

# Projet: ANOVA Phylogénétique

Présentation du Mercredi 10 Jan. 2024

---

Alizée Geffroy, Louis Lacoste, encadrés par Mélina Gallopin et Paul Bastide

M2 MathSV Université Paris-Saclay

1. Rappel du contexte et enjeux du projet
2. Conclusions et perspectives
3. Références et appendices



TODO Supprimer cette slide temporaire intro/contexte: biologique avec l'exemple de Chen (mettre l'arbre) + figure de l'article ? -¿ trouver les gènes différentiellement exprimés

Il existe déjà des méthodes statistiques pour cette problématique (EVEmodel ? State of the Art)

Transition avec le pourquoi du projet, trouver d'autres méthodes statistiques, adaptées de méthodes classiques qui pourraient bien marcher

Méthode pas par nous : 1 slide par tiret - Reprendre la forme matricielle de l'ANOVA phylo (mettre en rouge les diffs) - Présenter le MB qui évolue sur l'arbre + lien matrice K - Mettre la statistique de test (mettre en rouge la projection (donc diffs))

Transition vers notre travail - Mettre la formule avec erreur de mesure avec justification de l'ajout de l'erreur de mesure, formule transfo  $V_\lambda$ , pointer la limite qui est l'erreur dûe à l'estimation du  $\lambda$  Méthode par nous :



Setterthwaite : préciser que c'est nos calculs à partir de résultats sur

- comment obtenir la stat de test pour anova phylo (Cholesky) - en quoi c'est un modèle mixte pour Satterthwaite ? - calcul de la Hessienne optim vs formule analytique, mettre formule analytique



## **Rappel du contexte et enjeux du projet**

---

- Un arbre phylo avec plusieurs espèces
- Un trait quantitatif présent chez ces espèces
- Représenté par un paramètre  $\mu$

Typiquement un gène dont on mesure l'expression. Dans **Gomez-Mestre, Pyron, and Wiens (2012)** ces méthodes sont utilisées pour répondre à des questions d'évolution et d'ordre d'apparition de caractères chez les *Anoures*.



# ANOVA vs ANOVA phylogénétique

$$\mathbf{Y} = \mathbf{X}\beta + \sigma\mathbf{E} \text{ où } \mathbf{X} = (1, 1_2, \dots, 1_K) \text{ et } \beta = (\mu_1, \beta_2, \dots, \beta_K)^T$$

Anova

$$\mathbf{E} \sim \mathcal{N}(0_n, \mathbf{Id})$$

Anova phylogénétique

$$\mathbf{E} \sim \mathcal{N}(0_n, \mathbf{V})$$

Estimateur du max. de vraisemblance

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$
$$\hat{\sigma}^2 = \frac{1}{n-p} (\mathbf{Y} - \mathbf{X}\hat{\beta})^T (\mathbf{Y} - \mathbf{X}\hat{\beta})$$

$$\hat{\beta}_{phylo} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{Y}$$
$$\hat{\sigma}_{phylo}^2 = \frac{1}{n-p} (\mathbf{Y} - \mathbf{X}\hat{\beta})^T \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\hat{\beta})$$

Objectif : Comparer les méthodes Anova et Anova phylogénétique

---

<sup>0</sup>Pour l'ANOVA Phylogénétique, nos références sont [bastideContinuousTraitEvolution](#) et [Bastide, Mariadassou, and Robin 2022](#) et pour l'ANOVA, [Bel et al. n.d.](#)



On forme des groupes **en lien avec la phylogénie.**

**Figure 1:** Arbre phylogénétique et groupes concordants

On forme des groupes qui ne sont **pas phylogénétiques.**

**Figure 2:** Arbre phylogénétique et groupes non concordants



**Figure 3:** Avec de la variabilité purement phylogénétique

**Figure 4:** Avec de la variabilité phylogénétique et d'erreur de mesure



# Ce qu'il se passe vraiment

L'observation se présente sous la forme suivante si réécrite en tant que modèle à effets mixtes :

$$\mathbf{Y} = \mathbf{X}\beta + \underbrace{\mathbf{E}}_{\mathbf{Z}u + \epsilon}$$

avec  $\mathbf{Z}u \sim \mathcal{N}(0, \sigma_{phylo}^2 \mathbf{V})$ ,  $\epsilon \sim \mathcal{N}(0, \sigma_{mesure}^2 \mathbf{Id})$  et

$$\text{Var}(\mathbf{E}) = \sigma_{phylo}^2 (\mathbf{V} + \lambda \mathbf{Id}) \text{ où } \lambda = \frac{\sigma_{mesure}^2}{\sigma_{phylo}^2}$$

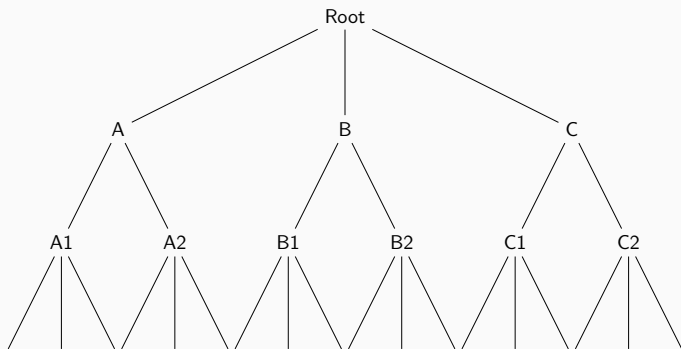
Les méthodes d'ANOVA phylogénétique telles qu'implémentées dans `phyloilm` estiment les paramètres  $\sigma_{mesure}^2$  d'erreur de mesure et  $\sigma_{phylo}^2$  de variabilité dûe à la phylogénie.

*Mais* une fois estimé, lors du test de Fisher les paramètres sont considérés comme si non estimés.

Le but ici est donc d'utiliser l'approximation de [Satterthwaite \(1946\)](#) afin de calculer les degrés de libertés du test de loi de Fisher à réaliser.

## **Conclusions et perspectives**

---



Le package `phylolimma`<sup>1</sup> permet de compléter un arbre existant en ajoutant les sous-branches au bout.

Nous voulons obtenir une méthode d'estimation des paramètres et l'implémenter sur ce type d'arbre.



<sup>1</sup>Disponible sur <https://github.com/pbastide/phylolimma/>

- Implémenter le test statistique correspondant à l'approximation de Satterthwaite.
- Implémenter le test de ratio de log-vraisemblance.
- Implémenter avec plusieurs individus par espèces

## **Objectif principal**

Trouver un test robuste et rapide, applicable à des milliers de données d'expressions de gènes mesurées dans une expérience RNAseq typique.



## Références et appendices

---

## References

---



Bastide, Paul, Mahendra Mariadassou, and Stéphane Robin (July 2022). “Modèles d’évolution de caractères continus”. In: Didier, Gilles and Stéphane Guindon. *Modèles et méthodes pour l’évolution biologique*. ISTE Group, pp. 47–85. ISBN: 978-1-78948-069-6. DOI: 10.51926/ISTE.9069.ch3. URL: <https://www.istegroup.com/fr/produit/modeles-et-methodes-pour-levolution-biologique/?/47495> (visited on 11/14/2023).



Bel, L et al. (n.d.). *Le Modèle Linéaire et ses Extensions*.





Gomez-Mestre, Ivan, Robert Alexander Pyron, and John J. Wiens (2012). “Phylogenetic Analyses Reveal Unexpected Patterns in the Evolution of Reproductive Modes in Frogs”. In: *Evolution* 66.12, pp. 3687–3700. ISSN: 1558-5646. DOI: 10.1111/j.1558-5646.2012.01715.x. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1558-5646.2012.01715.x> (visited on 11/13/2023).



Satterthwaite, F. E. (Dec. 1946). “An Approximate Distribution of Estimates of Variance Components”. In: *Biometrics Bulletin* 2.6, p. 110. ISSN: 00994987. DOI: 10.2307/3002019. JSTOR: 10.2307/3002019. URL: <https://www.jstor.org/stable/10.2307/3002019?origin=crossref> (visited on 01/08/2024).



Le code pour les simulations est disponible sur notre dépôt GitHub :

`https:  
//github.com/Polarolouis/anova-phylogenetique-projet-msv/`