Un cadre d'inférence bayésien pour l'étude de l'évolution de traits quantitatifs viraux

Paul Bastide [1], Guy Baele [1], Marc Suchard [2,3,4] & Philippe Lemey [1]

[1] *Department of Microbiology and Immunology, Rega Institute, KU Leuven – University of Leuven, Leuven, Belgium.* {*paul.bastide, guy.baele, philippe.lemey*}*@kuleuven.be*

[2] *Department of Biostatistics, Jonathan and Karin Fielding School of Public Health, University of California, Los Angeles, United States. msuchard@ucla.edu*

[3] *Department of Biomathematics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, United States.*

[4] *Department of Human Genetics, David Geffen School of Medicine at UCLA, Universtiy of California, Los Angeles, United States.*

Résumé. Au cours d'une épidémie, certains pathogènes viraux subissent une évolution rapide, si bien que les pressions évolutives liées à leur propagation se reflètent dans leur génome. Retrouver les traces moléculaires de ces processus de transmission est l'un des objectifs traditionnels du champ de la phylodynamique. L'évolution de traits quantitatifs viraux, comme la position géographique ou la virulence, a été relativement moins étudiée. Les méthodes comparatives phylogénétiques ont pour but d'étudier la distribution de traits quantitatifs au sein d'un ensemble d'organismes non indépendants, liés par une histoire évolutive partagée. Conditionnellement à cette histoire, les traits observés peuvent être vus comme le résultat d'un processus stochastique courant sur les branches d'un arbre phylogénétique. On décrit ici un cadre d'inférence bayésienne pour l'étude de ces modèles d'évolution. Celui-ci repose sur une méthode de type MCMC, rendue possible grâce, d'une part, à une procédure d'échantillonnage non biaisée de l'espace contraint des paramètres du modèle, et, d'autre part, à une méthode de calcul efficace de la vraisemblance, qui exploite la structure arborescente du problème. La méthode proposée peut s'appliquer à une large gamme de processus stochastiques gaussiens, rendant possible une modélisation fine des divers processus biologiques mis en œuvre. Elle est implémentée au sein du logiciel d'inférence phylogénétique BEAST, et permet, à titre d'exemple, de jeter un regard nouveau sur le problème de l'héritabilité de la virulence du VIH.

Mots-clés. Méthodes bayésiennes, Statistique computationnelle, Statistique des processus, Biostatistique.

Abstract. During the course of an epidemic, many viral pathogens are known to evolve rapidly, leaving an imprint of the pattern of spread in their genomes. Uncovering the molecular footprint of this transmission process is a key goal of phylodynamic inference. Less focus has been put on the evolution of quantitative traits of viruses, such as geographical location or virulence. The goal of Phylogenetic Comparative Methods is to account for a shared evolutionary history among a set of non-independent samples. Conditioning on such an history, the observed traits can be seen as the result of a stochastic

process running on the branches of a phylogenetic tree. We propose a Bayesian inference framework for the study of this flexible model. Using a MCMC based method, it relies on the efficient sampling of the constrained parameters of the model, and takes advantage of the tree structure for fast likelihood computations. It encompasses a wide family of Gaussian processes, allowing for fine-grained modelling of trait evolution of various biological systems. We implemented this new approach in the phylogenetic software BEAST, and applied it to the study of heritability of virulence in HIV.

Keywords. Bayesian methods, Computational statistics, Stochastic processes, Biostatistics.

1 Modèles d'évolution de traits quantitatifs

Inférence phylogénétique bayésienne. La phylodynamique s'intéresse aux processus épidémiologiques et environnementaux qui influencent l'évolution de pathogènes au cours d'une épidémie. Lors du suivi de la maladie, on recueille typiquement deux types de données au cours de l'échantillonnage : les séquences moléculaires datées $\mathbf{S}$, et un certain nombre de traits quantitatifs $\mathbf{Y}$ liés aux caractéristiques du virus, comme la position géographique ou la virulence. Le but de l'inférence statistique est alors d'obtenir des informations sur l'arbre phylogénétique $\mathcal{T}$ liant les n espèces échantillonnées entre elles, ainsi que sur les paramètres $\boldsymbol{\theta}$ des processus d'évolution des caractères quantitatifs. Dans une perspective bayesienne, on cherche à obtenir des informations sur la loi a posteriori $p(\boldsymbol{\theta}, \mathcal{T} \mid \mathbf{Y}, \mathbf{S})$. L'hypothèse fondamentale que l'on fait ici est que, conditionnellement à l'arbre $\mathcal{T}$, les séquences $\mathbf{S}$ et les traits $\mathbf{Y}$ sont **indépendants**. Cette indépendance conditionnelle permet de séparer l'inférence en deux problèmes distincts, en écrivant :

$$p(\boldsymbol{\theta}, \mathcal{T} \mid \mathbf{Y}, \mathbf{S}) \propto p(\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T})\, p(\boldsymbol{\theta}, \mathcal{T}) = p(\mathbf{Y} \mid \boldsymbol{\theta}, \mathcal{T})\, p(\boldsymbol{\theta}) \times p(\mathbf{S} \mid \mathcal{T})\, p(\mathcal{T}) \quad (1)$$

Le deuxième terme de ce produit porte sur l'inférence phylogénétique classique à partir des séquences moléculaires. Ce problème a reçu beaucoup d'attention ces dernières années, et il existe de nombreux modèles et méthodes bayésiennes pour l'étudier (Suchard et al., 2018). On s'intéresse ici à l'étude du premier terme du produit, portant sur les traits quantitatifs. L'écriture ci-dessus nous permet de supposer dans toute la suite, sans perte de généralité, que l'on raisonne conditionnellement à un arbre $\mathcal{T}$ fixé.

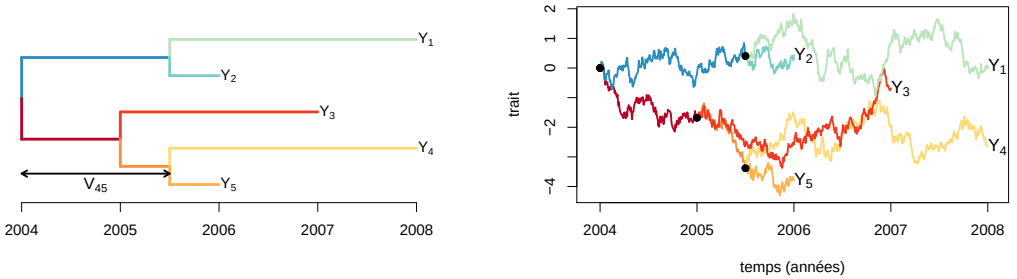
Méthode Comparatives Phylogénétiques (MCP). Le but des MCP est précisément d'étudier les traits quantitatifs d'organismes non indépendants, liés par une histoire évolutive commune, représentée par un arbre phylogénétique. Les MCP sont principalement développées en écologie et macro-évolution (Pennell & Harmon, 2013), mais leur utilisation a également été proposée plus récemment pour l'étude de pathogènes à évolution

rapide (Alizon et al., 2010). À arbre fixé, ces méthodes modélisent l'évolution des traits quantitatifs au cours du temps par un processus stochastique courant sur les branches de l'arbre (voir figure 1). Lors d'une spéciation (à un embranchement), le processus se divise en deux processus indépendants et de même loi. Le processus stochastique est choisi pour refléter les caractéristiques biologiques de l'évolution du système.

Mouvement brownien (MB) multivarié. Le MB de dimension p dépend de deux paramètres : la valeur ancestrale $\boldsymbol{\mu}$, et la matrice de variance $\mathbf{R}$. La covariance entre les traits k et l aux feuilles i et j est le simple produit entre le temps d'évolution partagé V_{ij} (voir fig. 1), et la covariance R_{kl} : $\mathbb{C}\mathrm{ov}\left[Y_{ik}; Y_{jl}\right] = V_{ij}R_{kl}$. La loi jointe de la matrice d'observations $\mathbf{Y}$ (taille $n \times p$) peut ainsi s'écrire comme une gaussienne matricielle :

$$\mathbf{Y} \sim \mathcal{MN}\left(\boldsymbol{\mu}\mathbf{1}_p^T, \mathbf{V}, \mathbf{R}\right). \tag{2}$$

Cette factorisation de la variance, qui découle de l'indépendance des incréments du MB, en rend l'inférence plus aisée. Cependant, cette même indépendance rend le MB inapte à modéliser des phénomènes de sélections qui peuvent influencer l'évolution des traits.



(a) Arbre calibré en temps. V_{45} est le temps d'évolution partagé entre les espèces 4 et 5.

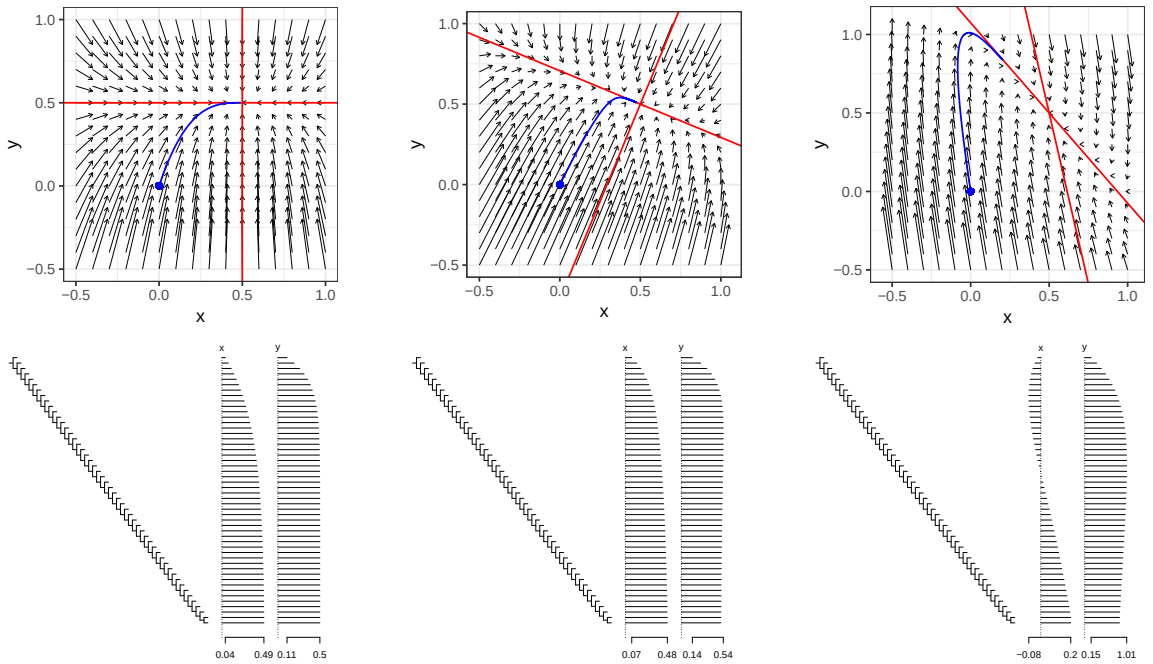
(b) Variation du trait en fonction du temps. Seule la valeur aux feuilles est observée.

FIGURE 1 – Arbre phylogénétique liant les organismes mesurées (gauche), et modélisation de l'évolution d'un trait par un MB (droite). Les couleurs des deux figures sont assorties.

Ornstein-Uhlenbeck (OU) multivarié. L'OU présente un mécanisme de rappel vers une valeur centrale $\boldsymbol{\beta}$, interprétée comme la valeur optimale des traits dans un environnement donné. La dynamique de rappel est contrôlée par la matrice de force de sélection $\mathbf{A}$, que l'on suppose dans toute la suite diagonalisable dans $\mathbb{R}$ avec des valeurs propres strictement positives ($\mathbf{A} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{-1}$ avec $\lambda_k > 0$). Les traits évoluent au cours du temps d'après l'équation différentielle stochastique : $\mathrm{d}\mathbf{X}_t = \mathbf{A}(\boldsymbol{\beta} - \mathbf{X})\mathrm{d}t + \mathbf{R}^{1/2}\mathrm{d}\mathbf{B}_t$. Des exemples de la partie déterministe de l'équation pour différentes formes de $\mathbf{A}$ sont présentés figure 2. La loi jointe des observations $\mathbf{Y}$ est gaussienne, mais ne peut pas se factoriser comme pour le MB. Les moments des vecteurs-lignes $\mathbf{Y}_i^T$ sont donnés par :

$$\mathbb{E}\left[\mathbf{Y}_i\right] = \boldsymbol{\mu}e^{-\mathbf{A}V_{ii}} + \boldsymbol{\beta}(1 - e^{-\mathbf{A}V_{ii}}) \qquad \mathbb{C}\mathrm{ov}\left[\mathbf{Y}_i; \mathbf{Y}_j\right] = \mathbf{P}\left[\mathbf{W}_{ij} \odot \mathbf{P}^{-1}\mathbf{R}\mathbf{P}^{-T}\right]\mathbf{P}^T \tag{3}$$

où $\odot$ est le produit de Hadamard, et $\mathbf{W}_{ij} = \left[\frac{1}{\lambda_k+\lambda_l} e^{-\lambda_k V_{ii}} e^{-\lambda_l V_{jj}} \left(e^{(\lambda_k+\lambda_l)V_{ij}} - 1\right)\right]_{1\leq k,l\leq p}$.



(a) **A** diagonale. (b) **A** symmétrique. (c) **A** diagonalisable dans $\mathbb{R}$.

FIGURE 2 – Dynamiques de rappel à la moyenne pour différentes formes de **A**. Première ligne : champ de vecteurs induit par l'équation différentielle déterministe. La base **P** est représentée en rouge. La trajectoire de $\boldsymbol{\mu} = (0, 0)$ à $\boldsymbol{\beta} = (0.5, 0.5)$ est tracée en bleu. Seconde ligne : effet sur l'espérance de traits échantillonnés aux feuilles d'un arbre parfaitement étagé. La partie stochastique n'est pas représentée.

2 Inférence bayésienne des paramètres

Métropolis-Hasting dans un espace contraint. En utilisant la décomposition (1), on cherche à échantillonner les paramètres $\boldsymbol{\theta} = (\boldsymbol{\mu}, \mathbf{R}, \boldsymbol{\beta}, \mathbf{A})$ dans la loi a posteriori. Grâce à sa loi factorisée (2), il est possible pour le MB d'utiliser un échantillonneur de Gibbs, en prenant des lois a priori conjuguées pour les paramètres, à savoir une loi inverse-Wishart pour **R**, et, conditionnellement à celle-ci, une loi normale pour $\boldsymbol{\mu}$. Un tel échantillonneur, qui a l'avantage de ne proposer automatiquement que des matrices symétriques définies positives, n'est pas possible pour l'OU, dont la loi est plus complexe. Pour échantillonner dans l'espace contraint des paramètres, on peut alors utiliser une transformation qui met cet espace en difféomorphisme avec $\mathbb{R}^q$, où q est le nombre de paramètres libres. Il suffit ensuite de corriger le ratio d'acceptation de Métropolis-Hasting par le jacobien de cette transformation. On présente dans la suite deux transformations pour **R** et **A**.

Échantillonnage de la variance. On utilise ici une version modifiée de la transformation proposée par Lewandowski et al. (2009), dont on donne une présentation originale, inspirée par la méthode utilisée dans Stan (2017). La première étape est de décomposer la variance en une matrice de corrélation et une matrice diagonale de variance : $\mathbf{R} = \mathbf{DCD}$, où $\mathbf{C}$ est symétrique définie positive d'éléments diagonaux tous égaux à 1. On utilise ensuite la décomposition de Cholesky de $\mathbf{C} = \mathbf{WW}^T$, avec $\mathbf{W}$ triangulaire supérieure, d'éléments diagonaux strictement positifs, ce qui assure l'unicité. Il est alors facile de voir que chaque colonne de $\mathbf{W}$ est de norme euclidienne égale à 1. Échantillonner $\mathbf{W}$ revient donc à échantillonner $p-1$ vecteurs de dimensions $k = 1, \ldots, p-1$ dans la boule unitaire ouverte. Le cœur de la transformation dite LKJ est alors le difféomorphisme entre le pavé $]-1\,,1[^k$ et la boule unitaire euclidienne, donnée, pour $1 \leq i \leq k$, par : $\mathbf{LKJ}_i(\mathbf{z}) = z_i \prod_{k=1}^{i-1} \sqrt{1-z_k^2}$ (où le produit vide est pris égal à 1). Une transformation de Fisher (ou arc tangente hyperbolique) permet enfin de se ramener à $\mathbb{R}^k$ pour chaque vecteur-colonne. Pour compléter cette transformation, on peut alors mettre comme loi a priori sur chaque vecteur-colonne une loi bêta sphérique, ce qui revient à utiliser une loi dite LKJ sur la matrice $\mathbf{C}$.

Échantillonnage de la force de sélection. On utilise la décomposition $\mathbf{A} = \mathbf{P\Lambda P}^{-1}$, en imposant les conditions suivantes pour assurer l'unicité : (a) les valeurs propres sont distinctes deux à deux et classées dans l'ordre croissant ($\lambda_k < \lambda_l$ pour tout $1 \leq k < l \leq p$) et (b) les vecteurs propres sont de norme 1, et de dernière composante strictement positive. La condition (a) est respectée facilement en échantillonnant le logarithme des différences de valeurs propres. Pour la condition (b), on peut utiliser la même transformation LKJ que précédemment, qui lie la boule unitaire à $\mathbb{R}^{p-1}$.

Calcul efficace de la vraisemblance. L'algorithme de calcul de la vraisemblance repose sur l'observation que, pour l'OU comme pour le MB, la loi conditionnelle $\mathbf{X}_i \mid \mathbf{X}_{\mathrm{pa}(i)}$ d'un nœud de l'arbre i sachant son père $\mathrm{pa}(i)$ est une gaussienne $\mathcal{N}\left(\mathbf{q}_i\mathbf{X}_{\mathrm{pa}(i)} + \mathbf{r}_i,\ \mathbf{\Sigma}_i\right)$, où $\mathbf{q}_i$ est une matrice inversible de taille $p \times p$, et $\mathbf{\Sigma}_i$ une matrice de variance. Pour un MB, on a par exemple : $\mathbf{q}_i = \mathbf{I}_p$, $\mathbf{r}_i = \mathbf{0}_p$ et $\mathbf{\Sigma}_i = \ell_i\mathbf{R}$, où ℓ_i est la longueur de la branche entre i et son père. Cette écriture permet de calculer la vraisemblance en un seul parcours de l'arbre, des feuilles vers la racine (Bastide et al., 2018). Un grand nombre de processus gaussiens, comme le MB ou l'OU intégrés, peuvent entrer dans ce cadre général. De plus, cette écriture permet de prendre en compte facilement des erreurs de mesures ou autres variations environnementales influençant les espèces mesurées de manière indépendantes, en ajoutant simplement une couche d'incertitude gaussienne aux feuilles de l'arbre.

Implémentation. On propose une implémentation efficace de cette procédure dans le logiciel d'inférence phylogénétique BEAST (Suchard et al., 2018). L'intégration de la méthode à cette plateforme permet de bénéficier de tous les outils disponibles dans celle-ci, et en particulier d'intégrer sur l'espace des arbres, en utilisant la décomposition (1).

3 Application à l'étude de la virulence du VIH

Présentation du problème. Les MCP ont été proposées pour étudier la virulence du VIH, telle que mesurée par le taux de déclin des cellules CD4 et par la charge virale de plateau (*set point viral load*, Alizon et al., 2010). Une virulence trop faible ou trop forte pouvant nuire à la propagation du virus, on s'attend à ce que celle-ci suive un OU. En phylogénie, l'héritabilité est définie comme le ratio de la variance strictement liée au modèle de propagation du trait sur l'arbre sur la variance totale, incluant les variations environnementales indépendantes. Les études précédentes de la questions ont donné des estimations très variables, allant de 50 à 6% (Blanquart et al., 2017).

Résultats préliminaires. Appliquée au jeu de données présenté dans Blanquart et al. (2017), la méthode donne des résultats mitigés. Une procédure d'estimation de la vraisemblance marginale nous permet de préférer l'OU au MB pour modéliser l'évolution de la virulence. On trouve cependant que les intervalles de crédibilité des différents paramètres, y compris l'héritabilité, sont très large, en accord avec la littérature. Ce peu de résolution pose la question de la définition même de l'héritabilité, pour laquelle un consensus manque encore pour ce type de méthodes appliquées en virologie (Mitov & Stadler, 2018).

Bibliographie

Alizon S, von Wyl V, Stadler T, et al. (18 co-authors). 2010. Phylogenetic approach reveals that virus genotype largely determines HIV set-point viral load. *PLoS Pathogens*. 6.

Bastide P, Ané C, Robin S, Mariadassou M. 2018. Inference of Adaptive Shifts for Multivariate Correlated Traits. *Systematic Biology*. 67 :662–680.

Blanquart F, Wymant C, Cornelissen M, et al. (30 co-authors). 2017. Viral genetic variation accounts for a third of variability in HIV-1 set-point viral load in Europe. *PLoS Biology*. 15 :1–26.

Lewandowski D, Kurowicka D, Joe H. 2009. Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*. 100 :1989–2001.

Mitov V, Stadler T. 2018. A Practical Guide to Estimating the Heritability of Pathogen Traits. *Molecular Biology and Evolution*. 35 :756–772.

Pennell MW, Harmon LJ. 2013. An integrative view of phylogenetic comparative methods : connections to population genetics, community ecology, and paleobiology. *Annals of the New York Academy of Sciences*. 1289 :90–105.

Stan DT. 2017. Stan Modeling Language : User's Guide and Reference Manual v 2.17.0.

Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*. 4 :1–5.