MATHÉMATIQUES ET SCIENCES HUMAINES

ALAIN GUÉNOCHE
STÉPHANE GRANDCOLAS

Approximations par arbre d'une distance partielle

Mathématiques et sciences humaines, tome 146 (1999), p. 51-64
<http://www.numdam.org/item?id=MSH_1999__146__51_0>

Math. Inf. Sci. hum., (37e année, n° 146, *1999*, pp. 51-64)

APPROXIMATIONS PAR ARBRE D'UNE DISTANCE PARTIELLE

Alain GUÉNOCHE[1], Stéphane GRANDCOLAS[1]

RÉSUMÉ — *En classification par arbre, on cherche à ajuster une dissimilarité donnée par une distance d'arbre. Mais bien souvent, surtout par comparaison de séquences biologiques, les valeurs obtenues sont peu fiables, voire indéterminées. On a alors une distance partielle qui n'est pas définie pour toute paire. Dans ce cas, on peut soit développer une méthode spécifique qui n'utilise que les valeurs disponibles, soit estimer les valeurs manquantes et utiliser une méthode classique pour reconstruire l'arbre. Cet article présente deux méthodes de ce type et les compare à l'aide de simulations sur des distances d'arbre partielles et bruitées.*

MOTS CLÉS — Reconstruction d'arbre, Distance partielle, Méthode séquentielle.

ABSTRACT — Tree adjustments for partial distances

In tree clustering, we try to approximate a given dissimilarity matrice by a tree distance. In some cases, especially when comparing biological séquences, some dissimilarity values cannot be evaluated and we get some partial dissimilarity with undefined values. In that case one can develop a sequential method to reconstruct a valued tree or evaluate the missing values using a tree model. This paper introduces two methods of this kind and compare them simulating noisy partial tree dissimilarities.

KEY WORDS — Tree Reconstruction, Partial Distance, Sequential Method.

En classification on s'intéresse à la représentation d'une distance D sur un ensemble X par un X-arbre (X est l'ensemble des feuilles, les arêtes sont pondérées par des valeurs positives ou nulles, et la longueur des chemins entre feuilles approxime la distance D) (Barthélemy & Guénoche [1988]). Suivant les domaines d'application, les nœuds (ou sommets internes) correspondent à des catégories (psychologie cognitive), à des ancêtres communs (évolution moléculaire, filiation de textes) ou tout simplement à des classes d'objets (archéologie). Il est bien connu que cette représentation est exacte si et seulement si D est une distance d'arbre, c'est-à-dire que D vérifie la *Condition des quatre points* (Zaretskii [1965], Buneman [1971]) : Pour tout $x, y, z, t \in X$,

$$D(x,yi)+D(z,t) \leq \text{Max}\{D(x,z) + D(y,t)\ ,\ D(x,t) + D(y,z)\ \}.$$

En d'autres termes, pour tout quadruplet, parmi les trois sommes $D(x,y) + D(z,t)$, $D(x,z) + D(y,t)$, $D(x,t) + D(y,z)$, les deux plus grandes sont égales.

[1] Laboratoire d'Informatique de Marseille, Université de la Méditerrannée, 163 avenue de Luminy, 13009 Marseille, e-mail : guenoche@lim.univ-mrs.fr.

Généralement dans tous ces domaines, à commencer par celui de la reconstruction phylogénétique, le modèle arboré s'impose, mais les indices de distance utilisés ne donnent pas exactement des distances d'arbre ; ce ne sont même pas nécessairement des distances, mais des dissimilarités. De plus, on ne dispose souvent que d'une information incomplète sur les éléments comparés, en particulier pour les séquences biologiques, et on est en présence de valeurs indéterminées (ou peu fiables). Par exemple, dans la version actuelle de la base Hovergen des gènes homologues de vertébrés (Duret, Mouchiroud, Gouy [1994]), il y a 20 % de familles contenant des séquences incomplètes. On a alors une *dissimilarité partielle*, notée par la suite Δ, à partir de laquelle on souhaite reconstruire un X-arbre.

Dans cette situation expérimentale, on a essentiellement deux possibilités. La première consiste à développer une méthode spécifique qui n'utilise que les valeurs de distance disponibles ; c'est la *Méthode Séquentielle Parcimonieuse* que nous présenterons tout d'abord. Dans la seconde, on commence par estimer les valeurs manquantes avant d'utiliser une méthode classique, *Neighbor Joining* de Saitou & Nei [1987], pour reconstruire l'arbre.

Pour finir, nous comparons ces deux approches à l'aide de simulations. Pour cela on tire au hasard un X-arbre T, donc une distance d'arbre D que l'on bruite pour obtenir une dissimilarité partielle Δ. A l'aide de l'une ou l'autre méthode, on reconstruit un X-arbre valué θ que l'on compare à T suivant des critères métriques et topologiques.

1. ARBRES ET MÉTHODES SÉQUENTIELLES DE RECONSTRUCTION

Comme la condition des quatre points le laisse entendre, toute distance sur trois points est arborée, c'est-à-dire représentable sur un arbre. Pour tout triplet $\{x,y,z\}$ les trois valeurs de distances $\Delta(x,y)$, $\Delta(y,z)$ et $\Delta(z,x)$ forment un triangle métrique si et seulement si ces trois valeurs vérifient la fameuse inégalité triangulaire. Ce triangle, habituellement représenté dans le plan euclidien, peut aussi être représenté de façon exacte par un arbre.

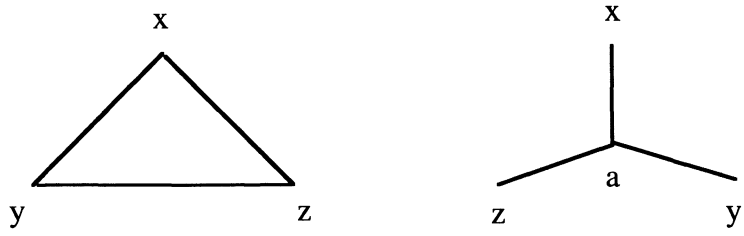

Figure 1 : Toute distance sur 3 points est une distance d'arbre

Si Δ est une distance sur $X = \{x,y,z\}$, alors il existe un X-arbre unique, appelé 3-étoile, dont les arêtes (x,a), (y,a) et (z,a) sont de longueurs positives ou nulles :

$L(x) = \frac{1}{2}\ [\Delta(x,y) + \Delta(x,z) - \Delta(y,z)]$, $L(y) = \frac{1}{2}\ [\Delta(y,x) + \Delta(y,z) - \Delta(x,z)]$ et $L(z) = \frac{1}{2}\ [\Delta(z,x) + \Delta(z,y) - \Delta(x,y)]$. Par la suite nous appellerons ces formules *les équations du triangle.*

Cette propriété à été utilisée très tôt pour la reconstruction d'un arbre à partir d'une distance supposée proche d'une distance d'arbre. Les premiers algorithmes sont des méthodes séquentielles dont le principe a été proposé par Farris [1972] et repris par Waterman et al. [1977]. Dans les méthodes séquentielles, à chaque étape on dispose d'un

Y-arbre, dans lequel Y désigne le sous-ensemble des sommets placés, et on ajoute un nouvel élément z. Pour cela, on détermine la position d'un nouveau nœud $a(z)$ dans l'arbre courant qui est le point d'accrochage de l'arête $(z,a(z))$ qui est ajoutée. Il y a essentiellement deux stratégies : soit on place z par rapport aux arêtes de l'arbre, soit on le place par rapport aux chemins entre paires d'éléments de Y. Dans un cas comme dans l'autre, il faut sélectionner l'arête ou le chemin qui minimise l'écart $E(z)$ entre z et le Y-arbre courant, c'est-à-dire la longueur de l'arête $(z,a(z))$; ce sont les équations du triangle qui permettent de mesurer cet écart.

Dans le cas du positionnement par rapport aux arêtes, là encore il y a deux choix. Soit on estime les valeurs de distance aux extrémités de chaque arête, c'est-à-dire que l'on étend la matrice de distance aux nœuds – c'est la solution retenue par Waterman –, soit on tient compte de toutes les valeurs entre éléments de Y situés de part et d'autre de chaque arête, solution adoptée par Makarenkov et Leclerc [1999].

Du point de vue de la complexité, il semble plus efficace de travailler par rapport aux arêtes, puisqu'elles sont en nombre $O(n)$, alors que les chemins sont en $O(n^2)$. Certes, mais nous verrons que, dans le premier cas, pour chaque arête l'évaluation de l'écart à l'arbre courant d'un sommet non placé est en $O(n^2)$, ce qui conduit à des itérations en $O(n^3)$ et donc un algorithme en $O(n^4)$. Alors que la procédure que nous développons, bien que se plaçant par rapport aux chemins, parce qu'elle exploite une stratégie similaire à l'algorithme de Prim [1957] pour les arbres couvrants de longueur minimum, est en $O(n^3)$ au total.

1.1. Méthode séquentielle parcimonieuse

Cette méthode doit beaucoup à celle des Triangles (Leclerc [1995], puis Leclerc & Makarenkov [1998]), basée sur la construction d'un 2-arbre, c'est-à-dire un ensemble de $(n\text{-}2)$ triangles sur X connexes pour la relation d'adjacence – deux triangles sont adjacents s'ils ont un côté commun. Son intérêt est double : d'une part elle montre l'équivalence entre un 2-arbre et le X-arbre résultant et, d'autre part, que les $2n\text{-}3$ valeurs de distances retenues dans le 2-arbre sont exactement représentées dans le X-arbre correspondant. En collaboration avec L. Duret, nous avons tout d'abord essayé d'étendre cette méthode au cas des distances partielles (Guénoche, Leclerc [1998]). Finalement la recherche d'une meilleure efficacité nous a conduit à adopter une méthode purement séquentielle, dans laquelle la notion de 2-arbre n'a plus lieu d'être. Néanmoins beaucoup des idées, et des solutions retenues ci-après, viennent de cette tentative de prolongement de la méthode des Triangles.

Initialement, on part d'un arbre réduit à 3 éléments $\{x,y,z\}$. On choisit ceux qui correspondent à un triangle métrique de périmètre minimum, c'est-à-dire à la 3-étoile de longueur minimum, et l'on pose $Y := \{x,y,z\}$. On a donc 3 chemins dans l'arbre courant noté θ. Pour tout autre élément w, notons $L_{xy}(w)$ la distance de w au chemin $[x,y]$ donnée par les équations du triangle, et $E(w)$ l'écart de w à Y calculé comme :

$$E(w) := \text{Min } \{L_{xy}(w), L_{xz}(w), L_{yz}(w)\}.$$

Dans notre méthode, on place à chaque itération l'élément z d'écart minimum tel que :

$$E(z) \leq \text{Min}_{w \in X \setminus Y} E(w).$$

Ce choix est motivé par deux observations : d'une part c'est le plus efficace, au vu des simulations que nous décrivons ultérieurement ; d'autre part, c'est l'application du principe de parcimonie dans les méthodes de distance. Ce principe, cher aux phylogénistes, veut que l'arbre *juste* soit celui qui demande le moins de mutations pour expliquer la diversité observée de nos jours et donc, si les distances sont proportionnelles au nombre de mutations, de construire un arbre de longueur minimum. Au même titre que NJ, qui sélectionne le groupement retenu à chaque itération comme celui qui promet la plus petite arête entre ce groupement et le reste de l'arbre – donc la plus petite arête interne possible – nous ajoutons dans l'arbre la plus petite des arêtes externes possibles.

Pour un choix efficace du sommet ajouté, on mémorise les valeurs des écarts des sommets non placés, ainsi que les chemins qui réalisent ces écarts. Une fois qu'on a choisi z, on détermine la position dans l'arbre θ du point d'accrochage $a(z)$; il est sur une arête (a,b) déterminée en parcourant sur le chemin $[x,y]$ la longueur $L(x)$ depuis x ou, ce qui revient au même si ce chemin est bien de longueur $\Delta(x,y)$, la longueur $L(y)$ depuis y. Dans la procédure d'expansion de l'arbre courant, on supprime l'arête (a,b) que l'on remplace par $(a,a(z))$ et $(b,a(z))$ avec des longueurs calculées à l'aide du seul triplet $\{z,x,y\}$. Pour $(z,a(z))$, on recalcule sa longueur de façon à prendre en compte plus d'information et à éviter une procédure qui, par essence, sous-estime sa longueur. Mais on ne remet en cause ni l'arête qui découle du choix de z, donc la paire (x,y), ni la position de $a(z)$.

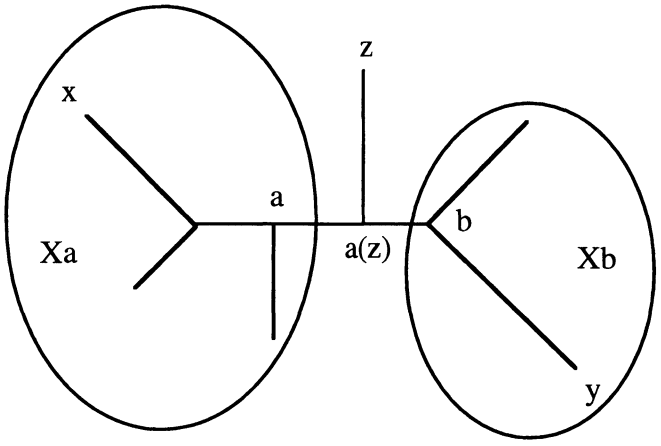

Figure 2 : Placer z dans l'arbre courant θ

L'arête (a,b) partitionne les éléments placés en deux classes, Xa et Xb, constituées des éléments situés de part et d'autre de (a,b) dans θ. La longueur $L(z)$ de l'arête $(z,a(z))$ peut être calculée par rapport à toute paire de sommets pris un dans chaque classe – si Δ est une distance d'arbre, toutes les longueurs seront identiques. Afin d'obtenir une certaine stabilité par rapport à des déviations éventuelles, on calcule $L(z)$ comme une moyenne, d'où une procédure en $O(n^2)$:

$$L(z) = \frac{1}{|Xa|.|Xb|} \sum_{x\in Xa,\, y\in Xb} \frac{1}{2} [D(z,x) + D(z,y) - D(x,y)].$$

Une fois cette longueur calculée, on estime les longueurs de chemins entre z et tous les éléments placés, c'est-à-dire que l'on calcule pour tout x dans Y la valeur de $D_\theta(x,z)$ et l'on pose $\Delta(x,z) := D_\theta(x,z)$. Ceci est nécessaire pour deux raisons :

- Pour mettre à jour l'écart, et le meilleur chemin, de tout sommet w non placé, il faut calculer $L_{xz}(w)$ donc connaître, si $\Delta(w,x)$ et $\Delta(w,z)$ sont connues, la valeur $\Delta(x,z)$. Si l'écart de w à un quelconque chemin $[x,z]$ d'extrémité z est plus petit que $E(w)$, alors :

$$E(w) := \text{Min}_{x \in Y} \Delta(w,x) + \Delta(w,z) - \Delta(x,z)$$

- Pour placer ultérieurement un nœud $a(w)$ sur le chemin $[x,z]$, il faut que la distance $\Delta(x,z)$ soit l'exacte longueur du chemin $[x,z]$ dans l'arbre courant θ. Si cette valeur est indéterminée, ceci est nécessaire pour placer w si ses seules distances connues le sont par rapport à x et z. Si elle est connue, elle n'est pas nécessairement exactement représentée ; le chemin entre x et z dans l'arbre n'est pas de longueur $\Delta(x,z)$. Le placement ultérieur de $a(w)$ sur ce chemin dépendrait alors du point d'où l'on part. Donc, dans la mesure où l'élément z est ajouté, les valeurs $\Delta(x,z)$ doivent être mises à jour.

Cet algorithme permet de mettre en évidence quelques propriétés de la reconstruction d'arbres à partir de distances partielles. Toutes les explications développées ci-dessus tiennent lieu de démonstration.

PROPOSITION 1. Cet algorithme calcule un X-arbre à partir d'une distance partielle Δ si et seulement si il existe un ordre sur X tel que pour tout élément z, il y a au moins deux éléments x et y à distances de z connues, et placés avant lui.

Comme pour la méthode des Triangles, il faut commencer par deux triangles adjacents. On en déduit de même qu'il faut au moins $2n$-3 valeurs de distances, que le graphe support de cette distance doit être 2-connexe et que tous les sommets soient de degré ≥ 2.

PROPOSITION 2. Si Δ est une distance d'arbre complète représentable sur T, cet algorithme reconstruit l'arbre T.

Mais si Δ est une distance d'arbre partielle, le rattachement d'un élément z sur l'arête (a,b) ne sera conforme à T que s'il existe au moins un élément de Xa et un élément de Xb dont les distances à z sont connues.

1.2. Complexité

On admettra qu'établir un chemin dans un arbre à n feuilles est en $O(n)$.

L'initialisation par un triangle de périmètre minimum est en $O(n^3)$. Le calcul des écarts aux trois chemins est en $O(n)$.

A chaque étape :

- Le choix de z est en $O(n)$; il se place par rapport à $[x,y]$.
- le positionnement du nouveau nœud sur $[x,y]$ est en $O(n)$; il permet de déterminer l'arête (a,b).
- Le partitionnement en Xa et Xb est en $O(n^2)$, ainsi que le calcul de $L(z)$.
- Pour chaque x placé
  - le calcul de $D_\theta(z,x)$ est en $O(n)$.
- Pour chaque w non placé
  - l'écart à l'arbre courant est mis a jour en $O(n)$.

A chaque étape la mise à jour de Δ et de E sont en $O(n^2)$, donc la méthode séquentielle est en $O(n^3)$ au total.

2. ÉVALUATION DES VALEURS MANQUANTES

L'autre approche étudiée ici consiste à estimer les valeurs manquantes sous l'hypothèse que l'on est proche d'une distance d'arbre.

2.1. HISTORIQUEMENT PARLANT

Dans la lignée des travaux de De Soete [1984], cette estimation des valeurs manquantes a été largement étudiée en phylogénie dans plusieurs articles de J. Landry, F.J. Lapointe et J.A.W. Kirsch ; nous nous référons à celui de 1996. Les auteurs comparent deux types d'évaluations ; l'une correspond au modèle ultramétrique (pour tout triplet d'éléments, les deux plus grandes distances sont égales) et l'autre au modèle plus général des arbres à distances additives (nos X-arbres). Soit $\Delta(u,v)$ une valeur inconnue que l'on veut estimer par Δuv.

Selon le premier modèle, puisque le triangle $\{x,u,v\}$ est ultramétrique, les deux plus grandes distances sont égales et donc $\Delta(u,v) \leq \text{Max}\ \{\Delta(x,u),\ \Delta(x,v)\}$. Ils adoptent donc la formule :

$$\Delta uv\ := \text{Min}_{x \in X}\ \text{Max}\ \{\Delta(x,u),\ \Delta(x,vI)\} \qquad (1)$$

Selon le second modèle, pour tout quadruplet $\{x,y,u,v\}$ dans lequel $\Delta(u,v)$ est la seule valeur indéterminée, on applique la Condition des quatre points et on obtient

$$\Delta uv\ := \text{Min}_{x \neq y \in X}\ [\text{Max}\ \{\Delta(x,u)+\Delta(y,v),\ \Delta(x,v)+\Delta(y,u)\} - \Delta(x,y)] \qquad (2)$$

Après simulations sur des données choisies, ils observent que l'évaluation suivant le second modèle est meilleur que selon le premier et que, si l'on a un grand nombre de valeurs manquantes, il n'est pas toujours faisable de compléter une distance avec les seules estimations du modèle d'arbre additif. Finalement, ils préconisent d'utiliser itérativement ce modèle tant qu'il produit de nouvelles estimations et de recourir au modèle ultramétrique quand on est bloqué, puisque celui-ci ne demande que deux distances connues par valeur manquante.

2.2. UN POINT DE VUE PLUS JUSTE

Les formules ci-dessus sont facilement améliorables. Revenons au cas où $\Delta(u,v)$ est la seule valeur inconnue sur $\{x,y,u,v\}$ et notons A et B les deux sommes calculables.

$$A\ := \Delta(x,u) + \Delta(y,v) \quad \text{et} \quad B\ := \Delta(x,v) + \Delta(y,u).$$

Conformément à la Condition des quatre points, la troisième somme inconnue $\Delta(x,y) + \Delta(u,v)$ n'est égale à $\text{Max}\{A,B\}$ que si A et B ne sont pas égales ; on obtient alors une *évaluation par quadruplet* de $\Delta(u,v)$ selon la même formule que précédemment :

$$\Delta uv\ := \text{Max}\{A,B\} - \Delta(x,y)$$

Mais si $A = B$, on ne peut rien dire de $\Delta(u,v)$ et appliquer systématiquement la formule (2) conduit à surestimer $\Delta(u,v)$ quand u et v sont deux feuilles adjacentes au même nœud – on dit que u et v sont *frères* (siblings). Dans ce cas, $\Delta(u,v)$ n'apparaît jamais dans une des deux plus grandes sommes. Aucune paire $\{x,y\}$ ne rend A et B différentes et $\Delta(u,v)$ ne peut être évaluée par quadruplet. Le recours au modèle ultramétrique de la formule (1) n'est pas plus satisfaisant, car il n'est pas non plus applicable. On aura alors recours aux équations du triangle pour une *évaluation par chemin*.

Pour contrôler ces estimations, on commencera par calculer une borne inférieure et une borne supérieure de chaque valeur inconnue. Comme nous calculons des valeurs de distance, elles doivent vérifier les inégalités triangulaires, et donc :

$$\mathrm{Max}_{x\neq u,\ x\neq v} \mid \Delta(x,u) - \Delta(x,v) \mid \leq \Delta uv \leq \mathrm{Min}_{x\neq u,\ x\neq v} \Delta(x,u) + \Delta(x,v).$$

Quand il y a un fort taux de valeurs manquantes, ces bornes ne sont pas toujours calculables, et il faut veiller, soit à ne pas utiliser les bornes indéfinies, soit à les mettre à jour au fil des estimations, soit à refaire le calcul avant les estimations par chemin.

2.2.1. Évaluation par quadruplets

Dans notre algorithme, on considère les valeurs manquantes $\Delta(u,v)$ dans l'ordre du nombre de paires $\{x,y\}$ qui permettent de les estimer, c'est-à-dire du nombre de quadruplets $\{x,y,u,v\}$ dans lesquels $\Delta(u,v)$ est la seule valeur manquante. Pour chacun, la question est de décider si A et B sont suffisamment proches pour être considérées comme les deux plus grandes sommes. Si oui, on passe au quadruplet suivant et sinon, on somme les quantités $\mathrm{Max}\{A,B\}$-$\Delta(x,y)$ pour en faire une moyenne.

Dans le cas d'une distance plus ou moins proche d'une distance d'arbre, il peut être délicat de décider si les deux sommes calculables sont égales ou non. On commencera donc, pour tout quadruplet $\{x,y,z,t\}$ pour lequel il n'y a pas de valeurs manquantes, par calculer les trois sommes notées sans ambiguïté Smin, Smed et Smax. On peut alors déterminer le facteur f qui rendrait les deux plus grandes sommes égales, c'est-à-dire tel que $(1+f)$ Smed = $(1-f)$ Smax. Le facteur final Fact est calculé comme le maximum, sur l'ensemble des quadruplets sans distances inconnues, des valeurs de

$$f = \frac{\mathrm{Smax} - \mathrm{Smed}}{\mathrm{Smax} + \mathrm{Smed}}\ .$$

Maintenant, pour toute paire (u,v), on décidera que A et B sont les deux plus grandes sommes si et seulement si :

$$(1 + \mathrm{Fact})\ \mathrm{Min}\{A,B\} \geq (1 - \mathrm{Fact})\ \mathrm{Max}\{A,B\}.$$

S'il n'y a pas de quadruplets pour évaluer ce facteur Fact, qui est d'autant plus grand que l'on s'écarte d'une distance d'arbre, on appliquera directement la stratégie d'évaluation par les chemins. On effectue donc plusieurs parcours de la liste des valeurs inconnues, tant que leur nombre décroît. S'il en reste, on passe à l'évaluation par chemins.

2.2.2. Évaluation par chemins

Nous avons vu que les équations du triangle permettent d'estimer l'écart de u à l'arbre, noté Lu. On a :

$$Lu := \frac{1}{2} \quad \text{Min}_{x \neq y \in X} [\Delta(u,x) + \Delta(u,y) - \Delta(x,y)].$$

Ce que l'on peut dire, si u et v sont frères, c'est que $\Delta(u,v) \leq Lu + Lv$. En fait, si u et v sont frères, on a certainement la topologie de la partie gauche de la Figure 3, mais on ne peut estimer que les chemins de la partie droite. Dans $Lu+Lv$ on compte deux fois l'arête qui sépare $\{u,v\}$ de $\{x,y\}$.

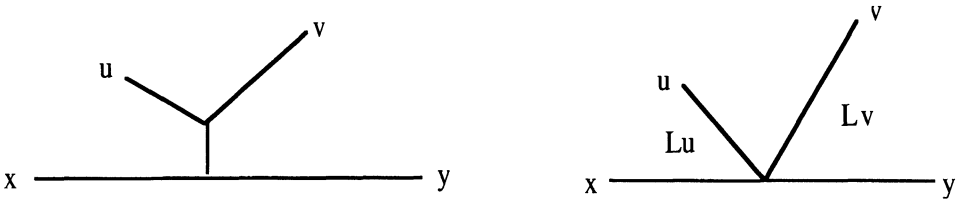

Figure 3

Si l'on veut que $Lu+Lv$ soit mesurée par rapport à la même paire (x,y), il faut qu'il n'y ait qu'une seule valeur manquante sur $\{x,y,u,v\}$. Suivant le nombre de valeurs inconnues, ce n'est pas toujours réalisable et donc nous procéderons à l'évaluation de Lu et Lv indépendamment l'un de l'autre. Ainsi, l'estimation de Lu est toujours possible si et seulement si il existe un triangle $\{x,y,u\}$ dont les trois valeurs de distances sont connues. Cette condition est un peu plus restrictive que celle de la Proposition 1, qui n'exige pas que $\Delta(x,y)$ soit définie, puisqu'elle est remplacée par $D(x,y)$, la longueur du chemin qui les sépare dans l'arbre courant.

Pour minimiser cette part excessive de distance entre u et v, on applique la formule

$$Duv := \frac{2}{3}(Lu + Lv)$$

dans laquelle 2/3 est fixé arbitrairement, en rapport avec le paramètre choisi lors des simulations qui veut que les arêtes internes soient deux fois plus courtes que les arêtes externes. De fait on traite de la même façon les "vrais" frères et les paires d'éléments dont la distance reste inconnue ; ce sont des "faux frères". Dans ce dernier cas, Duv est sous-estimée, comme le montre l'exemple ci-dessous.

Soit Δ une distance d'arbre représentée dans la Figure 4, dans laquelle $\Delta(u,v)$, $\Delta(u,z)$, $\Delta(v,y)$ sont les valeurs manquantes.

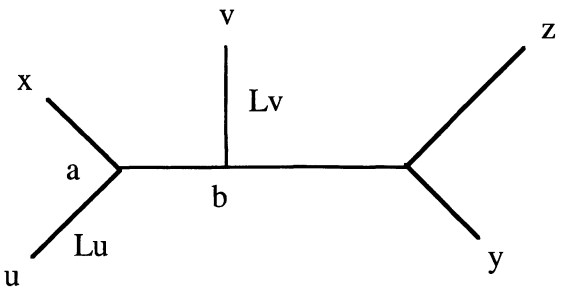

Figure 4

La valeur de $\Delta(u,v)$ ne peut être évaluée par quadruplet. Les éléments u et v sont donc des faux frères et l'on estime Lu par rapport au chemin $[x,y]$ et Lv par rapport à $[x,z]$, tous deux de façon correcte. Mais à l'évidence, dans Duv, il manque la longueur de l'arête (a,b).

2.3. ALGORITHME

```
Fact := -1 : NbIncon := 0
Pour tout (x,y)
    Si Δ(x,y) est inconnue Alors
        NbIncon := NbIncon + 1
        Dmin(x,y) := -1 ; Dmax(x,y) := + ∞
        NbP(x,y) := 0 ; /* Nombre de paires permettant d'évaluer Δ(x,y) */
Fin de Pour tout

/*Encadrement métrique*/
Pour tout (x,y) telle que Δ(x,y) est inconnue
    Pour tout z tel que Δ(z,x) et Δ(z,y) sont connues
        Min := | Δ(z,x) - Δ(z,y) |
        Si Min > Dmin(x,y) Alors Dmin(x,y) := Min
        Max := Δ(z,x) + Δ(z,y)
        Si Max < Dmax(x,y) Alors Dmax(x,y) := Max
    Fin Pour z
Fin de Pour (x,y)

/* Estimation du facteur */
Pour tout quadruplet {x,y,z,t}
    Si toutes les distances sont connues
        Calculer les 3 sommes Smin ≤ Smed ≤ Smax
        F := (Smax - Smed) / (Smax + Smed)
        Si F > Fact  Alors Fact := F
     Sinon si une seule valeur de distance Δ(x,y) est inconnue
        NbP(x,y) := NbP(x,y)+1
Fin de Pour {x,y,z,t}

Si Fact ≥ 0 Faire
    On ordonne les valeurs manquantes suivant l'ordre décroissant de NbP
    Répéter tant que NbIncon décroît
        Pour toute valeur manquante Δ(u,v)
            Duv := 0 ; Np := 0
            Pour toute paire (x,y)  telle que Δ(x,y) est connue
                Si Δ(x,u), Δ(x,v), Δ(y,u) et Δ(y,v) sont connues
                    A := Δ(x,u) + Δ(y,v)
                    B := Δ(x,v) + Δ(y,u)
                    Si  | A - B | / (A + B) ≥ Fact
                        Duv := Duv + Max {A,B} - Δ(x,y)
                        Np := Np + 1
            Fin pour (x,y)
        Si Np > 0 Alors
            Δ(u,v) := Duv / Np
            Si Δ(u,v) < Dmin(u,v) Alors Δ(u,v) := Dmin(u,v)
```

```
                Si Δ(u,v) > Dmax(u,v) Alors Δ(u,v) := Dmax(u,v)
                NbIncon := NbIncon - 1
            Fin de valeur manquante
        Fin de Répéter

Si NbIncon >0  /* Il ne reste que des frères */
        Pour tout élément z
            L(z) := + ∞
            Pour toute paire (x,y) telle que Δ(x,y), Δ(x,z) et Δ(y,z) sont connues
                Lz := Δ(x,z) + Δ(y,z) - Δ(x,y)
                If Lz ≤ L(z) Alors L(z) := Lz
            Fin Pour tout
        Fin de Pour
        Pour toute valeur manquante Δ(u,v)
            Δ(u,v) := (L(u) + L(v)) / 3
            Si Δ(u,v) < Dmin(u,v) Alors Δ(u,v) := Dmin(u,v)
            Si Δ(u,v) > Dmax(u,v) Alors Δ(u,v) := Dmax(u,v)
        Fin de valeurs manquantes
```

Il est clair que l'évaluation des valeurs inconnues est de complexité $O(n^4)$ en général et, si le nombre de valeurs manquantes est un paramètre m, à cause de leur mise en ordre en fonction du nombre décroissant des paires permettant de les estimer, l'évaluation est en $O(m[n^2+\log m])$.

3. SIMULATIONS

Nous nous sommes placés dans le cadre de la reconstruction à partir de distances partielles ; nous partons donc d'un tableau de distances. On commence par calculer une dissimilarité proche d'une distance d'arbre, pour que le modèle arboré soit justifié :

- On tire un X-arbre ($|X| = 20$) au hasard, noté T ; sa topologie tout d'abord, c'est-à-dire la liste de ses arêtes que l'on obtient par subdivisions successives de parties de X, jusqu'aux singletons. Ensuite on pondère les arêtes de façon aléatoire en donnant aux arêtes externes, celles qui aboutissent aux feuilles, des longueurs deux fois plus grandes que les arêtes internes – ce rapport correspond à des problèmes moyennement difficiles de reconstruction. Il suffit après d'établir la distance d'arbre D correspondante, par sommation des longueurs des arêtes le long de tous les chemins entre feuilles. Pour pouvoir interpréter l'écart quadratique, on norme D de façon que la valeur moyenne soit égale à 100.
- Ensuite on bruite cette distance selon un paramètre τ. Pour chaque valeur $D(x,y)$, on tire au hasard une valeur ε, telle que $0 \leq \varepsilon \leq \tau$, et on applique la formule

$$\Delta(x,y) := (1 \pm \varepsilon) * D(x,y)$$

  dans laquelle le $\pm$ est aussi le résultat d'un tirage aléatoire équiprobable ; la moitié des valeurs est augmentée, l'autre moitié diminuée d'un pourcentage borné par τ. Ainsi les variables de bruit sont indépendantes et d'espérance nulle.
- Enfin on considère un pourcentage ι de valeurs comme indéterminées. Les paires (u,v) correspondant à ces valeurs sont aussi tirées au hasard, en s'assurant que chaque élément de X a au moins 30 % de ses distances évaluées.

À partir de Δ on applique une méthode de reconstruction et on obtient un X-arbre θ, et la distance d'arbre D_θ associée, que l'on compare avec les données initiales. Longtemps on a considéré comme essentiel le point de vue *métrique*, et l'on pensait que l'arbre calculé était d'autant meilleur que l'écart quadratique entre D et D_θ était petit. De nos jours, on accorde plus d'importance à des critères *topologiques*, qui ne sont fonction que des structures des sous-arbres partiels à quatre feuilles et des bipartitions induites par les arêtes internes de l'arbre. De plus, à topologie fixée, il est facile d'ajuster au sens des moindres carrés les longueurs des arêtes, donc la distance d'arbre (Gascuel [1997]).

Par la suite, nous mesurons trois critères et donnons leurs valeurs moyennes sur 100 essais.

- l'écart quadratique moyen entre D et D_θ :

$$Eq := \frac{2}{n(n-1)} \sum_{x \neq y \in X} [\, D(x,y) - D_\theta(x,y)]^2.$$

  On notera que ce sont les distances d'arbre que l'on compare et que l'écart quadratique entre D et D_θ est généralement plus faible qu'entre Δ et D_θ. Il ne faut donc pas utiliser les valeurs moyennes indiquées ci-dessous comme l'écart moyen entre une dissimilarité donnée et une distance d'arbre.

- le pourcentage de quadruplets bien représentés, noté *quad.* Admettons que dans T on ait la topologie $xy|zt$. Si on a la même dans θ, on compte 1 point ; si on a la topologie non résolue, c'est-à-dire sans arête interne, on compte $\frac{1}{2}$ point. Si on a une autre topologie, on compte 0. Maintenant si dans T on a la topologie non résolue et que dans θ on a une topologie résolue quelconque, on compte $\frac{1}{2}$ point et si on trouve également la topologie non résolue, on compte 1 point. Donc ce n'est pas tout à fait un pourcentage mais, compte tenu de la façon de générer les distances, il y a très peu de quadruplets non résolus dans T, et les méthodes utilisées tendent à construire des arbres binaires.

- le pourcentage de bipartitions de T, dont aucune des classes n'est un singleton (non trivial splits), retrouvées dans θ. Ce sont celles qui correspondent aux arêtes internes des deux arbres. Comme les arbres sont binaires, ils ont tous les deux n-3 bipartitions de ce type. Le complément à 1 de ce pourcentage est la valeur de la distance de Robinson-Foulds [1981], ou distance de la différence symétrique entre les ensembles de bipartitions.

3.1. RÉSULTATS DE LA MÉTHODE SÉQUENTIELLE PARCIMONIEUSE

τ	ι	0 %	10 %	20 %	30 %	40 %	50 %
0 %	Eq	0	13	32	73	102	151
	quad	1	.99	.98	.95	.93	.89
	Bp	1	.94	.87	.76	.66	.54
5 %	Eq	18	42	54	96	129	189
	quad	1	.99	.97	.95	.91	.86
	Bp	1	.94	.86	.78	.66	.53

Eq	69	78	103	136	177	250
10 % quad	1	.98	.97	.94	.89	.83
Bp	.99	.93	.84	.74	.63	.50
Eq	141	145	175	216	263	315
15 % quad	.98	.97	.94	.90	.85	.79
Bp	.95	.89	.80	.70	.56	.44
Eq	210	230	260	292	335	387
20 % quad	.96	.94	.90	.88	.82	.76
Bp	.90	.83	.71	.63	.50	.39
Eq	302	319	339	365	435	464
25 % quad	.93	.91	.87	.82	.77	.72
Bp	.81	.73	.64	.53	.44	.34

Si l'on considère comme acceptables les reconstructions dans lesquelles 90 % des quadruplets sont correctement représentés, ainsi que 70 % des bipartitions obtenues, il ne faut jamais dépasser 30 % de valeurs inconnues si le taux de bruit est inférieur à 20 %. Pour un taux supérieur à 20 %, on tombe rapidement à 20 % puis 10 % de valeurs inconnues.

3.2. Résultats de la méthode NJ sur les distances complétées

Les valeurs manquantes de distance sont donc estimées par l'algorithme décrit ci-dessus. Quand le tableau est complet, on applique la méthode NJ qui était considérée, jusqu'à récemment, comme la plus efficace pour les reconstructions d'arbres en biologie moléculaire. Elle a également l'avantage d'être efficace, puisque, avec la variante de Studier & Keppler [1988], elle est en $O(n^3)$.

τ \ ι	0 %	10 %	20 %	30 %	40 %	50 %
Eq	0	0	2	4	11	19
0 % quad	1	1	1	1	.99	.99
Bp	1	1	.99	.98	.95	.91
Eq	2	3	10	17	44	116
5 % quad	1	1	.99	.99	.97	.91
Bp	1	.99	.97	.94	.86	.70
Eq	6	9	20	35	94	204
10 % quad	1	1	.99	.98	.94	.86
Bp	1	.98	.95	.90	.77	.59
Eq	13	23	43	76	162	345
15 % quad	1	.99	.98	.95	.89	.78
Bp	.99	.95	.92	.81	.65	.47
Eq	25	38	71	140	270	478
20 % quad	.99	.98	.96	.91	.84	.75
Bp	.98	.92	.84	.71	.55	.40
Eq	40	65	111	202	372	629
25 % quad	.99	.97	.93	.88	.79	.69
Bp	.94	.88	.77	.65	.47	.34

Soulignons tout d'abord que les évaluations des valeurs manquantes sont nettement meilleures que celles proposées par Lapointe et Kirsch [1996]. Nous avons calculé la moyenne des valeurs absolues des écarts entre les valeurs considérées comme inconnues et les valeurs estimées ; les nôtres sont de 3 à 6 fois plus faibles.

4. CONCLUSIONS

Nul n'est besoin de comparer les deux tableaux bien longtemps pour constater que la stratégie d'évaluation des valeurs inconnues suivie de NJ donne de meilleurs résultats que la méthode séquentielle. Cette dernière ne devient meilleure que pour les combinaisons extrêmes de bruit et d'incertitude, quand $\tau \geq .20$ et $\iota \geq .50$. Nous n'avons pas mesuré pour des valeurs supérieures, car de toute façon les résultats sont médiocres ; des erreurs sur les distances supérieures à 20 %, moins de 80 % de quadruplets correctement représentés et moins de 50 % de bipartitions retrouvées. Il devient alors illusoire de considérer l'arbre obtenu comme crédible.

On remarquera également que le critère métrique se comporte exactement comme les critères topologiques. Quand l'écart quadratique pour NJ est plus petit, les taux de quadruplets et de bipartitions sont plus élevés. Donc pour améliorer encore ces résultats, il faudrait calculer de meilleurs estimations.

Enfin, il pourrait être intéressant de détecter, préalablement aux évaluations, les vrais et les faux frères. Ceci pourrait être réalisé en utilisant la propriété de base de la méthode de Dispersion (Leclerc [1986]) à savoir que si u et v sont frères, pour toute paire {x,y} on devrait avoir

$$\Delta(u,x) - \Delta(v,x) \sim \Delta(u,y) - \Delta(v,y).$$

Enfin le lecteur attentif aura remarqué que nous n'utilisons pas la borne supérieure de $\Delta(u,v)$ offerte chaque fois que $A = B$; on a alors $\Delta(u,v) \leq \text{Max}\ \{A,B\} - \Delta(x,y)$. Nous avons testé cette possibilité, mais les simulations donnent de moins bons résultats quand on applique cette formule pour restreindre l'intervalle de variation de Δuv.

REMERCIEMENTS — Les auteurs tiennent à remercier Laurent Duret (LGBP, Lyon) et Bruno Leclerc (CAMS, Paris) avec lesquels ce travail a débuté, ainsi qu'Irène Charon (ENST, Paris) pour ses fructueuses suggestions lors d'une première présentation de ce travail au colloque ROADEF'99 à Autran.

BIBLIOGRAPHIE

BARTHÉLEMY, J.P., GUÉNOCHE, A., *Les arbres et les représentations des proximités*, Collection "Méthodes et Programmes", Masson, 1988, *Trees and Proximity Representations*, J. Wiley, 1991.

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

DE SOETE, G., "Ultrametric tree representations of incomplete dissimilarity data", *J. of Classification*, 1, (1984), 235-242.

DE SOETE, G., "Additive-tree representations of incomplete dissimilarity data", *Qual. Quantity,* 18, (1984), 387-393.

DURET, L., MOUCHIROUD, D., GOUY, M., "Hovergen : a database of homologous vertebrate genes", *Nucleic Acids Res.*, 22, (1994), 2360-2365.

GASCUEL, O., "Concerning the NJ Algorithm and its Unweighted Version, UNJ", *Mathematical Hierarchies and Biology*, B. Mirkin et al. (Eds.), DIMACS Series Discrete Mathematics and Theoretical Computer Science 37, AMS, (1997), 149-170.

GUÉNOCHE, A., LECLERC, B., "La méthode des triangles pour reconstruire un arbre à partir de distances incomplètes", *Actes des Journées de la Société Francophone de Classification*, Agro-Montpellier, (1998), 117-120.

GUÉNOCHE, A., LECLERC, B., "The triangles method to build phylogenetic trees from incomplete distance matrices", soumis à publication, (1998), 18 p.

LAPOINTE, F.J., KIRSCH, J.A.W., "Estimating phylogenies from lacunose distances matrices : Additive is superior to Ultrametric estimation", *Molecular Biology Evolution*, 13(6), (1996), 266-284.

LECLERC, B., "La méthode de dispersion", communication personnelle (1986). Voir Barthélemy, J.-P. & Guénoche, A., p. 74 (1988) & p. 73 (1991).

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

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

MAKARENKOV, V., LECLERC, B., "The fitting of a tree metric to a given dissimilarity with the weighted least squares crterion", *Journal of Classification*, (1999), 223-249.

ROBINSON, D.R., FOULDS, L.R., "Comparison of phylogenetic trees", *Mathematical Biosciences,* 53, (1981), 131-147.

SAITOU, N., NEI, M., "The neighbor-joining method: a new method for reconstructing phylogenetic trees", *Molecular Biology Evolution*, 4, (1987), 406-425.

STUDIER, J.A., KEPPLER, K.J., "A note on the neighbor-joining method of Saitou and Nei", *Molecular Biology Evolution* , 5, (1988), 729-731.

ZARETSKII, K., "Construction d'un arbre sur la base d'un ensemble de distances entre ses feuilles" (en russe), *Uspekhi Mat. Nauk.*, 20, (1965), 90-92.

WATERMAN, M.S., SMITH, T.F., SINGH, M., BEYER, W.A., "Additive Evolutionary Trees", *Journal of Theoretical Biology*, 64, (1977), 199-213.