Processus d'évolution réticulée : tests de signal phylogénétique

Cécile Ané[1,2], Paul Bastide[3,4], Mahendra Mariadassou[4], Stéphane Robin[3] & Claudia Solís-Lemus[1]

[1] *Department of Statistics, University of Wisconsin-Madison, WI, 53706, USA*
[2] *Department of Botany, University of Wisconsin-Madison, WI, 53706, USA*
[3] *MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France*
[4] *MaIAGE, INRA, Université Paris-Saclay, 78352 Jouy-en-Josas, France*
cecile.ane@wisc.edu, paul.bastide@agroparistech.fr, mahendra.mariadassou@inra.fr, stephane.robin@agroparistech.fr, solislemus@wisc.edu

Résumé. Les méthodes comparatives phylogénétiques (en anglais, PCM, pour *Phylogenetic Comparative Methods*) ont pour but d'étudier la distribution de traits quantitatifs au seins d'un ensemble d'espèces, en prenant en compte les relations de parentés qui existent entre elles. Ces relations sont représentées de manière classique par un arbre phylogénétique. Cependant, ces arbres, qui supposent une transmission verticale du patrimoine génétique d'une génération à l'autre, ne tiennent pas compte des événements d'hybridations, ou de transferts de gènes horizontaux, qui peuvent modifier la filiation de certaines espèces. Ce type d'événement peut être fréquent pour certains groupes de plantes, ou de bactéries. On a alors recours à un *réseau phylogénétique*, dans lequel certaines branches horizontales sont ajoutées à la structure arborescente pour représenter ces événements. On peut alors voir les traits observés aux feuilles, pour les espèces actuelles, comme le résultat d'un mouvement Brownien courant sur le réseau considéré. Ce modèle induit une structure de variance-covariance pour les traits observés, qu'il est possible d'utiliser pour des analyses statistiques subséquentes, comme la régression phylogénétique, ou la reconstruction d'états ancestraux. Ces outils, ainsi que le calcul efficace de la matrice de variance grâce à un algorithme récursif, ont été implémentés de manière flexible dans le paquet PhyloNetworks, sur julia. Devant ce nouveau modèle, il est naturel de se poser la question de l'impact de la structure de parenté sur les données observées. Ceci peut être fait grâce à un test statistique, comparant par exemple le modèle induit par un arbre simple à celui induit par un réseau. Nous présenterons une étude de puissance d'un tel test sur des données simulées.

Mots-clés. Réseau phylogénétiques, Processus stochastiques

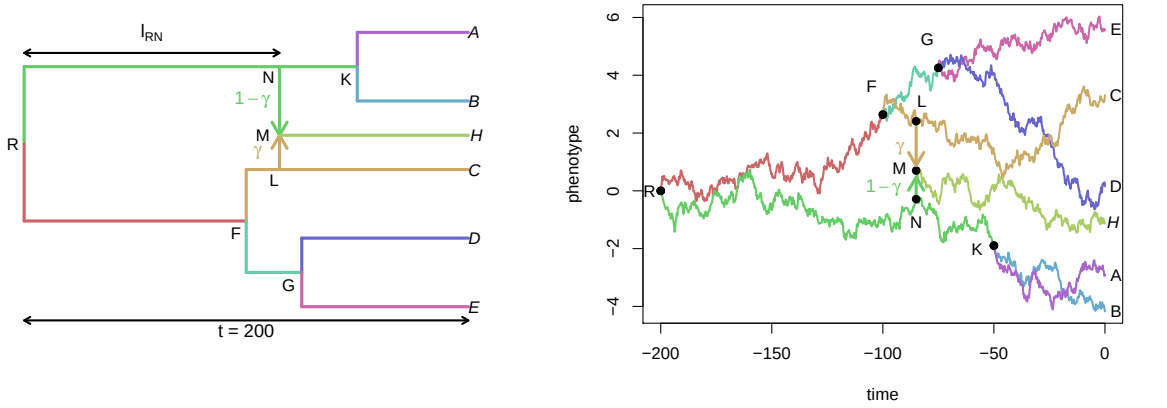
Abstract. The goal of Phylogenetic Comparative Methods (PCM) is to study the distribution of quantitative traits among related species. Species are usually seen as related through a phylogenetic tree. However, some events, such as hybridization, or horizontal gene transfers, can change substantially the relations between species, and are not taken

into account by a tree-shaped evolution, which assumes a vertical transmission of the genetic material from one generation to the other. This kind of events can be frequent for some groups of plants, or bacterial organisms. We can take them into account by adding some horizontal edges to the tree, transforming it into a *phylogenetic network*. The observed traits for extant species can then be seen as the result of a Brownian process running on the branches of this phylogenetic network. Such a model induces a given variance-covariance structure for the observed traits, that we can take into account for downstream statistical analysis, such as phylogenetic regression, or ancestral state reconstruction. Along with an efficient computation of the variance matrix through a recursive algorithm, those tools have been implemented in the julia package PhyloNetworks. This new model of evolution naturally raises the question of the impact of this structure on the observed data. A statistical test can be designed to compare the model induced by a simple tree, to the one induced by a network. We will present a study of the statistical power of such a test on some simulated data.

Keywords. Phylogenetic networks, Stochastic processes

1 Modèles d'évolution réticulée

Réseau phylogénétique. Les liens de parentés entre espèces sont représentées de manière classique par un arbre phylogénétique, qui, s'il est calibré en temps, représente l'histoire évolutive d'un groupe d'organismes. Cependant, cette représentation arborescente ne tient pas compte des événements d'hybridations, ou de transferts de gènes horizontaux, qui peuvent modifier substantiellement les relations de filiation entre les espèces présentes. On a alors recours à un *réseau phylogénétique* pour représenter ces liens. Un réseau phylogénétique est un graphe acyclique dirigé et raciné, dont les feuilles représentent les espèces actuelles observées, et les nœuds internes des espèces ancestrales. Les nœuds internes peuvent avoir un seul parent (filiation arborescente) ou bien deux parents (hybridation). Le réseau est calibré en temps, si bien que la longueur des branches arborescentes représente un temps évolutif. L'événement d'hybridation étant instantané, les branches menant à un nœud hybride sont supposées de longueur nulle. Un paramètre γ leur est cependant associé, représentant une proportion de patrimoine génétique transmis par chacun des deux parents. Un exemple de réseau phylogénétique présentant un seul événement d'hybridation est présenté figure 1-a. Sur cet exemple, le nœud hybride M a hérité d'une proportion γ de ses gènes de l'espèce L, et le reste $1-\gamma$ de l'espèce N. Plusieurs méthodes d'inférence ont été développées ces dernières années (voir par exemple Yu *et al.* (2014), Solís-Lemus et Ané (2016)), et ce type de réseaux phylogénétiques commence à être disponible pour un certain nombre de groupes d'espèces. Dans toute la suite, on suppose que le réseau est connu et fixé.



(a) Réseau phylogénétique. Le nœud L est hybride. ℓ_{RN} est la longueur de la branche allant de R à N. t est le temps total d'évolution.

(b) Variation du trait en fonction du temps. Seule la valeur du processus aux feuilles est observée.

FIGURE 1 – Réseau phylogénétique d'un ensemble d'espèces contemporaines, et modélisation de l'évolution d'un caractère par un mouvement Brownien.

Évolution d'un trait. On cherche ici à modéliser l'évolution au cours du temps d'un trait quantitatif, tel que la taille moyenne d'une espèce. On utilise pour cela un mouvement Brownien (BM) courant sur les branches du réseau phylogénétique liant les espèces entre elles (voir figure 1-b). Le processus est défini de la manière suivante :

- Sur une branche donnée, le trait évolue au cours du temps suivant un mouvement Brownien.
- Lors d'une spéciation (nœud arborescent), le processus se divise en deux Browniens indépendants, partants du même point et avec les mêmes paramètres, courant chacun sur une des deux branches filles.
- Lors d'une hybridation, le trait hybride est obtenu en faisant la moyenne pondérée par le coefficient γ des traits de ses deux parents, puis évolue suivant un Brownien indépendant, et avec les mêmes paramètres.

Ce type de modèle permet de capturer l'évolution de traits quantitatifs multi-loci (dont l'expression dépend de beaucoup de gènes répartis sur tout le génome), non soumis à une quelconque sélection.

2 Méthodes Comparatives Phylogénétiques

Problème. En écologie évolutive, le but des méthodes comparatives phylogénétiques (PCM) est l'étude de traits quantitatifs au seins d'une population d'espèces ayant une histoire évolutive commune. Les traits observés ne sont alors pas indépendants, et leur structure de corrélation dépend de leurs liens phylogénétiques, au travers d'un modèle d'évolution, tel que celui décrit ci-dessus. Sur des arbres, partant de l'article fondateur de

Felsenstein (1985), ce type de méthodes a reçu beaucoup d'attention ces dernières années, avec des complexifications croissantes du type de processus stochastique d'évolution considéré (pour une revue, voir Pennell et Harmon (2013)). La nouveauté est ici le caractère réticulé de l'histoire évolutive considérée.

Matrice de Variance. Pour un trait qui évolue suivant le modèle ci-dessus, il est possible d'écrire la matrice de variance covariance liant les observations au feuilles. La covariance entre les traits Y_i et Y_j de deux espèces i et j aux feuilles du réseau phylogénétique s'écrit (Pickrell et Pritchard (2012)) :

$$\mathbb{C}\mathrm{ov}\left[Y_i; Y_j\right] = \sigma^2 V_{ij} = \sigma^2 \sum_{\substack{p_i \in \mathcal{P}_i \\ p_j \in \mathcal{P}_j}} \left(\prod_{e \in p_i} \gamma_e\right)\left(\prod_{e \in p_j} \gamma_e\right) \sum_{e \in p_i \cap p_j} \ell_e$$

où σ^2 est la variance du mouvement Brownien, $\mathcal{P}_i$ est l'ensemble des chemins allant de la racine au nœud i, et, pour une arrête e, γ_e est le coefficient de transmission génétique ($\gamma_e = 1$ pour toutes les arrêtes arborescentes), et ℓ_e est la longueur de l'arrête, en temps phylogénétique.

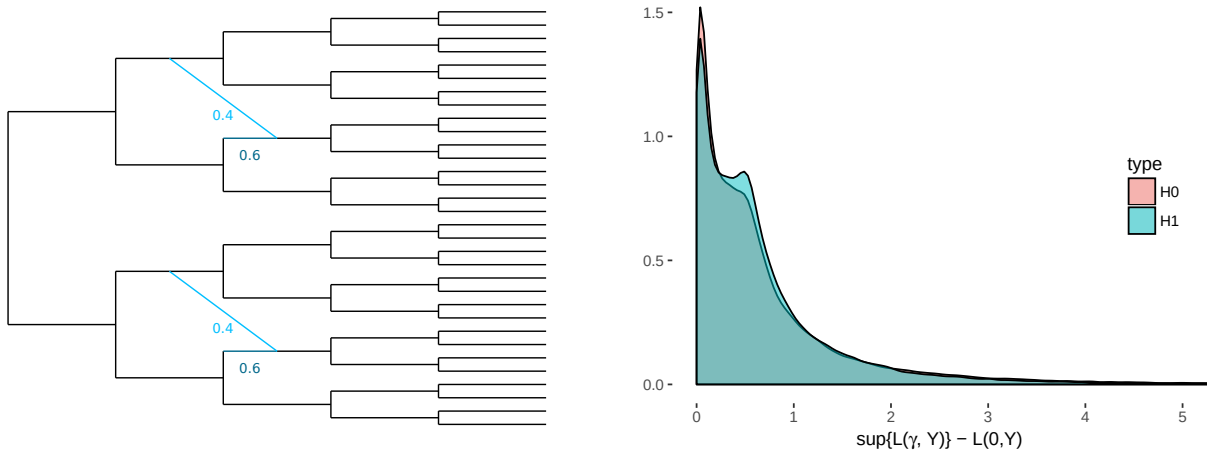
On montre qu'il est possible de trier les nœuds du réseau de telle sorte à ce que la matrice $\mathbf{V}$ puisse être calculée récursivement en un parcours du réseau, depuis la racine jusques aux feuilles.

Régression Phylogénétique. Le réseau étant fixé, il est possible de calculer la matrice $\mathbf{V}$. Elle peut alors servir de matrice de structure de corrélation des résidus dans un modèle de régression linéaire du trait considéré. Elle permet également de faire de la reconstruction d'états ancestraux, suivant les techniques classiques. Ces outils statistiques ont été implémentés de manière efficace et flexible au seins du paquet PhyloNetworks sur le langage de programmation julia (`https://github.com/crsl4/PhyloNetworks.jl`).

3 Tests de signal phylogénétique

Test de Pagel. Lorsque l'on a affaire à un trait mesuré sur un ensemble d'espèces, une question classique est de chercher à déterminer à quel point celui-ci est héritable. En terme de distribution aux feuilles, cela revient à se demander si le modèle induit par le réseau phylogénétique est meilleur qu'un modèle nul dans lequel tous les traits seraient indépendants. Ceci peut être fait grâce à une adaptation du test basé sur la transformation λ de Pagel (1999), telle que décrite pour un arbre. Il s'agit de multiplier toutes les branches internes par un facteur λ, tout en modifiant les longueurs des branches terminales de telle sorte à ce que le réseau obtenu garde la même hauteur totale. Lorsque $\lambda = 1$, le réseau n'est pas modifié, et le signal phylogénétique est considéré comme fort. Si,

au contraire, $\lambda = 0$, le réseau est réduit à un arbre-étoile, et toutes les espèces observées sont indépendantes. Il s'agit alors de tester $\lambda = 0$ contre $\lambda > 0$.



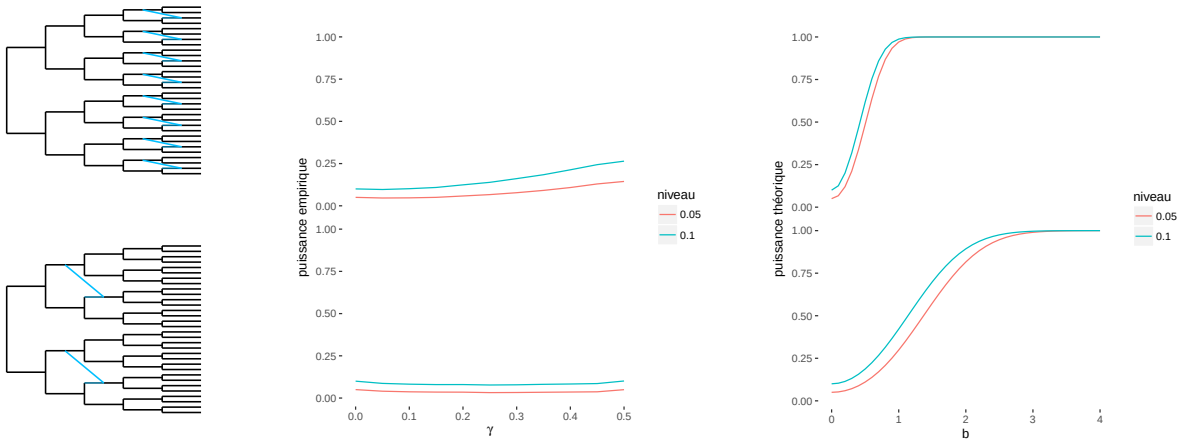
(a) Réseau phylogénétique comportant 32 feuilles, et deux événements d'hybridation avec $\gamma = 0.4$.

(b) Densité estimée de $\sup_{\gamma \neq 0}\{L(\gamma, Y)\} - L(0, Y)$, pour des données Y simulées suivant $H0$ ou $H1$.

FIGURE 2 – Densité estimée de la différence entre la log-vraissemblance maximisée en γ et la log-vraissemblance pour $\gamma = 0$, pour 100000 jeux de données simulées sur le réseau présenté à gauche, suivant un Brownien, soit sur l'arbre majeur ($H0 : \gamma = 0$) soit sur le réseau ($H1 : \gamma = 0.4$), avec une variance $\sigma^2 = 1$. La statistique est peu discriminante.

Test d'hybridation. De manière plus fine, on peut également chercher à savoir si les hybridations ont réellement eu un impact sur le trait considéré. Pour cela, on peut comparer le modèle induit par le réseau contre un modèle induit par l'*arbre majeur* issu de ce réseau, c'est-à-dire l'arbre obtenu en liant les espèces hybrides avec celui de leur parent duquel elles ont reçu la plus grande proportion de matériel génétique, et en supprimant leur relations avec leur autre parent. Lorsqu'il n'y a qu'un seul événement d'hybridation dans l'arbre, et que l'on pose, par convention $\gamma < 0.5$ pour cet événement, cela revient ainsi à tester $\gamma = 0$ contre $\gamma \neq 0$. Les deux modèles étant plus proches que dans le test précédent, on s'attend à ce que la puissance d'un tel test soit moindre (voir figures 2 et 3-b).

Ce test d'hybridation sur la variance s'avérant très peu puissant, on s'orientera plutôt vers un test d'hétérosis sur la moyenne. L'hétérosis, ou vigueur hybride, est un phénomène bien connu en génétique, qui rend possible la naissance d'un hybride ayant un caractère exeptionellement grand (ou petit) par rapport à ses deux parents. Dans notre modèle, le trait hybride est alors obtenu comme la moyenne pondérée des traits espèces parentes, comme précédemment, plus un saut d'une valeur b. On montre qu'il est alors possible de ré-écrire le problème sous la forme d'un modèle linéaire à effets fixes, et ainsi de replacer la



(a) Topologies de réseaux utilisées.

(b) Puissance empirique du test de variance (100000 simulations).

(c) Puissance théorique du test d'hétérosis (γ fixé à 0.4).

FIGURE 3 – Puissance empirique ou théorique des tests de détection d'hybridation ou d'hétérosis, en fonction de γ ou b, pour deux réseaux symmétriques de topologies différentes, et un niveau de test de 5 ou 10%. Si elle ne s'accompagne pas d'un événement d'hétérosis, la puissance de détection d'une hybridation est très faible.

question dans un cadre statistique bien connu. Ceci permet d'écrire un test sur la moyenne dont la puissance est meilleure, et augmente en fonction de b (voir Figure 3-c).

Bibliographie

[1] J. Felsenstein : Phylogenies and the Comparative Method. *The American Naturalist*, 125(1):1–15, 1985.

[2] M. Pagel : The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic biology*, 48(3):612–622, 1999.

[3] M. W. Pennell et L. J. Harmon : An integrative view of phylogenetic comparative methods : connections to population genetics, community ecology, and paleobiology. *Annals of the New York Academy of Sciences*, 1289(1):90–105, 2013.

[4] J. K. Pickrell et J. K. Pritchard : Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genetics*, 8(11):e1002967, 2012.

[5] C. Solís-Lemus et C. Ané : Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genetics*, 12(3):e1005896, 2016.

[6] Y. Yu, J. Dong, K. J. Liu et L. Nakhleh : Maximum likelihood inference of reticulate evolutionary histories. *PNAS*, 111(46):16448–16453, 2014.