

As a means of pre-processing the experimental data, our aim here is to correct each graph of this threshold family so that it will indeed be chordal, a process called *triangulation*.

The most classical way of correcting a non-chordal graph is called *minimal triangulation*. It is well studied ([3, 4, 6, 7, 27]) and consists in adding an inclusion-minimal set of edges to the graph in order to make it chordal. For a given graph with n vertices and m edges, computing such a minimal triangulation can be done in $O(nm)$ time.

Adding edges to a graph of a threshold family means *lowering* the thresholds of the corresponding edges. Biologists, however, estimate that in the context of constructing a phylogeny from DNA sequences, the thresholds obtained by experimentations tend to be rather too low than too high, the

argument being that the number of mutations, represented by the distance between taxa, is in fact higher than what can be observed experimentally (see [26]).

We have thus examined the problem of correcting a graph which fails to be chordal by *removing* edges rather than adding them, thereby computing a *maximal chordal subgraph* or *maximal subtriangulation* of each of the graphs of the threshold family. The maximal subtriangulation problem has been somewhat less studied than minimal triangulation [1, 15–17, 30], but there exist several algorithms [1, 16, 30] which compute a maximal subtriangulation in $O(\Delta m)$ time, where Δ is the maximum degree in the graph.

Our contribution here is a greedy algorithm which will propose a corrected matrix which *increases* the value of any edge on which an anomaly is detected. Our process adds the edges one by one, beginning with an independent set and ending with a clique, in an order as compatible as possible with the input matrix, maintaining throughout a chordal graph at each edge-addition step. This process defines as a side-effect maximal subtriangulations of the graphs defined by the input matrix, but with a better complexity than if the chordal subgraphs were computed separately for each graph of the threshold family.

Our process relies on a new characterization of the edges which can be added to a chordal graph without loosing chordality, which in turn yields a new edge-composition scheme on edges which characterizes chordal graphs.

The paper is organized as follows: we give some previous results and definitions in Sect. 2; in Sect. 3, we present and study the threshold family of graphs defined by an additive matrix, and define an edge-addition construction scheme for the class of chordal graphs; Sect. 4 contains our proposed algorithm; in Sect. 5, we give some experimental insights then go on to conclude with some open questions and perspectives.

2 Preliminaries

We will first give some definitions and properties which will be useful in the rest of the paper.

2.1 Additive matrices

A **dissimilarity** on a finite set X is a function δ: $X^2 \to \mathbb{R}^+$ such that $\forall x, y \in X$, $\delta(x, y) = \delta(y, x)$. A dissimilarity is represented by a pairwise comparison symmetric matrix. A **distance** is a dissimilarity such that $\forall x, y \in X$, $\delta(x, y) = 0 \iff x = y$ and $\forall x, y, z \in X, \delta(x, y) + \delta(y, z) \geq \delta(x, z)$.

A **phylogeny** or **evolutionary tree** is an unrooted binary tree with all edges weighted with positive values. We will denote by $\mathcal{L}$ the set of leaves representing the set of taxa. Figure 1 gives a simple example of such a tree.

For $a, b \in \mathcal{L}$, we will denote by $d(a, b)$ the length of the ab-path from a to b in the phylogeny, which gives the evolutionary distance between a and b. This distance is called

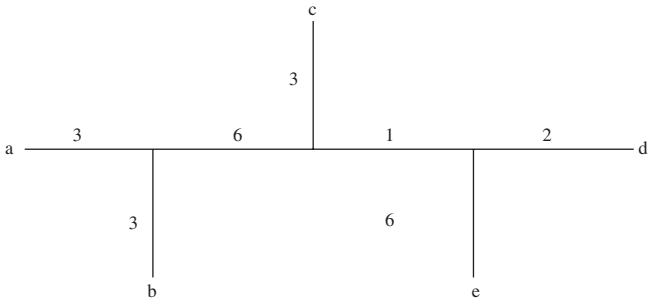

Fig. 1 A phylogeny T

an **additive distance** and the associated matrix on $\mathcal{L} \times \mathcal{L}$ is called an **additive matrix**. Note that an additive matrix is a special kind of dissimilarity matrix.

Additive matrices are well-studied and the following property, called the Quadrangular Inequality, characterizes them:

Characterization 2.1 [2] *A distance matrix M on a set of taxa is additive iff for any quadruple $\{a, b, c, d\}$ of taxa, from the 3 sums $d(a, b) + d(c, d)$, $d(a, c) + d(b, d)$ and $d(a, d) + d(b, c)$, the two largest are equal.*

The set of values of a dissimilarity matrix M can be ordered from 0 (as $M[x, x] = 0$) to the maximal value. This defines a number of different **thresholds**: $0, 1, \ldots, k$, in increasing order. An **ordinal matrix** of a dissimilarity matrix is thus defined as the matrix obtained by replacing each dissimilarity value by its threshold. We will denote by θ the function giving the threshold rank corresponding to a dissimilarity.

Example 2.2 *The phylogeny from Fig. 1 yields the following dissimilarity and ordinal matrices:*

M	a	b	c	d	e
a	0	6	12	12	16
b		0	12	12	16
c			0	6	10
d				0	8

The dissimilarity matrix M of T

W	a	b	c	d	e
a	0	1	4	4	5
b		0	4	4	5
c			0	1	3
d				0	2

The ordinal matrix W of T

Note that M is an additive matrix.

The corresponding 6 thresholds are the following: $\theta(0) = 0$, $\theta(6) = 1$, $\theta(8) = 2$, $\theta(10) = 3$, $\theta(12) = 4$, $\theta(16) = 5$. Dissimilarity values are: $\theta^{-1}(0) = 0, \theta^{-1}(1) = 6, \theta^{-1}(2) = 8$, $\theta^{-1}(3) = 10$, $\theta^{-1}(4) = 12$, $\theta^{-1}(5) = 16$.

2.2 Chordal graphs and triangulations

A graph $G = (V, E)$ is said to be **chordal** or **triangulated** if it contains no chordless cycle on more than 3 vertices. We will

need the following tree-oriented characterization for chordal graphs, due independently to Walter, Buneman and Gavril:

Characterization 2.3 [14,19,29] *A graph is chordal iff it is the intersection graph of a family of subtrees of a tree.*

Graph inclusion: If $G = (V, E)$ is a graph and $G' = (V, E')$ is another graph on the same vertex set, we will write $G \subseteq G'$ iff $E \subseteq E'$ and $G \subset G'$ iff $E \subset E'$ ($\subset$ denotes strict inclusion).

In [27], Rose, Tarjan and Lueker gave the following definition of minimal triangulation:

Definition 2.4 [27] *If $G = (V, E)$ is a non-chordal graph, a chordal graph $H = (V, E + F)$ is said to be a* minimal triangulation *of G if $\forall F' \subset F$, graph $(V, E + F')$ fails to be chordal.*

In the same paper, they proved that only one edge needs to be removed and the resulting graph tested:

Theorem 2.5 [27] *Let $G = (V, E)$ be a graph, let $H = (V, E + F)$ be a chordal graph; H is a minimal triangulation of G iff $\forall f \in F$, graph $(V, (E + (F \setminus \{f\})))$ fails to be chordal.*

This result relies on the following Lemma from the same paper, which ensures that, given two chordal graphs which are mutually inclusive, there is an ordering on the edges which need to be added to the smaller graph which will maintain chordality at each edge-addition step.

Lemma 2.6 [27] *Let $G_1 = (V, E_1)$ be a chordal graph, let $G_2 = (V, E_2)$ be a chordal graph such that $G_1 \subset G_2$. Then $\exists f \in E_2 \setminus E_1$ such that $G' = (V, E_2 \setminus \{f\})$ is chordal.*

We would like to point out that the property described by Lemma 2.6 is far from trivial; it fails to hold for hole-free graphs (graphs with no chordless cycle with length strictly greater than 4), as illustrated by the counterexample below, and it is not known whether it holds for weakly chordal graphs (graphs with no hole, and no hole in the complement).

Counterexample 2.7 *Graphs G_1 and G_2 in Fig. 2 are hole-free graphs. G_2 can be obtained from G_1 by adding edges ac and df, but $G_1 + \{ac\}$ and $G_1 + \{df\}$ are not hole-free graphs.*

Maximal subtriangulation was to the best of our knowledge introduced in 1983 by Erdös and Laskar ([17]) in view of removing a minimum number of edges in order to make a graph chordal. Maximal subtriangulation is defined in a fashion similar to minimal triangulation:

Definition 2.8 *Let $G = (V, E)$ be a non-chordal graph, let $H = (V, E \setminus F)$ be a chordal graph. We will say that H is a* maximal sub-triangulation *of G if $\forall F' \subset F$, $(V, (E \setminus F) + F')$ fails to be chordal.*

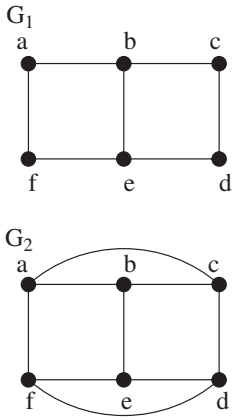

Fig. 2 Graphs of Example 2.7

3 Maintaining a family of chordal graphs

3.1 The threshold family of graphs defined by a dissimilarity matrix

We will use the ordinal matrix associated with a dissimilarity matrix to define the corresponding threshold family of graphs:

Definition 3.1 *Let $\mathcal{A}$ be a set of taxa with dissimilarity matrix M; let W be the corresponding ordinal matrix, on thresholds $0, 1, \ldots, k$. We will define a family of graphs $G_0 \subset G_1 \subset \ldots \subset G_k$, called the* ***threshold family of graphs*** *associated with W (and thus with M), with $G_i = (V, E_i)$, $V = \mathcal{A}$, and $ab \in E_i$ iff $W_{\mathcal{A}}[a, b] \leq i$.*

Remark 3.2 *The threshold family of graphs defined above should not be confused with threshold graphs, which are defined in correlation with integer programming; threshold graphs are characterized as being chordal, with a chordal complement, and P_4-free (see[20] and [12]), which is definitely not the general case for the graphs of a threshold family as defined by Definition 3.1.*

Note that G_0 is an independent set (a graph with no edges) and that G_k is a clique (a graph with all possible edges).

The threshold matrix W induces on the set of edges $V \times V$ a preorder relation $\mathcal{R}$: $ab\,\mathcal{R}\,cd$ iff $W[a, b] \leq W[c, d]$. $\mathcal{R}$ defines an *ordered partition of the edges* of G_k; each class F_i of edges is defined by $F_i = E_i - E_{i-1} = \{xy \mid W[x, y] = i\}$. Graph G_i is obtained from graph G_{i-1} by adding set of edges F_i. $\mathcal{R}$ defines a total ordering on these classes, with $F_i < F_j$ iff $i < j$.

Example 3.3 *Figure 3 illustrates the family of non-trivial graphs constructed from the matrix of Example 2.2. Ordered partition on the edges: $\{ab, cd\} < \{de\} < \{ce\} < \{ac, ad, bc, bd\} < \{ae, be\}$*

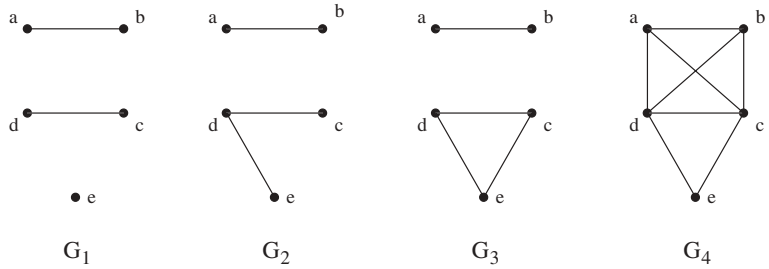

Fig. 3 Graphs $G_1 \subset G_2 \subset G_3 \subset G_4$ representing matrix of Example 2.2. G_0 is an independent set and G_5 is a clique

Property 3.4 *If M is an additive matrix then the threshold family of graphs defined by M is a family of chordal graphs.*

Proof Let T be the phylogeny associated with an additive matrix M, let G_i be the graph corresponding to threshold $\mathtt{i} \in [\mathtt{0..k}]$. It is easy to add extra internal nodes to T in order to obtain a tree T' where there is a node at mid-distance between any pair $\{a, b\}$ of vertices. Let us now consider the family of subtrees of T' defined by: for each leaf x, T'_x is the subtree containing all nodes at distance $\theta^{-1}(\mathtt{i})/2$ or less from x; G_i is the intersection graph of this family. By virtue of Characterization 2.3, G_i is chordal. □

The converse of Property 3.4 fails to be true: there are ordinal matrices which represent a chordal threshold family of graphs, but for which there is no corresponding additive matrix.

Counterexample 3.5 *An ordinal matrix W which can be associated with no dissimilarity matrix:*

W	a	b	c	d
a	0	3	2	1
b		0	4	4
c			0	1

W is associated with a threshold family of chordal graphs, but this matrix is the ordinal matrix of no additive matrix: for any dissimilarity δ of which W is the ordinal matrix, we have $\delta(a,c)+\delta(b,d) = \theta^{-1}(2)+\theta^{-1}(4)$, but $\delta(a,b)+\delta(c,d) = \theta^{-1}(3)+\theta^{-1}(1)$ and $\delta(a,d)+\delta(b,c) = \theta^{-1}(1)+\theta^{-1}(4)$, so $\delta(a,c)+\delta(b,d)$ is strictly greater then the other two sums, which contradicts the Quadrangular Inequality 2.1.

3.2 Preconditionning matrices

Experimental results show that not only do the dissimilarity matrices biologists have to work with fail to be additive, but the corresponding graphs very often fail to be chordal.

As we have stated in our introduction, our goal is to precondition a matrix into describing a family of chordal graphs, while dealing with threshold values which are too low. As a result of Counterexample 3.5, forcing a threshold family into a chordal family will not in general be sufficient to ensure that the matrix becomes additive; however, it may well be an important first step in error recovery.

Example 3.6 *Modified dissimilarity matrix of Example 2.2, with "incorrect" values for ad and bc, which have been lowered from 12 to 8. This describes a family (H_i) of graphs. H_2 and H_3 fail to be chordal; they are presented in Fig. 4.*

	a	b	c	d	e
a	0	6	12	8	16
b	/	0	8	12	16
c	/	/	0	6	10
d	/	/	/	0	8

3.3 An edge-addition composition scheme for chordal graphs

In order to compute a threshold family of graphs which are chordal and such that each graph G'_i of the new family is a subgraph of the corresponding original graph G_i, we will construct clique G_k from independent set G_0 by adding at each step an inclusion-maximal set of edges which maintains chordality.

The problem of maintaining a chordal graph while adding edges has been examined by Ibarra ([25]), who has obtained good results on the queries as to whether inserting or deleting an edge preserves a chordal graph. He uses clique trees (see [10] for a solid introduction of this concept) as an intermediate representation, and uses his work as an illustration of the power of this representation of a chordal graph. He derives a characterization of the edges which can be inserted or deleted, expressed in terms of clique trees.

We will propose yet a different approach, for which we will need the notion of 2-pair, defined by Hayward, Hoàng and Maffray to characterize weakly chordal graphs.

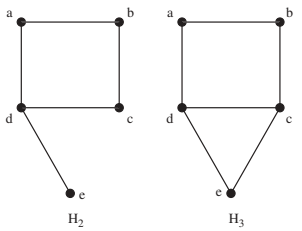

Fig. 4 Graphs H_2 and H_3 of Example 3.6

Definition 3.7 [22] *A pair $\{a, b\}$ of non-adjacent vertices is called a* 2-pair *iff every chordless path from a to b is of length exactly 2.*

Theorem 3.8 *Let G_1 be a chordal graph, let $\{a, b\}$ be a pair of non-adjacent vertices of G_1, let G_2 be the graph obtained from G_1 by adding edge ab; then G_2 is chordal iff $\{a, b\}$ is a 2-pair of G_1.*

Proof Let G_1 be a chordal graph, let $\{a, b\}$ be a pair of non-adjacent vertices of G_1, let G_2 be the graph obtained from G_1 by adding edge ab, let $\mu = ax_1x_2...x_kb$ be a longest chordless path from a to b in G_1. In G_2, $ax_1x_2...x_kba$ will be a chordless path on more than 3 vertices iff μ is of length greater than 2, that is iff $\{a, b\}$ fails to be a 2-pair of G_1. □

An additional property which is vital to our problem is ensuring that we are able to move from one graph of the threshold family to the next.

Property 3.9 *Let G_1 be a chordal graph, let G_2 be a chordal graph such that $G_1 \subset G_2$. Then G_2 can be obtained from G_1 by repeatedly adding an edge between the two vertices forming a 2-pair.*

Proof Let G_1 be a chordal graph, let G_2 be a chordal graph such that $G_1 \subset G_2$. By Lemma 2.6, $\exists ab \in E_2 \setminus E_1$ such that $(V, E_2 \setminus \{ab\})$ is chordal. By Theorem 3.8, $\{a, b\}$ is a 2-pair of $G_2 \setminus \{ab\}$. If we repeat this until we obtain graph G_1, we will have constructed (in reverse) a 2-pair edge addition ordering which enables us to construct G_2 from G_1. □

We use Theorem 3.8 to propose the following composition scheme for chordal graphs, which starts with an independent set and constructs the desired chordal graph by an edge-addition process. This is well-adapted to our problem, as we want to start with an independent set, scan a succession of mutually inclusive chordal graphs, and end with a clique, which is also chordal.

Composition Scheme 3.10 *A graph on n vertices is chordal iff it can be constructed by starting with an independent set on n vertices, and by adding at each step an edge between the two vertices forming a 2-pair.*

Remark 3.11 *Classical composition schemes for chordal graphs are vertex-addition schemes, such as starting with a clique and adding a simplicial vertex at each step. Very recent work by Berry, Heggernes and Villanger ([6]) gives a much more general process for adding a vertex v to a chordal graph, by characterizing the edges incident to v which must be added along with an edge vw in order to maintain chordality.*

4 An additive data pre-processing algorithm

4.1 Algorithmic strategy

We now propose an algorithm based on Composition Scheme 3.10, which uses as input a dissimilarity matrix M and outputs a dissimilarity matrix M' defining a threshold family of chordal graphs, and which raises all the thresholds it modifies.

Our algorithm starts with an independent set of vertices (graph G_0), and at each step i will construct graph G_i from graph G_{i-1} by adding as many edges as possible. The algorithm at step i repeatedly chooses, from a set of candidate pairs, a pair of vertices which allows to maintain a chordal graph.

At the beginning of step i, a candidate pair is defined as any pair $\{a, b\}$ such that $M[a, b] \leq \theta^{-1}(i)$ and ab is not an edge of G_{i-1}.

In order to remain as close as possible to the original matrix, we will give priority to the candidate pairs which correspond to the smallest threshold. We will implement this by using a FIFO queue; at each step, i, the new candidate pairs are added to the queue, and the algorithm then repeatedly chooses the first pair of the queue which is a 2-pair of the current graph, and adds it to the current graph G_i under construction.

By Property 3.9, at the end of the algorithm, the FIFO queue is empty and every edge has been given a threshold in the corrected matrix M' obtained.

4.2 Algorithm

Algorithm ADD-SUB-TRI
Input: A dissimilarity matrix M on n taxa, with threshold 0, ..., k.
Output: A dissimilarity matrix M', such that every graph in the threshold family is chordal.
Initialization:
G_0 is an independent set on n vertices;
Create an empty FIFO queue Q;
begin
For i = 1 **to** k−1 **do**
  $G_i \leftarrow G_{i-1}$;
  Compute the set F_i of pairs $\{a, b\}$
    such that $M[a, b] = \theta^{-1}(\mathtt{i})$;
  Add F_i to Q;
  Repeat
    Scan Q and remove the first pair ab
      which is a 2-pair;
    Add edge ab to graph G_i;
    $M'[a, b] \leftarrow \theta^{-1}(\mathtt{i})$;
    Until Q contains no 2-pair of G_i;
Give all remaining edges in Q value $\theta^{-1}(\mathtt{k})$ in M';
Add all remaining edges in Q to G_{k-1} to form G_k, a clique on n vertices.
end

Example 4.1 *On the "incorrect" matrix given in Example 3.6, at step* 2*, adding edge bc after adding edge ad would induce a 4-cycle abcd . We will add edge bc at step* 4*, after edge bd, thus raising the value of bc from* 8 *back to its "normal" value,* 12*. Note that edge ad has not been corrected.*

As a consequence of Composition Scheme 3.10, Algorithm ADD-SUB-TRI computes a threshold family of graphs

of which each member is a maximal sub-triangulation of the corresponding graph of the original matrix.

4.3 Complexity analysis

In [28], Spinrad and Sritharan propose an algorithm which repeatedly adds a 2-pair to the graph; they use a data structure which maintains the "2-pair structure" of the graph, which costs $O(n^2)$ to update for each edge addition. As there are $O(n^2)$ edges to process, using this 2-pair structure, our global complexity is thus $O(n^4)$.

Note that if we computed a maximal sub-triangulation in $O(\Delta m)$ time for each of the $O(n^2)$ graphs of the threshold family, this would cost $O(n^5)$.

5 Experimental results

We have implemented Algorithm ADD-SUB-TRI and run it on both experimental and artificial data. In a first step, we wanted to evaluate on experimental phylogenetic data how far this data is from data yielding a threshold family of chordal graphs. In a second step, since biologists do not have at their disposal exact phylogenetic data, we have run simulations in order to estimate the behaviour of Algorithm ADD-SUB-TRI. The resulting code is a C++ package enabling various options.

5.1 Weak/strong proximity of experimental threshold family with triangulated family

We first ran Algorithm ADD-SUB-TRI on real data to measure how "distant" a matrix obtained experimentally can be from a triangulated distance matrix.

We selected experimental data where thresholds are expected to be too low, with data sets ranging from being very close to phylogenetic, to others considered very unpure. We present results for five data matrices, ranging from size 11 to 57 (quite current sizes for phylogeny reconstruction) and dealing with plants and bacteria, with different so-called "noise" levels. Data sets 1, 2, 4 and 5 come from the databases of the French National Agronomical Research Institute (INRA). Data sets 1, 2, 4 and 5 respectively deal with data about fodder plants, sunflower species, wheat taxa and mildew races for the sunflower plant. Data sets 4 and 5 are known to be rather unpure. Data set 3 is a comparison between three genes of the TAT system (a system enabling folded protein transport through membranes) of several species of bacteria and is considered as having a high noise-level, and as such is analyzed extensively in [18] as having a questionable tree-structure.

Table 1 gives[1] the characteristics of the threshold families of graphs for these five matrices. These matrices present a percentage of triangulated graphs varying widely, from 4% to 96%. The families of graphs from experimental matrices 1 and 2 present a high percentage of triangulated graphs. As could be expected, data sets 4 and 5 conversely present a high proportion of non-triangulated graphs.

Table 2 gives results obtained by an execution of Algorithm ADD-SUB-TRI. These yield a criterion for evaluating the "distance" of a graph from a triangulated graph, by counting the percentage of edges which have been raised from some given threshold i to a higher threshold $i + j$. As the tree model was accepted for matrix 3 and is true for matrices 4 and 5, we can conclude that experimental data may be far from additive tree distances and even far from yielding a family of triangulated graphs.

5.2 Simulations

Next, we achieved a more thorough examination of how close the output of Algorithm ADD-SUB-TRI was to the original phylogeny. Since biologists do not have exact data on hand, we have run our algorithm on computer-generated additive matrices where some dissimilarity values were then artificially decreased.

Our experimental protocol is the following: we begin by randomly generating an additive matrix A; then we randomly generate from A a biased matrix B obtained by decreasing some of the dissimilarities in A; finally, we run Algorithm ADD-SUB-TRI on B, resulting in matrix T. We control the bias of B versus A with two parameters: the percentage of biased dissimilarities and the maximal amplitude of decrease for biased dissimilarities. Thus the average dissimilarity computed on B matrix differs from that computed on A, and so are we authorized to refer to the artificial modification as a **bias**.

Our aim is to compare couples (A, B) versus couples (A, T) to see whether matrix T is nearer to A than B is. To do this we use the metric and topological criteria described in [18]. Among these, we first checked that the arboricity criterion is improved by our method.

Arboricity criterion

The arboricity criterion is a measure of the tendance for any dissimilarity matrix D to be represented by an additive matrix:

$Arb(D) = \frac{1}{\binom{n}{4}} \mid \{x, y, z, t\}\ such\ that\ S_{\max} - S_{\text{med}} < S_{\text{med}} - S_{\min} \mid$ where $S_{\min}$, S_{med} and $S_{\max}$ respectively denote the three sums involved in the Characterization of additive matrices (see characterization 2.1), sorted in increasing order. In the ideal case (additive matrix M), S_{med} and $S_{\max}$ are equal, so Arb(M) scores its maximal value (1). For any dissimilarity matrix D, the more numerous quadruplets with S_{med} nearer to $S_{\max}$ than to $S_{\min}$ are the higher Arb(D) is.

Combining high, medium and low values for the two parameters **p** and **a** controlling the bias leads to the results in Table 3. Since our algorithm forces dissimilarity B towards a triangulated dissimilarity T (a necessary condition for an additive matrix) the improvement for arboricity was expected. Closely examining the improvement quality with

[1] The tables are presented at the end of the paper.

regard to bias parameters **p** and **a** shows the following: in the range 40–90% for parameter **p**, for constant value of **p**, the higher parameter **a** is, the more arboricity is "restored" by our algorithm.

Metric criteria

We systematically compute the following criteria:

- the average of deformations between two dissimilarity matrices D_1 and D_2:
  $\mathrm{Def}(D_1, D_2) = \frac{2}{n(n-1)} \Sigma \mid D_1(x, y) - D_2(x, y) \mid$,
- the Kruskal stress:
  $KS(D_1, D_2) = \sqrt{\frac{\Sigma[D_1(x,y)-D_2(x,y)]^2}{\Sigma {D_1}^2(x,y)}}$.

As expected in view of Garetta and Guénoche's results in [18], Def and KS are highly correlated (0.958) so we will only mention the behaviour of KS. Statistics on parameter KS in Table 4 lead to the conclusion of the efficiency of Algorithm ADD-SUB-TRI, with the strikingly similar tendance observed for arboricity criterion $\mathcal{ARB}$: again, in the range 40–90% for parameter **p**, the higher bias **a** is, the better T is.

Topological criterion

Characterization 888 for additive matrices states that for any quadruple $x,\ y,\ z,\ t$ in a phylogeny, from the three sums of two values which can be computed with the six available distances, the greater sums are equal. The smallest sum indicates the topology for quadruple $x,\ y,\ z,\ t$, that is whether x is associated with y or z or t on one side of an internal branch in the phylogenetic tree. Thus a topological criterion compiling the percentage of quadruplets sharing the same topology for dissimilarities D_1 and D_2 is computed as follows:

$wrq(D_1, D_2) = 1/\binom{n}{4} \mid \{x, y, z, t\}$ having same topology according to D_1 *and* D_2 $\mid$.

If D_1 is the underlying additive matrix, this topological criterion represents the percentage of well represented quadruplets (wrq). In this case, the higher wrq is (≤ 1), the more accurate D_2 is. The results obtained are presented in Tabel 5. In conclusion all three criteria indicate that Algorithm ADD-SUB-TRI seems a promising preprocessing method, a good first step in error recovery. However, its accuracy must be enforced so that it stays close enough to an aditive matrix.

6 Conclusion, perspectives and open questions

Regarding the complexity of Algorithm ADD-SUB-TRI as presented in this paper, we use data structures from [28], which deals with the problem of maintaining a two-pair structure in an arbitrary graph, whereas we deal with chordal graphs only. We believe that for chordal graphs this complexity should be improved, especially so since in a chordal graph there are many two-pairs which are not disrupted by an edge-addition, so that it may not be necessary to update the two-pair structure at every edge-addition step.

We also feel that in many biological data preprocessing problems, such as those dealing with biochip data, it may be interesting to maintain a chordal graph, but not necessarily by systematically lowering existing thresholds; it may be interesting to use the process described in [6] to allow the user to choose at each step whether to lower or to raise the thresholds, depending on how many modifications this causes.

References

1. Balas E (1986) A fast algorithm for finding an edge-maximal subgraph with a TR-formative coloring. Discrete Appl Math 15:123–134
2. Barthélémy J-P, Guénoche A (1991) Trees and proximity representations. Wiley (eds), New York
3. Berry A (1999) A wide-range efficient algorithm for minimal triangulation. In: Proceedings of tenth annual ACM-SIAM symposium on discrete algorithms (SODA'99), 860–861
4. Berry A, Blair J, Heggernes P (2002) Maximum cardinality search for computing minimal triangulations. In: Kucera L (ed) Graph theoretical concepts in computer science – WG 2002, LNCS 2573, Springer Berlin Heidelberg New York 1–12
5. Berry A, Bordat J-P, Heggernes P (2000) Recognizing weakly triangulated graphs by edge separability. Nordic J Comput 7:164–177
6. Berry A, Heggernes P, Villanger Y (2003) An on-line incremental approach for dynamically maintaining chordal graphs. Research Report LIMOS: RR 2003-04
7. Berry A, Bordat J-P, Heggernes P, Simonet G, Villanger Y (2003) A wide-range algorithm for minimal triangulation from an arbitrary ordering. Technical report reports in informatics 243, University of Bergen (Norway); Research Report LIMOS: RR 2003-02. J Algorithms (submitted)
8. Berry A, Sigayret A, Sinoquet C (2002) Towards improving phylogeny reconstruction with combinatorial-based constraints on an underlying family of graphs. Research Report LIMOS: RR 02-103
9. Berry V, Gascuel O (2000) Inferring evolutionary trees with strong combinatorial evidence. Theor Comput Sci 240 2:271–298
10. Blair JRS, Peyton B (1993) An introduction to chordal graphs and clique trees. Graph Theory Sparse Matrix Comput 56:1–29
11. Bonnot F, Guénoche A, Perrier X (1996) Properties of an order distance associated with a tree distance. In: Diday E et al (eds) Proceedings of OSDA'95 (Ordinal and Symbolic Data Analysis), Springer Berlin Heidelberg New York 252–261
12. Brandstädt A, Le VB, Spinrad J (1999) Graph classes – a survey. SIAM monographs on discrete mathematics and applications
13. Buneman P (1971) The recovery of trees from measures of dissimilarity. Mathematics in the archeological and historical sciences. Edinburgh University Press, 387–395
14. Buneman P (1974) A characterization of rigid circuit graphs. Discrete Math 9:205–212
15. Coleman TF (1988) A chordal preconditioner for large-scale optimization. Appl Math 40:265–287
16. Dearing PM, Shier DR, Warner DD (1988) Maximal chordal subgraphs. Discrete Appl Math 20:181–190
17. Erdös P, Laskar R (1983) On maximum chordal subgraph. Cong Numerantium 39:367–373
18. Garetta H, Guénoche A (2001) How confident can we be that a tree representation is good? (Quelle confiance accorder à une représentation arborée?). In: Gascuel O, Sagot M-F (eds) Proceedings of JOBIM 2000, LNCS, vol 2066 Springer Berlin Heidelberg, New York pp 45–56
19. Gàvril F (1974) The intersection graphs of subtrees of trees are exactly the chordal graphs. J Comb Theory B, 16:47–56
20. Golumbic MC (1980) Algorithmic graph theory and perfect graphs. Academic Press New York
21. Guénoche A (1998) Ordinal properties of tree distances. Discrete Appl Math 192:103–117
22. Hayward R, Hoàng C, Maffray F (1989) Optimizing weakly triangulated graphs. Graphs Comb 5:339–349
23. Hein J (1989) An optimal algorithm to reconstruct trees from additive distance data. Bull Math Biol 51(5):597–603

24. Huson D, Nettles S, Warnow T (1999) Obtaining highly accurate topology estimates of evolutionary trees from very short sequences. In: Proceedings of RECOMB'99, Lyon (France), 198–207
25. Ibarra L (2000) Fully dynamic algorithms for chordal graphs and split graphs. Technical report, University of Victoria DCS-262-IR
26. Kearney P, Hayward R, Meijer H (1997) Inferring evolutionary trees from ordinal data. In: Proceedings of eighth annual ACM-SIAM symposium on Discrete Algorithms (SODA'97) 418–426
27. Rose D, Tarjan RE, Lueker G (1976) Algorithmic aspects of vertex elimination on graphs. SIAM J Comput 5:146–160
28. Spinrad J, Sritharan R (1995) Algorithms for weakly triangulated graphs. Discrete Appl Math 59:181–191
29. Walter JR (1978) Representations of Chordal Graphs as Subtrees of a Tree. J Graph Theory 2:265–267
30. Xue J (1994) Edge-maximal triangulated subgraphs and heuristics for the maximum clique problem. Networks 24:109–120

Appendix

Tables for section 5

Table 1 Study of the threshold families of graphs associated with experimental dissimilarity matrices – All data sets are phylogenetic data. Matrices 1 and 2 have a low noise level; matrices 3, 4 and 5 present a high noise level

Data set	1	2	3	4	5
size of the dissimilarity matrix M (number of taxa)	17	41	15	11	57
size of the threshold family F of graphs associated with matrix M (number of thresholds)	82	81	88	54	97
number of triangulated graphs in F	79	71	45	17	4
percentage of triangulated graphs	96.3	87.7	51.1	31.5	4.1

Table 2 Influence of Algorithm ADD-SUB-TRI on experimental dissimilarity matrices (see Table 1)

Data set	1	2	3	4	5
number of cells for the upper triangular matrix induced from the symmetric dissimilarity matrix	136	820	105	55	1596
number of increased cells	4	49	18	11	889
percentage of increased cells	2.9	6.0	17.1	20.0	55.7

Table 3 Statistics on arboricity criterion improvement. The table compiles averages computed from 100 dissimilarity matrices B and T of size 20 (A: additive matrix, B: biased matrix computed from A, T: output matrix for *ADD – SUB – TRI* run on B) for $\mathcal{ARBi}(B,T) = Arb(T) - Arb(B)/1 - Arb(B)$ (percentage of arboricity "restored"). **p** is the percentage of cells which were decreased from matrix A to matrix B. **a** is the maximal decrease rate (computed from the maximal difference of cell values in matrix A)

		p								
		10	20	30	40	50	60	70	80	90
a	10	4.0	3.6	4.6	4.3	4.7	4.1	3.9	3.4	1.8
	30	6.8	11.0	13.8	14.8	15.7	15.0	13.3	13.5	11.6
	50	0.7	9.4	13.7	16.8	18.3	18.4	18.3	18.9	15.4

Table 4 Statistics on the improvement of the metric Kruskal stress criterion. The table compiles averages computed from 100 triples (A, B, T) (A: additive matrix (size 20), B: biased matrix computed from A, T: output matrix for *ADD – SUB – TRI* run on B) for $\mathcal{KSi}(A,B,T) = KS(A,B) - KS(A,T)/KS(A,B)$ (percentage of deformation corrected). Bias parameters **p** and **a** are described in Table 3

		p								
		10	20	30	40	50	60	70	80	90
a	10	3.4	4.0	4.1	4.3	4.0	3.8	3.3	3.1	1.7
	30	9.3	11.5	14.1	15.0	15.7	14.2	13.0	12.2	10.0
	50	7.3	10.2	12.9	15.1	17.4	17.3	17.6	17.5	15.0

Table 5 Statistics on the improvement of the topological criterion counting taxon quadruplets having the same topology as in the reference matrix A.The table compiles averages computed from 100 triples (A, B, T) (A: additive matrix (size 20), B: biased matrix computed from A, T: output matrix for *ADD – SUB – TRI* run on B) for $\mathcal{WRQi}(A,B,T) = wrq(A,T) - wrq(A,B)/1 - wrq(A,B)$ (percentage of common topology rate corrected). Bias parameters **p** and **a** are described in Table 3.

		p								
		10	20	30	40	50	60	70	80	90
a	10	3.7	2.3	5.6	3.2	3.0	3.2	3.1	3.5	1.8
	30	7.3	9.8	11.9	12.2	12.4	11.4	9.9	8.5	7.7
	50	0.9	1.6	1.7	4.0	6.2	4.3	4.1	5.6	2.2