SDM: Une Méthode de Distance Rapide pour les Etudes de Phylogénomique

Alexis Criscuolo, Vincent Berry, Emmanuel Douzery, Olivier Gascuel

SDM: une méthode de distance rapide pour les études de phylogénomique.

Alexis Criscuolo[1,2], Vincent Berry[2], Emmanuel J. P. Douzery[1] et Olivier Gascuel[2]

[1] Groupe Phylogénie Moléculaire. ISEM, Université Montpellier 2, CC 064, 34095 MONTPELLIER Cedex 05

[2] Equipe Méthodes et Algorithmes pour la Bioinformatique. LIRMM (CNRS, Université Montpellier 2), 161 rue Ada, 34392 Montpellier Cedex 05

Contact : criscuol@lirmm.fr

Résumé

*Les études de phylogénomique se proposent de reconstruire la phylogénie d'un ensemble de taxons en utilisant un grand nombre de gènes homologues. Les données, de tailles "génomiques", imposent des méthodes rapides. Dans un tel contexte, les méthodes de distance constituent une approche de choix, qu'il s'agisse de réaliser des études exploratoires, ou bien de construire un premier arbre qui sera raffiné ensuite par une approche plus lourde de type maximum de vraisemblance (ML). Néanmoins, une distance évolutive estimée directement à partir des gènes concaténés induit généralement un signal topologique perturbé. Nous proposons ici une nouvelle méthode, nommée Super Distance Matrix (*SDM*), consistant à combiner une collection de matrices de distances évolutives obtenues à partir de chaque gène en une seule supermatrice de distance. Cette supermatrice est ensuite utilisée pour reconstruire un arbre à l'aide d'une méthode de distance classique. Le principe consiste à déformer les matrices sources sans modifier leur message topologique, de manière à minimiser leur éloignement réciproque au sens de l'écart quadratique. Une difficulté est que les matrices sources correspondent à des ensembles de taxons qui ne sont que partiellement recouvrants. Nous montrons que ce problème s'exprime comme la minimisation d'un critère quadratique sous contraintes linéaires, ce qui revient à résoudre un système linéaire. La résolution de ce système creux a une complexité pratique de l'ordre de $n^a k^a$, où n représente le nombre de taxons, k le nombre de matrices et $a < 2$, ce qui permet d'obtenir très rapidement la supermatrice de distance souhaitée. Nous étudions les performances de* SDM *à l'aide de simulations. Plusieurs utilisations de* SDM *sont envisagées, de l'étude exploratoire rapide à des approches plus lourdes en temps calculs. Nous montrons que* SDM *constitue une alternative pertinente à la méthode standard "Matrix Representation with Parsimony" (*MRP*), en particulier lorsque les matrices sont peu recouvrantes. Nous montrons également que* SDM *construit un excellent arbre de départ pour une approche basée sur le critère ML, qui permet à la fois de réduire les temps calculs et de gagner en précision. Nous analysons à l'aide de* SDM *le jeu de données moléculaires de Gatesy et al. [14] composé de quarante-huit gènes et soixante-quinze mammifères. Les résultats inférés par SDM indiquent une très forte hétérogénéité des vitesses d'évolution dans cette collection de gènes et confirment les résultats théoriques obtenus par simulations.*

Mots clés : *phylogénomique, distances évolutives, superarbre, supermatrice, MRP, total evidence*

1 Introduction

Les études de phylogénomique, consistant à inférer des phylogénies en utilisant un grand nombre de gènes, est une tendance très actuelle bénificiant de l'accroissement de la quantité de gènes séquencés [7, 27, 8, 4, 14]. Une des principales difficultés de la phylogénomique est la nécessité de développer des méthodes rapides pour traiter ces grandes collections de gènes. Une deuxième difficulté est l'apparition de données manquantes lorsque l'on considère de tels jeux de données. Certains gènes ou certaines espèces sont en effet peu représentés dans les bases de données génomiques. Pour traiter ce problème, trois types de techniques de combinaison de gènes ont été proposés, nommés combinaisons basse, haute et moyenne ([35] ch. 7) :

– la combinaison basse consiste à concaténer les différents gènes afin d'obtenir un unique alignement sur lequel on applique une méthode standard de reconstruction phylogénétique. Cette technique, dite

de *total evidence*, est handicapée par l'apparition de données manquantes plus ou moins bien gérées suivant la méthode de reconstruction utilisée [18, 19, 41, 42]. De plus, les gènes évoluant sous des contraintes différentes, l'hétérogénéité des vitesses et des modes d'évolution représente une difficulté supplémentaire [28, 43].
– La combinaison haute amalgame en un seul arbre l'information topologique contenue dans l'ensemble des phylogénies inférées à partir de chaque gène. L'absence de certains gènes implique que les différentes phylogénies sont définies sur des ensembles partiellement recouvrants de taxons. Cette généralisation du consensus d'arbres est appelée problème du superarbre. Parmi les nombreuses solutions proposées, la méthode *Matrix Representation with Parsimony* (MRP) [2, 29] est l'une des plus utilisées. Elle consiste à encoder l'information topologique de chaque arbre en une matrice de caractères binaires. Le superarbre est alors l'arbre le plus parcimonieux inféré à partir de cette matrice.
– La combinaison moyenne consiste à passer par une étape d'interprétation des gènes qui ne soit pas les arbres eux-mêmes. La méthode TREE-PUZZLE [36], consistant à produire, à partir de chaque gène, toutes les topologies de quadruplet les plus vraisemblables avant de les combiner en un seul arbre, représente un bon exemple de ce type de combinaison.

Ces trois approches s'appuient principalement sur deux critères : le maximum de parcimonie (MP) et le maximum de vraisemblance (ML). Or, un des enjeux de la phylogénomique est la reconstruction de phylogénies définies sur un très grand nombre d'espèces à partir de très nombreux gènes. En conséquence, il est nécessaire de pouvoir traiter des données de plus en plus importantes par des méthodes ayant des temps d'exécution très rapides. Les méthodes de distance ont toujours été appréciées dans ce but. Si n est le nombre de taxons, les algorithmes NJ [37] et BIONJ [13] en $O(n^3)$ ainsi que FASTME [6], présentant une complexité de l'ordre de n^2 en pratique, sont parmi les plus rapides. Cette rapidité fait qu'ils sont fréquemment utilisés dans des études exploratoires. Ces méthodes servent aussi de point de départ à des méthodes optimisant des critères plus coûteux en temps d'exécution. Le programme PHYML [16] est un très bon exemple de ce mode opérationnel pour le critère ML.

Paradoxalement, très peu de méthodes de distance ont été utilisées dans le cadre de la combinaison de matrices de distance. Une première approche, la technique du *Average Consensus Supertree* (ACS) [20], a été proposée par Lapointe et Cucumel. Détaillée dans [21], elle s'applique pour la combinaison de seulement deux matrices sources. L'idée est de standardiser chaque matrice puis d'en faire la moyenne pour produire une supermatrice de distance. La standardisation consiste à diviser chaque matrice par la plus grande distance qu'elle contient ou bien par la plus grande distance commune à ces deux matrices. Lapointe et Levasseur suggèrent d'investir d'autres voies pour la combinaison de plus de deux matrices [21]. Notons également qu'il est toujours possible d'estimer directement une matrice de distance à partir de la supermatrice de caractères. Les données manquantes sont gérées en utilisant la solution MISSDIST=IGNORE de PAUP* [38] consistant, lorsqu'on estime la distance évolutive entre deux séquences, à ne considérer que les sites ne présentant aucun caractère manquant. Néanmoins, comme nous le verrons par la suite, cette technique, appelée *Distance-based Total Evidence* (DTE) fonctionne mal en général, principalement à cause de l'hétérogénéité des taux d'évolution entre gènes.

Nous proposons ici une nouvelle solution pour obtenir une supermatrice de distance représentant au mieux les distances évolutives et l'information topologique de vastes collections de matrices. Cette nouvelle méthode, appelée *Super Distance Matrix* (SDM), consiste à déformer les matrices sources sans modifier l'information topologique contenue dans chacune d'elles, de manière à minimiser leur éloignement réciproque. Plusieurs simulations montrent que cette méthode présente un très bon compromis entre fiabilité et rapidité pour inférer une phylogénie à partir d'une collection de gènes. Nous montrons également que SDM, initialement présentée comme méthode de combinaison moyenne, est une alternative efficace dans différents scénarios de combinaison haute et basse.

Dans la suite, nous décrivons cette nouvelle méthode (partie 2). Nous la comparons à d'autres techniques de combinaison de gènes à l'aide de simulations (partie 3). Nous concluons (partie 4) par une étude phylogénétique des mammifères placentaires sur un jeu de données moléculaires qui a été utilisé dans le cadre de la combinaison basse par Gatesy *et al.* [14].

2 La méthode SDM

2.1 Notations et définitions

Soit $\mathcal{C} = \{(\Delta_{ij}^1), (\Delta_{ij}^2), ..., (\Delta_{ij}^p), ..., (\Delta_{ij}^k)\}$ une collection de k matrices de distance. La valeur Δ_{ij}^p est la distance évolutive entre les taxons i et j pour le gène p. On note $\mathcal{L}_p$ l'ensemble des taxons définissant la matrice de distance (Δ_{ij}^p), n_p le nombre de taxons contenus dans $\mathcal{L}_p$, n le nombre de taxons contenus dans $\mathcal{L} = \bigcup_p \mathcal{L}_p$ et k_{ij} le nombre d'apparitions de la paire de taxons ij dans la collection $\mathcal{C}$, *c.-à-d.* $|\{p : \{i, j\} \subset \mathcal{L}_p\}|$. On pose également :

$$\tilde{\mathcal{L}}_p = \{i \in \mathcal{L}_p : \exists j \in \mathcal{L}_p - \{i\}\ ,\ k_{ij} > 1\}\ , \qquad \tilde{n}_p = \left|\tilde{\mathcal{L}}_p\right|\ , \qquad \tilde{\mathcal{L}} = \bigcup_{1 \leq p \leq k} \tilde{\mathcal{L}}_p \qquad \text{et} \qquad \tilde{n} = \left|\tilde{\mathcal{L}}\right|.$$

Notre méthode consiste à rapprocher (au sens des écarts quadratiques) chaque matrice (Δ_{ij}^p) des autres sans déformer l'information topologique qu'elle contient. Les taxons de $\tilde{\mathcal{L}}$ sont ceux sur lesquels on s'appuie pour comparer les différentes distances communes à plusieurs matrices.

2.2 Méthode

Soit (Δ_{ij}^p) une matrice de distance additive (*c.-à-d.* équivalente à un arbre valué de topologie T^p). Il est connu ([1] p. 218) que la multiplication par un facteur $\alpha_p > 0$ pour obtenir la nouvelle matrice $(\alpha_p \Delta_{ij}^p)$ ne modifie pas la topologie T^p. Cette opération revient à multiplier chaque longueur de branche de T^p par α_p. D'une manière similaire, on peut également montrer que l'ajout d'une constante $a_x^p \geq 0$ à chacune des $n_p - 1$ distances non-nulles correspondant au taxon x ne modifie pas non plus la topologie T^p. Cette opération est équivalente à allonger d'une longueur a_x^p la branche externe correspondant au taxon x. Elle peut se généraliser sur l'ensemble des taxons pour obtenir la nouvelle matrice $(\Delta_{ij}^p + a_i^p + a_j^p)$.
Si la matrice (Δ_{ij}^p) n'est pas additive, il faut avoir recours à des algorithmes de reconstruction phylogénétique pour inférer une topologie T^p. Un des algorithmes les plus connus, l'algorithme NJ [33, 37], possède la propriété d'inférer la même topologie T^p à partir de la matrice (Δ_{ij}^p) et de la matrice déformée $(\alpha_p \Delta_{ij}^p + a_i^p + a_j^p)$ [12]. On montre facilement que les algorithmes implémentés dans FASTME [6] possèdent également cette propriété. Les algorithmes de moindres carrés, *e.g.* FITCH du package PHYLIP [11] et MW [26] du logiciel TREX [24], sont quant à eux très peu sensibles à la multiplication par un facteur ou à l'ajout de constantes.
Les différentes standardisations décrites dans [21] montrent que ACS utilise la propriété d'invariance à la multiplication d'une matrice par un facteur positif. La méthode SDM utilise aussi la propriété d'invariance des différentes matrices sources à l'ajout de constantes. Elle cherche à déformer les k matrices (Δ_{ij}^p) en les multipliant par un facteur α_p et en ajoutant des constantes a_i^p afin de minimiser leur éloignement réciproque. Pour cela, contrairement à ACS, elle s'appuie sur l'ensemble des distances communes entre les matrices de $\mathcal{C}$ pour lesquelles on a $k_{ij} > 1$. Ainsi, pour chaque paire ij, on calcule les valeurs optimales α_p et a_i^p permettant de minimiser le terme de variance :

$$V_{ij} = \sum_{\substack{1 \leq p \leq k \\ \mathcal{L}_p \supset \{ij\}}} w_p (\alpha_p \Delta_{ij}^p + a_i^p + a_j^p - \overline{\Delta}_{ij})^2 \tag{1}$$

où $\overline{\Delta}_{ij}$ est la moyenne pondérée sur chaque matrice $p = 1, 2, ..., k$ des différentes valeurs déformées :

$$\overline{\Delta}_{ij} = \frac{1}{W_{ij}} \sum_{\substack{1 \leq p \leq k \\ \mathcal{L}_p \supset \{ij\}}} w_p (\alpha_p \Delta_{ij}^p + a_i^p + a_j^p) \qquad \text{avec} \qquad W_{ij} = \sum_{\substack{1 \leq p \leq k \\ \mathcal{L}_p \supset \{ij\}}} w_p. \tag{2}$$

La pondération de chaque matrice par un coefficient w_p permet d'appliquer une valeur de confiance aux données sources. Typiquement, la variance d'une distance étant d'autant plus petite qu'elle est inférée à partir de séquences longues, on peut associer à w_p la longueur ℓ_p de la séquence à partir de laquelle la matrice (Δ_{ij}^p) a été inférée. On peut aussi donner à chaque gène un poids identique et indépendant du nombre $\tilde{n}_p$ de taxons qu'il contient en utilisant comme pondération $\ell_p / (\tilde{n}_p\ (\tilde{n}_p - 1))$. Une solution

intermédiaire consiste à utiliser la pondération $\ell_p/\tilde{n}_p$. On pourrait aussi facilement étendre SDM en utilisant comme pondération les variances (dépendantes de la paire ij) des distances de chaque matrice source.
L'opération de minimisation de V_{ij} s'appliquant pour toutes les paires de taxons $ij \in \tilde{\mathcal{L}}$, la méthode SDM consiste donc à minimiser la somme de variances :

$$\sum_{\substack{i \neq j \\ k_{ij} > 1}} V_{ij}. \tag{3}$$

Différentes contraintes linéaires sur ces variables sont associées à la minimisation du critère (3). Les facteurs α_p cherchent à compenser la vitesse d'évolution propre à chaque gène de manière similaire aux *gene-specific rate models* suggérés par Yang [43]. La contrainte (4), identique à celle du *proportional model* de Pupko *et al.* [28], force les coefficients α_p à être égaux à 1 en moyenne :

$$\sum_{1 \leq p \leq k} \alpha_p = k. \tag{4}$$

Elle permet de standardiser les matrices sources et évite la solution triviale $\alpha_p = a_i^p = 0,\ \forall p,\ \forall i$.
Les branches externes d'une phylogénie sont généralement plus longues que les branches internes, et l'essentiel de la variance d'une distance entre deux espèces est supportée par les deux branches externes correspondantes. De plus, Lapointe et Levasseur ont remarqué qu'une forte hétérogénéité dans les longueurs de branches et dans les vitesses d'évolution de chaque donnée source altérait la qualité d'inférence topologique obtenue via ACS [21]. Les différentes variables a_i^p cherchent donc à harmoniser les différentes longueurs de branches externes. La contrainte (5) force, pour chaque taxon i, la somme des différents a_i^p à être nulle et interdit un trop grand allongement des longueurs de branches externes correspondant au taxon i :

$$\sum_{p : i \in \tilde{\mathcal{L}}_p} a_i^p = 0,\ \forall i \in \tilde{\mathcal{L}}. \tag{5}$$

La contrainte (6) oblige l'annulation de la somme des a_i^p pour chaque matrice (Δ_{ij}^p) et empêche l'*étoilisation* et l'étouffement du signal topologique par un allongement global de toutes les branches externes d'un arbre p par rapport aux branches internes :

$$\sum_{i \in \tilde{\mathcal{L}}_p} a_i^p = 0,\ \forall p = 1, 2, ..., k-1. \tag{6}$$

La contrainte $\sum_{i \in \tilde{\mathcal{L}}_k} a_i^k = 0$ est inutile car elle induit une dépendance linéaire entre contraintes et, par conséquent, une infinité de solutions.
La minimisation du critère (3) revient à en calculer la dérivée partielle pour chacune des $k + \sum_{1 \leq p \leq k} \tilde{n}_p$ variables α_p et a_i^p. En associant un multiplicateur de Lagrange à chacune des $\tilde{n} + k - 1$ contraintes linéaires, on obtient un système linéaire défini sur $O(\tilde{n}k)$ variables. La résolution de ce système s'effectue en $O(\tilde{n}^3 k^3)$, ce qui est théoriquement équivalent au temps d'exécution de l'algorithme NJ sur une matrice de $\tilde{n}k$ taxons. Néanmoins, le système linéaire étant très creux, la complexité pratique de l'étape de résolution est bien plus rapide en utilisant une librairie adéquate (la librairie MTJ, disponible à l'URL `https://mtj.dev.java.net/`, a été choisie pour notre implémentation). Nous avons généré une centaine de collections de matrices avec $k = 4, 8, 12, 16$ et $n \in [50, 500]$ et mesuré le temps t d'exécution de SDM. Considérant que t est de la forme $b(nk)^a$, nous avons estimé la droite de régression $\log(t) = a \log(nk) + \log(b)$. Nous avons observé que la puissance a est toujours inférieure à 2. Ainsi, en pratique, la méthode SDM peut être considérée comme une méthode quadratique. Elle est à même de traiter des données correspondant à de très grandes valeurs nk. Par exemple, quelques minutes ont suffi pour des matrices générées avec $n = 500$ et $k = 20$, sur un PC Pentium IV 1.8Go (1Go RAM).

2.3 Reconstruction phylogénétique

On obtient une phylogénie en appliquant directement une méthode de distance sur la supermatrice de distance SDM. Néanmoins, comme pour ACS, des distances manquantes peuvent y apparaître suivant le taux de recouvrement des taxons des matrices sources. Deux types de méthodes sont utilisés pour inférer un arbre à partir d'une telle supermatrice de distance.
La méthode indirecte consiste à estimer les valeurs manquantes à l'aide d'un algorithme de complétion ultramétrique, additive [5] ou par quadruplets [17]. Le logiciel TREX propose plusieurs algorithmes de complétion avant d'inférer un arbre avec l'algorithme NJ.
La méthode directe consiste à appliquer un algorithme de moindres carrés pondérés en associant une variance infinie aux distances absentes ([39] p.449). L'algorithme MWMODIF du logiciel TREX ou le logiciel FITCH du package PHYLIP permettent d'appliquer cette méthode.
Une combinaison des deux est possible avec l'algorithme MW* [25] du logiciel TREX qui applique un algorithme de complétion ultramétrique ou additive selon la densité des valeurs manquantes puis infère un arbre par la méthode de moindres carrés pondérés MW [26].
De telles supermatrices de distance incomplètes sont tout de même rarement obtenues à partir des collections de gènes couramment étudiées (par ex. [4, 14, 23]). En effet, certains gènes sont séquencés pour un très grand nombre d'espèces et impliquent l'existence d'au moins une distance entre chaque paire de taxons, tels le cytochrome *b* [14, 23] ou les gènes des ARN ribosomiques 12S et 16S [23]. Dans ce cas, la supermatrice ne présente pas de distances manquantes et on peut appliquer directement une méthode rapide comme NJ ou FASTME.

3 Simulations

Différentes simulations ont été effectuées afin d'évaluer les performances de SDM. Le but de ces simulations est de comparer la capacité d'inférence de la bonne topologie et le temps d'exécution des techniques standards de combinaison basse, moyenne et haute. Dans les trois cas, un scénario impliquant l'utilisation de SDM est proposé. Nous décrivons dans ce qui suit les processus de génération des données puis d'inférence et de comparaison des phylogénies. Chaque processus a été répété 500 fois pour chaque valeur de $k = 2, 4, 6, ..., 20$.

Génération des arbres modèles Un arbre modèle défini sur $n = 48$ taxons a été généré à l'aide du logiciel R8S [34] suivant le processus de Yule-Harding. L'arbre modèle ultramétrique enraciné UT ainsi obtenu respecte l'hypothèse de l'horloge moléculaire. Une déviation est créée par rapport à cette hypothèse en multipliant chaque branche par $(1 + X)$ où X suit une loi exponentielle de paramètre $\mu = 0.2/(0.001 + U)$, U suivant une loi uniforme [16]. Soit tbl la longueur totale de cet arbre. On obtient l'arbre non-ultramétrique T de longueur totale 1 en divisant chaque longueur de branche par tbl. On génère, à partir de T, k arbres T^p en multipliant chaque branche par $0.4 + 8.6V_p$ où V_p suit une loi uniforme. La valeur de V_p est propre à chaque arbre T^p. Les k arbres sources T^p possèdent ainsi la même topologie que l'arbre T et un taux d'évolution propre variant uniformément entre 0.4 et 9.0, ce qui induit des taux relatifs allant de 1 à 20 ($\approx 9.0/0.4$) dans les cas extrêmes.

Génération des données Une collection de k alignements de $n = 48$ séquences sur ℓ_p sites a été générée à l'aide du logiciel SEQ-GEN [30] à partir de chaque arbre T^p suivant le modèle K2P. La valeur de ℓ_p est tirée uniformément entre 200 et 1000 sites et est propre à chaque arbre T^p.
Pour chaque alignement, certains taxons ont été aléatoirement supprimés avec une probabilité de 25% ou 75% de délétion [10]. Un recouvrement d'au moins quatre taxons entre chaque paire de matrices a été néanmoins préservé afin de conserver une histoire évolutive et une information topologique significatives communes entre chacun des k gènes.

Inférence des phylogénies et des superarbres Nous avons utilisé la méthode de combinaison moyenne SDM afin d'inférer un arbre à partir de la collection de gènes ainsi générés. Nous avons aussi appliqué les techniques standards de combinaison haute et basse. Dans ces deux derniers types de combinaison,

nous avons également utilisé la méthode SDM afin d'observer ses performances face aux techniques standards.

• *Combinaison moyenne :*

A partir des k sous-alignements, une collection $\mathcal{C}$ de k matrices de distance a été estimée suivant le modèle K2P. Une supermatrice de distance SDM a été calculée à partir de la collection $\mathcal{C}$. Chaque matrice (Δ_{ij}^p) a été pondérée dans l'équation (1) par la taille ℓ_p des séquences correspondantes. Une phylogénie a ensuite été inférée avec le programme FITCH, à même de traiter les valeurs manquantes dans la supermatrice SDM. Nous appellerons SDM+FITCH ce scénario de reconstruction phylogénétique.

• *Combinaison basse :*

Une supermatrice de caractères a été construite par concaténation des k gènes partiellement délétés. Une phylogénie a été inférée à partir de cette supermatrice par le logiciel PHYML suivant le modèle K2P. Ce logiciel, en effectuant des réarrangements topologiques à partir d'un arbre de départ, recherche la phylogénie optimisant le critère ML. Nous appellerons BIONJ+PHYML le scénario consistant à laisser l'option standard inférant un arbre de départ par l'algorithme BIONJ appliqué sur la matrice de distance calculée directement suivant la procédure DTE. Le scénario SDM+FITCH+PHYML consiste à utiliser l'arbre SDM+FITCH comme arbre de départ.

• *Combinaison haute :*

Une collection de k phylogénies ML a été inférée à partir des k gènes partiellement délétés, à l'aide du logiciel PHYML. Nous avons combiné ces arbres suivant la technique standard MRP en construisant une matrice d'encodage binaire des différentes topologies. Les arbres MP ont été inférés par le logiciel TNT [15], connu pour sa rapidité, à partir de cette matrice binaire. L'approche standard considère le superarbre MRP comme étant le consensus strict de tous les arbres les plus parcimonieux retrouvés. Nous appellerons PHYML+MRP ce scénario de construction de superarbre.

La collection de k phylogénies étant composée d'arbres valués, elle est équivalente à une collection de matrices additives (*c.-à-d.* chaque distance d'une matrice (Δ_{ij}^p) est égale à la somme des longueurs des branches formant le chemin entre chaque paire de taxons dans l'arbre p). Cette collection de matrices de distance a été traitée suivant le même processus que SDM+FITCH. Nous appellerons PHYML+SDM+FITCH ce scénario de construction de superarbre.

Comparaison des phylogénies synthétiques et des superarbres avec les arbres modèles Nous avons comparé les performances des différents scénarios d'inférence en utilisant la distance de quadruplets d_q [9] comme critère topologique entre les phylogénies inférées $\hat{T}$ et l'arbre modèle T. Cette distance a été choisie pour sa finesse, en comparaison avec la distance plus classique dite de *Robinson & Foulds* [3, 31]. Elle mesure le nombre de 4-arbres (*c.-à-d.* arbres définis sur 4 feuilles) résolus qui sont présents dans un arbre mais pas dans l'autre. Nous l'avons normalisée par $2\mathrm{C}_n^4$ afin de la situer dans l'intervalle $[0, 1]$, où 0 correspond à deux arbres de même topologie. Cette distance d_q est la somme de deux types d'erreur : l'erreur de type 1 (e_1) correspondant aux 4-arbres résolus inférés dans $\hat{T}$ non-existants dans T et l'erreur de type 2 (e_2) correspondant aux 4-arbres résolus de T non-inférés dans $\hat{T}$. Typiquement, la présence de noeuds internes non-résolus dans $\hat{T}$ augmente la valeur de e_2 et diminue celle de e_1.

Pour chaque arbre inféré, toutes les branches de longueur inférieure à 0.0001 ont été transformées en multifurcation.

Résultats Pour chaque $k = 2, 4, 6, ..., 20$ et chacun des cinq scénarios d'inférence, la moyenne des 500 valeurs d_q obtenues a été calculée et reproduite sous forme de graphiques dans la Figure 1. Ces graphiques représentent les valeurs d_q moyennes observées en fonction de k pour 25% et 75% de délétion.

Les temps d'exécution moyens (en secondes sur un PC Pentium IV 1.8Go et 1Go RAM) des principaux programmes utilisés lors des simulations sont reportés dans le Tableau 1 pour $k = 2, 10, 20$ et pour 25% et 75% de délétion. Dix jeux de données ont été générés avec $n = 96$ taxons suivant le processus décrit précédemment, avec les mêmes valeurs de k et les mêmes taux de délétion. Chaque scénario d'inférence a été appliqué sur ces données afin de mesurer les temps d'exécution moyens qui sont reportés dans le Tableau 1.

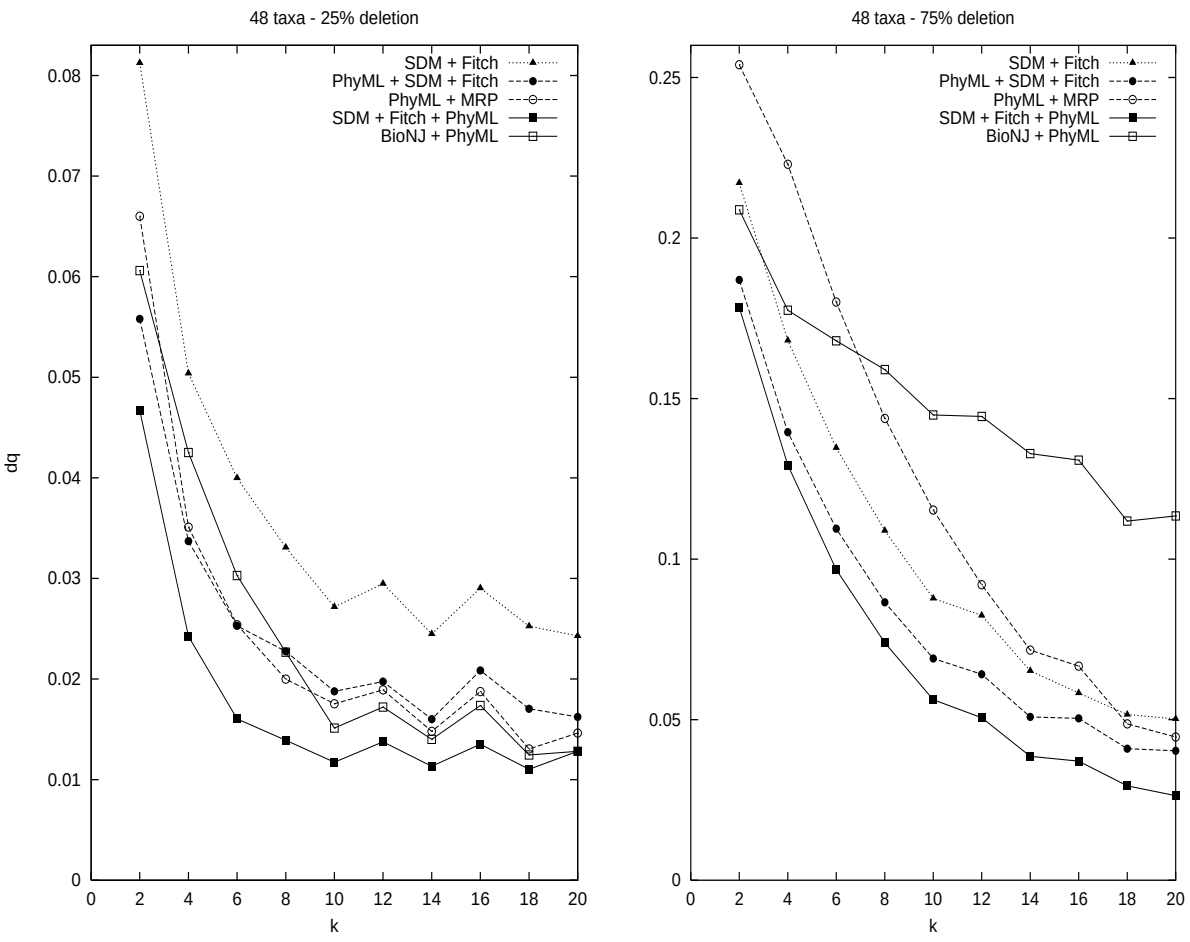

FIG. 1 – **Performance des différentes méthodes d'inférence pour 25% et 75% de délétions des taxons.** *k : nombre de gènes utilisés dans la reconstruction. d_q : distance de quadruplets entre l'arbre modèle et l'arbre inféré. Triangles : combinaison moyenne. Cercles : combinaison haute. Carrés : combinaison basse. A noter la différence d'échelle en ordonnée des deux graphiques.*

4 Discussion

4.1 Précision topologique

Comme on s'y attend, toutes les courbes des graphiques de la Figure 1 sont décroissantes. Les arbres modèles T sont d'autant mieux retrouvés (*i. e.* leur distance d_q avec $\hat{T}$ diminue) que le nombre k de matrices sources augmente. Comme attendu aussi, les phylogénies et les superarbres inférés sont d'autant plus proches des arbres modèles T que le taux de délétion des taxons est faible.

La méthode de combinaison moyenne SDM+FITCH correspond aux valeurs d_q les plus élevées pour 25% de délétion. Par contre, pour $k < 20$ et 75% de délétion, elle présente des valeurs d_q plus petites que celles correspondant à la technique standard de combinaison haute PHYML+MRP. Néanmoins, ce phénomène s'atténue avec l'augmentation du nombre de gènes.

Dans le cadre de la combinaison haute, la méthode PHYML+SDM+FITCH présente des résultats très proches de ceux de PHYML+MRP pour le taux de délétion de 25% et meilleurs pour 75%. En effet, les superarbres PHYML+MRP présentent des erreurs e_2 toujours beaucoup plus importantes que les erreurs e_1. Ceci est dû au fait que la méthode MRP renvoie le consensus strict des arbres les plus parcimonieux obtenus à partir de la matrice binaire d'encodage des arbres sources. Ce consensus strict implique l'apparition de multifurcations dans le superarbre, d'autant plus nombreuses que le taux de recouvrement des taxons des arbres sources est faible. Par exemple, dans nos simulations, pour $k = 10$ et 75% de délétion, les superarbres inférés par PHYML+MRP contiennent en moyenne 17% de quadruplets irrésolus.

								total evidence						MRP					
		SDM			FITCH			PHYML (*départ* BIONJ)			PHYML (*départ* SDM+FITCH)			$k \times$ PHYML			TNT		
	k	2	10	20	2	10	20	2	10	20	2	10	20	2	10	20	2	10	20
48 taxons	25%	<1	2	9	18	23	23	186	901	1223	89	430	808	38	143	267	4	12	23
	75%	<1	<1	2	1	17	23	45	2474	4104	21	988	2134	6	37	86	1	7	15
96 taxons	25%	1	10	48	335	497	491	432	1248	2108	405	846	1275	43	230	451	12	38	71
	75%	<1	1	4	20	348	486	134	3162	6325	93	1877	3796	11	50	101	3	24	46

TAB. 1 – **Temps d'exécution des différentes méthodes.** *Les valeurs correspondent au temps moyen d'exécution en secondes. Les temps de calcul des matrices de distance et de caractères ne sont pas inclus car négligeables.*

Pour chaque valeur de k et chaque taux de délétion, la méthode de combinaison basse BIONJ+PHYML présente des valeurs d_q moyennes plus élevées que la méthode SDM+FITCH+PHYML. Une matrice de distance calculée à partir d'une supermatrice de caractères est souvent peu représentative du signal phylogénétique à cause de l'hétérogénéité des vitesses d'évolution de chaque gène ainsi que l'existence de données manquantes. En effet, lorsque deux taxons partagent des gènes lents, ils sont estimés proches, alors que si leurs gènes communs sont rapides, ils sont prédits éloignés. La faiblesse de cette matrice de distance explique les mauvaises performances de BIONJ pour inférer un arbre à partir d'une supermatrice de caractères et implique les mauvais résultats de la méthode BIONJ+PHYML. En effet, lorsque l'arbre de départ dans PHYML est trop éloigné de l'arbre modèle, la phylogénie renvoyée peut seulement correspondre à un optimum local. Les arbres SDM+FITCH présentant de très bonnes valeurs d_q moyennes (par exemple, les arbres BIONJ construits à partir de matrices de distance DTE présente des valeurs d_q moyennes de 0.1 pour $k = 10$ et 25% de délétion et autour de 0.5 pour toutes les valeurs de k et 75% de délétion), ils se révèlent être d'excellents points de départ au sein du scénario SDM+FITCH+PHYML.

4.2 Temps d'exécution

Les temps d'exécution moyens du Tableau 1 montrent que la méthode de combinaison moyenne SDM+FITCH est parmi les plus rapides. Par exemple, avec 96 taxons, $k = 20$ et 25% de délétion, le scénario d'inférence SDM+FITCH dure 539s, BIONJ+PHYML nécessite 3383s et PHYML+MRP nécessite 522s. Néanmoins, des méthodes de distance plus rapides peuvent être utilisées sur des supermatrices de distance ne présentant aucune valeur manquante. Ainsi, dans le cadre des 25% de délétion sur nos jeux de données simulées, toutes les supermatrices de distance sont complètes pour $k > 14$. Nous leur avons donc appliqué l'algorithme FASTME qui a permis d'obtenir des résultats similaires à FITCH en un temps toujours inférieur à une seconde. Ainsi, avec 96 taxons, $k = 20$ et 25% de délétion, le scénario d'inférence SDM+FASTME nécessite 50s environ, contre 539 lorsque l'on utilise FITCH. Dans ce cas biologiquement fréquent (*cf.* Partie 5), SDM+FASTME est de loin le scénario d'inférence le plus rapide, d'un facteur 10 à 100 avec 96 taxons, et ce facteur augmente avec le nombre de taxons.
D'autre part, comme les arbres inférés par SDM+FITCH sont proches de l'arbre modèle T, on constate une nette amélioration du temps de calcul du logiciel PHYML quand on l'exécute avec SDM+FITCH comme arbre de départ. Ainsi, avec 96 taxons, $k = 20$ et 75% de délétion, le scénario d'inférence SDM+FITCH+PHYML nécessite 4286s alors que 6326s sont nécessaires à BIONJ+PHYML, soit un gain relatif de 50% environ.

5 Applications

Dans le but d'illustrer les performances de la méthode SDM, nous avons étudié un jeu de données déjà utilisé dans un contexte de combinaison basse par Gatesy *et al.* [14]. Nous montrons que les bonnes performances de SDM en termes de rapidité et de recouvrement topologique en font un outil efficace pour des études phylogénétiques sur un jeu de données réelles de grande taille. Nous montrons également que dans cette collection, les gènes présentent une très grande hétérogénité de taux d'évolution, ce qui rend une approche de type SDM indispensable dès que l'on veut obtenir des distances évolutives qui fassent sens.

5.1 Inférence d'arbres

Ce jeu de données était composé à l'origine de cinquante-sept sources de caractères comprenant trois ensembles de données morphologiques et les séquences de cinq protéines, un transposon, trente-trois gènes nucléaires et quinze gènes d'ADN mitochondrial. Pour notre étude, nous n'avons conservé que les séquences d'ADN. Nous avons donc considéré un jeu de données de quarante-huit gènes définis sur 37018 sites et soixante-quinze mammifères placentaires dont sept Afrothériens qui ont servi de groupe externe pour l'enracinement. Comme illustré dans [14], cette collection de gènes présente une forte hétérogénéité de l'échantillonage taxonomique et correspond à un taux de 68% de caractères manquants.
Nous avons appliqué les cinq scénarios d'inférence décrits dans nos simulations sur cette collection de gènes. Néanmoins, dans le scénatio de combinaison moyenne, nous avons estimé les matrices de distance suivant le modèle GTR [32] et nous avons choisi de pondérer le terme de variance (1) de la méthode SDM par $\ell_p/\tilde{n}_p$ afin de compenser l'hétérogénéité des tailles des différentes matrices de distance. De plus, le gène du cytochrome *b* étant présent pour l'ensemble des soixante-quinze taxons, la supermatrice de distance inférée par SDM ne contient pas de distance manquante et nous avons appliqué le programme FASTME à la place de FITCH. Dans les scénarios de combinaison basse et haute, nous avons appliqué le programme PHYML suivant le modèle GTR avec huit catégories de vitesse de substitution ainsi qu'une proportion de site invariables et un paramètre de loi gamma estimés.
Afin de mesurer la variabilité des résultats, nous avons généré dix jeux de données analogues par un processus de bootstrap non-paramétrique sur chacun des quarante-huit gènes. Nous avons appliqué les cinq scénarios d'inférence sur ces dix collections de gènes puis estimé la vraisemblance de chaque arbre à l'aide du logiciel PHYML afin d'obtenir d'autres points de mesure.
Nous avons également appliqué l'algorithme BIONJ sur la matrice de distance estimée suivant la procédure DTE en utilisant le modèle GTR à partir de la supermatrice de caractères initiale et des dix réplicats. Nous avons estimé la vraisemblance de ces onze phylogénies afin de vérifier que ce schéma d'inférence est bien pénalisé par la mauvaise qualité de la matrice de distance initiale, en comparaison avec l'approche SDM.

5.2 Résultats

A partir de la collection de gènes initiale, nous avons constaté que les scénarios SDM+FASTME+PHYML et BIONJ+PHYML ont permis d'inférer le même arbre en des temps d'exécution respectifs d'environ 6h et 6h30. Cette phylogénie est représentée dans la Figure 2 et sa vraisemblance est de -333106. Cet arbre est partiellement en accord avec la topologie retrouvée par Gatesy *et al.* (de vraisemblance -333244) mais présente une autre interprétation des branchements de certains grands groupes. En effet, nous trouvons que l'ensemble Pholidotes + Carnivores forme un clade proche parent des Périssodactyles et que l'ensemble Camélidés + Tayassuidés + Suidés forme un groupe paraphylétique. La position basale des Camélidés au sein des Cétartiodactyles, suivie par le groupe Suidés + Tayassuidés, a déjà été proposée et discutée par Madsen *et al.* [22] et Waddell *et al.* [40]. De plus, les embranchements correspondant ayant un faible support de bootstrap dans la topologie retrouvée par Gatesy *et al.*, l'arbre de la Figure 2 constitue donc une alternative vraisemblable (biologiquement et mathématiquement).
Le deuxième meilleur arbre a été inféré en environ 1h20 par le scénario PHYML+MRP et présente une vraisemblance de -333182. Pour un temps d'exécution similaire, l'arbre obtenu par le scénario PHYML+SDM+FASTME présente une vraisemblance de -333275.
L'arbre obtenu par SDM+FASTME présente une vraisemblance de -333380. Le temps d'exécution de ce scénario a été d'environ 30s. La méthode SDM a permis d'obtenir des valeurs α_p variant de 0.26 à 2.88. Ils présentent une médiane de 0.83 et des valeurs de quartiles de 0.56 et 1.23. Les paramètres α_p étant inversement proportionnels à la vitesse d'évolution, ce jeu de données est donc composé de gènes avec des vitesses d'évolution très hétérogènes (par exemple, $\alpha_p = 2.88$ correspond au gène le plus lent ZFX et $\alpha_p = 0.26$ au gène le plus rapide ATP8). On s'attend donc à ce que la matrice de distance estimée dans le scénario DTE contienne une information topologique très perturbée. En effet, l'arbre obtenu par DTE (inféré en quelques secondes) présente une vraisemblance de -336467.

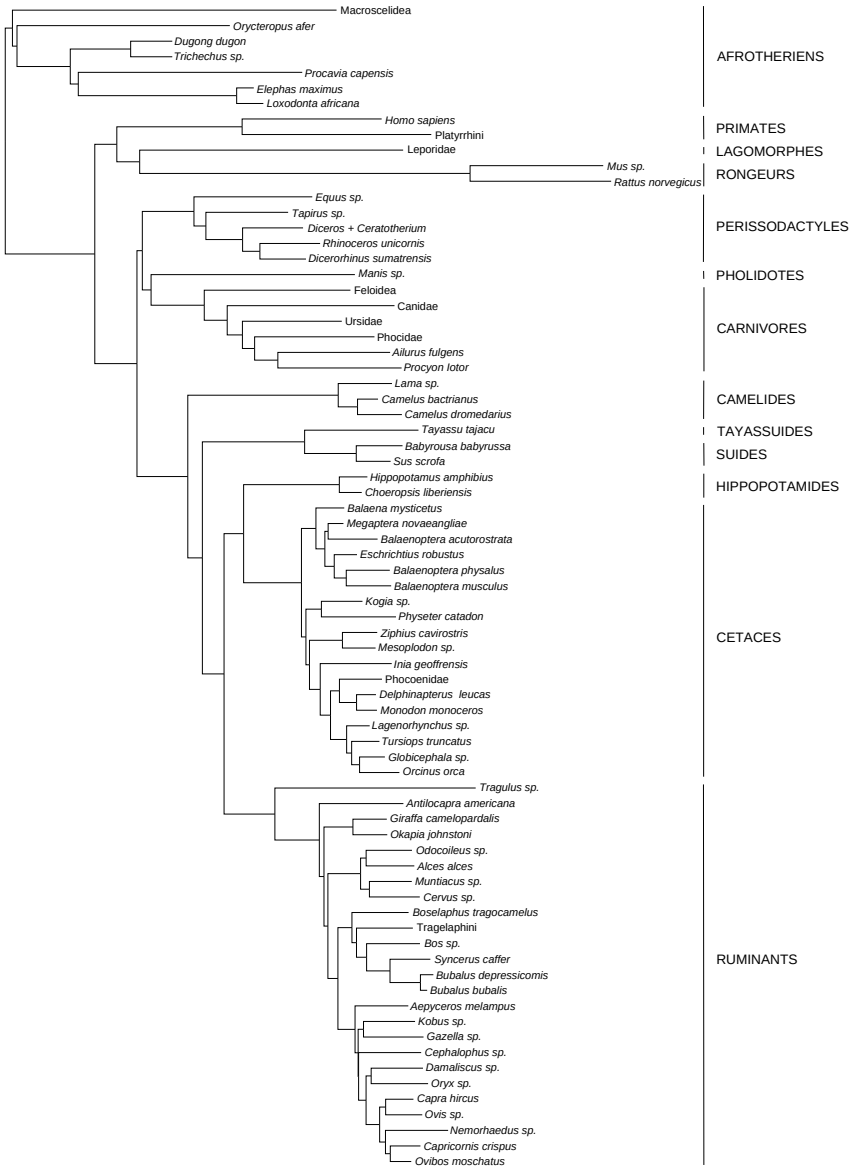

FIG. 2 – **Phylogénie inférée par la méthode** SDM+FASTME+PHYML**.**

Afin d'estimer la stabilité de ces résultats, pour chacun des dix réplicats que nous avons générés par bootstrap, nous avons numéroté chacun des six scénarios d'inférence dans l'ordre décroissant de la vraisemblance des différents arbres inférés. Ainsi le scénario renvoyant l'arbre le plus vraisemblable (*resp.* le moins vraisemblable) aura le rang 1 (*resp.* le rang 6). Pour chaque scénario, nous avons calculé la moyenne des dix valeurs de rang correspondantes. Les scénarios de combinaison basse SDM+FASTME+PHYML, BIONJ+PHYML et DTE présentent des rangs moyens respectifs de 1.25, 1.75 et 6.0. Les scénarios de combinaison haute PHYML+SDM+FASTME et PHYML+MRP présentent des rangs moyens respectifs de 3.5 et 3.8. La méthode de combinaison moyenne SDM+FASTME présente un rang moyen de 4.7. On retrouve là des résultats en plein accord avec les simulations :

– le scénario SDM+FASTME infère un arbre de bien meilleure qualité que l'autre approche de distance DTE correspondant à la première étape de BIONJ+PHYML.
– Les techniques de combinaison basse effectuées par une maximisation du critère ML offrent les meilleures performances. De plus, partant d'un arbre initial préférable, le scénario SDM+FASTME+PHYML marque un avantage net sur BIONJ+PHYML. On notera cependant que

le logiciel PHYML remanie l'arbre inféré par DTE de manière importante puisque ce dernier est systématiquement le moins vraisemblable et qu'il passe au deuxième voire au premier rang une fois modifié par PHYML.
– En combinaison haute, le scénario avec SDM semble légèrement plus performant que PHYML+MRP mais d'autres expériences seraient nécessaires pour vérifier que ce résultat est significatif.

6 Conclusion

Nous avons présenté une nouvelle méthode, SDM, permettant de combiner différentes matrices de distance en une unique supermatrice de distance. Nous avons montré, par le moyen de diverses simulations et par une application sur un jeu de données réel, que SDM, associé aux logiciels FITCH ou FASTME, permet d'obtenir des résultats équivalents ou meilleurs que la plus utilisée des méthodes de combinaison de données, MRP, en particulier lorsque ces dernières sont très incomplètes. De par sa rapidité et sa capacité de recouvrement de la topologie de l'arbre vrai, elle constitue aussi un excellent point de départ pour des algorithmes d'optimisation du critère ML.
Néanmoins, ces premiers résultats pourraient être améliorés en proposant, par exemple, de nouveaux types de pondération au sein de SDM ou de nouvelles contraintes linéaires. Le développement d'algorithmes de reconstruction phylogénétique rapides et dédiés aux supermatrices de distance contenant des distances manquantes est un autre axe de recherche à développer.

Notre implémentation de la méthode SDM, en JAVA 1.4 pour plus de portabilité, est disponible à l'URL : `http://www.lirmm.fr/~criscuol/soft/sdm`.

Références

[1] J.P. Barthélemy and A. Guénoche. *Les arbres et les représentations des proximités*. Masson, 1988.

[2] B.R. Baum. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon*, 41 :3–10, 1992.

[3] M. Bourque. *Arbres de Steiner et réseaux dont varie l'emplagement de certains sommets*. PhD thesis, Departement d'informatique et de recherche opérationnelle, Université de Montréal, 1978.

[4] V. Daubin, M. Gouy, and G. Perrière. A phylogenomic approach to bacterial phylogeny : evidence of a core of genes sharing a common history. *Genome Research*, 12(7) :1080–1090, 2002.

[5] G. De Soete. Additive tree representations of incomplete dissimilarity data. *Quality and Quantity*, 18 :387–393, 1984.

[6] R. Desper and O. Gascuel. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of Computational Biology*, 19(5) :687–705, 2002.

[7] G. Devulder, M. Pérouse de Montclos, and J.P. Flandrois. A multigene approach to phylogenetic analysis using the genus Mycobacterium as a model. *International Journal of Systematic and Evolutionary Microbiology*, 55 :293–302, 2005.

[8] J.A. Eisen and C.M. Fraser. Phylogenomics : Intersection of evolution and genomics. *Science*, 300 :1706–1707, 2003.

[9] G.F. Estabrook, F.R. McMorris, and C.A. Meacham. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Systematic Zoology*, 34 :193–200, 1985.

[10] O. Eulenstein, D. Chen, J.G. Burleigh, D. Fernandez-Baca, and M.J. Sanderson. Performance of flip supertree construction with a heuristic algorithm. *Systematic Biology*, 53(2) :299–308, 2004.

[11] J. Felsenstein. PHYLIP : Phylogeny inference package, version 3.6. *Distributed by the author. University of Washington, Seattle, Washington*, 1993.

[12] O. Gascuel. A note on Sattath and Tversky's, Saitou ans Nei's and Studier and Keppler's algorithms for inferring phylogenies from evolutionary distances. *Molecular Biology and Evolution*, 11(6) :961–963, 1994.

[13] O. Gascuel. BIONJ : an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14 :685–695, 1997.

[14] J. Gatesy, C. Matthee, R. DeSalle, and C. Hayashi. Resolution of a supertree/supermatrix paradox. *Systematic Biology*, 51(4) :652–664, 2002.

[15] P. Goloboff, J. Farris, and K. Nixon. TNT : Tree analysis using new technology. *Distributed by the authors*, 2003.

[16] S. Guindon and O. Gascuel. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5) :696–704, 2003.

[17] A. Guénoche and S. Grandcolas. Approximations par arbre d'une distance partielle [tree adjustments for partial distances]. *Mathématiques, Informatique et Sciences humaines*, 146 :51–64, 1999.

[18] J. P. Huelsenbeck. When are fossils better than extant taxa in phylogenetic analysis? *Systematic Zoology*, 40(4) :458–469, 2001.

[19] M. Kearney. Fragmentary taxa, missing data, and ambiguity : mistaken assumptions and conclusions. *Systematic Biology*, 51 :369–381, 2002.

[20] F.-J. Lapointe and G. Cucumel. The average consensus procedure : combination of weighted trees containing identical or overlapping sets of taxa. *Systematic Biology*, 46(2) :306–312, 1997.

[21] F.-J. Lapointe and C. Levasseur. Everything you always wanted to know about the average consensus, and more. In O.R.P. Bininda-Emonds, editor, *Phylogenetic supertrees : Combining information to reveal the tree of life*. Kluwer Academic, Dordrecht, The Nederlands, 2004.

[22] O. Madsen, M. Scally, C.J. Douady, D.J. Kao, R.W. DeBry, R. Adkins, H.M. Amrine, M.J. Stanhope, W.W. de Jong, and M.S. Springer. Parallel adaptive radiations in two major clades of placental mammals. *Nature*, 409(6820) :610–614, 2001.

[23] S.A. Mahon. A molecular supertree of the Artiodactyla. In O.R.P. Bininda-Emonds, editor, *Phylogenetic supertrees : Combining information to reveal the tree of life*. Kluwer Academic, Dordrecht, The Nederlands, 2004.

[24] V. Makarenkov. T-REX : reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics*, 17(7) :664–668, 2001.

[25] V. Makarenkov and F.-J. Lapointe. A weighted least-squares approach for inferring phylogenies from incomplete distance matrices. *Bioinformatics*, 20 :2113–2121, 2004.

[26] V. Makarenkov and B. Leclerc. An algorithm for the fitting of a phylogenetic tree according to a weighted least-squares criterion. *Journal of classification*, 16(1) :3–26, 1999.

[27] H. Philippe, E.A. Snell, E. Bapteste, P. Lopez, P.W.H. Holland, and D. Casane. Phylogenomics of eukaryotes : Impact of missing data on large alignments. *Molecular Biology and Evolution*, 21(9) :1740–1752, 2004.

[28] T. Pupko, D. Huchon, Y. Cao, N. Okada, and M. Hasegawa. Combining multiple data sets in a likelihood analysis : which models are the best ? *Molecular Biology and Evolution*, 19(12) :2294–2307, 2002.

[29] M.A. Ragan. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution*, 1 :53–58, 1992.

[30] A. Rambaut and N.C. Grassly. SEQ-GEN : an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences*, 13 :235–238, 1997.

[31] D. Robinson and L. Foulds. Comparison of weighted labeled trees. *Lectures Notes in Mathematics*, 748 :119–126, 1979.

[32] R. Rodriguez, J.L. Oliver, A. Marin, and J.R. Medina. The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology*, 142 :485–501, 1990.

[33] N. Saitou and M. Nei. The neighbor-joining method : a new method for reconstructing phylogenetic trees. *Molecular Biology Evolution*, 4 :406–425, 1987.

[34] M.J. Sanderson. Inferring absolute rates of molecular evolution and divergence times in the absence of molecular clock. *Bioinformatics*, 19 :301 – 302, 2003.

[35] H.A. Schmidt. *Phylogenetic Trees from Large Datasets*. PhD thesis, Düsseldorf, Germany, 2003.

[36] H.A. Schmidt, K. Strimmer, M. Vingron, and A. von Haeseler. TREE-PUZZLE : Maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 18 :502–504, 2002.

[37] J.A. Studier and K.J. Keppler. A note on the neighbor-joining method of Saitou and Nei. *Molecular Biology and Evolution*, 5 :729–731, 1988.

[38] D.L. Swofford. *PAUP* :Phylogenetic Analysis using Parsimony (*and other methods), version 10*, Sinauer, Sunderland, Massachussetts edition, 2002.

[39] D.L. Swofford, G.J. Olsen, P.J. Waddell, and D.M. Hillis. Phylogenetic inference. In *Molecular Systematics*, Massachussets, Sinauer Associates, 1996. Hillis, D.M. and Moritz, C. and Mable, B.K.

[40] P.J. Waddell and S. Shelley. Evaluating placental inter-ordinal phylogenies with novel sequences including RAG1, γ-fibriogen, ND6, and mt-tRNA, plus MCMC-driven nucleotide, amino acid, and codon models. *Molecular Phylogenetics and Evolution*, 28 :197–224, 2003.

[41] J.J. Wiens. Does adding characters with missing data increase or decrease phylogenetic accuracy? *Systematic Biology*, 47 :625–640, 1998.

[42] J.J. Wiens and T.W. Reeder. Combining data sets with different numbers of taxa for phylogenetic analysis. *Systematic Biology*, 44 :548–558, 1995.

[43] Z. Yang. Maximum-likelihood models for combined analysis of multiple sequence data. *Journal of Molecular Evolution*, 42 :587–596, 1996.