BIOINFORMATICS

Vol. 20 no. 13 2004, pages 2113–2121
doi:10.1093/bioinformatics/bth211

A weighted least-squares approach for inferring phylogenies from incomplete distance matrices

Vladimir Makarenkov[1,2,] and François-Joseph Lapointe[3]*

[1]Département d'Informatique, Université du Québec à Montréal, C.P. 8888, Succ. Centre-Ville, Montréal, Québec, Canada H3C 3P8, [2]Institute of Control Sciences, 65 Profsoyuznaya, Moscow 117806, Russia and [3]Département de Sciences Biologiques, Université de Montréal, C.P. 6128, Succ. Centre-ville, Montréal, Québec, Canada H3C 3J7

Received on April 4, 2003; revised on February 20, 2004; accepted on February 20, 2004
Advance Access publication April 1, 2004

ABSTRACT

Motivation: The problem of phylogenetic inference from datasets including incomplete or uncertain entries is among the most relevant issues in systematic biology. In this paper, we propose a new method for reconstructing phylogenetic trees from partial distance matrices. The new method combines the usage of the four-point condition and the ultrametric inequality with a weighted least-squares approximation to solve the problem of missing entries. It can be applied to infer phylogenies from evolutionary data including some missing or uncertain information, for instance, when observed nucleotide or protein sequences contain gaps or missing entries.

Results: In a number of simulations involving incomplete datasets, the proposed method outperformed the well-known Ultrametric and Additive procedures. Generally, the new method also outperformed all the other competing approaches including Triangle and Fitch which is the most popular least-squares method for reconstructing phylogenies. We illustrate the usefulness of the introduced method by analyzing two well-known phylogenies derived from complete mammalian mtDNA sequences. Some interesting theoretical results concerning the NP-hardness of the ordinary and weighted least-squares fitting of a phylogenetic tree to a partial distance matrix are also established.

Availability: The T-Rex package including this method is freely available for download at http://www.info.uqam.ca/~makarenv/trex.html

Contact: makarenkov.vladimir@uqam.ca

INTRODUCTION

In systematic biology, incomplete datasets can arise in a variety of situations that can be caused by the lack of biological material, the imprecision of experimental methods or a combination of unpredictable factors. For one, molecular sequences of different genes may contain gaps or missing entries, which makes them hardly comparable and difficult to align. Moreover, experimental techniques like DNA-hybridization (Werman *et al.*, 1996), comparative serology (Maxson and Maxson, 1990) or microarray hybridization (Troyanskaya *et al.*, 2001) are limited in terms of pairwise comparisons and often led to incomplete distance matrices. Also, the combination of partially overlapping phylogenetic trees derived from different sources can produce incomplete matrices that are then used for the computation of supertrees (Bininda-Emonds *et al.*, 2002). While some maximum-likelihood and parsimony methods can handle missing information for estimating trees (Wiens, 1998), the reconstruction of phylogenies from distance matrices, usually requires complete matrices (Swofford *et al.*, 1996). Indeed, the most popular distance-based methods such as Neighbor-Joining (Saitou and Nei, 1987), BioNJ (Gascuel, 1997), or UPGMA (Michener and Sokal, 1957) cannot be carried out unless a complete matrix of pairwise distances among all species is available.

Different approaches have been proposed to solve the challenging problem of inferring phylogenies from partial distance matrices. Whereas indirect methods rely on the estimation of missing cells prior to phylogenetic reconstruction using the mathematical properties of ultrametrics or tree metrics, the direct approach allows to construct a phylogenetic tree directly from a partial distance matrix using a specific tree-building algorithm. Two procedures, reported in Tables 1 and 2, can be defined to estimate missing entries in a partial distance matrix **D** on a finite set X of taxa using either the ultrametric inequality [as was proposed by De Soete (1984) and then by Lapointe and Kirsch (1995)]:

$$d(i,j) \leq \text{Max}\{d(i,k); d(j,k)\}, \text{ for all } i, j \text{ and } k \text{ in } X, \quad (1)$$

*To whom correspondence should be addressed.

Table 1. Ultrametric procedure using the ultrametric inequality (1) to complete a partial distance d

Ultrametric procedure

Input: Partial distance d on the set X of n taxa.
Output: Complete or partial distance d on the set X of n taxa.
1. Count the Number_of_Missing_Entries in d.
2. **do**
  do for each pair of taxa ij, such that $d(i,j)$ is a missing entry of d
    MinMax is set to the maximum known entry of d
    do for each taxa k, such that $d(i,k)$ and $d(j,k)$ are known entries of d
      Max = $Max(d(i,k);d(j,k))$
      if (Max < MinMax) **then** MinMax=Max
    end do
    if there is at least one known pair of entries $d(i,k)$ and $d(j,k)$ **then**
      $d(i,j)$ = MinMax
      Number_of_Missing_Entries=Number_of_Missing_Entries −1
  end do
end do
3. **if** Number_of_Missing_Entries has changed **then go to** Point 2.

Table 2. Additive procedure using the four-point condition (2) to complete a partial distance d

Additive procedure

Input: Partial distance d on the set X of n taxa.
Output: Complete or partial distance d on the set X of n taxa.
1. Count the Number_of_Missing_Entries in d.
2. **do**
  do for each pair of taxa ij, such that $d(i,j)$ is a missing entry of d
    MinMax is set to the maximum known entry of d
    do for each pair of taxa k and l, such that $d(i,k)$, $d(j,k)$, $d(i,l)$, $d(j,l)$, and $d(k,l)$ are known entries
      Max = $Max(d(i,k)+d(j,l);d(i,l)+d(j,k))-d(k,l)$
      if (Max < MinMax) **then** MinMax=Max
    end do
    if, at least once, five entries $d(i,k)$, $d(j,k)$, $d(i,l)$, $d(j,l)$, and $d(k,l)$ are known **then**
      $d(i,j)$ = MinMax
      Number_of_Missing_Entries=Number_of_Missing_Entries −1
  end do
end do
3. **if** Number_of_Missing_Entries has changed **then go to** Point 2.

or the additive inequality, i.e. the four-point condition [as was proposed by Landry *et al.* (1996) and Landry and Lapointe (1997)]:

$$d(i,j)+d(k,l) \leq \text{Max}\{d(i,k)+d(j,l);d(i,l)+d(j,k)\}, \quad \text{for all } i,j,k \text{ and } l \text{ in } X. \tag{2}$$

For ultrametric distances (1), a missing value in a triangle is equal to the greatest of the two others if and only if they are different. However, if the two available distances are equal, one cannot estimate the missing value. Similarly for additive distances (2), the four-point condition proposes a value if and only if the two available sums are not equal.

On the other hand, three tree-building algorithms allowing missing cells in distance matrices have been proposed. The Triangle method (Guénoche and Leclerc, 2001; see also Guénoche and Grandcolas, 2000) relies on an iterative procedure having some interesting combinatorial properties, the Fitch program from the PHYLIP package (Felsenstein, 1997) uses a weighted least-squares optimization for the tree topology rearrangement and the MW method (Makarenkov and Leclerc, 1999) is also based on a weighted least-squares criterion.

The main objective of this paper is to present a new efficient method for inferring phylogenies from incomplete distance matrices. This method is compared with the Ultrametric (Table 1) and Additive (Table 2) estimation procedures as well as to the Fitch and Triangle direct reconstruction algorithms. Monte Carlo simulations have been carried out to assess the performance of the new method using two phylogenies derived from complete mammalian mtDNA sequences (see Cao *et al.*, 1998; Reyes *et al.*, 2000; Li *et al.*, 2001). Our results show that the method introduced in this paper, along with Fitch, are usually the most accurate for inferring phylogenies from incomplete distance matrices.

METHODS

Fitting a phylogenetic tree to a partial distance matrix

For the sake of mathematical convenience, the discussion in this section is conducted in terms of a dissimilarity. 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 $x, y \in X$.

Two computational problems are considered in this study. The first one is defined as follows: let **D** be a partial dissimilarity matrix on the set X of n taxa. The least-squares criterion consists in minimizing the following function:

$$Q = \sum_{i,j \in X} [d(i,j)-\delta(i,j)]^2, \tag{3}$$

where a tree metric $\delta(i,j)$ is an estimate of a known entry $d(i,j)$ in **D**.

In the next paragraph, we will show how the above-stated problem can be solved by defining the following computational problem: let **W** be a matrix of weights associated with a partial dissimilarity matrix **D** on the set X of n taxa. The weighted least-squares criterion consists in minimizing the following function:

$$Q_w = \sum_{i,j \in X} w(i,j)[d(i,j)-\delta(i,j)]^2, \tag{4}$$

where the sum is taken over all existing pairs of entries in **D**. We will prove that both optimization problems described by

Equations (3) and (4) are NP-hard. Therefore, the solutions of the optimization problems 3 and 4 are not 'likely' to be found in polynomial time. Efficient heuristic algorithms have to be developed to solve them.

Problems of phylogenetic inference are often stated as optimization problems. To prove their NP-hardness, one has to consider decision problems associated with them. An optimization problem is at least as hard as the associated decision problem and is usually harder.

Fitting additive trees to a partial dissimilarity (FAT_PD)

Instance: Partial dissimilarity matrix **D** on the set X of n taxa; non-negative integer k.

Question: Is there a tree metric δ such that:

$$\sum_{i,j\in X} [d(i,j) - \delta(i,j)]^2 \leq k, \qquad (5)$$

where the sum is taken over all existing pairs of entries in **D**.

Fitting additive trees to a partial dissimilarity with weights (FAT_PDW)

Instance: Partial dissimilarity matrix **D** on the set X of n taxa, matrix of weights **W** on X, non-negative integer k.

Question: Is there a tree metric δ such that:

$$\sum_{i,j\in X} w(i,j)[d(i,j) - \delta(i,j)]^2 \leq k, \qquad (6)$$

where the sum is taken over all existing pairs of entries in **D**.

The decision problem 5 is associated with the optimization problem 3, whereas the decision problem 6 is associated with the optimization problem 4 (for more details, see Barthélemy and Brucker, 2001; Day, 1987). In the seminal paper, Day (1987) defined the FAT decision problem, which is slightly different from FAT_PD: a complete dissimilarity matrix and a positive integer k were considered in FAT. It is easy to see that allowing k to take a 0 value does not complicate the FAT problem. When k is set to 0, the FAT problem is equivalent to the following question: is d a tree metric? This question can be answered in a polynomial time. However, in FAT_PD the case $k = 0$ is not trivial and should be considered.

THEOREM 1. *(Farach et al., 1995). The following decision problem (Matrix Completion to Additive, MCA) is NP-complete: given a partial dissimilarity d on a finite set X, is there a tree metric extending d to all pairs of elements of X?*

This theorem was first formulated but not proved, due to space limitation, in Farach *et al*. (1995). For the technical proof of Theorem 1 the reader is referred to Chepoi and Fichet (2000).

THEOREM 2. *The problem of an optimal least-squares fitting of a tree metric to a partial dissimilarity [Equation (3)] is NP-hard.*

Proof. First, we have to prove that the decision problem FAT_PD associated with the optimization problem 3 is contained in NP, i.e. any claimed solution can be verified in polynomial time. There exist polynomial time algorithms allowing one to check that a given dissimilarity d is a tree metric [see for instance an $O(n^2)$ algorithm by Makarenkov and Leclerc (1997)]. The correctness of Inequality 5 can be also verified in polynomial time for a given tree metric δ.

Second, we have to show that an NP-complete problem can be reduced to an instance of FAT_PD in polynomial time. Consider the MCA problem defined in Theorem 1. The MCA problem is equivalent to the following one: is there a tree metric δ, such that:

$$\sum_{i,j\in X} [d(i,j) - \delta(i,j)]^2 \leq 0, \qquad (7)$$

where the sum is taken over all pairs of known values $d(i,j)$. Indeed, the problem described by Equation (7) is an instance of FAT_PD. Thus, the decision problem FAT_PD is NP-complete and the associated optimization problem 3 is NP-hard.

THEOREM 3. *The problem of an optimal weighted least-squares fitting of a tree metric to a partial dissimilarity [Equation (4)] is NP-hard.*

Proof. The decision problem FAT_PDW associated with the optimization problem 4 is obviously in NP. Similar to the previous theorem, we can exhibit a polynomial time algorithm allowing one to check that a given dissimilarity is a tree metric and that Inequality 6 is satisfied.

Second, we have to show that an NP-complete problem can be reduced in polynomial time to an instance of FAT_PDW. According to Theorem 2, the problem FAT_PD is NP-complete. Indeed, FAT_PD is an instance of FAT_PDW in which all the values $w(i,j)$ of the weight matrix **W** are set to 1. Thus, the decision problem FAT_PDW is NP-complete and the associated optimization problem 4 is NP-hard.

Description of the new method

The method described in this paper takes advantages of the properties of indirect and direct approaches. It uses both the Ultrametric [Equation (1)] and the Additive [Equation (2)] estimation procedures, followed by a weighted least-squares fitting algorithm to infer phylogenetic trees from partial distance matrices. The new method, called MW*, is an extension of the Method of Weights (MW) introduced in Makarenkov and Leclerc (1999). MW was originally developed to build phylogenies from complete distance matrices using different weighting schemes. The first attempts to use it with

incomplete matrices were made by Levasseur *et al*. (2000, 2003). In these studies, binary weights were used to distinguish missing cells (weight of zero) from known entries (weight of one) in a partial distance matrix. The results of simulations showed, however, that MW was not always accurate in such instances. The MW* method is an attempt to improve on these results.

The new method proceeds in two main steps. In Step A, either the Ultrametric procedure (Table 1) or the Additive procedure (Table 2), is carried out to fill missing entries in a partial distance matrix. It is worth noting that the Ultrametric and the Additive procedures defined in Tables 1 and 2 do not always permit to obtain a complete distance matrix. For instance, they are unable to proceed when only one known value exists by row and by column of a given distance matrix. Moreover, when a complete distance matrix can be obtained by applying the Ultrametric or Additive procedure, the resulting distance is not necessarily an ultrametric or additive distance; in general, the returned distance does not verify the ultrametric inequality or the four-point condition. The smallest possible value used in the Additive and Ultrametric procedures (Tables 1 and 2) generally provides better results than the average or the greatest values; indeed the minimax option was proved to be efficient by Landry and Lapointe (1997) and Makarenkov (2002). In Step B, a stepwise addition procedure using a weighted least-squares criterion is carried out to complete the tree-building process.

The procedure, Ultrametric or Additive, to apply in Step A depends on the dimension of the given partial distance matrix and on the percentage of missing entries. The use of ultrametric estimates is recommended for small distance matrices with high percentages of missing distances. It is worth noting that the performances of the Additive and Ultrametric procedures significantly depend on the matrix dimension n (Makarenkov, 2001b). While increasing n, the performance of the four-point condition compared with the ultrametric inequality should also increase. This has been confirmed by a number of simulation studies including the current one (for details, see also Makarenkov, 2001b, 2002) as well as by the following theoretical considerations: for a given ratio of missing distances α, varying from 0 to 1, the mean number of estimates of a given unknown distance obtained using the ultrametric inequality is $(1-\alpha)^2 * (n-2)$, whereas the mean number of estimates obtained using the four-point condition is $(1-\alpha)^5 * (n-2)(n-3)/2$. Therefore, for a given missing ratio α, more estimates of a particular missing value are expected to be found using the four-point condition when $n > 3 + 2/(1-\alpha)^3$. For example, with 30% of missing distances, the four-point condition should be preferred when $n > 9$. However, as shown by Makarenkov (2001b), the following thresholds can be defined depending on the matrix size: the Additive procedure should be used with <20% of missing entries for matrices of size (8×8), <30% for matrices of size (16×16) and <40% for matrices of size (24×24). In cases where Ultrametric and Additive procedures yield similar results, one should use the latter because, in general, phylogenetic trees do not satisfy the molecular clock hypothesis characterizing ultrametric trees.

Let **D** be a symmetric partial distance matrix on the set X of n taxa. The MW* method proceeds as follows:

Step A. The Ultrametric procedure (Table 1) is applied to estimate missing cells in **D**, if the distance matrix has more than a pre-fixed percentage of missing entries (this percentage should be chosen depending on the matrix dimension, see the discussion above); otherwise the Additive procedure (Table 2) is applied. In the T-Rex program (Makarenkov, 2001a), the thresholds suggested in the previous paragraph are used to decide which technique, Ultrametric or Additive, will be carried out. Both procedures can stop when no more estimations can be made using Equations (1) or (2). A weight matrix**W** associated with **D** is then computed as follows:

$$w(i,j) = \begin{cases} 1, \text{ if } d(i,j) \text{ is a known distance in} \\ \quad \textbf{D} \text{ prior to Step A,} \\ 1/2, \text{ if } d(i,j) \text{ is an estimated distance in} \\ \quad \textbf{D} \text{ in Step A,} \\ 0, \text{ if } d(i,j) \text{ is missing in } \textbf{D} \text{ following Step A.} \end{cases} \quad (8)$$

The main reason for choosing the weight of 0.5 for an estimated entry $d(i,j)$ is based on simulations; the other possible weights which we tried were 0.25, 0.75 and p/N, where p is the number of the iteration in which the distance $d(i,j)$ was estimated and N is the total number of iterations in the Ultrametric or Additive procedures. Note that the Ultrametric and Additive procedures are not combined here, either the Ultrametric or Additive strategy is used at this step. The time complexity of one loop of the Ultrametric procedure is $O(n^3)$ and of the Additive procedure is $O(n^4)$. Thus, the time complexity of Step A is $O(kn^3)$ for the Ultrametric and $O(kn^4)$ for the Additive procedure, where k is a number of loops necessary for completing a partial distance matrix. In the simulation study described below, k never exceeded 5 for 34×34 distance matrices even with 50% of entries missing. In practice, the time complexity of Step A is $O(n^3)$ when the Ultrametric procedure is carried out and $O(n^4)$ when the Additive procedure is used.

Step B. The weighted least-squares criterion Q_w consists in minimizing the function defined by Equation (4), where the function δ is a tree metric associated with a phylogenetic tree T; thus, δ satisfies the four-point condition. Using the matrices **D** and **W** computed in Step A, the tree T can be obtained by applying a stepwise addition procedure:
Step 1 (Fig. 1a). The taxa i and j are chosen, such that $d(i,j)$ is a known distance in **D**. The corresponding tree T_2 comprises only the edge ij of length $d(i,j)$.

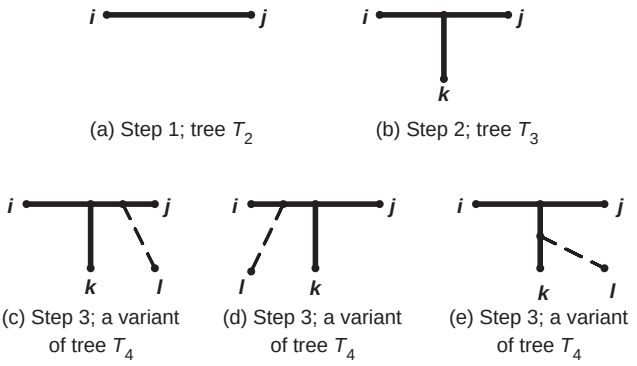

Fig. 1. The first three steps of the stepwise addition procedure used to infer a phylogenetic tree from a partial distance matrix.

Step 2 (Fig. 1b) and 3 (Fig. 1c, d and e). A third taxon k can now be placed into the tree. This taxon $k \in X - \{i, j\}$ is chosen to maximize the sum of weights $w(i,k) + w(j,k)$. If two or more taxa yield this maximum, the taxon k providing the smallest value of the weighted least-squares function Q_w is selected. This taxon is not always unique, however. If two or more taxa yield the minimum of the objective function Q_w, the one that has the greatest possible score over all taxa in X, defined by Equation (9), is selected for addition to T_2.

$$\text{Score}(k, X) = \sum_{l \in X} w(l,k). \tag{9}$$

A fourth taxon l is then placed into the tree at the Step 3. This taxon $l \in X - \{i, j, k\}$ is chosen to maximize the following sum $w(i,l) + w(j,l) + w(k,l)$. If two or more taxa yield the maximum of this sum, the taxon l providing the smallest value of the weighted least-squares criterion Q_w and, if necessary, the greatest possible score over all taxa in X, is selected.

Step p (with $p < n$). Let T_p be a phylogenetic tree with p leaves constructed at the previous steps. The leaves of T_p are associated with p taxa from X. Among the $n - p$ taxa in X that are not represented by the leaves of T_p, the next taxon $p + 1$ is selected to maximize the following score function:

$$\text{Score}(p+1, L) = \sum_{l \in L(T_p)} w(l, p+1), \tag{10}$$

where $L(T_p)$ is the set of leaves of the tree T_p. As in the previous steps, if two or more taxa yield the maximum score [Equation (10)], the taxon $p + 1$ providing the smallest value of the weighted least-squares criterion Q_w [Equation (4)] and, if necessary, the greatest possible score over all taxa in X [Equation (9)], is selected. Thus, the score function [Equation (10)] allows to select first the taxa whose values are the most certain. The exact location of the new leaf $p + 1$ in T_{p+1} and the lengths of the three new edges are found by a weighted least-squares procedure (see Makarenkov and Leclerc, 1999). The best grafting point of the new leaf on each edge in T_p, according to the objective function 4, is determined and, then, the location providing the overall minimum of the objective function over all edges of T_p is retained for grafting. When the location of the new leaf $p+1$ is not unique, the optimization procedure has to select one possible location from a set of possible ones. Note that the latter case cannot take place when the next leaf to be added to the tree T_p has no missing distances to the other leaves added to T_p in the previous steps.

The time complexity of the Step B is $O(n^3)$ for a given partial distance matrix **D** of size $(n \times n)$. Such a low-time complexity is due to a number of computational tricks used in the implementation of the MW procedure (for more details, see Makarenkov and Leclerc, 1999). In the simulations presented below, we carried out this procedure for all possible pairs of taxa ij selected at Step 1 of the algorithm; this exhaustive strategy increases the algorithmic time complexity up to $O(n^5)$, but it often enables a substantial improvement in fit. The only limitation for the new method is that the given distance matrix should not have rows or columns entirely filled with missing distances. As will be proven in the Results section, the use of weights makes the MW* procedure more efficient than a simple combination of Ultrametric or Additive techniques and MW method.

RESULTS

Application of MW* to whole genome phylogenies

Two phylogenies built from complete mammalian mtDNA sequences (see Cao *et al*., 1998; Reyes *et al*., 2000; Li *et al*., 2001) were analyzed to assess the relative performance of the new strategy relatively to four other methods for inferring phylogenetic trees from partial distance matrices. To do so, we used the information-based distances computed by Li *et al*. (2001) and available at www.math.uwaterloo.ca/~mli/distance.html. The first phylogeny (Fig. 2a) depicts evolutionary relationships among 20 species representing three main groups of placental mammals (Cao *et al*., 1999; Li *et al*., 2001). This tree inferred with the MW method (Makarenkov and Leclerc, 1999) is identical to the Neighbor-Joining (NJ) tree (Saitou and Nei, 1987) computed from the same data. The second phylogeny (Fig. 2b) illustrates evolutionary relationships among 34 taxa, including 19 species from Figure 2a and 15 additional taxa. In this case, the tree obtained with the MW method differs from the NJ tree and from the phylogenies derived with maximum-likelihood and minimum evolution methods (see Reyes *et al*., 2000). However, it is very similar to the consensus tree in Li *et al*. (2001), except for the position of the cat, dog and rabbit.

Monte Carlo simulations were carried out with the phylogenies in Figure 2. In a series of experiments, the accuracy of three direct and two indirect methods of phylogenetic inference was evaluated in presence of missing distances. The direct methods considered were Triangle (Guénoche and

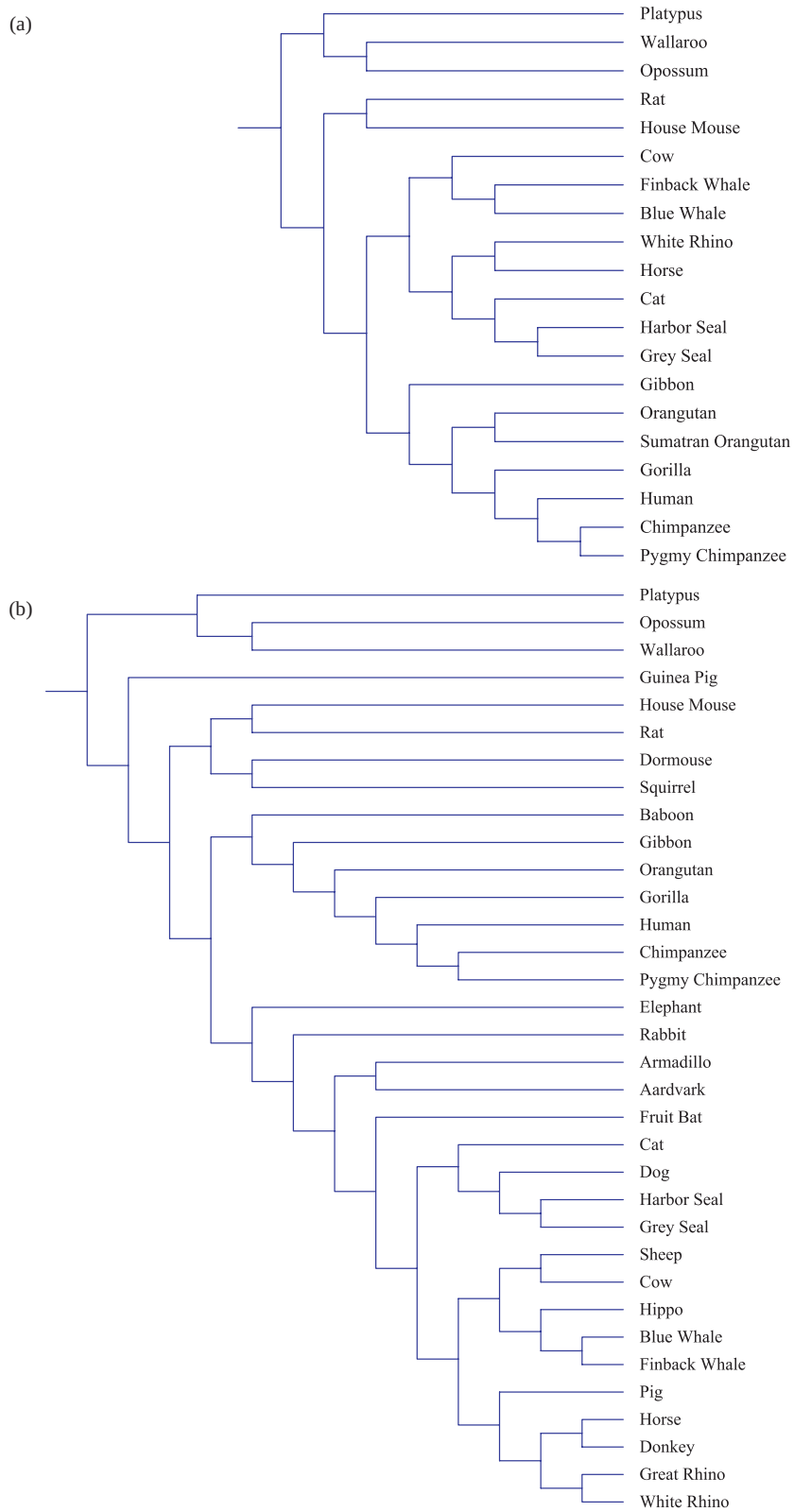

Fig. 2. Phylogenetic tree built from the complete mammalian mtDNA sequences of the species analyzed in (**a**) Cao *et al*. (1998) and (**b**) Reyes *et al*. (2000). These trees were inferred with the MW method of Makarenkov and Leclerc (1999), using the information-based distances from Li *et al*. (2001).

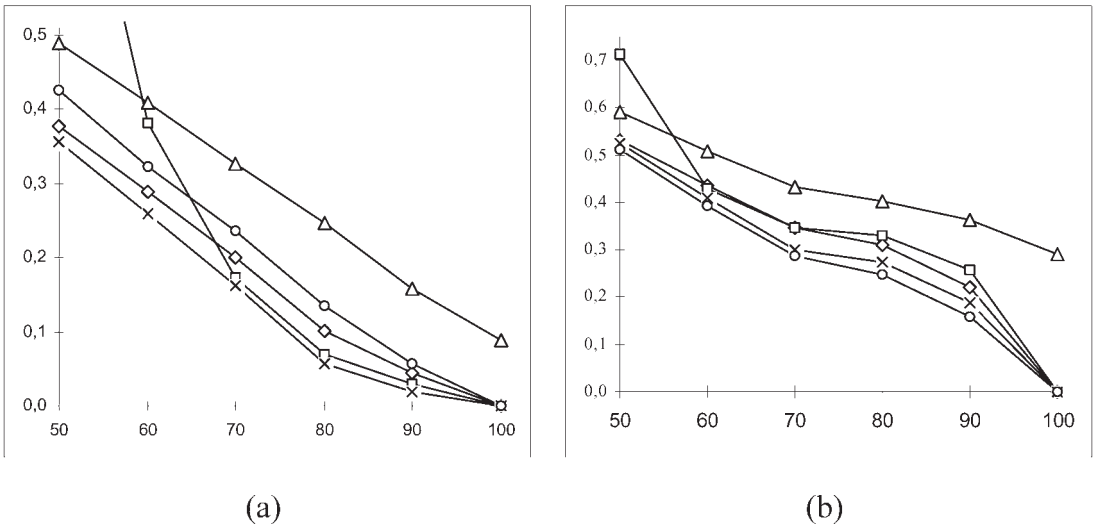

Fig. 3. (**a**, **b**) Topological recovery values obtained for different percentages of missing entries by the five competing methods (Triangle, open triangles; Ultrametric, open diamonds; Additive, open squares; Fitch, open circles; and MW*, multiplication symbol). The distance matrix computed by Li *et al.* (2001) were used in the simulations. The abscissa represents the percentage of known entries in the distance matrix; the ordinate represents the RF topological distance between the correct trees in Figure 2a and b, respectively, and the trees derived from partial distances using the above mentioned methods. Lower RF values indicate a better recovery of the correct tree.

Leclerc, 2001), Fitch (Felsenstein, 1997) and MW*. The Fitch program was used with the replicates option: the number of replicates was set to 0 for missing entries and to 1 for known ones. The indirect estimation approaches compared were the Ultrametric (De Soete, 1984) and Additive (Landry *et al*., 1996) procedures. Because the Ultrametric and Additive procedures do not always provide a tree metric, they were followed by the MW method with all weights set to 1. Moreover, as was discussed above, the Ultrametric and Additive procedures do not always allow one to obtain a complete distance matrix. In our simulation study, all considered partial distance matrices were generated in such a way that no missing values were left in them after the application of the Ultrametric or Additive procedure. Thus, the MW algorithm were always applied to a complete distance matrix.

Inferring phylogenies from incomplete distance matrices

The experiment involved random deletion of pre-fixed numbers of entries from **D**. For each case ($n = 20$ and $n = 34$), 100 replicates of a partial distance matrix were generated with different percentages of missing values ranging from 0 to 50%, and phylogenies were estimated from the partial distances matrices. The topological recovery of the five methods was then quantified with the Robinson and Foulds (RF) distance (see Robinson and Foulds, 1981; Makarenkov and Leclerc, 2000) computed between the correct trees (Fig. 2a and b) and the trees obtained from partial distances. This measure of tree similarity is equal to the minimum number of elementary operations, consisting of merging and splitting vertices, necessary to transform one tree into another. Lower RF values indicate a better recovery of the correct tree; the RF distance equals 0 when the recovery is perfect. To compare the results obtained for trees of different sizes, the computed RF distances were normalized by its largest possible value, which is $2n - 6$ for two binary trees with n leaves. The simulation results obtained for the two complete genome phylogenies (Fig. 2a and b) are presented in Figure 3. In each case, the mean RF values computed over 100 replicates are reported.

The results of the simulations show that the MW* method provides better results in terms of topological recovery for the (20×20) distance matrix compared with the four other algorithms (Fig. 3a). For the (20×20) matrix, the new method is particularly good when partial distance matrices contain 50–90% of known entries. However, when analyzing the (34×34) distance matrix the Fitch algorithm slightly outperforms MW* (Fig. 3b). Because of a large number of the tree topologies being examined by Fitch, it provides better results for large distance matrices but remains the slowest of the five competing strategies. However, when analyzing the (20×20) distance matrix, the Fitch procedure is among the worst ones, sitting behind MW*, Additive and Ultrametric. For the (20×20) distance matrix, the second best approach is the indirect estimation based on the Additive procedure, but its performance decreases rapidly with the increase in the percentage of missing cells. Thus, with $>30\%$ of missing entries for the (20×20) matrix and $>40\%$ of missing entries for the (34×34) matrix, the Ultrametric procedure outperforms the Additive procedure. The worst of the five methods compared is clearly Triangle, which never recovered

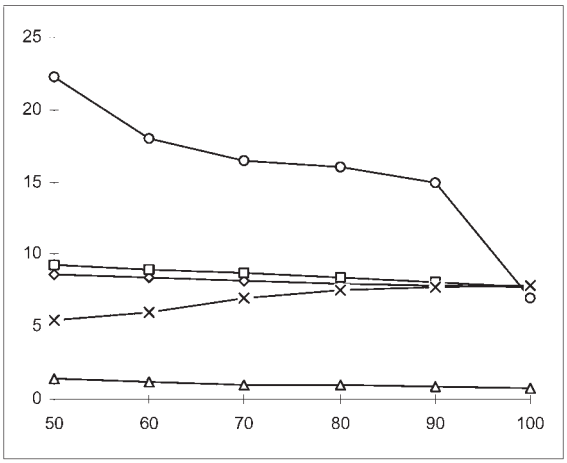

Fig. 4. Average computational time, for 100 datasets, required by the five tree inferring methods (Triangle, open triangles; Ultrametric, open diamonds; Additive, open squares; Fitch, open circles; and MW*, multiplication symbols) to process (34 × 34) partial distance matrices with different percentages of missing entries. The partial distance matrices were generated by removing entries from the original complete distance matrix provided by Li *et al.* (2001). The abscissa represents the percentage of known entries in the distance matrix; the ordinate represents the computational time, in seconds, taken to infer the tree.

the correct trees, even for complete distance matrices (Fig. 3a and b).

Figure 4 illustrates the average computational time required by the five tree inferring methods considered in this study to process (34 × 34) datasets with different percentages of missing entries. The experiments were carried out using a Ciara computer equipped with Intel Pentium IV (CPU 1.6 GHz) processor. For the (20 × 20) distance matrices considered in this study, the computing time of the five method was not graphed because it was similar for the majority of methods: the Triangle method took <0.2 s on average, Fitch (version 3.573c used with global optimization option) 1.8 s, and MW* and Ultrametric and Additive procedures followed by MW* ~2 s. The difference appears when analyzing (34×34) matrices. We can conclude that the Fitch computing time highly depends on the percentage of missing values. This time increases, on average, from 7.1 s for a complete (34 × 34) distance matrix to 22.3 s for a partial distance matrix with 50% of missing entries. Fitch is slightly faster than MW* for small datasets, but it is much slower than MW* for big data matrices, especially with big percentage of missing entries. It seems that the global optimization in Fitch works better for big distance matrices. The Triangle method is very fast but not very reliable. As explained above, in the simulations presented in this paper we carried out MW* for all possible pairs of taxa *ij* selected at the first step of the algorithm; this exhaustive strategy increases the algorithmic time complexity up to $O(n^5)$. MW* works faster when the percentage of missing entries increases. This paradox is due to the fact that at its first step MW* considers only pairs of taxa *ij* such that $d(i, j)$ is a known entry of the given partial distance matrix **D**; while increasing the percentage of missing entries, the number of taxa available at the first step of the algorithm (and so the number of iterations of the algorithm) decreases. The Additive and Ultrametric procedures (Tables 1 and 2) are very fast, but because they were followed by MW their computing time was very close to that of MW*.

CONCLUSIONS

We compared the relative performances of five different methods of phylogenetic inference intended to deal with incomplete distance matrices. We proved that the problem of fitting a phylogenetic tree to a partial distance matrix is NP-hard for both the ordinary least-squares and weighted least-squares models. Thus, to solve these problems new efficient heuristics should be proposed and tested through simulations. We described the MW* method which is based on a combination of indirect and direct tree reconstruction approaches. Simulation studies showed that the new method, along with Fitch, provides the best tree recovery among the competing approaches. As such, this technique can be useful to derive phylogenies from distance matrices including incomplete or uncertain entries; this problem is among the most relevant issues in systematic biology.

Two more applications of the MW* method can also be considered. First, MW* can be used to combine trees bearing overlapping sets of leaves in a supertree setting. Indeed, a supertree defined for all species can be obtained by combining the submatrices representing the partially overlapping subtrees. All the five methods considered in this paper can be useful for building supertrees from the gene distance data that are not affected by horizontal gene transfer events. However, the analysis of the supertree problem necessitates simulations involving genes with different rates of evolution. Further experiments will be needed to compare MW* with other methods of supertree construction (see Bininda-Emonds *et al.*, 2002). Second, the weighted least-squares approach employed in this paper can also be applied to phylogenetic inference from complete distance matrices. For instance, the weight matrix can be used to indicate either the precision of distance measurements or the confidence levels of distance values. Future simulations will be necessary to compare the performance of MW* with other distance methods of phylogenetic inference [e.g. NJ: Saitou and Nei (1987); BioNJ: Gascuel (1997); Fitch: Felsenstein (1997)] in case of sequence data with different percentages of gaps and missing bases.

ACKNOWLEDGEMENTS

The authors are grateful to the software developer Philippe Casgrain for his contribution to programming the T-Rex package. The authors are also thankful to Drs Jean-Pierre

Barthélemy, Bernard Fichet, Olivier Gascuel and Bruno Leclerc as well as to two anonymous referees for their helpful comments. This research was supported by the Natural Sciences and Engineering Research Council of Canada research grants to V.M. (OGP 249644) and to F.-J.L. (OGP 155251).

REFERENCES

Barthélemy,J.P. and Brucker,F. (2001) NP-hard approximations problems in overlapping clustering. *J. Classif.*, **18**, 159–183.

Bininda-Emonds,O.R.P., Gittleman,J.L. and Steel,M.A. (2002) The (super)tree of life: procedures, problems and prospects. *Ann. Rev. Ecol. Syst.*, **33**, 265–289.

Cao,Y., Janke,A., Waddell,P.J., Westerman,M., Takenaka,O., Murata,S., Okada,N., Pääbo,S. and Hasegawa,M. (1998) Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. *J. Mol. Evol.*, **47**, 307–322.

Chepoi,V. and Fichet,B. (2000) L-infinite approximation via subdominants. *J. Math. Psychol.*, **44**, 600–616.

Day,W.H.E. (1987) Computational complexity of inferring phylogenies from dissimilarity matrices. *Bull. Math. Biol.*, **49**, 461–467.

De Soete,G. (1984) Additive-tree representations of incomplete dissimilarity data. *Qual. Quant.*, **18**, 387–393.

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

Felsenstein,J. (1997) An alternating least-squares approach to inferring phylogenies from pairwise distances. *Syst. Zool.*, **46**, 101–111.

Gascuel,O. (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.*, **14**, 685–695.

Guénoche,A. and Grandcolas,S. (2000) Estimating missing values in tree distances, In Kiers,H.A.L *et al*. (eds), *Data Analysis, Classification and Related Methods*, In *Proceedings of the IFCS'2000*. Springer, Berlin, pp. 143–148.

Guénoche,A. and Leclerc,B. (2001) The triangle method to build X-trees from incomplete distance matrices. *RAIRO Oper. Res.*, **35**, 283–300.

Landry,P.A., Lapointe,F.-J. and Kirsch,J.A.W. (1996) Estimating phylogenies from distance matrices: additive is superior to ultrametric estimation. *Mol. Biol. Evol.*, **13**, 818–823.

Landry,P.-A. and Lapointe,F.-J. (1997) Estimation of missing distances in path-length matrices: problems and solutions. In Mirkin,B., McMorris,F.R., Roberts,F.S. and Rzhetsky,A. (eds), *Mathematical Hierarchies and Biology*. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. American Mathematical Society, Providence, pp. 209–224.

Lapointe,F.-J. and Kirsch,J.A.W. (1995) Estimating phylogenies from lacunose distance matrices, with special reference to DNA hybridization data. *Mol. Biol. Evol.*, **12**, 266–284.

Levasseur,C., Landry,P.A. and Lapointe,F.-J. (2000) Estimating trees from incomplete distance matrices: a comparison of two methods. In Kiers,H.A.L., Rasson,J.-P., Groenen,P.J.F. and Schader,M. (eds), *Data Analysis, Classification and Related Methods*. Springer-Verlag, Berlin, pp. 149–154.

Levasseur,C., Landry,P.A., Makarenkov,V, Kirsch,J.A.W. and Lapointe,F.-J. (2003) Incomplete distance matrices, supertrees and bat phylogeny. *Mol. Phylogenet. Evol.*, **27**, 239–246.

Li,M., Badger,J.H., Xin,C., Kwong,S., Kearney,P. and Zhang,H. (2001) An information based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, **17**, 149–154.

Makarenkov,V. and Leclerc,B. (1997), Circular orders of tree metrics, and their uses for the reconstruction and fitting of phylogenetic trees. In Mirkin,B., McMorris,F.R., Roberts,F. and Rzhetsky,A. (eds), *Mathematical Hierarchies and Biology*, *DIMACS—Series in Discrete Mathematics and Theoretical Computer Science*. American Mathematical Society, Providence, RI, pp. 183–208.

Makarenkov,V. and Leclerc,B. (1999) An algorithm for the fitting of a tree metric according to a weighted least-squares criterion. *J. Classif.*, **16**, 3–26.

Makarenkov,V. and Leclerc,B. (2000) Comparison of additive trees using circular orders. *J. Comput. Biol.*, **7**, 731–744.

Makarenkov,V. (2001a) T-Rex: reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics*, **17**, 664–668.

Makarenkov,V. (2001b) Une nouvelle méthode efficace pour la reconstruction des arbres additifs à partir des matrices de distances incomplètes. In *Proceedings of the 8-ièmes Rencontres de la Société Francophone de Classification*, Université de Antilles-Guyane, Pointe-à-Pitre, Guadeloupe, pp. 238–244.

Makarenkov,V. (2002) Comparison of four methods for inferring phylogenetic trees from incomplete dissimilarity matrices. In Jajuga,K. Sokolowski,A. and Bock,H.-H. (eds), *Classification, Clustering, and Data Analysis*. Springer-Verlag, Berlin, pp. 371–378.

Maxson,L.R. and Maxson,R.D. (1990) Proteins II: immunological techniques. In Hillis,D.H. and Moritz,C. (eds), *Molecular Systematics*. Sinaeur Associates, Sunderland, MA, pp. 127–155.

Michener,C.D. and Sokal,R.R. (1957) A quantitative approach to a problem in classification. *Evolution*, **11**, 130–162.

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.

Reyes,A., Gissi,C., Pesole,G., Catzeflis,F.M. and Saccone,C. (2000) Where do rodents fit? Evidence from the complete mitochondrial genome of *Sciurus vulgaris*. *Mol. Biol. Evol.*, **17**, 979–983.

Robinson,D.F. and Foulds,L.R. (1981) Comparison of phylogenetic trees. *Math. Biosci.*, **53**, 131–147.

Swofford,D.L., Olsen,G.J., Waddell,P.J. and Hillis,D.M. (1996) Phylogenetic inference. In Hillis,D.M., Moritz,C. and Mable,B.K. (eds), *Molecular Systematics*, 2nd edn. Sinaeur Associates, Sunderland, MA, pp. 407–514.

Troyanskaya,O., Cantor,M., Sherlock,G., Brown,P., Hastie,T., Tibshirani,R., Botstein,D. and Altman,R.B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Werman,S.D., Springer,M.S. and Britten, R.J. (1996) Nucleic Acids I: DNA–DNA Hybridization. In Hillis,D.M., Moritz,C. and Mable,B.K. (eds), *Molecular Systematics*, 2nd edn. Sinaeur Associates, Sunderland, MA, pp. 169–203.

Wiens,J.J. (1998). Does adding characters with missing data increase or decrease phylogenetic accuracy? *Syst. Biol.*, **47**, 625–640.