Une nouvelle méthode efficace pour la reconstruction des arbres additifs à partir des matrices de distances incomplètes

Vladimir Makarenkov

Département d'Informatique, Université du Québec à Montréal, C.P. 8888, succ. Centre-ville, Montréal, Québec H3C 3P8, Canada et Institute of Contol Sciences, 65 Profsoyuznaya, Moscou 117806, Russie.
e-mail : makarenkov.vladimir@uqam.ca

Résumé.

Des jeux de données incomplètes peuvent apparaître dans plusieurs situations pratiques. Par exemple, ils sont très courants en biologie moléculaire et, plus précisément, en phylogénétique. Cependant, une vaste majorité des techniques de reconstruction des arbres additifs (ou phylogénétiques) ne peuvent pas être exécutées sans qu'une matrice de distances complète entre les objets (ou espèces) observés soit disponible.

D'autre côté, le problème de la reconstruction des arbres additifs à partir des matrices de distances incomplètes est connu comme un problème très délicat. Dans cet article, nous introduisons une nouvelle méthode de reconstruction des arbres additifs qui permet de traiter des matrices de distances partielles. La nouvelle méthode, qui est basée sur une technique d'approximation par les moindres carrés, permet d'obtenir des meilleures performances que des méthodes existantes largement employées, qui sont souvent basées sur les propriétés de la condition ultramétrique ou de la condition des quatre points. La méthode proposée est inspirée par l'algorithme *MW* de Makarenkov et Leclerc [MAK 99]. Une fonction supplémentaire, nommée fonction de score, est utilisée pour prendre en compte des valeurs manquantes présentes dans la matrice de distances. La nouvelle méthode décrite dans cette article fait partie du logiciel T-Rex disponible pour téléchargement sur : http://biol10.biol.umontreal.ca/makarenv/.

MOTS-CLES : arbre additif, arbre phylogénétique, distance d'arbre, matrice de distances incomplète

1. Introduction

La présence des entrées manquantes dans une matrice de distances est un phénomène courant lorsque des jeux de données réels sont considérés. Les données incomplètes sont souvent présentes en biologie évolutive où la tâche est de retrouver un arbre d'évolution permettant d'établir des relations de proximités entre les espèces observées. Il existe plusieurs méthodes efficaces pour la reconstruction d'un arbre ultramétrique ou additif à partir des matrices de distances complètes, comme par exemple dans le cas additif, la méthode *ADDTREE* de Sattath et Tversky [SAT 77], *Neighbor-joining* de Saitou et Nei [SAI 87] ou *Unweighted neighbor-joining* de Gascuel [GAS 97]. Cependant, le problème de la reconstruction des arbres ultramétriques et additifs à partir des matrices de distances contenant des valeurs manquantes, bien qu'il soit d'actualité, n'a pas été bien investigué dans la littérature scientifique. Dans cet article nous serons premièrement intéressés par l'inférence des arbres additifs à partir des matrices incomplètes, bien que le problème de reconstruction des arbres ultramétriques soit aussi fort intéressant pour la classification mathématique.

Présentement, il existe dans la littérature deux types de méthodes de reconstruction des arbres additifs à partir des matrices incomplètes. Il s'agit des méthodes procédant par l'estimation *directe* et *indirecte* des valeurs manquantes. Les méthodes indirectes se basent sur les estimations des valeurs manquantes avant que la procédure de la reconstruction des arbres additifs soit lancée. Une fois toutes les valeurs manquantes sont estimées, n'importe quelle des méthodes d'ajustement d'arbre à une matrice de distance complète pourra être appliquée. Les méthodes directes, quant à elles, procèdent par l'inférence directe de la structure arborescente en utilisant une technique particulière de reconstruction. En ce qui concerne les méthodes indirectes, ici nous devons mentionner les travaux de De Soete [DES 84] et de Landry *et al.* [LAN 96]. Ces auteurs ont montré

comment obtenir des estimations des valeurs manquantes dans une matrice de distance en employant soit la *condition ultramétrique* :

$$d(i,j) \leq \max \{d(i,k); d(j,k)\}, \text{ pour tous } i, j, \text{ et } k,$$

soit la *condition des quatre points* (Zaretskii [ZAR 95] et Buneman [BUN 71]) :

$$d(i,j) + d(k,l) \leq \max \{d(i,k) + d(j,l); d(i,l) + d(j,k)\}, \text{ pour tous } i, j, k, \text{ et } l.$$

De Soete [DES 84] et de Landry *et al.* [LAN 96] ont montré comment estimer les entrées incomlètes à travers des combinaisons des valeurs existantes en utilisant les propriétés de la condition ultramétrique et de la condition des quatre points.

En ce qui concerne les méthodes directes, deux méthodes de reconstruction des arbres additifs à partir des matrices incomplètes ont été récemment proposées dans la littérature scientifique. Il s'agit ici de la *méthode des triangles* de Guénoche et Grandcolas [GUE 99] et de la méthode *MW* de Makarenkov et Leclerc [MAK 99]. La dernière méthode, basée sur une procédure d'optimisation par les moindres carrés, n'a pas été spécialement conçue pour le traitement des matrices incomplètes. Elle permet d'inférer un arbre additif en tenant compte de deux matrices : la matrice de distances et la matrice de poids. Lorsque la matrice de distances contient des valeurs manquantes, les poids 0 peuvent être associés à ces valeurs, alors que les poids 1 sont associés à des valeurs existantes. Une telle stratégie d'utilisation de la fonction des poids a mené à une méthode originale de traitement des données incomplètes (voir Levasseur *et al.* [LEV 00]). Toutefois, ce n'était qu'une première tentative d'usage de la procédure *MW* pour la résolution de ce problème. Cette première version de la méthode était moins performante que la *méthode des triangles* en terme de reconstruction de la topologie arborescente [LEV 00].

Dans cet article nous introduirons une nouvelle méthode, appelée *MW modifiée*, qui est beaucoup plus efficace pour la reconstruction des arbres additifs à partir des matrices incomplètes. Cette méthode, s'inspirant fortement de la procédure d'optimisation *MW* [MAK 99], n'utilisera cependant qu'une seule matrice, qui sera celle des distances ; la matrice des poids ne sera pas considérée. Une investigation exhaustive de la capacité de la nouvelle méthode de retrouver la vraie structure arborescente a été effectuée sur des matrices de distance d'arbre de différentes tailles. Les simulations Monte Carlo ont été effectuées sur les matrices incluant le différent pourcentage de valeurs manquantes. Les performances de la nouvelle méthode ont été évaluées et comparées à celles de la *méthode des triangles* [GUE 99], de la *procédure ultramétrique* [DES 84] et de la *procédure additive* [LAN 96]. *MW modifiée* a généralement permis d'obtenir de meilleures performances que les méthodes existantes mentionnées selon un critère métrique et un critère topologique mesurés.

2. Description de la nouvelle méthode

Maintenant, nous présentons la méthode d'ajustement *MW modifiée* qui pourrait s'appliquer à des données de types variés (métrique, dissimilarité, tableau symétrique de données contenant des valeurs négatives). Dans cette étude, nous avons considéré les matrices de distances (ou de dissimilarité) symétriques et non négatives. Rappelons que notre but est d'obtenir un arbre additif à partir d'une telle matrice de distances. La complexité algorithmique de notre méthode peut varier, entre $O(n^3)$ et $O(n^5)$ pour une matrice de dimension nxn, en fonction de la précision souhaitée. Lorsqu'elle est supérieure à $O(n^3)$, nous obtenons plusieurs arbres additifs distincts, et c'est à l'utilisateur de notre logiciel de choisir le meilleur arbre selon son critère de préférence. Dans ce papier nous avons considéré deux mesures classiques, qui sont largement employées en classification pour comparer des performances des méthodes d'ajustement d'arbres. Ces mesures sont le critère des moindres carrés (critère métrique) et la distance de Robinson et Foulds [ROB 81] (critère topologique). Notre procédure d'ajustement consiste à retrouver une topologie arborescente appropriée en partant d'un arbre contenant deux sommets et une arête. L'arbre entier est construit en ajoutant une nouvelle feuille à l'arbre courant à chaque pas de l'algorithme. L'ajout de la nouvelle feuille se fera en fonction de la matrice de distances donnée et d'une fonction spéciale introduite, nommée fonction de scores.

Soit **D** la matrice de distances symétriques de dimension nxn, sur l'ensemble X de n éléments. On suppose que certaines des entrées de cette matrice sont manquantes. Le critère des moindres carrés, qui sera notre critère de base, consiste à minimiser la fonction suivante :

$$Q = \Sigma_{i,j \in X,\ i<j} (d(i,j) - \delta(i,j))^2,$$

où $\delta(i,j)$ est une estimation de l'*entrée existante* $d(i,j)$ de **D**. La fonction δ est une *distance d'arbre* correspondant à un arbre additif et vérifiant donc la condition des quatre points.

Soit *Score* (*i, j*) la fonction de score définie comme suit pour une paire d'éléments *i, j* de l'ensemble *X* :

$$Score(i,j) = \begin{cases} 1, & si\ \ d(i,j)\ \ est\ \ une\ \ entrée\ \ présente\ \ dans\ \ \mathbf{D} \\ 0, & si\ \ d(i,j)\ \ est\ \ une\ \ entrée\ \ absente\ \ dans\ \ \mathbf{D} \end{cases}.$$

Voici comment nous proposons de construire un arbre additif, prenant en considération des entrées manquantes dans **D** :
Pas 1. On prend *i* et *j*, tels que $d(i,j)$ est une entrée présente de la matrice **D** ; l'arbre T_2 aura *ij* pour unique arête, de longueur $d(i,j)$.
Pas 2. À cette étape on a à sélectionner le troisième élément à placer dans l'arbre. L'élément *k*, tel que $k \in X - \{i, j\}$ est choisi de façon à maximiser la fonction suivante :

$$Score\ (i, k) + Score\ (j, k)$$

L'idée principale est de choisir un élément *k*, pour lequel il a le plus d'entrées $d(i,k)$ et $d(j,k)$ présentes dans **D**. L'arbre T_3 aura donc trois arêtes et trois feuilles, qui seront respectivement *i*, *j* et *k*. La longueur optimale des arêtes de T_3 sera définie au moyen de la procédure *MW* (voir [MAK 99]).
Pas *p*, (*p* < *n*). Soit T_p l'arbre additif à *p* feuilles que nous avons construit lors des pas précédents. Les feuilles de T_p correspondent au *p* éléments de *X*. Parmi les *n* - *p* éléments de *X* qui ne sont pas encore placés dans l'arbre nous avons à trouver un qui conviendra le mieux pour le prochain placement. L'élément *p* + 1à placer dans l'arbre est choisi de façon à maximiser la fonction suivante :

$$\sum_{l \in L(T_p)} Score(l, p+1),$$

où $L(T_p)$ est l'ensemble des feuilles de l'arbre T_p. Comme à tous les pas précédents, le choix d'emplacement de la nouvelle feuille *p* + 1 et la longueur de nouvelles arêtes seront déterminés par les moindres carrés (en appliquant la procédure *MW*).
Au pas *n*-1, un arbre additif complet sera construit. Les *n* feuilles de cet arbre vont représenter les *n* éléments de l'ensemble *X*.

La complexité algorithmique de la procédure décrite ci-dessus est $O(n^3)$. Pour améliorer la qualité d'ajustement nous pouvons exécuter cette procédure en considérant plusieurs différentes paires d'éléments *i* et *j* au premier pas. Comme le nombre maximal de différentes paires à considérer à la première étape est $n(n-1)/2$, la complexité algorithmique d'une telle stratégie exhaustive sera donc $O(n^5)$. C'est cette dernière stratégie, qui a été choisie pour les tests de la nouvelle méthode dont les résultats sont présentés dans la section suivante. Une seule restriction s'applique à notre approche : la matrice de distances donnée doit contenir au moins une entrée existante par ligne et par colonne.

3. Comparaison des performances des quatre méthodes

Les performances de la nouvelle méthode *MW modifiée* ont été évaluées et comparées à celles de la *méthode des triangles* [GUE 99], de la *procédure ultramétrique* [DES 84] et de la *procédure additive* [LAN 96]. L'évaluation des résultats a été effectuée selon une stratégie analogue à celle proposée par Pruzansky *et al*. [PRU 82]. Chaque matrice de dissimilarité à tester a été obtenue comme suit :
- On a généré aléatoirement un arbre binaire comprenant *n* feuilles et 2*n*-3 arêtes, où *n* = 8, 16 et 24.
- Les longueurs des arêtes de cet arbre ont été tirées aléatoirement suivant une loi uniforme sur [0;1].
- La matrice initiale de distance d'arbre a été réduite de façon à avoir une variance de 1.
- Pour chaque dimension des données *n*, 100 arbres aléatoires ont été engendrés. Les résultats obtenus et illustrés sur les Figures 1 et 2 sont donc issus de 300 jeux de donnée s différents.
- De 50 à 0 pourcent de données ont ensuite été enlevées des matrices générées pour former les matrices partielles avec le différent pourcentage de valeurs manquantes.

Pour estimer la qualité des ajustements, nous avons mesuré, sur chacune des matrices de distances et pour chacune des méthodes considérées, les deux quantités suivantes :
1. *Le pourcentage de variance* entre la distance de départ et la distance d'arbre obtenue. C'est un critère métrique classique largement employé pour ce type de tests. La formule exacte utilisée ici est celle proposée dans [PRU 82] :

$$\%\text{Var} = 100\% \times \left(1 - \frac{\sum_{i<j} (d(i,j) - \delta(i,j))^2}{\sum_{i<j} (d(i,j) - m(d))^2}\right),$$

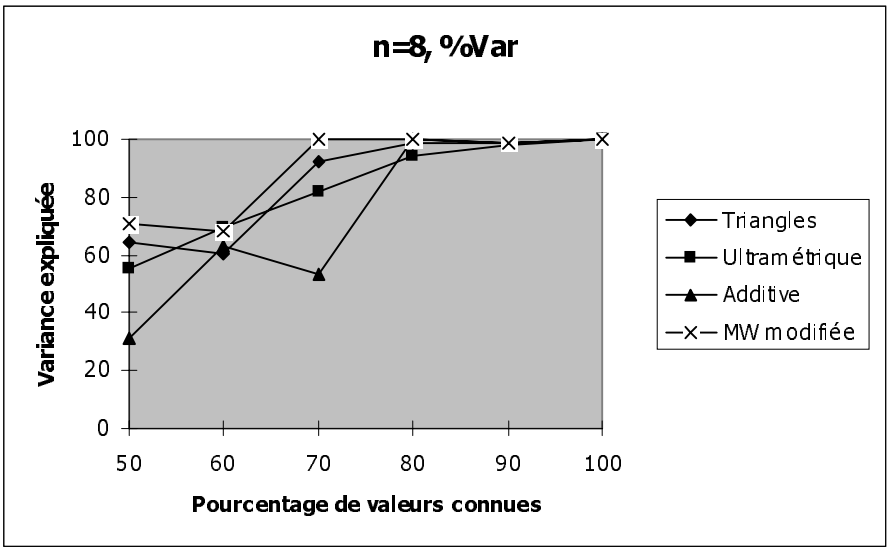
où $m(d)$ est la moyenne des valeurs de la distance initiale et $\delta(i,j)$ est la valeur de la distance d'arbre retrouvée associée à la paire i,j.
2. *La distance topologique de Robinson et Foulds* [ROB 81] entre le vrai arbre généré et l'arbre solution correspondant à la distance d'arbre obtenue δ. Cette distance, qui est souvent utilisée pour comparer les topologies de deux arbres additifs (voir par exemple Saitou et Nei [SAI 87] ou Makarenkov et Legendre [MAK 01]), est le nombre minimum d'opérations nécessaires (unifications ou séparations de noeuds) pour transformer une structure d'arbre en une autre. Robinson et Foulds ont montré que cette distance est aussi égale au nombre de bipartitions, ou scissions d'après Buneman [BUN 71], qui sont présentes dans un arbre et absentes dans l'autre. En calculant la distance topologique, nous avons toujours supposé qu'une arête de longueur nulle induit une bipartition. Une fois la distance de Robinson et Foulds obtenue, nous l'avons normalisé par sa valeur maximum égale à $2n$-6. Rappelons que la distance topologique entre deux arbres est égale à zéro si est seulement si leur topologies sont identiques.

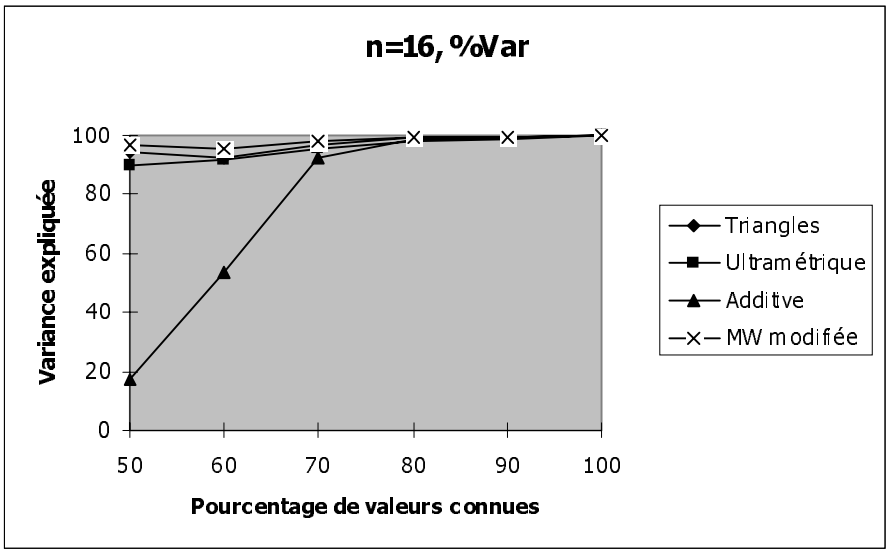
Comme la procédures *ultramétrique* de De Soete [DES 84] et la *procédure additive* de Landry *et al.* [LAN 96] permettent seulement de remplacer les valeurs manquantes par des valeurs calculées comme des combinaisons des valeurs existante, sans garantir l'obtention d'une matrice de distance d'arbre, ces deux procédures doivent être suivies par une méthode de reconstruction des arbres additifs fonctionnant à partir des matrices complètes. Par conséquent, pour compléter les deux procédures en question, nous avons utilisé la méthode *MW* de [MAK 99] pour inférer la topologie finale de l'arbre additif. Quant à la *méthode des triangles* de Guénoche et Grandcolas [GUE 99], elle permet de reconstruire directement la topologie arborescente à partir d'une matrice incomplète. Cependant pour être compétitive, cette méthode, qui n'a pas été spécialement conçue pour optimiser le critère des moindres carrés, doit être complétée par une procédure d'ajustement des longueurs des arêtes sur un arbre de topologie fixée (voir par exemple [BAR 88]). Cette dernière procédure a donc été employée dans notre étude pour optimiser les résultats fournis par la *méthode des triangles.*

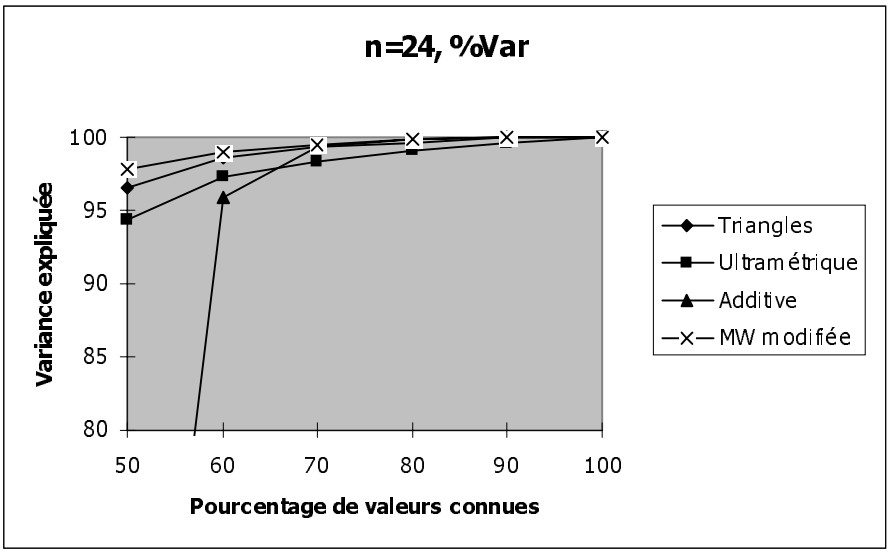
Les performances des quatre méthodes considérées dans cet article ont été évaluées en terme de pourcentage de variance expliquée et de distance topologique de Robinson et Foulds. Les résultats obtenus par ces quatre stratégies sont illustrés sur les Figures 1 et 2 ci-dessous. En analysant les courbes, qui décrivent le comportement des méthodes, nous pouvons constater que la nouvelle approche *MW modifiée* se montre la plus performante dans la plupart des situations, peu importe le pourcentage de valeurs manquantes. Notre méthode fournit généralement les plus grandes valeurs du pourcentage de variance expliquée et les plus petites de la distance topologique de Robinson et Foulds normalisée.

Quant au pourcentage de variance expliquée, ce sont les méthodes *MW modifiée* et celle *des triangles* qui se montrent les plus performantes (Figure 1). Remarquons aussi que de très faibles résultats ont été obtenus par la *procédure additive* pour des petits pourcentages des valeurs connues (de 50 à 70 pourcent). La tendance générale suivante peut être observée : plus grande est la dimension de la matrice de distances, plus grand est le pourcentage de variance expliquée par les méthodes.

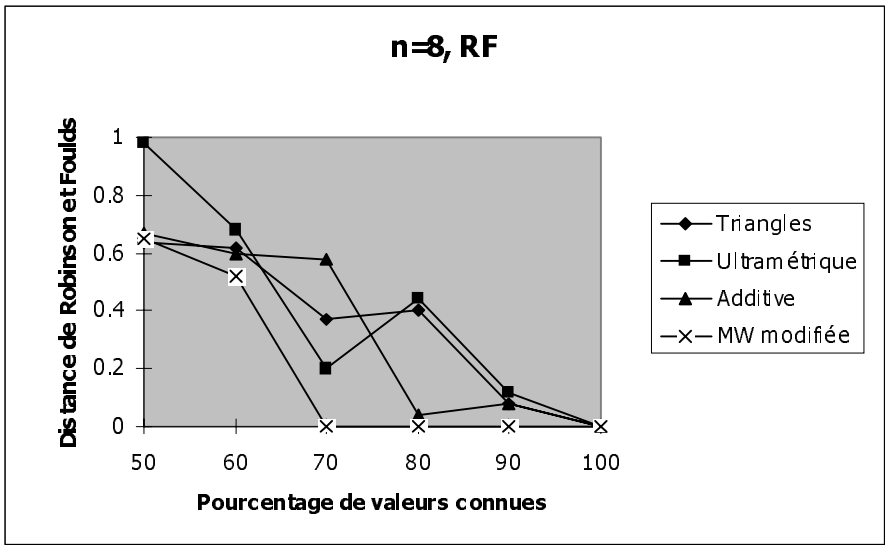
Quant au recouvrement de la topologie de l'arbre d'origine (Figure 2) c'est la méthode *MW modifiée*, qui se montre nettement supérieure aux autres. La *procédure additive* fonctionne bien lorsque nous avons un grand pourcentage des valeurs connues (80 pourcent et plus), mais elle devient très instable quand le nombre de valeurs manquante augmente. La *méthode des triangles* et la *procédure ultramétrique* montrent un comportement semblable dans la plupart des situations en terme de distance topologique. Cependant, la *procédure ultramétrique* fonctionne très mal dans le cas des petites matrices de distances comprenant beaucoup de valeurs manquantes.

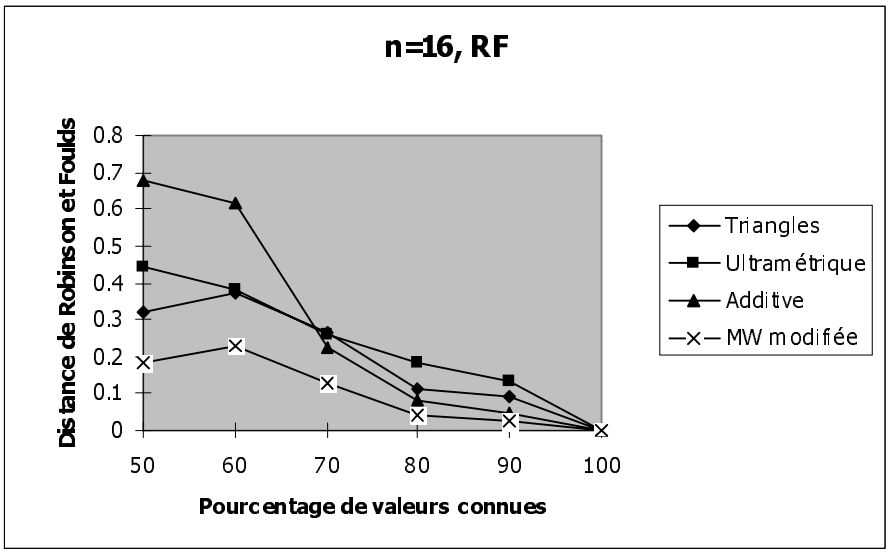


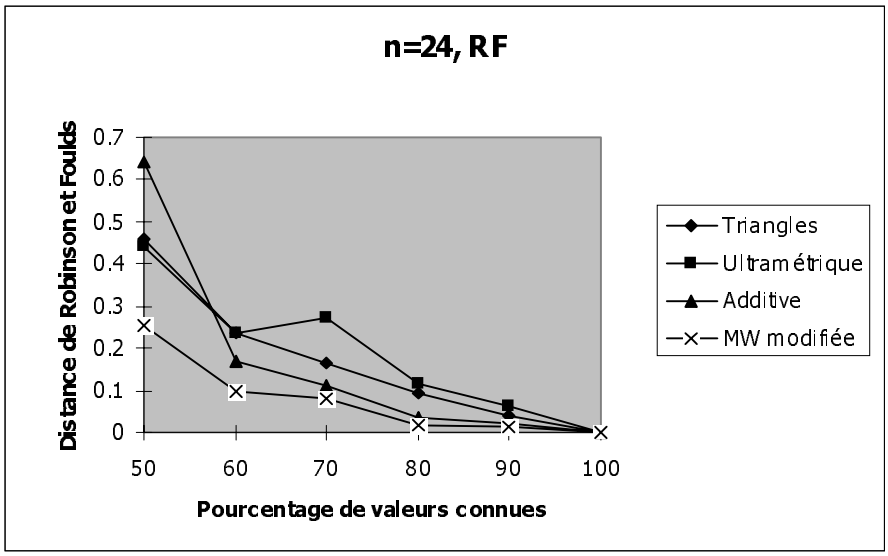




Figures 1. *Les performances des quatre méthodes de reconstruction pour le pourcentage de variance expliquée. Les simulations ont été effectuées pour les matrices de distances de tailles 8x8, 16x16 et 24x24. Plus grand est le pourcentage de variance expliquée par une méthode, plus performante elle est.*







Figures 2. *Les performances des quatre méthodes de reconstruction pour la distance topologique de Robinson et Foulds entre l'arbre d'origine et l'arbre construit. Les simulations ont été effectuées pour les matrices de distances de tailles 8x8, 16x16 et 24x24. Plus petite est la valeur de la distance topologique obtenue par une méthode, plus performante elle est.*

4. Remerciements

Je remercie Alain Guénoche, Pierre-Alexandre Landry et François-Joseph Lapointe pour la mise en ma disposition de leurs logiciels réalisant les méthodes exposées dans cet article.

5. Bibliographie

[BAR 88] BARTHELEMY J.P., GUENOCHE A. *Les arbres et les représentations des proximités*, Paris, Masson, 1988.

[BUN 71] BUNEMAN P. « The Recovery of Trees from Measures of Dissimilarité », in *Mathematics in Archaeological and Historical Sciences*, eds. F.R. Hodson, D.G. Kendall and P. Tautu, Edinburgh, Edinburgh University Press, 1971, p. 387-395.

[DES 84] DE SOETE G. « Additive-Tree Representations of Incomplete Dissimilarity Data », *Quality and Quantity*, 18, 1984, p. 387-393.

[GAS 97] GASCUEL,O. « Concerning the NJ algorithm and its unweighted version UNJ ». Dans *Mathematical hierarchies and Biology (B. Mirkin, F.R. McMorris, F. Roberts, A. Rzhetsky, eds.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Amer. Math. Soc., Providence, RI*, 1997, p. 149-171.

[GUE 99] GUENOCHE A. GRANDCOLAS, S. « Approximation par arbre d'une distance partielle ». *Mathématiques, Informatique et Sciences humaines*, 146, 1999, p. 51-64.

[LAN 96] LANDRY P. A., LAPOINTE F.-J., KIRSCH J. A. W. « Estimating phylogenies from distance matrices: additive is superior to ultrametric estimation ». *Molecular Biology and Evolution*, 13 (6), 1996, p. 818-823.

[LEV 00] LEVASSEUR C., LANDRY P. A., LAPOINTE F.-J. « Estimating trees from incomplete distance matrices: a comparison of two methods ». *Data analysis, Classification and Related Methods* (H. A.L. Kiers, J.-P. Rasson, P.J.F. Groenen, M. Schader, eds), 2000, p. 149-154.

[MAK 01] MAKARENKOV V. « T-Rex : reconstructing and visualizing phylogenetic trees and reticulation networks ». *Bioinformatics*, 17 (7), 2001, p. 664-668.

[MAK 01] MAKARENKOV V., LEGENDRE P. « Optimal Variable Weighting for Ultrametric and Additive Tree Clustering and *K*-means Partitioning: Methods and Software ». *Journal of Classification*, 18(2), 2001, p. 245-271.

[MAK 99] MAKARENKOV V., LECLERC B. « An algorithm for the fitting of a tree metric according to a weighted least-squares criterion ». *Journal of Classification*, 16, 1999, p. 3-26.

[SAI 87] SAITOU N., NEI M. « The neighbor-joining method: a new method for reconstructing phylogenetic trees ». *Molecular Biology and Evolution*, 4, 1987, p. 406-425.

[SAT 77] SATTATH S., TVERSKY A. « Additive similarity trees ». *Psychometrika* , 42, 1977, p. 319-345.

[PRU 82] PRUZANSKY S., TVERSKY A., CARROLL J. D. « Spatial Versus Tree Representations of Proximity Data ». *Psychometrika*, 47, 1982, p. 3-19.

[ROB 81] ROBINSON D. R., FOULDS L. R. « Comparison of phylogenetic trees ». *Mathematical Biosciences*, 53, 1981, p. 131-147.

[ZAR 65] ZARETSKII K. « Construction of a tree on the basis of a set of distances between its leaves ». *Uspekhi Mat. Nauk*. 20, 1965, p. 90-92.