Processus avec sauts sur arbres : détection de changements adaptatifs

Paul Bastide[1,2], Stéphane Robin [1] & Mahendra Mariadassou [2]

[1] *UMR 518 AgroParisTech-INRA; 16 rue Claude Bernard 75231 Paris*
[2] *MaIAGE INRA; Bât. 233 Domaine de Vilvert 78352 Jouy en Josas*
paul.bastide@agroparistech.fr, stephane.robin@agroparistech.fr,
mahendra.mariadassou@jouy.inra.fr

Résumé. En écologie comparative et évolutive, les traits quantitatifs d'un jeu d'espèces peuvent être vus comme le résultat d'un processus stochastique courant le long d'un arbre phylogénétique. Cette modélisation permet de prendre en compte les corrélations produites par une histoire évolutive partagée. Le processus stochastique est choisi afin de de capturer les mécanismes qui gouvernent l'évolution d'un trait. Les écologues préfèrent ainsi le processus d'Orstein-Uhlenbeck (OU) au Mouvement Brownien (BM), plus simple mais moins réaliste. Le processus OU modélise la sélection naturelle s'opérant sur un trait par un mécanisme de rappel vers une valeur centrale, interprétée comme optimale dans un environnement donné. On s'intéresse ici à des changements de niche évolutive qui auraient entraîné un changement abrupt dans la valeur de cet optimum, et dont il s'agit de retrouver la position sur l'arbre. À partir des mesures d'un trait pour un jeu d'espèces liées par un arbre phylogénétique connu, on se propose de construire, d'étudier, et d'implémenter efficacement un modèle à données incomplètes permettant d'inférer simultanément la position des sauts et la valeur des paramètres. Les sauts sur l'arbre induisent naturellement une classification des espèces actuelles en groupes cohérents avec la phylogénie et définis par une même valeur de trait. Au vu des données, seule cette classification est identifiable, ce qui pose problème pour la localisation exacte des sauts sur l'arbre. On se propose alors de dénombrer, d'une part, les allocations non-identifiables équivalentes, et, d'autre part, les solutions distinctes identifiables. Cette dernière quantité nous sert alors à calibrer une pénalité de sélection de modèle.

Mots-clés. Écologie Comparative, Ornstein-Uhlenbeck, Expectation-Maximization, Phylogénie

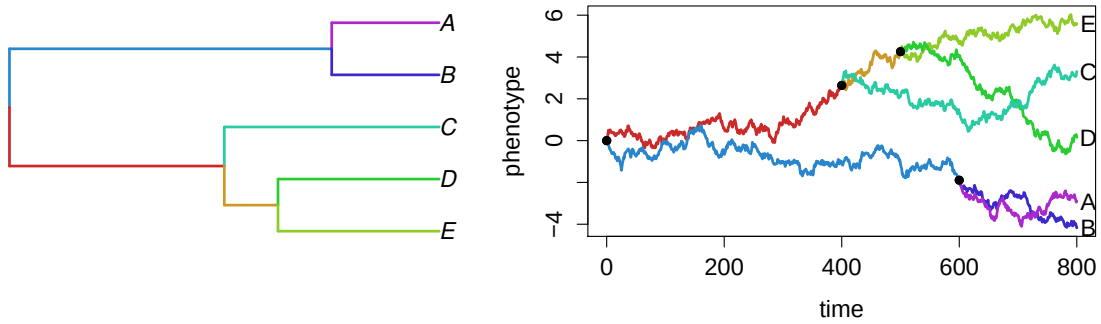
Abstract. In comparative and evolutive ecology, quantitative traits measured on related species can be seen as the outcome of a stochastic process running on a phylogenetic tree. This modeling can account for correlations between species that have a shared evolutionary history. The chosen process must be able to capture the mechanisms of a trait evolution. Ecologists hence prefer the Orstein-Uhlenbeck (OU) process to the simpler but less realistic Brownian Motion (BM). This OU process has a tendency to revert to a central value, interpreted in ecology as the optimal value of a trait in a given environment. We want to detect the shifts that occurred in this central value, allowing us to reconstruct

the history of the changes of ecological niches along the tree. Given measures of a trait on related species and a phylogenetic tree of those species, we aim at build and use an incomplete-data model to infer the parameters of the stochastic process, via an automatic detection and characterization of the shifts. Extant species, that are the inheritors of this shifted history of traits, are naturally classified into clusters that share a same probability distribution, and that are coherent with their phylogeny. Given the data we have, only this clustering is identifiable, and the position of the shifts on the tree is known only to a equivalence relationship. It is however possible, first, to enumerate all the non-identifiable equivalent allocations of the shifts and, second, to count all the different identifiable solutions to the problem when the number of change-points is fixed. This last quantity allows us to devise an adequate penalty for a model selection procedure.

Keywords. Comparative Ecologie, Ornstein-Uhlenbeck, Expectation-Maximization, Phylogeny

1 Définition du problème et modélisation

Contexte. Notre objectif est ici d'étudier les traces laissées par des changements adaptatifs brutaux sur les valeurs actuelles des caractères d'un ensemble d'espèces liées par un arbre phylogénétique. Ces *sauts* ancestraux sont provoqués par des changements de niches écologiques, induites par exemple par des changements climatiques. Dans toute la suite, on suppose que l'arbre phylogénétique est donné, fixe, et calibré en temps, donc ultramétrique. On s'intéresse au cas où un seul caractère est mesuré aux feuilles de l'arbre. L'arbre considéré possède n feuilles et m noeuds internes (où $m = n - 1$ si l'arbre est binaire). On numérote ainsi les branches de 1 (pour la branche à la racine) à $m + n$.



(a) Arbre phylogénétique.

(b) Variation du trait en fonction du temps.

Figure 1: Arbre phylogénétique d'un ensemble d'espèces contemporaines, et modélisation de l'évolution d'un caractère par un processus stochastique (ici, un mouvement brownien).

Modélisation de l'évolution par un processus stochastique. La valeur des traits des différentes espèces évolue dans le temps suivant un processus stochastique. Tout

événement de spéciation conduit à la création de deux processus indépendants et partants du même point (voir figure 1). Le processus le plus simple est le Mouvement Brownien (MB), introduit dans ce cadre par Felsenstein (1985). Il ne représente cependant qu'un bruit pur, inapte à capturer des phénomènes de sélection. Pour prendre en compte ce phénomène, on a recours à un processus d'Orstein-Uhlenbeck (OU), comme proposé initialement par Hansen (1997). Suivant ce modèle, l'évolution d'un trait le long d'une lignée est définie par l'équation différentielle stochastique suivante :

$$dW_t = -\alpha(W_t - \beta)dt + \sigma dB_t$$

où les paramètres sont définis comme suit :

- β est l'*optimum primaire* du trait, défini de manière mécanique par la niche écologique dans laquelle évolue l'espèce à un instant donné. C'est la valeur de trait sélectionnée.
- W_t est l'*optimum secondaire* du trait, distinct de l'optimum primaire en raison de perturbations locales et temporaires, notamment dues au bruit génétique. On suppose qu'il est donné par la moyenne du trait sur l'ensemble des individus vivants de l'espèce.
- σdB_t représente les variations stochastiques browniennes du trait, de variance σ^2.
- α est un paramètre de rappel vers l'optimum primaire, ou *force de sélection*. On définit le *temps de demie-vie phylogénétique* $t_{1/2} = \ln(2)/\alpha$ comme le temps nécessaire pour que la moyenne du processus soit à mi-chemin entre sa valeur initiale et sa valeur optimale (Hansen, 1997). Il peut être comparé avec la durée totale sur lequel le processus évolue, c'est-à-dire la hauteur t_a de l'arbre.

Dans ce modèle, un *saut adaptatif* est défini comme un changement d'intensité δ dans la valeur de l'optimum primaire β, qui passe alors de β à $\beta + \delta$. On fait l'hypothèse que les paramètres α et σ sont constants et identiques dans toutes les lignées.

Paramétrisation des sauts. On suppose la présence d'exactement K sauts dans la valeur de β, qui ont tous lieu au début d'une branche, soit juste après un événement de spéciation (voir figure 2). On note $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)$ le vecteur donnant le numéro des branches où ont lieu les sauts, d'intensités $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_K)$. En notant β_0 la valeur de l'optimum primaire initial (à la racine), celle à un nœud j est donnée par : $\beta_j = \beta_0 + \sum_{i \in \mathrm{Par}(j)} \sum_k \mathbb{I}\{\tau_k = b_i\}\delta_k$, où b_i désigne la branche se terminant au nœud i, et $\mathrm{Par}(j)$ l'ensemble des ancêtres de j. On définit de plus le vecteur $\boldsymbol{\Delta}$, de taille $m+n$, des sauts sur les branches de l'arbre, tel que $\Delta_{\tau_k} = \delta_k$ pour tout $1 \leq k \leq K$. Ce vecteur a K composantes non nulles. On suppose que la racine est dans l'état stationnaire du processus initial, à savoir une gaussienne de moyenne β_0 et de variance $\gamma^2 = \frac{\sigma^2}{2\alpha}$. On note $\boldsymbol{\theta} = (\alpha, \sigma^2, \beta_0, \boldsymbol{\tau}, \boldsymbol{\delta})$ le vecteur des paramètres à inférer.

Modèle à variables latentes. On distingue les données observées $\mathbf{Y} = (Y_1, \ldots, Y_n)$, qui sont les valeurs du trait mesurées pour les n espèces actuelles aux feuilles, des variables latentes $\mathbf{Z} = (Z_1, \ldots, Z_m)$, qui sont les valeurs du trait aux m nœuds internes. On note

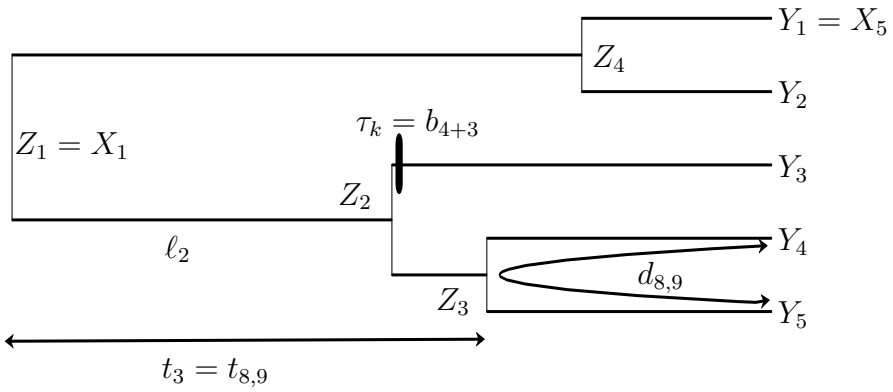


Figure 2: Structure des données et notations.

$\mathbf{X} = (\mathbf{Z}, \mathbf{Y})$ le jeu de données complété, qui est tel que : $\begin{cases} X_j = Z_j & \forall 1 \leq j \leq m \\ X_{m+i} = Y_i & \forall 1 \leq i \leq n \end{cases}$ (voir figure 2). On peut alors exprimer la loi de la valeur d'un trait à un nœud j sachant la valeur de son parent $\text{pa}(j)$ (ℓ_j est la longueur de la branche allant de $\text{pa}(j)$ à j) :

$$X_j | X_{\text{pa}(j)} \sim \mathcal{N}\left(X_{\text{pa}(j)} e^{-\alpha \ell_j} + \beta^j (1 - e^{-\alpha \ell_j}),\ \frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha \ell_j}) \right)$$

Modèle linéaire pour la loi marginale de Y. Pour un nœud j, $1 \leq j \leq n+m$, on note $\mathbf{e}_{\text{de}(j)}$ le vecteur de taille n, dont l'entré i vaut 1 si la feuille i descend du nœud j, et 0 sinon. La matrice $\mathbf{T}$ de taille $n \times (m+n)$ constituée par les vecteurs-colonnes $\mathbf{e}_{\text{de}\, j}$ représente la topologie de l'arbre. Le vecteur gaussien des données $\mathbf{Y}$ est donné par :

$$\mathbf{Y} = \beta_0 \mathbf{1} + \mathbf{T}\mathbf{V}\boldsymbol{\Delta} + \mathbf{E}$$

où $\mathbf{V} = \text{Diag}(1 - e^{-\alpha(t_a - t_{\text{pa}(j)})}; 1 \leq j \leq m+n)$ est une matrice diagonale d'actualisation, et $\mathbf{E} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, avec $\boldsymbol{\Sigma} = \frac{\sigma^2}{2\alpha} \left[e^{-\alpha d_{ij}} \right]_{1 \leq i,j \leq n}$, et d_{ij} la distance sur l'arbre entre i et j (voir figure 2).

2 Inférence des paramètres par un algorithme EM

De nombreux auteurs se sont attaqués à l'inférence des paramètres de ce modèle, qui intéresse particulièrement les écologues et évolutionnistes. On renvoie à Pennell et Harmon (2013) pour une revue de littérature, et à Ho et Ané (2013) pour une exposition des problèmes de consistances des estimateurs de certains paramètres, liés à la structure d'arbre sous-jacente. Butler et King (2004) se sont intéressés à l'inférence dans le cas où les sauts sont fixées *a priori* sur l'arbre. Uyeda et Harmon (2014) ont utilisé une approche

bayésienne pour détecter les sauts. La structure du modèle à variables latentes nous permet d'écrire un algorithme de type Expectation-Maximization (EM) pour inférer tous les paramètres en même temps par maximum de vraisemblance. Pour réaliser l'étape critique de l'initialisation, on utilise la formulation linéaire du problème, avec une régularisation de type LASSO. Les codes correspondants, implémentés sur R, sont disponibles sur GitHub (`https://github.com/pbastide/Phylogenetic-EM`).

3 Problèmes d'identifiabilité

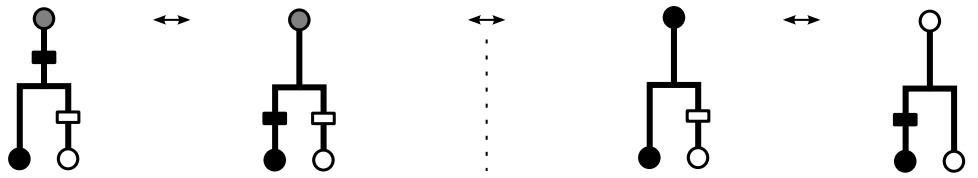


Figure 3: Quatre allocations équivalentes sur une fourche. Les deux de droite sont parcimonieuses.

Allocations équivalentes. Notre modélisation produit de manière naturelle un certain nombre de groupes d'espèces, chacun défini par une histoire évolutive propre et la distribution du trait associée. Dans le cas d'un OU sur un arbre ultramétrique, on montre que le problème d'allocation des sauts se ramène à un problème de coloration des nœuds de l'arbre, dans lequel chaque groupe correspond à une couleur. Plusieurs allocations peuvent cependant induire la même distribution aux feuilles, et ne sont donc pas distinguables. Comme on le voit figure 3, les sauts peuvent ainsi être changés de place, et l'on peut même en supprimer certains, quitte à modifier la valeur à la racine.

Allocations parcimonieuses. On règle le problème de la sur-paramétrisation en ne gardant que les solutions *parcimonieuses* (à droite figure 3). À classification des feuilles données, une telle coloration impose un nombre de changements minimum. Une modification des algorithmes classiques de Fitch et Sankoff (décrits dans le livre de Felseinstein (2004), chapitre 2) permet de compter et d'énumérer l'ensemble de ces configurations.

Ensemble des solutions identifiables à K sauts. En supposant que chaque saut produit une nouvelle couleur, cet ensemble est en bijection avec les classifications en $K+1$ groupes *compatible avec l'arbre*, c'est-à-dire obtenues par un processus de sauts. Un algorithme récursif nous permet de compter ces classifications. On constate ainsi que le nombre de modèles à K sauts *distincts* dépend en général de la topologie de l'arbre, sauf si ce dernier est binaire. Dans ce cas, ce nombre est donné par $\binom{2n-2-K}{K}$. Une procédure pour sélectionner le nombre de sauts, basée sur la méthode de Birgé et Massart (2001), est actuellement à l'étude. La pénalité envisagée dépend de ce cardinal.

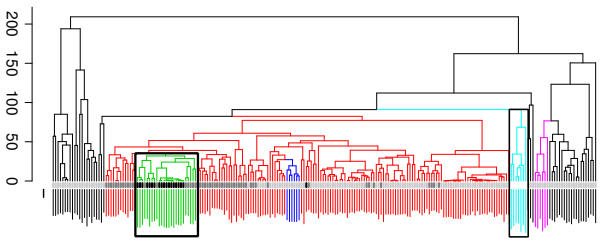


Figure 4: Arbre phylogénétique des chéloniens. Les habitats sont figurés par la frise aux feuilles : noir pour insulaire, blanc pour marin, gris clair pour eau douce, et gris foncé pour terrestre. Les caractères (en log) sont représentés par les traits prolongeant les feuilles, l'unité étant donnée à gauche. Les couleurs sur l'arbre correspondent à la classification produite par l'EM. L'arbre est étalonné en millions d'années.

4 Gigantisme insulaire chez les chéloniens

Les chéloniens sont une sous-classe de reptiles dont les seuls représentants actuels sont les tortues. Ils sont présents dans divers habitats partout dans le monde. Jaffe et al. (2011) ont recensé les tailles de carapaces pour 226 espèces de tortues, dont l'arbre phylogénétique est connu. Parmi les 6 groupes trouvés par notre algorithme (voir figure 4), on remarque que les espèces marines sont séparées dans un groupe à part (couleur cyan, cadre à droite), ainsi que les espèces insulaires (couleur verte, cadre à gauche), à quelques exceptions près, compatibles avec l'arbre. La valeur optimale pour ce dernier groupe est de 66 *cm*, contre 38 *cm* à l'origine, ce qui étaye bien l'hypothèse d'un gigantisme insulaire. Le temps de demi-vie trouvé est de 4.6 millions d'années, soit 2.3% de la hauteur de l'arbre, ce qui indique une forte sélection.

Bibliographie

[1] Butler, M. A. and King, A. A. (2004). Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. *The American Naturalist*, 164(6):683–695.
[2] Felsenstein, J. (1985). Phylogenies and the Comparative Method. *The American Naturalist*, 125(1):1–15.
[3] Felsenstein, J. (2004) Inferring Phylogenies. *Sinauer Associates*, Suderland, USA.
[4] Hansen, T. F. (1997). Stabilizing selection and the comparative analysis of adaptation. *Evolution*, 51(5):1341–1351.
[5] Ho, L. and Ané, C. (2013). Asymptotic Theory with Hierarchical Autocorrelation : Ornstein-Uhlenbeck Tree Models. *The Annals of Statistics*, 41(2):957–981.
[6] Jaffe, A. L., Slater, G. J. and Alfaro, M. E. (2011) The evolution of island gigantism and body size variation in tortoises and turtles. *Biology letters*.
[7] Pennell, M. W. and Harmon, L. J. (2013). An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. *Ann. N. Y. Acad. Sci.*, 1289:90–105.
[8] Uyeda, J. C. and Harmon, L. J. (2014). A Novel Bayesian Method for Inferring and Interpreting the Dynamics of Adaptive Landscapes from Phylogenetic Comparative Data. *Syst. Biol.*