An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications, and transfers

Jean-Philippe DOYON[1], Celine SCORNAVACCA[2], Gergely J. SZÖLLŐSI[3], Vincent Ranwez[4] and Vincent Berry[1]

[1] LIRMM, CNRS - Univ. Montpellier 2, France.
[2] Center for Bioinformatics (ZBIT), Tuebingen Univ., Germany.
[3] LBBE, CNRS - Univ. Lyon 1, France.
[4] ISEM, CNRS - Univ. Montpellier 2, France.

Abstract *(Motivation) Tree reconciliation is an approach that explains the discrepancies between two evolutionary trees by a number of events such as speciations, duplications, transfers and losses. It has important applications in ecology, biogeography and genomics, for instance to decipher relationships between homologous sequences. (Results) We provide a fast and exact reconciliation algorithm according to a parsimony criterion that considers duplication, transfer and loss events. We also present experimental results that give first insights on the conditions under which parsimony is able to accurately infer evolutionary scenarios involving such events. Over all, parsimony performs well under realistic cases, as well as for relatively high duplication and transfer rates. As expected, transfers are in general less accurately recovered than duplications. Availability:* `www.lirmm.fr/phylariane/`

Keywords reconciliation, gene and species trees, transfers, duplications, losses, parsimony.

Un algorithme de parcimonie efficace pour la réconciliation d'arbres de gènes/espèces avec pertes, duplications et transferts

Résumé *(Motivation) La réconciliation d'arbres est une approche qui permet d'expliquer les différences entre deux arbres évolutifs par le biais d'événements comme les spéciations, duplications, transferts et pertes de gènes. Cette approche est appliquée en écologie, en biogéographie et en génomique, par exemple pour étudier les relations entre séquences homologues. (Résultats) Nous proposons un algorithme de réconciliation efficace et exact, basé sur un critère de parcimonie et prenant à la fois en compte les duplications, les transferts et les pertes de gènes. Des résultats expérimentaux montrent que la parcimonie fonctionne bien dans des conditions réalistes, mais aussi dans le cas de taux de duplication et de transfert relativement élevés. Sans surprise, les transferts sont les événements les plus difficiles à inférer correctement.*

Mots-clefs réconciliation, arbres de gènes et d'espèces, transferts, duplications, pertes, parcimonie.

1 Introduction

L'histoire évolutive des organismes vivants est généralement représentée par un *arbre d'espèces* dont les nœuds internes représentent des événements de spéciations [5,20]. L'histoire évolutive d'un ensemble de séquences homologues dérivées d'une séquence ancestrale commune (*famille de gènes*) est elle aussi représentée par un arbre, on parle alors d'un *arbre de gènes*. Contrairement à un arbre d'espèces, un arbre de gènes résulte non seulement d'événement de spéciations, mais aussi de transferts, de duplications et de pertes de matériel génétique. Certains auteurs pensent que les transferts chez les procaryotes (et à proximité de l'ancêtre commun) sont si importants qu'un *réseau de la vie* est plus approprié qu'une simple arborescence [6,7]. Des études complémentaires semblent toutefois indiquer que les transferts n'oblitèrent pas complètement le signal évolutif de spéciation et qu'un arbre de la vie peut encore être discerné malgré le bruit qu'ils engendrent [5,13,20]. Même si ce débat n'est pas encore clos, il a d'ores et déjà engendré des progrès considérables. Par exemple, il est bien établi

que la détection de transferts par approche phylogénétique est plus fiable que par comparaisons de séquences [13,15,24]. L'approche phylogénétique la plus populaire est la *réconciliation d'arbres* et se base sur une comparaison détaillée d'un arbre de gènes avec un arbre d'espèces référent. Ce dernier n'est pas toujours connu mais peut être estimé de manière satisfaisante par des analyses phylogénomiques sur des séquences moléculaires de nombreux gènes ou des caractéristiques de génomes complets [15].

Les méthodes de réconciliation permettent d'expliquer les différences possibles entre un arbre de gènes et un arbre d'espèces suite à des événements de transfert, de duplication et de perte. Une réconciliation d'arbres plonge l'arbre de gènes dans l'arbre d'espèces, représenté par un ensembles de tubes, et associe chaque nœud interne de l'arbre de gènes à un événement évolutif particulier (i.e. spéciation, duplication, transfert ou perte) [18].

Les approches pour réconcilier un arbre de gènes G et un arbre d'espèces S se basent sur des modèles combinatoires [8,18,11,12,9] ou probabilistes [1,23]. Ces derniers intègrent un plus grand nombre de paramètres et offrent une meilleure représentation de l'évolution génomique que les modèles combinatoires, mais sont beaucoup plus exigeants en mémoire et temps de calcul. C'est pourquoi seuls les modèles combinatoires sont utilisables pour des études phylogénomiques, considérant régulièrement plusieurs dizaines de milliers de familles de gènes [19].

Cependant, au vue des avancées rapides des nouvelles technologies de séquençage qui permettent d'obtenir de nouveaux génomes complets (c.f. [2]) en peu de temps, même les méthodes combinatoires sont en voie de devenir trop lentes. Pour gérer les défis liés à ce déluge de données, nous proposons un modèle combinatoire de réconciliation qui considère tous les événements cités ci-dessus (spéciation, duplication, transfert et perte) et un algorithme de complexité beaucoup plus faible que ceux actuellement proposés.

Formellement, nous considérons le problème d'optimisation nommé *Réconciliation la Plus Parcimonieuse* (ou *MPR* [1]). Pour un arbre d'espèces S, un arbre de gènes G et des coûts associés aux événements de spéciation, de duplication, de transfert et de perte (respectivement notés $\mathbb{S}$, $\mathbb{D}$, $\mathbb{T}$, et $\mathbb{L}$), il s'agit de trouver une réconciliation de coût minimum. Le coût d'une réconciliation est la somme des coûts des événements induits par le plongement de G dans S.

Dès que l'on considère les transferts, le problème MPR est NP-complet, même dans le cas où l'on doit

[1] Cet acronyme est lié à l'intitulé anglophone du problème : "Most Parsimonious Reconciliation".

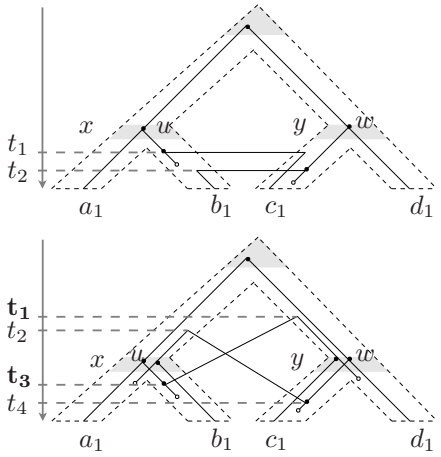


Fig. 1. Deux scénarios de réconciliation entre l'arbre de gènes G (traits pleins) et l'arbre d'espèces S (tubes). Ces scénarios induisent des pertes (feuilles marquées par le symbole $\circ$), des transferts (arêtes horizontales) et fournissent une version plus complète de G notée G^o. (En haut) Un scénario temporellement consistant. (En bas) Un scénario temporellement inconsistant : le transfert allant du donneur au temps t_3 vers le receveur au temps t_1 implique que u précède w. Parallèlement, l'autre transfert implique que $t_4 = t_2$, et donc que w précède u.

réconcilier un seul arbre de gènes binaire avec un arbre d'espèces binaire [12,22]. Ceci est directement lié au fait que les transferts induisent des contraintes chronologiques entre les nœuds de S qui sont difficiles à respecter. En effet, comme les transferts se passent horizontalement entre deux espèces vivant au même moment, ils imposent des contraintes temporelles entre deux nœuds de S (où l'un n'est pas ancêtre de l'autre) qui s'ajoutent aux contraintes initiales (un nœud est nécessairement plus ancien que ses descendants). Des inconsistances temporelles peuvent alors apparaître dans toute réconciliation ayant plusieurs transferts. En effet, ces transferts peuvent alors induire des contraintes temporelles entre les nœuds de S qui sont mutuellement incompatibles (cf Fig. 1).

Plusieurs approches ont été proposées pour surmonter la difficulté liée aux contraintes temporelles. Une première solution [10,12] est de définir à l'avance (par des moyens externes à la méthode de réconciliation) les paires de branches de S entre lesquelles les transferts sont autorisés. De nouvelles branches horizontales sont ajoutées pour connecter de telles paires de branches et le graphe obtenu de S est appelé *graphe d'espèces* $\mathcal{S}$. La réconciliation plonge l'arbre de gènes G non plus dans S mais dans $\mathcal{S}$ et une réconciliations la plus parcimonieuse se calcule en temps $O(|\mathcal{S}|^3 \cdot |G|)$. Cependant, calculer un graphe d'espèces induisant une réconciliation la plus parcimonieuse est un problème NP-complet [10]. Une approche plus prometteuse est de considérer une variante réaliste du problème MPR où l'arbre d'espèces S est daté. En

effet, il est généralement possible de dater les nœuds de S basé sur les séquences moléculaires et quelques points de calibrations. Plusieurs auteurs ont donc creusé cette piste qui, initialement proposée dans le cadre d'études de coévolution [16,3,17], est désormais reprise dans le contexte de la réconciliation d'arbre de gènes/arbre d'espèces [9,22]. La datation des nœuds de S permet d'assigner un intervalle de temps à chaque branche. Il est alors possible d'assurer la consistance individuelle de chaque transfert en vérifiant que la branche dite *donneuse* et celle dite *receveuse* ont des intervalles de temps dont l'intersection est non-vide (le transfert est dit temporellement et localement consistant). La variante du problème MPR respectant cette contrainte locale peut être résolue en $O(\max(|S| \cdot |G|)^3)$ par programmation dynamique [17]. Cependant, si deux transferts sont consistants de façon locale mais pas de façon conjointe (cf Fig. 1), alors la réconciliation n'est pas globablement consistante. De telles inconsistances peuvent être corrigées a posteriori en modifiant certains événements $\mathbb{T}$ [16,17], mais l'optimalité de la réconciliation obtenue n'est plus garantie et l'approche proposée n'est qu'une heuristique pour le MPR.

Une solution pour calculer une réconciliation globalement consistante est de subdiviser la période couverte par S en temps élémentaires, d'associer chacune de ses branches à un de ces temps et de permettre un transfert seulement entre un donneur et un receveur d'un même temps élémentaire. Cette approche permet, dans le cas d'arbres binaires, d'obtenir des algorithmes exacts pour résoudre le problème MPR. Les algorithmes proposés par [14] et [9] utilisent tout deux cette approche. Le premier a une complexité théorique en $O(|S|^4 \cdot |G|^4)$ tandis que le second est en $O(|S|^4 \cdot k^4 \cdot |G|)$, où k est le nombre de nœuds résultants de la subdivision de S ([2]). Ces complexités, bien que polynomiales, restent élevées et impliquent des temps de calcul importants.

Certains des algorithmes décrits ci-dessus s'appuient sur un modèle combinatoire de réconciliation issu de travaux se focalisant sur les duplications et pour lesquels chaque nœud de G est *couplé* avec un seul nœud de S. Toutefois, un tel couplage est insuffisant pour les transferts car il ne peut explicitement indiquer à la fois le donneur et le receveur d'un transfert immédiatement suivit d'une perte. Cette difficulté a conduit certains auteurs à ne considérer qu'une restriction du problème MPR qui néglige le coût des pertes [12,14,22].

[2] Selon ces auteurs, des modifications non-mentionnées dans le papier permettent d'obtenir une version en $O(|S|^2 \cdot k^2 \cdot |G|)$, mais dont la correction reste à montrer.

Étant donnés un arbre de gènes G et un arbre d'espèces S daté, nous présentons dans cet article un algorithme polynomial de réconciliation basé sur un modèle combinatoire où les quatre types d'événements évolutifs ($\mathbb{DTLS}$) sont considérés [3]. Contrairement aux approches existantes, notre algorithme gèrent correctement la combinaison d'événements $\mathbb{T}+\mathbb{L}$. Notre modèle s'appuie sur une subdivision S' de S similaire à celle introduite par [9,14,22] et permet de résoudre le MPR en $O(|S'| \cdot |G|)$. Nous explorons ensuite la question fondamentale suivante : *La parcimonie est elle un critère pertinent pour identifier le véritable scénario évolutif d'une famille de gènes?*

2 Méthodes

2.1 Définitions et notations basiques

Soit T un arbre où les ensembles de nœuds et de branches sont respectivement notés $V(T)$ et $E(T)$ et seulement ses feuilles sont étiquetées. $r(T)$, $L(T)$ et $\mathcal{L}(T)$ dénotent respectivement sa racine, l'ensemble de ses feuilles et l'ensemble des étiquettes de ses feuilles. Nous allons adopter la convention que la racine est au haut de l'arbre et ses feuilles dans le bas.

Une branche de T est dénotée $(u, v) \in E(T)$, où u est le père de v. Pour un nœud u de T, T_u dénote le sous-arbre de T enraciné en u, u_p est son père, (u_p, u) est la branche parent de u et $T_{(u_p,u)}$ dénote le sous-arbre de T enraciné avec la branche (u_p, u). Un nœud interne u de T a un ou deux fils, notés respectivement $\{u_1\}$ ou $\{u_1, u_2\}$. Il est important de souligner qu'un arbre T est non-ordonné et les deux fils u_1 et u_2 d'un nœud interne u de T sont interchangeables. Autrement dit, u_1 peut être arbitrairement sélectionné comme l'unique fils de u qui respecte une contrainte donnée. Pour deux nœuds u, u' de T, u' est dit un *descendant* (resp. strict) de u si u est sur l'unique chemin entre u' et $r(T)$ (resp. et $u \neq u'$).

Un nœud interne u de T est dit *artificiel* lorsqu'il a un seul fils. La *contraction* d'un nœud artificiel signifie que ce nœud est enlevé de l'arbre et que les deux branches adjacentes sont jointes. Un arbre T' est dit une *subdivision* d'un arbre T si la contraction récursive de tous les nœuds artificiels de T' donne T.

Un *arbre d'espèces* S est un arbre binaire tel que chaque élément de $\mathcal{L}(S)$ représente une espèce existante et étiquette exactement une feuille de S (il y a une bijection entre $L(S)$ et $\mathcal{L}(S)$). Un *arbre de gènes* G est un arbre binaire. Dorénavant, nous considérons un arbre d'espèces S et un arbre de gènes G tel que

[3] L'algorithme présenté considère un coût de spéciation null, mais il est facile de l'adapter pour un coût non null.

$\mathcal{L}(G) \subseteq \mathcal{L}(S)$ et $\mathcal{L} : L(G) \to L(S)$ dénote la fonction qui couple chaque feuille de G à l'unique feuille de S avec la même étiquette. Aussi, le terme arc réfère à une branche de G et le terme branche est pour S.

Dans le reste de l'article, nous assumons que de l'information temporelle est donnée pour l'arbre d'espèces S (c'est-à-dire qu'une période de temps est associée à chaque événement de spéciation) et que l'arbre S est ultramétrique.

Une fonction d'étiquetage temporel pour S est notée $\theta_S : V(S) \to \mathbb{R}$ et est telle que pour chaque feuille $x \in L(S)$, $\theta_S(x) = 0$, et pour chaque paire de nœuds $x, x' \in V(S)$, que x' soit un descendant strict de x implique $\theta_S(x') < \theta_S(x)$. Cet étiquetage temporel est interprété de la façon suivante: chaque feuille de S corresponds à une espèce contemporaine qui existe au temps présent $t = 0$ et chaque nœud interne corresponds à une espèce ancestrale qui a donné naissance à deux lignées au temps passé $t > 0$.

DÉFINITION 2.1. *Soit un arbre T et un sous-ensemble de feuillles $K \subseteq L(T)$.* L'arbre homéomorphique *de T qui connecte K est noté $T|_K$ et est le plus petit sous-arbre induit de T tel que $L(T|_K) = K$.*

Nous introduisons ci-dessous le concept d'un scénario d'évolution d'un gène débutant à $r(S)$ et évoluant dans S par des événements $\mathbb{DTLS}$. Un tel scénario génère un *arbre de gènes complet* noté G^o, où l'ensemble de feuilles est formé de gènes contemporains mais aussi de gènes perdus durant le scénario (cf Fig. 1 et Fig. 2). Formellement, $L(G^o) = L_{\mathbb{C}}(G^o) \cup L_{\mathbb{L}}(G^o)$, où $L_{\mathbb{C}}(G^o)$ et $L_{\mathbb{L}}(G^o)$ sont disjoints et correspondent respectivement aux gènes contemporains ($\mathbb{C}$) et perdus ($\mathbb{L}$).

DÉFINITION 2.2. *Soit un arbre de gènes observé G et un arbre d'espèces S, avec sa fonction d'étiquetage temporel θ_S. Un* scénario $\mathbb{DTLS}$ *pour G le long de S est noté (G^o, M, θ^o_G), où G^o est l'arbre de gènes complet, $M : V(G^o) \to V(S)$ couple chaque nœud u de G^o à un nœud de S et $\theta^o_G : V(G^o) \to [0, \theta_S(r(S))]$ associe chaque noeud u de G^o à une étiquette temporelle de S. Les événements $\mathbb{DTLS}$ correspondants et associés aux nœuds $u \in V(G^o)$ sont définis ci-dessous.*

1. *Si $M(u) = x$, $M(u_1) = x_1$ et $M(u_2) = x_2$, alors u est un événement $\mathbb{S}$.*
2. *Si $M(u) = M(u_1)$ et $M(u) = M(u_2)$, alors u est un événement $\mathbb{D}$.*
3. *Si u est une feuille de G^o qui n'est pas dans G, alors u est un événement $\mathbb{L}$*
4. *Si $M(u_1) = M(u) = x$, $M(u_2) = y$ et y n'est ni un ancêtre, ni un descendant de x, alors u est un événement $\mathbb{T}$, (x_p, x) et (y_p, y) correspondant respectivement aux branches donneuse et receveuse.*

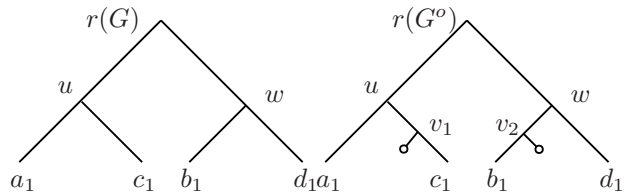


Fig. 2. (À gauche) Un arbre de gènes G avec quatre feuilles a_1, b_1, c_1, et d_1, appartenant respectivement aux espèces contemporaines A, B, C et D (cf Fig. 1). (À droite) Un arbre de gènes complet G^o.

Un scénario $\mathbb{DTLS}$ est dit consistant *si et seulement si les contraintes suivantes sont respectées. Premièrement, l'arbre de gènes homéomorphique $G^o|_{L_{\mathbb{C}}(G^o)}$ est G. Deuxièmement, pour un événement $\mathbb{T}$ tel que décrit ci-dessus (c'est-à-dire en (4)) $[\theta_S(x), \theta_S(x_p)] \cap [\theta_S(y), \theta_S(y_p)] \neq \emptyset$. Troisièmement, pour chaque arc $(u_p, u) \in E(G^o)$, $\theta^o_G(u_p) > \theta^o_G(u)$.*

Le coût d'un tel scénario est noté $Coût(G^o, M, \theta^o_G) = d\delta + t\tau + l\lambda$, où d, t, et l dénotent respectivement le nombre d'événements $\mathbb{D}$, $\mathbb{T}$ et $\mathbb{L}$, et δ, τ et λ sont leurs coûts respectifs.

Le problème d'optimization considéré dans cet article, nommé *MPR*, est defini ci-dessous :

Entrées. Un arbre d'espèces S avec une fonction d'étiquetage temporel $\theta_S : V(S) \to \mathbb{R}$, un arbre de gènes observé G, la fonction d'association entre feuilles $\mathcal{L} : L(G) \to L(S)$ et les trois coûts δ, τ et λ des événements $\mathbb{DTL}$.

Résultats. Un scénario $\mathbb{DTLS}$ consistant (G^o, M, θ^o_G) pour G le long de S qui minimise $Coût(G^o, M, \theta^o_G)$.

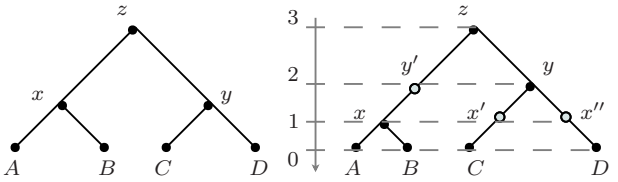
2.2 Un modèle de réconciliation efficace

Pour obtenir un modèle efficace, l'arbre d'espèces est subdivisé pour obtenir une discrétization du temps et permettre de calculer une réconciliation la plus parcimonieuse (de maniıère similaire à [3,23]).

DÉFINITION 2.3. *Pour un arbre (binaire) d'espèces S et une fonction d'étiquetage $\theta_S : V(S) \to \mathbb{R}$, soit S' la subdivision de S suivante : pour chaque nœud $x \in V(S) \setminus L(S)$ et chaque branche$(y_p, y) \in E(S)$ tel que $\theta_S(y_p) > \theta_S(x) > \theta_S(y)$, un nœud artificiel est inséré sur la branche (y_p, y) au temps $\theta_S(x)$. La subdivision nous permet de définir une fonction d'étiquetage temporelle pour S' en se basant seulement sur sa topologie: pour chaque $x \in V(S'), \theta'_{S'}(x)$ est le nombre de branches qui séparent x d'une de ses feuilles descendantes (toutes à la même distance).*

L'étiquetage temporelle d'une branche (x_p, x) de S' est noté $\theta'_{S'}(x_p, x) = \theta'_{S'}(x)$. Pour un temps t, $E_t(S') = \{(x_p, x) \in E(S') : \theta'_{S'}(x_p, x) = t\}$ dénote l'ensemble des branches de S' localisées au temps t.

Notre modèle de réconciliation défini ci-dessous se base sur les événements et groupes d'événements



(a) (b)

Fig. 3. (a) L'arbre d'espèces S et (b) sa subdivision S'. Les nœuds artificiels de S' sont représentés en gris et dénotés y', x' et x'', où $\theta'_{S'}(x) = \theta'_{S'}(x') = \theta'_{S'}(x'')$ et $\theta'_{S'}(y) = \theta'_{S'}(y')$.

$\mathbb{DTLS}$, en plus d'un événement dit "null" et noté $\varnothing$ (cf Fig. 4 et Fig. 5).

DÉFINITION 2.4. *Une* réconciliation *entre G et S est notée α et associe chaque arc $(u_p, u) \in E(G)$ à une séquence ordonnée de branches de la subdivision S' notée $\alpha(u_p, u)$. Dans cette séquence de ℓ éléments, $\alpha_i(u_p, u)$ dénote le i-ième élément pour $1 \leq i \leq \ell$. Chaque branche $\alpha_i(u_p, u)$, dénotée ci-dessous (x_p, x), respecte une et seulement une des contraintes suivantes (cf Fig. 4).*

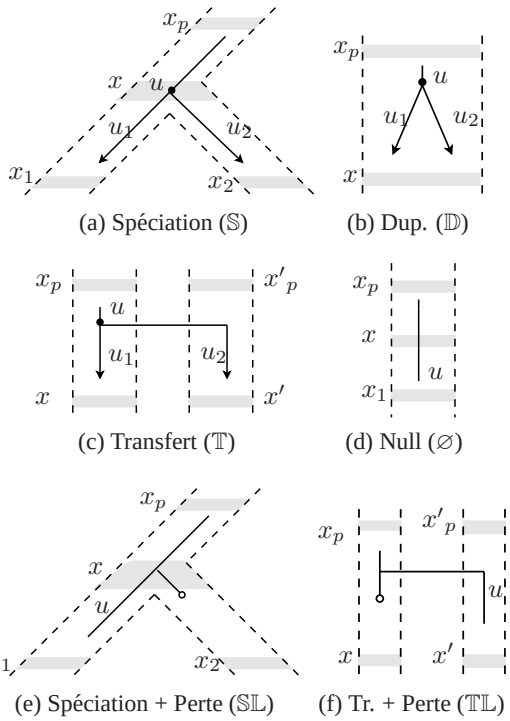
Premièrement, considérons que (x_p, x) est la dernière branche $\alpha_\ell(u_p, u)$ de la séquence. Si u est une feuille de G, alors x est l'unique feuille de S' ayant la même étiquette que u (c'est-à -dire que $x = \mathcal{L}(u)$) (Contrainte de couplage contemporain). Sinon, un des cas ci-dessous est vérifié.

- $\{\alpha_1(u, u_1), \alpha_1(u, u_2)\} = \{(x, x_1), (x, x_2)\}$ *(événement $\mathbb{S}$);*
- $\alpha_1(u, u_1)$ *et* $\alpha_1(u, u_2)$ *sont tous les deux égales à* (x_p, x) *(événement $\mathbb{D}$);*
- $\{\alpha_1(u, u_1), \alpha_1(u, u_2)\} = \{(x_p, x), (x'_p, x')\}$*, où* (x'_p, x') *est une branche de S' différente de* (x_p, x) *et localisée au temps* $\theta'_{S'}(x_p, x)$ *(événement $\mathbb{T}$);*

Si (x_p, x) n'est pas la dernière branche $\alpha_\ell(u_p, u)$ de la séquence, un des cas suivants est vérifié.

- *x est un nœud artificiel de S' avec un seul fils x_1, et la prochaine branche $\alpha_{i+1}(u_p, u)$ est (x, x_1) (événement $\varnothing$).*
- *x n'est pas un nœud artificiel et $\alpha_{i+1}(u_p, u) \in \{(x, x_1), (x, x_2)\}$ (événement $\mathbb{SL}$);*
- *$\alpha_i(u_p, u)$ et $\alpha_{i+1}(u_p, u)$ sont deux branches différentes localisées au temps t de S' (événement $\mathbb{TL}$).*

Pour une branche (u_p, u) de G, toute réconciliation *parcimonieuse* α entre G et S contient au plus un événement $\mathbb{TL}$ localisé à un temps donné de S'. Dans ce modèle on peut montrer qu'un événement $\mathbb{TL}$ doit être suivi par un événement différent ($\mathbb{S}$, $\mathbb{D}$, $\mathbb{T}$, $\varnothing$ ou $\mathbb{SL}$) (cf Propriété 2.5). Ceci permet de développer un algorithme efficace pour résoudre le MPR.



(a) Spéciation ($\mathbb{S}$) (b) Dup. ($\mathbb{D}$)

(c) Transfert ($\mathbb{T}$) (d) Null ($\varnothing$)

(e) Spéciation + Perte ($\mathbb{SL}$) (f) Tr. + Perte ($\mathbb{TL}$)

Fig. 4. Les six événements $\mathbb{DTLS}$ de la Déf. 2.4. L'arbre de gènes complet G^o est plongé dans la subdivision S' d'un arbre d'espèces S, où un arc de G est représenté par une ligne pleine, une branche de S' par un tube pointillé (zone blanche) et un nœud de S' par une zone grise. Six couplages sont possibles pour un arc (u_p, u) de G^o et une branche (x_p, x) de la séquence $\alpha(u_p, u)$, où le symbol ➤ sur (u_p, u) représente le cas où (x_p, x) est la première branche de la séquence dans S', le symbole • sur le nœud u_p représente le cas où (x_p, x) est la dernière branche de la séquence.

PROPRIÉTÉ 2.5. *Considérons une réconciliation parcimonieuse α entre G et S, un arc (u_p, u) de G et un temps t de S'. La séquence $\alpha(u_p, u)$ contient au plus deux branches localisées au temps t. S'il existe de telles branches, notées $\alpha_i(u_p, u)$ et $\alpha_j(u_p, u)$, alors elles sont adjacentes dans la séquence (c'est-à -dire $|i - j| = 1$).*

2.3 Un algorithme efficace pour MPR

Basé sur le modèle de réconciliation précédent, nous proposons dans cette section un algorithme poynomial en temps et en espace pour résoudre le problème MPR.

Considérons un arc $(u_p, u) \in E(G)$, une branche $(x_p, x) \in E(S')$ et un temps $t = \theta'_{S'}(x_p, x)$.

Notons par $Coût(u, x)$ le coût minimal parmi toutes les réconciliations entre $G_{(u_p,u)}$ et la forêt de sous-arbres de S' enracinés en une branche localisée au temps t, tel que (x_p, x) est la première branche dans la séquence associée à (u_p, u). $Coût(r(G), r(S'))$ correspond au coût minimal d'une réconciliation entre G et S. L'algorithme de programmation dynamique (voir le pseudo-code) remplit la matrice $Coût : V(G) \times$

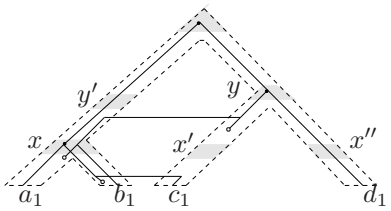


Fig. 5. Une réconciliation α pour un arbre G et une subdivision S' (voir Fig. 3), où chaque branche (resp. nœud) de G est représentée par une ligne pleine (resp. le symbole $\bullet$) et chaque branche (resp. nœud) de S' par un tube pointillé sur fond blanc (resp. gris). Le chemin de la branche (u, b_1) est $[(y, x'), (y', x'), (x, B)]$.

$V(S') \rightarrow \mathbb{N}$ avec deux boucles imbriquées: une qui visite tous les arcs de G selon un parcours de bas-en-haut et une qui visite tous les étiquettes temporelles de S' en débutant au temps présent $t = 0$ et en remontant progressivement le temps. Pour l'arc (u_p, u) et le temps t actuellement visités (respectivement aux lignes 3 et 4), deux boucles consécutives sur toutes les branches $(x_p, x) \in E_t(S')$ calculent le coût minimal de coupler (u_p, u) avec (x_p, x) selon les six événements ou groupe d'événements $\mathbb{S}$, $\mathbb{D}$, $\mathbb{T}$, $\varnothing$, $\mathbb{SL}$ et $\mathbb{TL}$ (cf Fig. 4). Pour une branche $(x_p, x) \in E_t(S')$, la première boucle (lignes 5 à 20) calcule le coût minimal pour les cinq premiers événements, la deuxième boucle (lignes 21 à 24) calcule ce coût pour les $\mathbb{TL}$s et $Coût(u, x)$ est le coût minimum des six événements.

PROPRIÉTÉ 2.6. *La taille de S' est bornée en dessous par $\Omega(n)$ et au dessus par $O(n^2)$.*

THÉORÈME 2.7. *L'Algorithme 1 résout le problème MPR en temps et en espace $\Theta(|S'|.|G)$.*

Algorithm 1 Calcule $Coût(r(G), r(S'))$.

1: Construire la subdivision S' de S de la façon décrite à la Définition 2.3
2: La matrice $Coût : V(G) \times V(S') \rightarrow \mathbb{N}$ est initialisée ci-dessous: si $u \in L(G)$, $x \in L(S')$ et $\mathcal{L}(u) = x$, alors $Coût(u, x) \leftarrow 0$. Sinon, $Coût(u, x) \leftarrow \infty$.
3: **pour tout** $(u_p, u) \in E(G)$ selon un parcours de bas-en-haut **faire**
4: **pour tout** $t \in \{0, 1, \ldots, \theta'_{S'}(r(S'))\}$ **faire**
5: **pour tout** $(x_p, x) \in E_t(S')$ **faire**
6: **si** $u \in L(G)$, $x \in L(S')$ et $\mathcal{L}(u) = x$ **alors**
7: Sauter les lignes 8 à 24 et se rendre à la prochaine itération de la boucle à la ligne 5 {Case de base}
8: $Coût_g \leftarrow \infty$, pour $g \in \{\mathbb{S}, \mathbb{D}, \mathbb{T}, \varnothing, \mathbb{SL}\}$
9: **si** u a deux enfants **alors**
10: **si** x a deux enfants **alors**
11: $Coût_{\mathbb{S}} \leftarrow \min\{Coût(u_1, x_1) + Coût(u_2, x_2), Coût(u_1, x_2) + Coût(u_2, x_1)\}$
12: $Coût_{\mathbb{D}} \leftarrow Coût(u_1, x) + Coût(u_2, x) + \delta$
13: $(y_p, y) \leftarrow MeilleurReceveur((u, u_1), (x_p, x))$
14: $(z_p, z) \leftarrow MeilleurReceveur((u, u_2), (x_p, x))$
15: $Coût_{\mathbb{T}} \leftarrow \min\{ Coût(u_1, x) + Coût(u_2, z), Coût(u_1, y) + Coût(u_2, x) \} + \tau$
si x a un seul enfant **alors** $Coût_{\varnothing} \leftarrow Coût(u, x_1)$
16: **si** u a deux enfants **alors**
17: $Coût_{\mathbb{SL}} \leftarrow \min\{Coût(u, x_1), Coût(u, x_2)\} + \lambda$
18: $Coût(u, x) \leftarrow \min\{Coût_g : g \in \{\mathbb{S}, \mathbb{D}, \mathbb{T}, \varnothing, \mathbb{SL}\}\}$
19: **pour tout** $(x_p, x) \in E_t(S')$ **faire**
20: $(x'_p, x') \leftarrow MeilleurReceveur((u_p, u), (x_p, x))$
21: $Coût_{\mathbb{TL}} \leftarrow Coût(u, x') + \tau + \lambda$
22: $Coût(u, x) \leftarrow \min\{ Coût(u, x), Coût_{\mathbb{TL}} \}$
23: **retourner** $Coût(r(G), r(S'))$

3 Résultats expérimentaux

Afin d'évaluer les performances de notre approche, nous avons généré un grand nombre de jeux de données, suivant un modèle d'évolution probabiliste incluant des événements de duplications, de transferts et de pertes. Ceci a permis de comparer les scénarios proposés par notre algorithme avec les scénarios évolutifs réels (connus puisque simulés).

3.1 Simulation des arbres d'espèces

Nous avons utilisé un processus de naissance et de mort (birth and death) pour générer aléatoirement 10 arbres d'espèces contenant chacun 100 taxons (logiciel PhyloGen [21] avec un ratio de naissance/mort fixé à 1.25). Ces arbres ont ensuite été normalisés afin qu'ils aient tous la même hauteur h.

3.2 Simulation des scénarios $\mathbb{DTL}$

À partir d'une seule copie d'un gène, présente à la racine d'un arbre S au temps $t = h$, nous avons généré des scénarios $\mathbb{DTL}$ en faisant évoluer cette copie selon un processus de Poisson caractérisé par trois paramètres : le taux de duplications (r_δ), le taux de transferts (r_λ) et le taux de pertes (r_τ). Dans le cas d'un transfert, le donneur est choisi uniformément parmi les autres gènes existants au moment du transfert. On obtient ainsi, pour chaque simulation, un arbre de gènes G^o et une réconciliation simulée α_R incluant la liste des événements $\mathbb{DTLS}$ à l'origine de G^o.

Csűrös and Miklós ont récemment publié une étude portant notamment sur l'ampleur relative des taux de duplications, de transferts et de pertes au sein des archébactéries [4]. Les auteurs de cette étude estiment que, dans ce clade, environ 23% des événements sont des duplications, 1% sont des acquisitions, et 76% sont des pertes. Ils observent également un taux approximatif de pertes de 1.5 pour un arbre d'hauteur unitaire. Nous avons choisi de nous appuyer sur ces résultats pour faire varier de manière réaliste les taux $\mathbb{D}$,$\mathbb{T}$ et $\mathbb{L}$ de deux manières différentes.

Dans le premier jeu de données, nommé ds_1, nous avons fixé le taux de perte r_λ à 0.7 et la hauteur des arbres d'espèces (h) à 1, et nous avons fait varier r_δ et

r_τ dans l'intervalle $[0.01, 0.35]$ avec un pas de 0.034 (soit 11 valeurs). Nous avons donc obtenu 11×11 ensembles de paramètres cohérents avec une évolution le long d'une échelle temporelle importante correspondant, par exemple, au phylum des bactéries ou à celui des archébactéries. En effet, le taux de pertes choisi est réaliste (suivant [4]) et nous ne faisons pas de suppositions sur le taux relatif de transferts et de pertes, la seule contrainte étant que $r_\delta + r_\tau \leq r_\lambda$. Pour chacun des 10 arbres d'espèces et des 121 ensembles de paramètres, nous avons généré 5 arbres de gènes, obtenant ainsi 6 050 arbres de gènes au total.

Dans le deuxième jeu de données, nommé ds_2, nous avons fixé le rapport $r_\lambda/(r_\lambda + r_\delta + r_\tau)$ à 0.7 [4]. L'objectif de ce second jeu de données est d'étudier la pertinence d'une approche de parcimonie pour différentes échelles temporelles (phylogénies profondes ou récentes). Nous avons donc fait varier la hauteur de S ($h = 0.2, 0.4, 0.8$ et 1.6) ainsi que le taux de transfert $r_\tau \in [0, 0.3]$ par pas de 0.03 (soit 11 valeurs) tout en imposant $r_\delta = 0.3 - r_\tau$. Chacune de ces 44 combinaisons de paramètres a été appliquée aux 10 arbres d'espèces, et nous avons dans chaque cas généré 20 arbres de gènes, obtenant 8 800 arbres de gènes au total pour ds_2.

3.3 Analyses et résultats

Pour chaque jeu de données, nous avons utilisé comme coût d'un événement $\mathbb{DTL}$ l'inverse du taux moyen de ce type d'événement au long du processus de simulation (par exemple, pour ds_1, nous avons fixé $\delta = 1/0.18$). Pour chaque couple d'arbres G et S , nous avons calculé une réconciliation parmi les plus parcimonieuses, nommée α_p, grâce à l'Algorithme 1.

Il faut noter qu'une réconciliation réelle α_R contient souvent des événements qui n'ont laissé aucune trace et qui ne peuvent donc en aucun cas être repérés par une approche de parcimonie. Des exemples de ce type d'événements sont les événements de duplications immédiatement suivis par une perte ou plusieurs pertes ou événements $\mathbb{TL}$ qui se passent d'affilée. Avant de comparer α_p avec le scénario évolutif réel, nous avons éliminé de α_R tous les événements de ce type que nous avons pu détecter. Nous avons ainsi obtenu une réconciliation α'_R.

Nous avons d'abord étudié les conditions dans lesquelles la parcimonie peut correctement estimer les événements $\mathbb{DTL}$ en comparant les coûts de α_p et α'_R : quand les deux coûts diffèrent de façon importante, la parcimonie n'est plus une approche souhaitable. Le surcoût relatif de α'_R par rapport à une réconciliation la plus parcimonieuse est défini ainsi :

$$Surcoût(\alpha'_R, \alpha_P) = \frac{Coût(\alpha'_R) - Coût(\alpha_P)}{Coût(\alpha_P)}.$$

Il faut noter que $Coût(\alpha'_R) = Coût(\alpha_P)$ n'implique pas $\alpha_P = \alpha'_R$ puisque plusieurs réconciliations plus parcimonieuses peuvent exister. La Fig. 6 montre l'ampleur du surcoût selon les taux r_δ et r_τ et la hauteur de l'arbre.

Dans la Fig. 6, on remarque que le surcoût reste très limité pour toutes les combinaisons de taux mais qu'il augmente sensiblement avec la hauteur de l'arbre. Ceci est probablement dû aux événements cachés de α_R qui sont encore présents dans α'_R.

Nous nous sommes ensuite penchés sur les conditions dans lesquelles la parcimonie retrouve correctement la position des événements $\mathbb{DTL}$ qui ont engendré G. Rappelons qu'une réconciliation α pour un arbre de gènes G définit les événements $\mathbb{DTL}$ associés aux nœuds et branches internes de G. Puisque la position des événements de duplications et transferts définit univoquement la position des pertes, nous nous sommes focalisés sur les événements $\mathbb{D}$ et $\mathbb{T}$.

Soit $\mathbb{D}_S(\alpha)$ le sous-ensemble de paires $(u, (x_p, x)) \in V(G) \setminus L(G) \times E(S)$ tel que u est une duplication localisée sur (x_p, x) selon α. Soit $\mathbb{T}_S(\alpha)$ le sous-ensemble de triplets $((u_p, u), (x_p, x), (y_p, y)) \in E(G) \times E(S)^2$ tel que (u_p, u) est transféré et (x_p, x) (resp. (y_p, y)) est le donneur (resp. receveur). Pour une réconciliation la plus parcimonieuse α_P, la précision avec laquelle elle retrouve les événements $\mathbb{D}$ et $\mathbb{T}$ de la réconciliation réelle α'_R est évaluée par les ratios de faux positifs/négatifs définis ci-dessous (où $\mathbb{E} \in \{\mathbb{D}, \mathbb{T}\}$).

$$FP_{\mathbb{E}}(\alpha'_R, \alpha_P) = \frac{|\mathbb{E}_S(\alpha_P) - \mathbb{E}_S(\alpha'_R)|}{|\mathbb{E}_S(\alpha_P)|}$$
$$FN_{\mathbb{E}}(\alpha'_R, \alpha_P) = \frac{|\mathbb{E}_S(\alpha_P) - \mathbb{E}_S(\alpha'_R)|}{|\mathbb{E}_S(\alpha'_R)|},$$

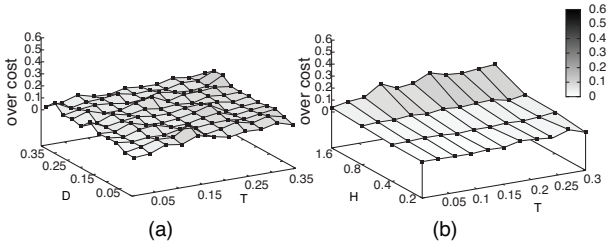


Fig. 6. Le surcoût relatif de α'_R en terme de coût de parcimonie par rapport à une réconciliation parmi les plus parcimonieuses, en faisant varier les taux de duplications et transferts et la hauteur de l'arbre, i.e., ds_1 (a) and ds_2 (b). Les valeurs élevées montrent les cas où il est inadéquat d'utiliser la parcimonie.

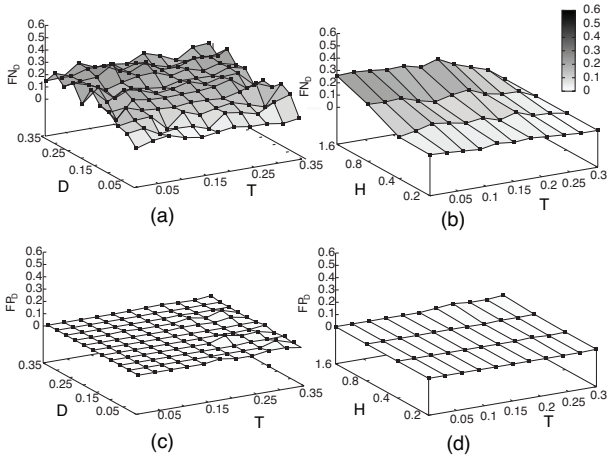


Fig. 7. Étude des conditions dans lesquelles la parcimonie retrouve précisément les événements $\mathbb{DTL}$. Ratios de faux négatifs (a-b) et faux positifs (c-d) pour les événement $\mathbb{D}$ en faisant varier r_δ, r_τ et la hauteur de l'arbre, pour ds_1 (a-c) et ds_2 (b-d).

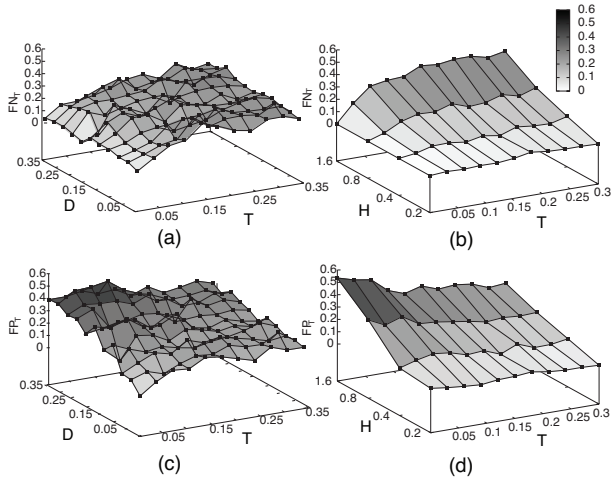


Fig. 8. Étude des conditions dans lesquelles la parcimonie retrouve précisément les événement $\mathbb{DTL}$. Ratios de faux négatifs (a-b) et faux positifs (c-d) pour les événement $\mathbb{T}$ en faisant varier r_δ, r_τ et la hauteur de l'arbre, pour ds_1 (a-c) et ds_2 (b-d).

Les Fig. 7 et Fig. 8 montrent l'évolution de ces ratios en faisant varier les taux de duplications et transferts et la hauteur de l'arbre. Dans la Fig. 7, on peut voir que $FP_{\mathbb{D}}$ est proche de zéro pour toutes les combinaisons de r_δ, r_τ et de hauteur de l'arbre. Cela veut dire que quasiment toutes les duplications inférées par notre algorithme sont présentes dans α'_R. Les valeurs très élevées de $FN_{\mathbb{D}}$ peuvent avoir plusieurs causes. Premièrement, α'_R peut contenir des événements de duplications qui ne peuvent pas être détectés. Deuxièmement, avoir fixé $\delta = \tau$ peut amener à inférer un événement de type $\mathbb{T}$ au lieu de $\mathbb{D}$ (ce qui expliquerait aussi le taux très élevé de $FP_{\mathbb{T}}$ en Fig. 8). Enfin, cela peut être dû au fait qu'on a choisi une réconciliation plus parcimonieuse parmi plusieurs existantes (cela expliquerait aussi le taux élevé de $FN_{\mathbb{T}}$).

Pour des arbres de 100 espèces et des taux faibles (resp. élevés), l'algorithme résout le MPR avec un temps moyen de 1.09 (resp. 1.38) secondes.

Remerciements

Nous remercions J.-F. Dufayard pour son aide avec les différents programmes de réconciliation, K. Gorbunov et V. Lyubetsky pour les discussions sur [9]. Ce travail est financé par le projet ANR-08-EMER-011.

Références

[1] L. Arvestad, A. C. Berglund, J. Lagergren, and B. Sennblad. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics*, 19 Suppl 1 :7–15, 2003.

[2] A. Bernal, U. Ear, and N. Kyrpides. Genomes OnLine Database (GOLD) : a monitor of genome projects world-wide. *Nucleic Acids Res.*, 29 :126–127, 2001.

[3] C. Conow, D. Fielder, Y. Ovadia, and R. Libeskind-Hadas. Jane : a new tool for the cophylogeny reconstruction problem. *Algorithms Mol Biol*, 5 :16, 2010.

[4] M. Csuros and I. Miklos. Streamlining and Large Ancestral Genomes in Archaea Inferred with a Phylogenetic Birth-and-Death Model. *Mol Biol Evol*, 26(9) :2087–2095, 2009.

[5] V. Daubin, N. A. Moran, and H. Ochman. Phylogenetics and the cohesion of bacterial genomes. *Science*, 301 :829–832, 2003.

[6] W. F. Doolittle. Phylogenetic classification and the universal tree. *Science*, 284 :2124–2129, 1999.

[7] N. Goldenfeld and C. Woese. Biology's next revolution. *Nature*, 445 :369, 2007.

[8] M. Goodman, J. Czelusniak, G. W. Moore, Romero A. Herrera, and G. Matsuda. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool.*, 28 :132–163, 1979.

[9] K. I. Gorbunov and V. A. Liubetsky. Reconstructing genes evolution along a species tree. *Mol. Biol. (Mosk.)*, 43 :946–958, 2009.

[10] P. Górecki. Reconciliation problems for duplication, loss and horizontal gene transfer. In *RECOMB*, 2004.

[11] R. Guigo, I. Muchnik, and T. F. Smith. Reconstruction of ancient molecular phylogeny. *Mol. Phylogenet. Evol.*, 6 :189–213, 1996.

[12] M. Hallett, J. Lagergren, and A. Tofigh. Simultaneous identification of duplications and lateral transfers. In *RECOMB '04*, pp. 347–356, New York, NY, USA, 2004. San Diego, California, USA, ACM.

[13] C. G. Kurland, B. Canback, and O. G. Berg. Horizontal gene transfer : a critical view. *Proc. Natl. Acad. Sci. U.S.A.*, 100 :9658–9662, 2003.

[14] R. Libeskind-Hadas and M. A. Charleston. On the computational complexity of the reticulate cophylogeny reconstruction problem. *JCB*, 16(1) :105–117, 2009. http ://dx.doi.org/10.1089/cmb.2008.0084.

[15] J. O. McInerney, J. A. Cotton, and D. Pisani. The prokaryotic tree of life : past, present and future ? *Trends Ecol. Evol.*, 23 :276–281, 2008.

[16] D. Merkle and M. Middendorf. Reconstruction of the cophylogenetic history of related phylogenetic trees with divergence timing information. *Theory Biosci*, 123(4) :277–299, 2005.

[17] D. Merkle, M. Middendorf, and N. Wieseke. A parameter-adaptive dynamic programming approach for inferring cophylogenies. *BMC Bioinformatics*, 11(Suppl 1) :S60, 2010.

[18] R. D. Page. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst. Biol.*, 43 :58–77, 1994.

[19] S. Penel, A. M. Arigon, J. F. Dufayard, A. S. Sertier, V. Daubin, L. Duret, M. Gouy, and G. Perriere. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*, 10 Suppl 6 :S3, 2009.

[20] P ; Puigbo, Y. Wolf, and E. Koonin. Search for a 'tree of life' in the thicket of the phylogenetic forest. *Journal of Biology*, 8(6) :59, 2009.

[21] A. Rambaut. Phylogen : phylogenetic tree simulator package, 2002.

[22] A. Tofigh, M. Hallett, and J. Lagergren. Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM TCBB*, 99, 2010.

[23] A. Tofigh, J. Sjöstrand, B. Sennblad, L. Arvestad, and J. Lagergren. Detecting LGTs using a novel probabilistic model integrating duplications, lgts, losses, rate variation, and sequence evolution, 2009.

[24] B. Vernot, M. Stolzer, A. Goldman, and D. Durand. Reconciliation with non-binary species trees. *J. Comput. Biol.*, 15 :981–1006, 2008.