Votez veto pour l'Arbre de la Vie

Vincent Berry, Vincent Ranwez, Pierre-Henri Fabre, Emmanuel Douzery

Votez veto pour l'Arbre de la Vie : la méthode *PhySIC* pour reconstruire des superarbres

Vincent Berry,[1] Vincent Ranwez,[2] Pierre-Henri Fabre[2] et Emmanuel J.P. Douzery[2]

[1] Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier - LIRMM - UMR 5506
161 rue Ada 34392 Montpellier Cedex 5 - France
vberry@lirmm.fr

[2] Institut des Sciences de l'Evolution (ISEM, UMR 5554 CNRS), Université Montpellier II, Place E. Bataillon - CC 064 - 34095 Montpellier Cedex 05
{ranwez,fabreph,douzery}@isem.univ-montp2.fr

Résumé *Phylogenetic methods are used to infer the evolutionary history of species. In the Tree of Life framework, heterogeneous character data and very large species sets are considered. Supertree methods have been developped to deal with such a situation. These methods combine source topologies, inferred from separate character sets, into a larger, so-called supertree. One of the main challenges met by supertree methods is the handling of topological conflicts arising among source trees. Some methods resolve these conflicts based on a* voting *approach : they rely on various criteria to decide which particular resolution, among conflicting ones, is to be kept in the supertree. Other methods follow a* veto *approach, proposing supertrees that do not favor any resolution among several conflicting ones. Compared to voting methods, they usually propose less resolved trees but avoid the problematic, implicit weighting of non-independent information contained in source trees. This article points out some desirable mathematical properties of* veto *supertrees called* induction *and* non-contradiction*. It also describes* PhySIC *(Phylogenetic Signal with Induction and non-Contradiction), an effective supertree method outputting supertrees that verify these properties. The method is available as a web service at* www.lirmm.fr/~vberry/TOL/physic.cgi*. The relevance of* PhySIC *is illustrated on a biological case study on primates where it is compared to MRP, the most widely used voting method.*

Keywords: Tree of Life, Supertree, Mathematical properties, Induction, Compatibility, Polynomial-time algorithms, MRP, Primates.

1 Introduction

1.1 Reconstruire des superarbres

L'évolution des organismes est communément décrite par des arbres phylogénétiques, dont les feuilles sont des espèces et les nœuds leurs ancêtres communs hypothétiques. Les phylogénies sont incontournables en biologie, et d'autant plus riches en informations qu'elles incorporent un grand nombre d'espèces. Elles nous renseignent ainsi sur la systématique, la génomique, et les profils de diversification des espèces. (e.g., [11]). La reconstruction de ces Arbres du Vivant se fonde actuellement sur au moins trois approches complémentaires [10,12] : l'étude de caractéres génomiques (présence-absence de gènes, intégration d'éléments transposables), les supermatrices, et les superarbres. Dans l'approche par supermatrice de caractères, les lots de données initiaux – c'est-à-dire les alignements de gènes (par exemple les ARNr) ou de protéines – sont concaténés, afin d'accumuler un maximum

d'information phylogénétique potentielle. L'arbre synthétique découle alors directement d'un critère d'optimisation appliqué à cette supermatrice, par exemple le maximum de parcimonie ou de vraisemblance [13,20]. Cependant, le chevauchement des espèces au sein des lots de données sources est partiel — peu de gènes ont été séquencés pour de nombreux taxons, et réciproquement, peu de taxons ont une large couverture génomique – ce qui nécessite la déclaration de données manquantes dans la supermatrice. L'approche par superarbre est quant à elle indirecte, et permet d'éviter la manipulation de données manquantes. Pour cela, ce ne sont pas les caractères originaux qui sont directement utilisés, mais les phylogénies sources qui en résultent. Ces dernières sont assemblées en un superarbre respectant autant que possible leurs topologies [5]. Cette approche est notamment privilégiée lorsque les données initiales sont hétérogènes, par exemple dans le cas de caractères sources d'origine morphologique et moléculaire, nucléotidique et protéique, ou encore représentant la présence-absence d'événements génomiques rares. Notons qu'une approche intermédiaire entre les supermatrices de caractères et les superarbres consiste à convertir les caractères initiaux en matrices de distances. Ces dernières peuvent alors être déformées puis associées en une supermatrice de distances, ensuite analysée par des algorithmes classiques de distances, avec ou sans méthode de complétion de distances manquantes [8,19].

1.2 Trois approches pour gérer les conflits topologiques

Depuis la publication originale du superarbre des Primates [23], les superarbres sont devenus de plus en plus populaires, à tel point qu'un livre entier vient de leur être consacré [5]. L'une des difficultés inhérente aux méthodes de superarbres est l'utilisation d'arbres sources incongruents, c'est-à-dire en désaccord sur la position phylogénétique de certain taxons ou clades. Selon leur manière de gérer ces topologies conflictuelles, les méthodes de superarbres se divisent en *trois* grands types. Dans le premier, les topologies sources incongruentes ne sont pas assemblées. Dans les deux autres types, les topologies sources sont toujours assemblées – quel que soit leur degré de congruence –, mais selon des philosophies distinctes. Les incongruences sont traitées par une procédure de *vote* dans le deuxième type, tandis qu'elles sont gérées par une procédure de *veto* dans le troisième.

Les approches pionnières de superarbres telles que BUILD [1] et le consensus strict [16] sont classées dans la première famille. Bien que constituant une étape importante de l'histoire des superarbres, Bininda-Emonds les désigne comme étant *"d'un usage limité. En effet la plupart des systématiciens le savent, les phylogénies sont généralement en conflit les unes avec les autres*" [5, p. 4]. D'éventuelles incongruences peuvent effectivement émerger lorsque le signal phylogénétique diffère d'un jeu de données initial à l'autre. Par exemple, lorsque les arbres sources sont des arbres de gènes, le signal phylogénétique principal peut être brouillé par les transferts horizontaux, hybridations, et autres duplications-pertes de gènes, ces événements étant d'importance relative différente selon que les organismes étudiés sont procaryotes ou eucaryotes.

Dans la deuxième famille de méthodes, dite de *vote*, une décision est prise en faveur de l'une ou l'autre des alternatives possibles concernant la position d'un (groupe de) taxon(s). Cette décision est prise sur la base d'un critère d'optimisation qui varie d'une méthode à l'autre. Ces méthodes sont supposées *résoudre* les conflits [30]. Une telle stratégie est adaptée aux études dont l'objectif est d'extraire le signal phylogénétique congruent d'un lot d'arbres sources contenant des clades de solidité variable. Pour cela, il est raisonnable de "faire voter" les topologies sources, et ainsi d'élire les clades candidats les plus soutenus. Dans ce contexte, l'approche la plus répandue est celle par Représentation Matricielle avec Parcimonie (MRP) [2]. Ici, les nœuds de chacune des topologies sources sont codés par des caractères binaires. Ce codage permet d'obtenir une matrice de caractères binaires sur laquelle un critère de parcimonie est appliqué afin d'obtenir un superarbre. Les conflits potentiels d'information qui apparaissent au sein des éléments de la matrice vont donc être résolus au sens du maximum

de parcimonie. Cette vision globale des topologies sources permet à MRP de générer de nouveaux clades – collectivement induits par les arbres d'entrée mais cependant absents de chacun d'entre eux pris individuellement – montrant en cela tout l'intérêt des superarbres. Malheureusement, MRP peut aussi proposer de nouveaux clades qui sont contredits par un ou plusieurs arbres sources [6, Fig. 3].

Dans la troisième famille de méthodes de superarbres, dite de *veto*, le message phylogénétique de chaque arbre source est respecté. Ainsi, un clade est retenu dans le superarbre si et seulement si les topologies sources sont unanimement en accord avec sa présence. Le superarbre ne peut donc posséder de clade auquel un des arbres sources pourrait s'opposer. Pour cette raison, de telles méthodes proposent des multifurcations au sein du superarbre [7] ou bien en soustraient les taxons problématiques [3]. Dans la terminologie de [30], ces méthodes *retirent* les conflits. La principale application des méthodes de superarbres de type *veto* est de construire l'Arbre de la Vie. Atteindre cet objectif requiert de partir de phylogénies sources bien établies quoique fragmentaires, pour obtenir des superarbres toujours plus grands, à la fois taxonomiquement représentatifs et fortement soutenus. Cependant, tous les arbres sources étant supposés également fiables, le superarbre reconstruit ne doit pas favoriser une topologie initiale au détriment d'une autre. Plusieurs méthodes de superarbres suivant cette approche par *veto* ont été récemment proposées. Toutes s'inspirent des méthodes de consensus qui opèrent sur des arbres possédant des feuilles identiques : consensus strict [16], consensus semi-strict [14], et sous-arbre d'accord maximum [3].

1.3 Des propriétés pertinentes pour la reconstruction de l'Arbre de la Vie

Steel et al. ont énoncé une liste de propriétés simples que toute méthode d'inférence de superarbres devrait vérifier [28]. De manière surprenante, les méthodes se basant sur des arbres sources non enracinés ne peuvent vérifier simultanément toutes ces propriétés. Cependant, cette limitation ne s'applique pas au cas d'arbres sources enracinés (*cf*. la méthode MinCut [26]).

Le présent article s'inscrit dans le cadre des méthodes de superarbres enracinés de type *veto*. Dans ce contexte, on souhaite que les méthodes de superarbres évitent les *résolutions arbitraires*. Ceci suggère deux propriétés. Le superarbre inferré ne doit pas favoriser une résolution plutôt qu'une autre lorsque plusieurs possibilités contradictoires existent : c'est la propriété de *non-contradiction* (c'est-à-dire de compatibilité). De plus, chaque information topologique de ce superarbre doit être induite, *i.e.* forcée, par un ou plusieurs arbres sources : c'est la propriété d'*induction*. Des propriétés souhaitables des superarbres ont été énoncées par Goloboff et Pol [14] en utilisant les *triplets enracinés* (voir Sect. 2.1). Nous montrons ici que les propriétés de Goloboff et Pol sont tantôt trop restrictives, tantôt trop permissives. Plus récemment, Grünewald et al. [17] ont fourni une autre caractérisation de propriétés de superarbres, en liaison avec l'absence de prise de décision arbitraire en regard des arbres sources. Malheureusement, il ne semble pas y avoir d'algorithme évident pour vérifier ces propriétés. Sur la base de divers exemples, on peut montrer que les méthodes de superarbres impliquant un *vote* – comme MRP et MinCut – ne respectent manifestement pas ces propriétés [14].

Nous utilisons ici une caractérisation approfondie des propriétés de *non-contradiction*, PC, et d'*induction*, PI, récemment proposée [24]. Cette formalisation permet de décider en temps polynomial si un superarbre vérifie PI et PC. Nous proposons ici un algorithme nommé *PhySIC– "PHYlogenetic Signal with Induction and non-Contradiction"* – qui reconstruit toujours un superarbre satisfaisant les propriétés PI et PC. La complexité polynomiale de cet algorithme conduit réguilèrement à des temps de calcul meilleurs que ceux obtenus par MRP. Enfin, nous illustrons le fonctionnement de *PhySIC* sur un cas d'étude biologique bien connu, la reconstruction du superarbre des Primates initiée par [23]. Nous utilisons ici des caractères sources empruntés à deux gènes nucléaires et à des éléments

génomiques transposables, et nous montrons que les méthodes *PhySIC* (*veto*) et MRP (*vote*) peuvent conduire à des superarbres de résolution globalement comparable.

2 Formalisation de propriétés souhaitables pour la construction de superarbres

Dans ce travail, nous nous plaçons dans le cadre de la construction d'Arbres de la Vie. Dans ce contexte, les arbres sources sont composés de clades fiables. Aussi demande-t-on que les méthodes aient la propriété de ne pas contredire les clades sources (*propriété de non-contradiction*, notée PC) et que tout clade du superarbre proposé soit présent ou induit par les arbres sources (*propiété d'induction*, notée PI). Cette section introduit le vocabulaire et les notations nécessaires pour définir de manière formelle PC et PI. Des exemples simples illustrent aussi la pertinence de ces deux propriétés et permettent de les comparer avec des propriétés proposées dans un contexte similaire.

2.1 Définitions et notations

Pour un arbre enraciné ayant trois feuilles a, b, c, il n'existe que trois topologies binaires possibles appelées triplets (enracinés) et notées $ab|c$, resp. $ac|b$, resp. $bc|a$ en fonction du groupe de feuilles le plus interne (ab, resp. ac, resp bc). Une topologie alternative est l'arbre étoile, i.e. l'arbre topologiquement non informatif constitué d'un seul noeud interne directement relié aux trois feuilles.

Etant donné un triplet t, on note $\bar{t}$ n'importe lequel des deux autres triplets ayant les mêmes feuilles. Un arbre T de plus de trois feuilles peut être représenté par l'ensemble des triplets homéomorphiques aux sous-arbres de T contenant trois feuilles [27]. Cette représentation est largement utilisée dans le contexte des superarbres. Dans la suite, l'ensemble de triplets équivalent à T sera noté $tr(T)$. Cette notation se généralise à un ensemble d'arbres $\mathcal{T}$: $tr(\mathcal{T}) = \bigcup_{T_i \in \mathcal{T}} tr(T_i)$. Notons qu'il est possible que $tr(\mathcal{T})$ contienne à la fois t et $\bar{t}$.

Définition 1 (Représentation et compatibilité) *Etant donné $\mathcal{R}$ un ensemble de triplets, un arbre T* représente *$\mathcal{R}$ ssi $\mathcal{R} \subseteq tr(T)$. Un ensemble de triplets $\mathcal{R}$ est* compatible *ssi il existe au moins un arbre qui le représente.*

Définition 2 (Induction - cas compatible) *Soit $\mathcal{R}$ un ensemble compatible de triplets. On dit que $\mathcal{R}$* induit *le triplet t ($\mathcal{R} \vdash t$) ssi $\mathcal{R} \cup \bar{t}$ est incompatible. Une définition alternative de $\mathcal{R} \vdash t$ consiste à dire que t doit être présent (i.e. $t \subseteq tr(T)$) dans chaque arbre T qui représente $\mathcal{R}$ [17].*

Définition 3 (Fermeture) *Soit $\mathcal{R}$ un ensemble de triplets, la* fermeture (closure) *de $\mathcal{R}$, notée $cl(\mathcal{R})$, est définie de la manière suivante :* $\mathbf{cl}(\mathcal{R}) = \{ab|c \text{ tel que } \mathcal{R} \vdash ab|c\}$.

La définition 2 peut être généralisée au cas où l'ensemble de triplets n'est pas forcément compatible.

Définition 4 (Induction - cas général) *Soit $\mathcal{R}$ un ensemble de triplets et t un triplet. On dit que $\mathcal{R}$* induit *t ($\mathcal{R} \vdash t$) ssi $\exists \mathcal{R}' \subseteq \mathcal{R}$ tel que $\mathcal{R}'$ est compatible et $\mathcal{R}' \vdash t$. L'ensemble des triplets induits par $\mathcal{R}$ sera noté $ind(\mathcal{R})$. Notons que lorsque $\mathcal{R}$ est compatible, $cl(\mathcal{R}) = ind(\mathcal{R})$.*

Définition 5 (Identification d'un arbre) *Soit $\mathcal{R}$ un ensemble compatible de triplets. On dit que $\mathcal{R}$* identifie *un arbre T ssi $cl(\mathcal{R}) = tr(T)$. Il est clair qu'un ensemble incompatible n'identifie donc aucun arbre. Notons que si $\mathcal{R}$ est un ensemble compatible de triplets alors il y a au moins un arbre qui représente $\mathcal{R}$, mais cela n'est pas suffisant pour assurer que $\mathcal{R}$* identifie *un arbre particulier.*

2.2 Propriétés de non-contradiction et d'induction

Dans le cas où $\mathcal{T}$ identifie un arbre T, T est un superarbre idéal pour représenter $\mathcal{T}$. En pratique il est peu fréquent qu'une collection d'arbres sources $\mathcal{T}$ soit compatible [5, p4] ; or, même dans de tels cas, elle n'identifie pas nécessairement un arbre particulier. Dans le cas général où $tr(\mathcal{T})$ n'est pas forcément compatible, il est cependant possible qu'un sous-ensemble $\mathcal{R}_T$ de $tr(\mathcal{T})$ identifie un arbre T. Dans ce cas, chaque information topologique de T est présente, directement ou de manière induite, dans $\mathcal{T}$. Ceci fait de T un bon candidat pour représenter $\mathcal{T}$. Néanmoins, pour que T soit un superarbre valable, dans le contexte de projets du type Arbre de la Vie, il est également souhaitable que ses triplets ne contredisent pas d'autres triplets de $tr(\mathcal{T})$ sur les mêmes ensembles de feuilles. Si tel est le cas, T représente alors un sous-ensemble consensuel d'informations topologiques présentes dans les arbres sources, tandis que les informations conflictuelles de $\mathcal{T}$ ne sont pas présentes dans T (soit du fait de multifurcations dans T soit du fait de l'absence de certains taxons).

Définition 6 ($\mathcal{R}_\mathbf{T}$) *Soit T un arbre, et $\mathcal{T}$ un ensemble d'arbres. On définie $\mathcal{R}_\mathbf{T}(\mathcal{T}) = \{ab|c \in tr(\mathcal{T})$ tel que $\{ab|c, ac|b, bc|a\} \cap tr(T) \neq \emptyset\}$. En l'absence d'ambiguïté sur $\mathcal{T}$ on notera simplement $\mathcal{R}_T$.*

Notons qu'il est possible que $\mathcal{R}_T(\mathcal{T})$ soit incompatible. C'est notamment le cas dès que T contient un triplet qui est résolu différemment dans deux arbres de $\mathcal{T}$. A l'aide de ces notations, il est maintenant possible de définir de manière formelle les propriétés évoquées ci-dessus.

Définition 7 *Soit $\mathcal{T}$ un ensemble d'arbres et T un superarbre, on dit que :*
- *T vérifie* **PI** *pour $\mathcal{T}$ ssi $\forall t \in tr(T)$, $\mathcal{R}_T \vdash t$.*
- *T vérifie* **PC** *pour $\mathcal{T}$ ssi $\forall t \in tr(T)$, $\mathcal{R}_T \nvdash \bar{t}$.*

PI et PC sont des propriétés pertinentes dans le sens où, dès qu'un arbre T vérifie à la fois PI et PC, cela garantit que $\mathcal{R}_T$ représente une partie de $tr(\mathcal{T})$ qui correspond exactement à un arbre.

Proposition 1 *Soit $\mathcal{T}$ un ensemble d'arbres, et T un superarbre, $\mathcal{R}_T$ identifie T (ie $cl(\mathcal{R}_\mathcal{T}) = tr(T)$) ssi T verifie PI et PC pour $\mathcal{T}$.*

La preuve de cette proposition ainsi que celles des autres résultats théoriques de cet article peuvent être trouvées dans le rapport de recherche [4].

Définition 8 (Contradiction directe) *Un arbre T* contredit directement *un ensemble de triplets $\mathcal{R}$ ssi $\exists t \in tr(T)$ tel que $\bar{t} \in \mathcal{R}$. En particulier, pour un ensemble d'arbres sources $\mathcal{T}$, si T contredit directement $\mathcal{R} = tr(\mathcal{T})$, alors on dit que T contredit $\mathcal{T}$.*

Lemme 1 ([24]) *Si T est un arbre qui ne contredit pas directement $\mathcal{T}$, alors T vérifie les trois propriétés suivantes :*
1. $\mathcal{R}_T \subseteq tr(T)$; 2. $\mathcal{R}_T$ est compatible ; 3. PC.

2.3 Liens avec d'autres propriétés souhaitables des superarbres

Des propriétés similaires à PI et PC sont décrites dans [14, p.519] : "le superarbre doit représenter $ab|c$ si ce triplet est présent dans un arbre source – ou induit par une combinaison d'arbres sources – et si les triplets $ac|b$ et $bc|a$ sont absents de tous les arbres sources et ne sont pas induits par une

combinaison d'arbres sources". Ces propriétés, dont la pertinence est soulignée par [17], peuvent être décrites dans notre formalisme comme suit :

$$- \ \mathbf{PI'} : \forall t \in T, tr(\mathcal{T}) \vdash t \qquad\qquad - \ \mathbf{PC'} : \forall t \in T, tr(\mathcal{T}) \nvdash \bar{t}.$$

Du fait que $tr(T) \subseteq tr(\mathcal{T})$, il est clair que PC'$\Rightarrow$ PC et PI $\Rightarrow$ PI'. Il est donc naturel de se demander quelle version de ces propriétés est la plus adaptée pour qualifier un superarbre. Les exemple ci-dessous montre un cas où PC' est trop restrictif (Fig. 1), et un autre où PI' est trop laxiste (Fig. 2). Ces exemples illustrent le fait que PI' et PC' ne sont pas forcément aussi pertinentes qu'elles le semblent au premier abord. En revanche, PI et PC ont le comportement escompté sur ces deux exemples. Quant à savoir s'il existe des propriétés encore plus discriminantes que PI et PC, la question reste ouverte.

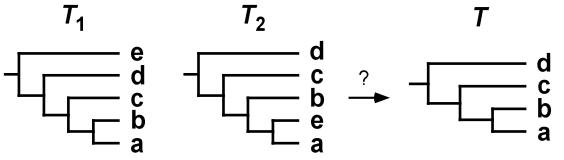

FIG. 1. Un ensemble d'arbres sources $\{T_1, T_2\}$ et un superarbre T pouvant être proposé par une méthode de type *veto*. Ce superarbre exclut le seul taxon problématique (e). T vérifie PI et PC, mais pas PC' ($\{ae|b, ac|e\} \vdash ac|b$). Cet exemple se généralise au cas plus fréquent où le *super*arbre contient plus de taxons que chacun des arbres sources.

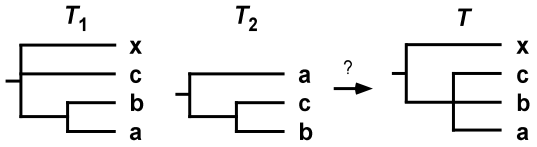

FIG. 2. Un exemple où la présence de contradiction entre les arbres sources ($ab|c$ dans T_1 et $bc|a$ dans T_2) conduit à inférer un clade arbitraire (x en dehors du clade $\{a, b, c\}$) dans le superarbre potentiel T. Pourtant T vérifie PI' et PC'. En revanche, T ne vérifie pas PI qui détecte ce problème.

Il n'existe, à notre connaissance, aucun algorithme capable de déterminer si PI' et PC' sont vérifiées par un arbre donné. En revanche, il est possible de déterminer si un superarbre T vérifie PC et PI en utilisant les algorithmes polynomiaux décrits dans [24].

3 *PhySIC* : un algorithme construisant un superarbre qui vérifie PI et PC

Les méthodes de superarbres qui suivent une approche de *vote* produisent rarement un arbre qui satisfait PC et PI (voir les exemples [6, Fig. 1] pour MRP et [14, Fig. 4] pour MinCut). Malheureusement, comme la plupart des méthodes de superarbres ne sont pas initialement conçues pour produire un arbre vérifiant PC et PI, cette approche peut conduire à écraser la majorité des arêtes qu'elles proposent, menant au bout du compte à un superarbre très peu résolu. Une approche alternative, a priori plus prometteuse, consiste donc à proposer une méthode qui construise *ab initio* un superarbre vérifiant PC et PI à partir des arbres sources. Dans cette section nous décrivons un tel algorithme, nommé *PhySIC*, obtenu sur la base de l'algorithme $Build_{PC}$ présenté dans [24].

Soit T un arbre, $L(T)$ dénote l'ensemble des feuilles de T, et soit $\mathcal{R}$ un ensemble de triplets (ou plus généralement d'arbres), on note $L(\mathcal{R}) = \bigcup_{t \in \mathcal{R}} L(t)$. Le *Graphe de Aho* pour $\mathcal{R}$ est le graphe ayant $L(\mathcal{R})$ comme sommets et une arête entre deux sommets a et b ssi il existe un triplet $ab|c \in \mathcal{R}$. On note $CC(G)$ l'ensemble des composantes connexes d'un graphe G et $v(C_i)$ l'ensemble des sommets d'une composante connexe C_i de G. La restriction de $\mathcal{R}$ aux sommets de C_i est $\mathcal{R}|v(C_i) = \{ab|c \in \mathcal{R}$ tel que $\{a, b, c\} \subseteq v(C_i)\}$. Nous rappelons ci-dessous l'algorithme $Build_{PC}$ de [24].

L'arbre T renvoyé par $Build_{PC}$ ne peut pas contenir de contradiction directe avec l'ensemble $\mathcal{R}$ sur lequel il est appliqué. En effet les triplets $ab|c$ présents dans T sont créés à la ligne `1` lorsque

Algorithme $Build_{PC}\ (\mathcal{R})$

Soit G le graphe de Aho pour $\mathcal{R}$
si $|CC(G)| = 1$ **alors** Renvoyer l'arbre étoile sur $v(G)$
sinon
  $\mathcal{C}_{PC} \leftarrow CC(G)$
  pour chaque $C_i \in C_{PC}$ **faire**
    si $|(\mathcal{R}|v(C_i))| = \emptyset$ **alors** $T_i \leftarrow$ l'arbre étoile sur $v(C_i)$
    sinon $T_i \leftarrow\ Build_{PC}\ (\mathcal{R}|v(C_i))$
1 Renvoyer l'arbre composé d'un noeud racine connecté à $T_1, T_2, ..., T_{|\mathcal{C}_{PC}|}$

a, b sont dans une même composante connexe C_i et c et dans une autre composante, soit C_j. Si $\mathcal{R}$ contenait $ac|b$ ou $bc|a$, alors C_i et C_j ne seraient pas des composantes connexes disjointes (car G contiendrait alors à cette étape une arête (a, c) ou (b, c)). $Build_{PC}$ garantit donc d'obtenir un arbre satisfaisant PC (lem. 1). Cependant, il produit généralement des arbres très peu résolus : lorsque G ne contient qu'une seule composante (en raison de conflits au sein de $\mathcal{R}$ sur $v(G)$), une multifurcation non-informative sur les taxons concernés est renvoyée. Dans les cas extrêmes, un tel conflit intervient dès la première étape de l'algorithme qui renvoie alors un arbre étoile.

Les conflits les plus simples entre triplets de $\mathcal{R}$ sont ceux concernant un triplet t tel que $t, \bar{t} \in \mathcal{R}$. De tels triplets ne pourront jamais être présents dans un arbre vérifiant PC. Les enlever de l'ensemble de triplets utilisés pour construire l'arbre peut donc amener à plus de résolution. Si l'on note $\mathcal{R}_{cd}$ l'ensemble des triplets t.q. $t, \bar{t} \in \mathcal{R}$ on peut donc se demander s'il n'est pas pertinent de considérer l'arbre obtenu par $Build_{PC}$ sur $\mathcal{R}' = \mathcal{R} - \mathcal{R}_{cd}$. L'arbre T' ainsi obtenu est généralement beaucoup plus résolu que T. Cependant, s'il vérifie PC par rapport à $\mathcal{R}'$, rien ne garantit qu'il vérifie cette propriété par rapport à $\mathcal{R}$. Il faut pour cela qu'il ne résolve aucun des triplets de $\mathcal{R}_{cd}$. Une solution pour s'en assurer est d'écraser chaque arête de T' qui résout un des triplets de $\mathcal{R}_{cd}$. On obtient ainsi un arbre T_{PC} qui est toujours au moins aussi résolu que T mais qui contient potentiellement plus d'arêtes que lui, puisque la présence d'une contradiction directe à la racine d'un clade n'empêche plus d'obtenir une résolution sur des sous-ensembles du clade.

Ces remarques nous conduisent à proposer l'algorithme $PhySIC_{PC}$ comme raffinement de $Build_{PC}$. Etant donné un (sous)-arbre T, on note $SousArb(T)$ l'ensemble des sous-arbres complets connectés à la racine de T.

L'ensemble $\mathcal{C}_{PC}$ contient les clades potentiels de l'arbre T_{PC} construit par $PhySIC_{PC}$. Dans le cas où G contient plusieurs composantes connexes, chacune correspond à un clade définitif de T_{PC}. Alternativement, G est connexe et on sait qu'il existe des conflits. Si ces conflits sont dus à des contradictions directes (i.e. dans $\mathcal{R}_{cd}$), alors G', le graphe de Aho sur $\mathcal{R} - \mathcal{R}_{cd}$ est non-connexe et on obtient de nouveaux clades potentiels (à l'inverse de ce que ferait $Build_{PC}(\mathcal{R})$). Si l'un de ces clades résout des triplets de $\mathcal{R}_{cd}$ (ligne 4), il ne peut être conservé sans que PC soit invalidée. Dans ce cas, il est donc partitionné en plusieurs sous-ensembles qui le remplacent dans $\mathcal{C}_{PC}$ (ligne 6), ce qui correspond à l'écrasement d'une arête dans l'arbre $Build_{PC}(\mathcal{R}')$, ici implicitement parcouru. Enfin, dans le cas où G' est connexe, cela signifie que le conflit topologique ne se limite pas à une contradiction directe, auquel cas $PhySIC_{PC}$ renvoie une multifourche.

Théorème 1 *Etant donné un ensemble de triplets $\mathcal{R}$ sur n feuilles, $PhySIC_{PC}$ renvoie un arbre T qui vérifie PC pour $\mathcal{R}$. La complexité de cet algorithme est en $O(n^4)$.*

L'algorithme $PhySIC_{PI}$ (voir pseudo-code) permet de transformer si nécessaire T_{PC} pour qu'il vérifie aussi PI. La différence entre cet algorithme et l'algorithme $Build_{PI}$ de [24] est du même ordre que la différence entre $PhySIC_{PC}$ et $Build_{PC}$ expliquée plus haut. Par hypothèse, l'arbre T_{PC} transmis à $PhySIC_{PI}$ vérifie PC. L'algorithme renvoie un arbre T obtenu par écrasement

```
Algorithme PhySIC_PC (R)
    Soit G le graphe de Aho pour R
    si |CC(G)| > 1 alors C_PC ← CC(G)
    sinon
        Soit R_cd l'ensemble des triplets t tel que t, t̄ ∈ R
        R' ← R − R_cd
        Soit G' le graphe de Aho pour R'
1       si G' est connexe alors C_PC ← v(G)
        sinon
            C_PC ← CC(G')
2           répéter
3               pour chaque ab|c ∈ R_cd faire
4                   si a, b ∈ C_i et c ∈ C_j (avec C_i, C_j ∈ C_PC et i ≠ j) alors
5                       Construire G'_i le graphe de Aho pour R'|v(C_i)
                        si G'_i est connexe alors C_PC ← (C_PC − {C_i}) ∪ v(C_i)
6                       sinon C_PC ← (C_PC − {C_i}) ∪ CC(G'_i)
            jusqu'à ce que C_PC ne change plus
    pour chaque C_i ∈ C_PC faire
        si |(R|v(C_i))| = 0 alors T_i ← l'arbre étoile sur v(C_i)
        sinon T_i ← PhySIC_PC(R|v(C_i))
7   Renvoyer l'arbre composé d'un noeud racine connecté à T_1, T_2, ..., T_|C_PC|
```

d'arêtes dans T_{PC}, autrement dit tel que $tr(T) \subseteq tr(T_{PC})$. Ceci montre que T_{PI} vérifie toujours PC. Ainsi, l'algorithme $Check_{PI}$ (appelé par $PhySIC_{PI}$) renverra toujours un arbre. Par ailleurs, si l'on remplace la ligne 2 de $Check_{PI}$ par `renvoyer ''erreur triplet non-induit''` on retrouve l'algorithme *Identifies* de [9]. Etant donné un ensemble de triplets $\mathcal{R}$, *Identifies*$(\mathcal{R})$ renvoie l'arbre T identifié par $\mathcal{R}$ ou une erreur si T n'identifie aucun arbre [9, Thm. 3.1.1]. Ce qui assure que lors d'un appel de $Check_{PI}(T, \mathcal{R})$ aucune arrête de T n'est écrasée ssi l'ensemble $\mathcal{R}$ identifie l'arbre T. Or, l'arbre T_{PI} renvoyé par $PhySIC_{PI}$ est tel que le dernier appel à $Check_{PI}$ ne l'a pas modifié (aucun écrasement d'arête). Cet arbre T_{PI} est donc identifié par $\mathcal{R}_{T_{PI}}$, autrement dit, il vérifie PI et PC (par la Prop. 1).

```
Algorithme PhySIC_PI (T, R)
    T_PI ← T
    répéter
        R_PI ← R_{T_PI}(R)
1       T_PI ← Check_PI (T_PI, R_PI)
    jusqu'à ce que T_PI ne change plus
    Renvoyer T_PI
```

```
Algorithme PhySIC (T)
    T_PC ← PhySIC_PC (tr(T))
    Renvoyer PhySIC_PI (T_PC, tr(T))
```

```
Algorithme Check_PI (T, R)
    si T est composé d'une seule feuille alors renvoyer T
    Soit G le graphe de Aho pour R
    si |CC(G)| = 1 alors renvoyer "erreur, R incompatible"
1   répéter
        pour chaque T_i ∈ SousArb(T) faire
            Soit G_i le graphe de Aho pour R|L(T_i)
            pour chaque T_j ∈ SousArb(T) t.q. T_i ≠ T_j faire
                Construire G_ij depuis G_i et R|(L(T_i) ∪ L(T_j))
                si G_ij n'est pas connexe alors
2                   Ecraser l'arête entre la racine de T et T_i
    jusqu'à ce que aucune arête de T ne soit plus écrasée
    pour chaque T_i ∈ SousArb(T) faire
        T'_i ← Check_PI (T_i, R|L(T_i))
    Renvoyer l'arbre composé d'un noeud racine connecté à
    T'_1, T'_2, ..., T'_|SousArb(T)|
```

Théorème 2 *Soit $\mathcal{R}$ un ensemble de triplets sur n feuilles et T un arbre sur $L(\mathcal{R})$ vérifiant PC pour $\mathcal{R}$. $PhySIC_{PI}(T, \mathcal{R})$ renvoit en $O(n^4)$ un arbre T_{PI} raffiné par T et vérifiant PC et PI pour $\mathcal{R}$.*

Dans le pseudo-code, l'ensemble de triplets $\mathcal{R}$ donné en entrée à $PhySIC_{PI}$ est considéré par celui-ci comme une collection d'arbres sources (notamment pour le calcul de $\mathcal{R}_{PI}$). L'algorithme *PhySIC* (voir pseudo-code) construit un superarbre pour une collection d'arbres sources $\mathcal{T}$, en enchaînant simplement les algorithmes $PhySIC_{PC}$ et $PhySIC_{PI}$.

Théorème 3 *Etant donné une collection $\mathcal{T}$ de k arbres sources sur n feuilles,* PhySIC *renvoie en $O(kn^3 + n^4)$ un arbre ayant $L(\mathcal{T})$ comme ensemble de feuilles et vérifiant PC et PI.*

4 Un cas d'étude biologique : le superarbre des Primates

4.1 Inférence des arbres sources à partir de gènes nucléaires et d'éléments SINE

Notre méthode est ici appliquée au cas d'étude emblématique concernant la reconstruction du superarbre des Primates. Pour cela, nous avons utilisé trois arbres sources issus de données hétérogènes. Deux d'entre eux ont été inférés à partir d'alignements de séquences d'IRBP (gène codant la protéine de liaison au rétinol de la matrice interphotoréceptrice) et d'ADRA2B (gène du récepteur α2b-adrénergique) [22,21]. La souris (*Mus*, Rongeurs) et le lapin (*Oryctolagus*, Lagomorphes) constituent le groupe externe. Le représentant des hominoïdes pour ADRA2B est le chimpanzé (*Pan*), dont la séquence est issue de la base de données génomiques ENSEMBL (`www.ensembl.org` : *Pan troglodytes*, accession ENSPTRG00000012224). Les phylogénies pour ADRA2B et IRBP ont été établies en maximum de vraisemblance par PHYML [18], version 2.4.4, sous un modèle d'évolution des séquences GTR+Γ+I. La solidité des nœuds a été mesurée par le même logiciel, suite à 100 réplications de bootstrap. Le troisième arbre source a été collecté pour les Strepsirrhiniens (*i.e.*, lémurs et galagos) à partir des caractères de présence-absence d'éléments transposables de type SINE ("*Short Interspersed Nuclear Elements*") dans le génome des Primates [25, Fig. 2]. Les 61 caractères monolocus mis en évidence par ces auteurs ont été soumis à une analyse de maximum de parcimonie. Pour cela, le logiciel PAUP* [29] a été employé dans sa version 4b10, et 100 réplications de bootstrap ont été conduites à l'aide d'une recherche heuristique avec réarrangements de branches de type TBR, répétée 10 fois avec ordre aléatoire d'entrée des taxons.

Dans notre approche par superarbre, nous n'avons pris en compte que les clades suffisamment fiables, comme attendu dans un contexte d'inférence de l'Arbre de la Vie. Pour cela, chacune des topologies sources est restreinte aux clades dont la valeur de bootstrap dépasse 50% (*cf.* aussi [10]), c'est-à-dire présents dans l'arbre de consensus majoritaire de bootstrap. Des superarbres de Primates ont alors été reconstruits, en utilisant les méthodes MRP et *PhySIC*. La méthode MRP est appliquée sur la représentation matricielle des 47 nœuds des trois arbres sources. Les arbres les plus parcimonieux sont ensuite obtenus par PAUP*, à l'aide d'une recherche heuristique avec réarrangements de branches de type TBR, et 1000 répétitions d'un ordre aléatoire d'entrée des taxons. Le superarbre MRP résulte du consensus strict des 864 arbres les plus parcimonieux. Pour la méthode *PhySIC*, nous fournissons aussi le superarbre intermédiaire proposé par $PhySIC_{PC}$, ce qui permet de montrer l'apport de chaque étape ($PhySIC_{PC}$ et $PhySIC_{PI}$) de la nouvelle méthode.

4.2 Vote et veto : impact sur le superarbre des Primates

Les arbres sources reconstruits à partir des caractères d'ADRA2B, IRBP, et SINE (Figure 3, *partie supérieure*) correspondent aux idées actuelles sur la phylogénie des Primates, au moins pour

les nœuds soutenus par l'analyse de bootstrap [15]. Les superarbres reconstruits en utilisant les méthodes *PhySIC*, et MRP (Figure 3, *partie inférieure*) s'accordent sur le fait qu'une dichotomie fondamentale au sein des Primates sépare les Strepsirrhiniens (rectangles noirs) des Haplorrhiniens (rectangles clairs). Les Strepsirrhiniens se scindent en Lorisiformes (loris et galagos) et en Lémuriformes (lémurs et *Daubentonia*). Les Haplorrhiniens se divisent quant à eux en Tarsiiformes (tarsiers) et Anthropoïdes. Ce dernier clade comporte à son tour les Primates du Nouveau Monde (Platyrrhiniens) et ceux de l'Ancien Monde (Catarrhiniens).

Au sein des Catarrhiniens, le superarbre $PhySIC_{PC}$ propose un regroupement hétérodoxe, l'homme avec le gibbon (*Hylobates*) versus le chimpanzé avec les deux cercopithécoïdes (*Cercopithecus* et *Macaca*). Cette situation incorrecte provient d'un échantillonnage taxonomique particulier entre les arbres sources : l'homme et le chimpanzé ne sont pas simultanément présents dans les topologies sources, le premier se groupant avec le gibbon (IRBP) et le second avec les cercopithécoïdes (ADRA2B). Cette situation est détectée par $PhySIC_{PI}$, la seconde étape de la méthode *PhySIC*, ainsi que par MRP. Ces deux méthodes ne proposent donc pas de clade arbitraire dans le superarbre pour les 5 catarrhiniens (Fig. 3).

Parmi les Platyrrhiniens, le conflit topologique au sein des 4 genres ici échantillonnés est détecté par $PhySIC_{PC}$, la première étape de *PhySIC*. Ce dernier propose donc une multifurcation dans le superarbre, tandis que la résolution *Ateles* groupe-frère de *Pithecia* + (*Cebus* + *Callithrix*) est proposée dans l'arbre MRP. Ceci reflète le vote de MRP en faveur du sous-clade de Platyrrhiniens proposé par ADRA2B qui est incompatible avec l'information topologique *Cebus* + *Ateles* proposée par IRBP. Pour MRP, c'est ADRA2B qui l'emporte sur IRBP, simplement en raison du degré de résolution des deux topologies sources correspondantes : la première comporte deux résolutions (i.e., deux nœuds) au sein des Platyrrhiniens contre une seule pour la seconde. Nous retrouvons ici le biais "taille de clade" auquel est sensible MRP, et qui implique que le vote en faveur d'un nœud présent dans une zone de conflit entre arbres sources aura d'autant plus de poids que ce nœud fera partie d'un clade de taille importante [6]. La méthode *PhySIC* que nous proposons ici, fondée sur le *veto*, résiste à ce conflit d'information des topologies sources (Fig. 3).

Enfin, au sein des Strepsirrhiniens, *Lepilemur* apparaît dans le superarbre intermédiaire $PhySIC_{PC}$ comme groupe-frère de tous les Lémuriformes excepté *Daubentonia*, alors que cette information topologique n'est pas présente dans le seul arbre source (SINE) pour lequel *Lepilemur* est échantillonné. Ce résultat s'explique par le fait que la restriction de la topologie source IRBP aux taxons a-b-c-x (Fig. 3) conduit à la situation décrite en Fig. 2. MRP n'est pas sensible à ce problème, et notre méthode *PhySIC* non plus, grâce à l'étape $PhySIC_{PI}$.

5 Conclusion et perspectives

Dans cet article nous proposons une nouvelle méthode de superarbres de type *veto* ayant de bonnes propriétés mathématiques. Cette méthode, nommée *PhySIC* (*"PHYlogenetic Signal with Induction and non-Contradiction"*), renvoie un superarbre dont les clades sont non seulement non-contredits par les topologies sources, mais qui plus est induits par ces topologies. *PhySIC* est un raffinement de la méthode *BioBuild* [24] qui permet d'obtenir des superarbres plus résolus. Le superarbre des primates obtenu par *PhySIC* en temps polynomial comporte une résolution comparable à celui inféré par MRP. Cet exemple illustre également la différence de philosophie entre les méthodes de types *vote* et *veto* en termes de clades proposés.

Afin de proposer un superarbre qui ne contredit aucune topologie source, la version actuelle de *PhySIC* propose des multifurcations aux endroits conflictuels et conserve tous les taxons initiaux.

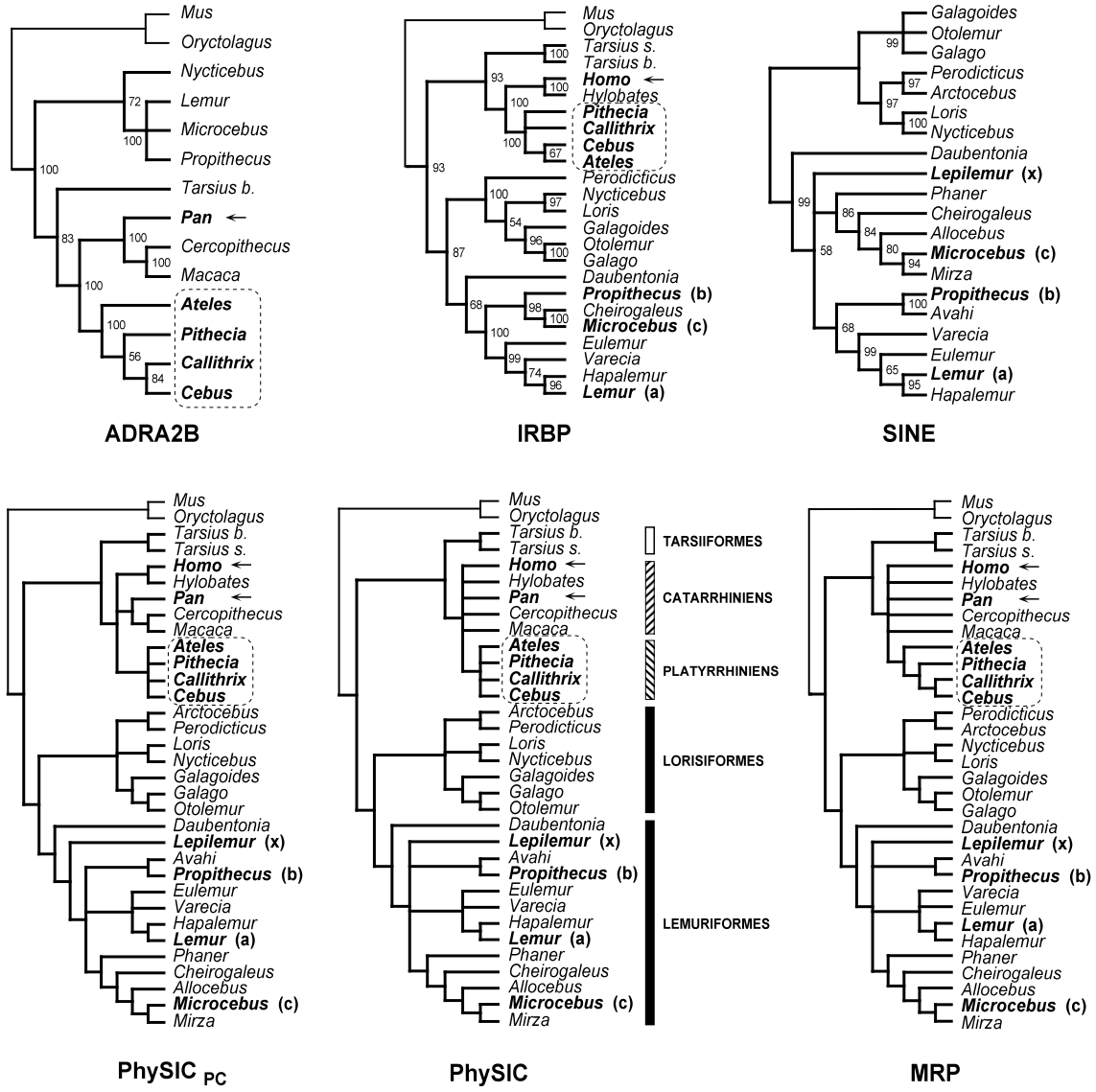

FIG. 3. *Moitié supérieure* : arbres de consensus majoritaire résultant de l'analyse par bootstrap des caractères d'ADRA2B et IRBP (en maximum de vraisemblance), et des éléments SINE (en maximum de parcimonie). Les pourcentages de bootstrap sont indiqués aux nœuds. Seuls les nœuds soutenus par des valeurs de bootstrap d'au moins 50% ont été retenus dans ces topologies sources. NB : Les deux espèces de *Tarsius* sont *T. bancanus* (*b*) et *T. syrichta* (*s*). *Moitié inférieure* : superarbres reconstruits à partir des trois topologies sources par les méthodes *PhySIC* et MRP, ainsi que l'arbre intermédiaire proposé par $PhySIC_{PC}$. Le cadre systématique des Primates est fourni sur le superarbre *PhySIC*. Les rectangles hachurés, blanc + hachurés, et noirs indiquent respectivement les Anthropoïdes (Catarrhiniens + Platyrrhiniens), les Haplorrhiniens, et les Strepsirrhiniens. Les branches fines conduisent au groupe externe. Les taxons en gras sont traités de manière différente par les trois algorithmes, et illustrent plusieurs situations : (i) les flèches pointant vers l'homme et le chimpanzé soulignent leur mauvais positionnement dans l'approche intermédiaire $PhySIC_{PC}$; (ii) les lettres a-b-c-x correspondent aux taxons que $PhySIC_{PC}$ agence de manière arbitraire (*cf.* Fig. 2) ; et (iii) les boîtes en pointillés contiennent les Platyrrhiniens pour lesquels MRP contredit la topologie source IRBP. Notons que *PhySIC* renvoie ici un superarbre deux fois plus résolu que celui renvoyé par la méthode *BioBuild* de [24], ce qui illustre la pertinence des raffinements décrits dans la section 3.

Dans le cas où plusieurs taxons ont une position très variable d'un arbre source à l'autre (par exemple en raison de transferts latéraux ou de la présence de paralogues), une autre approche est envisageable. Elle consiste à enlever les quelques taxons problématiques – mais seulement dans les arbres sources conflictuels – et à proposer un arbre plus résolu sur l'ensemble des taxons conservés [3]. Les développements futurs de *PhySIC* devront donc incorporer des études par simulation pour évaluer les performances de cette approche dans la reconstruction de l'Arbre de la Vie, et associer la suppression ciblée de taxons à l'introduction de multifurcations en vue de produire un superarbre aussi informatif que possible.

La méthode *PhySIC* est implémentée et disponible à l'adresse `www.lirmm.fr/~vberry/TOL/physic.cgi`.

Remerciements

Ce travail a bénéficié du soutien financier et institutionnel de l'ACI Informatique-Mathématique-Physique en Biologie Moléculaire [ACI IMP-Bio], et de l'Equipe-Projet multi-laboratoires CNRS-STIC "Méthodes informatiques pour la biologie moléculaire". Il représente la contribution N° 2006-035 de l'Institut des Sciences de l'Evolution de Montpellier (UMR 5554 - CNRS).

Références

[1] A. V. Aho, Y. Sagiv, T. G. Szymanski, and J. D. Ullman. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM J. Comp.*, 10(3) :405–421, 1981.

[2] B.R. Baum and M.A. Ragan. The MRP method. In O.R.P. Bininda-Emonds, editor, *Phylogenetic supertrees : combining information to reveal the Tree of Life*, pages 17–34. Kluwer, 2004.

[3] V. Berry and F. Nicolas. Maximum agreement and compatible supertrees. In S. C. Sahinalp, S. Muthukrishnan, and U. Dogrusoz, editors, *Proceedings of CPM*, volume 3109 of *LNCS*, pages 205–219, 2004.

[4] V. Berry, V. Ranwez, P.-H. Fabre, and E.J.P. Douzery. Vote ou veto pour la reconstruction des superarbres phylogénétiques. Technical Report 06021, LIRMM, 2006.

[5] O.R.P. Bininda-Emonds. *Phylogenetic supertrees (combining information to reveal the tree of life)*, volume 4 of *computational biology series*. Kluwer academic publishers, 2004.

[6] O.R.P. Bininda-Emonds and H.N. Bryant. Properties of matrix representation with parsimony analyses. *Syst. Biol.*, 47(3) :497–508, 1998.

[7] D. Bryant. A classification of consensus methods for phylogenies. In M. Janowitz, F.-J. Lapointe, F.R. McMorris, B. Mirkin, and F.S. Roberts, editors, *Bioconsensus*, DIMACS, pages 163–184. AMS, 2002.

[8] A. Criscuolo, V. Berry, E.J.P. Douzery, and O. Gascuel. SDM : une méthode de distance rapide pour les études de phylogénomique. In G. Perrière, A. Guénoche, and C. Geourjon, editors, *Journées Ouvertes Biologie Informatique Mathématiques*, pages 231–243. JOBIM, Lyon (France), 2005.

[9] P. Daniel. Supertree methods : some new approaches. Master's thesis, University of Canterbury, 2004.

[10] V. Daubin, M. Gouy, and G. Perrière. A phylogenomic approach to bacterial phylogeny : Evidence of a core of genes sharing a common history. *Genome Res.*, 12 :1080–1090, 2002.

[11] T.J. Davies, T.G. Barraclough, M.W. Chase, P.S. Soltis, D.E. Soltis, and V. Savolainen. Darwin's abominable mystery : Insights from a supertree of the angiosperms. *Proc. Natl. Acad. Sci. USA*, 101 :1904–1909, 2004.

[12] F. Delsuc, H. Brinkmann, and H. Philippe. Phylogenomics and the reconstruction of the tree of life. *Nature Rev. Genet.*, 6 :361–375, 2005.

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

[14] P. A. Goloboff and D. Pol. Semi-strict supertrees. *Cladistics*, 18(5) :514–525, 2002.

[15] M. Goodman, L.I. Grossman, and D.E. Wildman. Moving primate genomics beyond the chimpanzee genome. *Trends Genet.*, 21(9) :511–517, 2005.

[16] A. G. Gordon. Consensus supertrees : the synthesis of rooted trees containing overlapping sets of labelled leaves. *Journal of Classification*, 3 :335–348, 1986.

[17] S. Grunewald, M.A. Steel, and M.S. Swenson. Phylogenetic closure operations in phylogenetics. Submitted to Mathematical Biosciences, 2006.

[18] S. Guindon and O. Gascuel. A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood. *Syst. Biol.*, 52(5) :696–704, 2003.

[19] F.-J. Lapointe and G. Cucumel. The average consensus procedure : Combination of weighted trees containing identical or overlapping sets of taxa. *Syst. Biol.*, 46 :306–312, 1997.

[20] 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. *Mol. Biol. Evol.*, 9 :1740–1752, 2004.

[21] C. Poux, P. Chevret, D. Huchon, W.W. de Jong, and E.J.P. Douzery. Arrival and diversification of caviomorph rodents and platyrrhine primates in South America. *Syst. Biol.*, 55(2) :228–244, 2006.

[22] C. Poux and E.J.P. Douzery. Primate phylogeny, evolutionary rate variations, and divergence times : A contribution from the nuclear gene IRBP. *Am. J. Phys. Anthropol.*, 124 :1–16, 2004.

[23] A. Purvis. A composite estimate of primate phylogeny. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 348(1326) :405–421, 1995.

[24] V. Ranwez, V. Berry, and E.J.P. Douzery. Desirable properties of supertrees to build the Tree of Life. Technical report, Université de Montpellier II, 2006.

[25] C. Roos, J. Schmitz, and H. Zischler. Primate jumping genes elucidate strepsirrhine phylogeny. *Proc. Natl. Acad. Sci. USA*, 101(29) :10650–10654, 2004.

[26] C. Semple and M. Steel. A supertree method for rooted trees. *Discrete Appl. Math.*, 105 :147–158, 2000.

[27] C. Semple and M. Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, 2003.

[28] M. A. Steel, A. W. M. Dress, and S. Böcker. Simple but fundamental limitations on supertree and consensus tree methods. *Systematic Biology*, 49(2) :363–368, 2000.

[29] D.L. Swofford. *PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4*. Sinauer Associates, Sunderland, Massachusetts, 1998. version 4.0b2.

[30] J. L. Thorley and M. Wilkinson. A view of supertrees methods. In M. F. Janowitz, F.-J. Lapointe, F. R. McMorris, and F. S. Roberts, editors, *Bioconsensus*, volume 61 of *Discrete Mathematics and Theoretical Computer Science*, pages 185–194. DIMACS, 2003.