Les dates de divergence améliorent la qualité des superarbres

Yimin SHEN [1], Laurent BREHELIN [1], Emmanuel J. P. DOUZERY [2], Vincent BERRY [1,*]

[1] LIRMM, Université Montpellier 2, 161 rue Ada, 34392 Montpellier Cedex 5
[*] vberry@lirmm.fr PhylAriane ANR-08-EMER-011

[2] ISEM, Université Montpellier 2, 34095 MONTPELLIER Cedex 5

Mots-Clés *: Bioinformatique, superarbre, datation, classification non supervisée.*

En biologie évolutive, les arbres sont utilisés pour retracer l'évolution des espèces, regroupées en fonction des innovations qu'elles partagent. Les méthodes de *superarbres* consistent à représenter aussi bien que possible – au sens d'un critère combinatoire – l'information topologique contenue dans une collection d'arbres de gènes en un seul arbre synthétique, par exemple pour reconstruire l'arbre de la Vie. Cette inférence est souvent problématique car différents arbres de gènes pour un même ensemble d'espèces peuvent avoir des topologies différentes. S'ajoute à ces difficultés le fait que les arbres de gènes à assembler en un superarbre peuvent posséder des ensembles d'espèces différents, parfois insuffisamment chevauchants (voir figure 1-I).

Pour compenser ces problèmes, nous proposons d'utiliser des informations supplémentaires par rapport à celles habituellement utilisées : recouper entre elles les dates de divergences inférées séparément pour chaque arbre de gène (voir figure 1-II). La date de divergence de deux espèces est l'âge de leur ancêtre commun le plus récent. Nous utilisons une méthode de pénalisation des différences de taux d'évolution entre branches liées à un même nœud pour rendre les arbres de gènes ultramétriques, et obtenir les âges de chaque nœud interne [1].

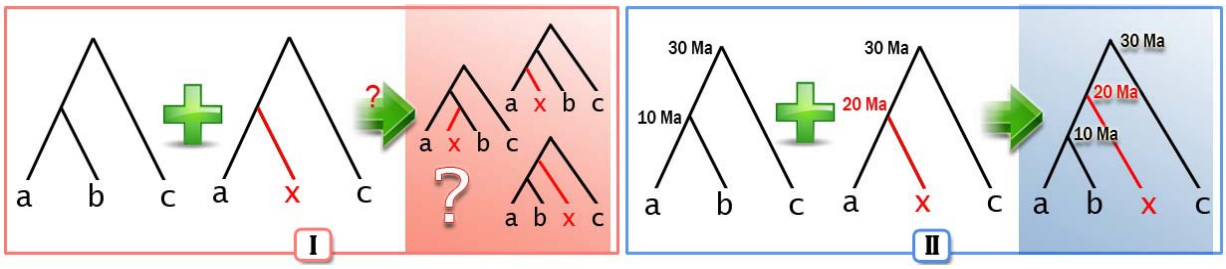

FIG. 1- La topologie des arbres sources ne suffit pas à elle seule pour connaître la position de certaines espèces (I). Prendre en compte les dates de divergence permet parfois de conclure sans ambiguïté (II).

Nous avons collecté un jeu d'arbres sources depuis la version actuelle de la base de données *OrthoMaM* [2] qui contient des marqueurs génomiques orthologues pour reconstruire les arbres phylogénétiques de mammifères. Un des enjeux de l'histoire évolutive des placentaires est la connaissance des relations de parenté entre ses quatre groupes majeurs — Euarchontoglires, Laurasiathériens, Afrothériens, et Xénarthres — notamment à l'aide de leur enracinement par les Marsupiaux. Nous avons sélectionné les arbres sources contenant au moins un représentant des cinq groupes cités ci-dessus à l'aide de *PhyloExplorer* [3].

Étant donnée une espèce de référence, nous relevons dans chaque arbre le temps d'évolution séparant cette espèce de deux autres espèces cibles. Grâce au test non-paramétrique de Wilcoxon, nous déterminons si les distributions de ces temps de divergences diffèrent significativement, relativement à une *p*-valeur choisie. Lorsque ce n'est pas le cas, nous en déduisons une information de placement relatif des espèces dans la structure du superarbre. La procédure est répétée pour toutes les paires d'espèces cibles, et un graphe est utilisé pour représenter les groupes d'espèces ayant des temps de divergence communs vis-à-vis de l'espèce de référence (voir figure 2-I). Les éventuelles incohérences sont ensuite corrigées grâce à une étape de partitionnement du graphe en cliques utilisant une variante de la distance de Czekanovski-Dice [4] (voir figure 2-II).

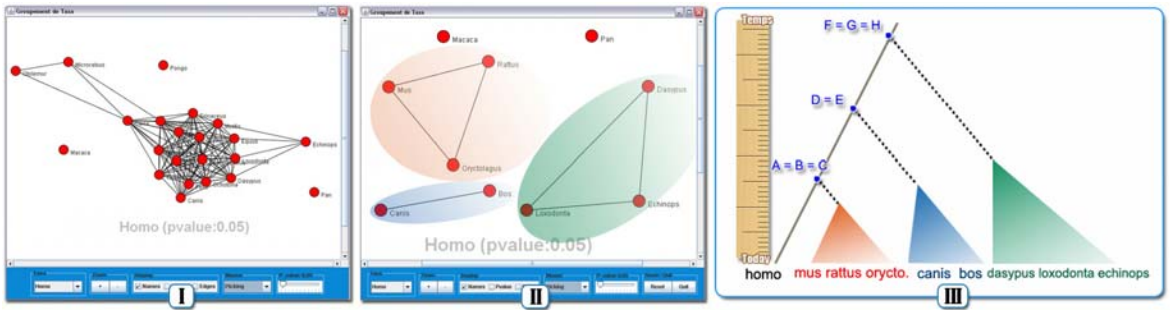

FIG. 2- Le graphe codant les informations de dates (I) est décomposé en cliques disjointes (II) puis traduit en informations topologiques (III).

MRP – la méthode d'inférence de superarbres la plus connue – encode les informations topologiques des arbres sources sous la forme d'une matrice de caractères binaires, depuis laquelle est inféré un superarbre en optimisant un critère combinatoire de parcimonie maximale. Les cliques obtenues depuis les informations de datation sont traduites sous formes d'arbres complémentaires (voir figure 2-III), encodés eux aussi dans la matrice MRP. Une comparaison du MRP classique au MRP utilisant additionnellement les informations de dates montre que de nouveaux groupes d'espèces – en majorité corrects – sont ainsi retrouvés. Ceci permet de résoudre la question biologique concernant l'agencement des grands groupes de mammifères placentaires (voir figure 3). Les expériences menées montrent qu'une *p*-valeur de 1% dans le test de Wilcoxon amène aux meilleurs résultats.

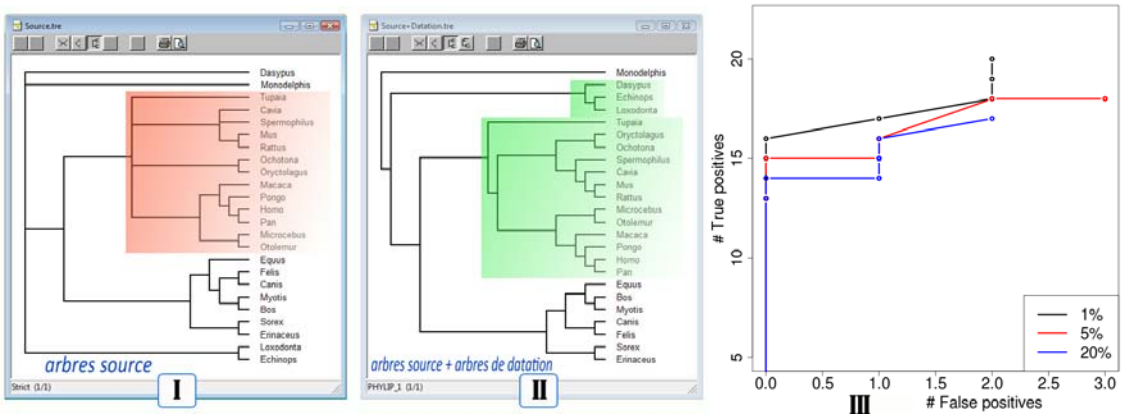

FIG. 3- L'utilisation des seuls arbres de gènes (I) donne des inférences moins complètes qu'en tenant aussi compte des datations (II). Courbe ROC montrant qu'une *p*-valeur de 1% produit les meilleurs résultats (III).

Ces travaux suggèrent que l'utilisation de dates de divergence entre espèces peut améliorer le pouvoir résolutif des méthodes de superarbres.

Références :

[1] Sanderson, M. J. 2003 r8s: *inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics,* 19:301–302.

[2] Ranwez V. et al., 2007 *OrthoMaM : A database of orthologous genomic markers for placental mammal phylogenetics. BMC Evolutionary Biology,* 7 : 241

[3] Ranwez V. et al., 2009 *PhyloExplorer: a web server to validate, explore and query phylogenetic trees. BMC Evolutionary Biology,* 9:108

[4] Angelelli et al. 2008, *Two local dissimilarity measures for weighted graphs with application to protein interaction networks. Advances in Data Analysis and Classification,* 2:3-16.