

Rapport de Projet : ANOVA Phylogénétique

Alizée Geffroy

Louis Lacoste

5 mars 2024

Table des matières

1	Introduction	3
2	Méthodes	4
2.1	L'ANOVA	4
2.2	L'ANOVA phylogénétique	4
2.3	Approximation de Satterthwaite	4
3	Méthodologie	5
3.1	Simulations	6
4	Données	6
5	Résultats	6
6	Discussion et conclusion	6

1 Introduction

Ici contexte biologique, les données de [GOMEZ-MESTRE, PYRON et WIENS 2012](#), les données de Paul et Mélina etc.

Avec l'avènement des données massives de génomiques, transcriptomiques, protéomiques etc, il y a besoin de techniques statistiques robustes et passant à l'échelle permettant de mener à bien l'analyse. Format des données : arbres phylogénétiques, données génétiques. Arbres avec des petites branches : plusieurs individus par espèces avec chacun leurs données → problème biologique

Deux sujets différents écologie et transcriptomique mais une même méthode.

Pour données [CHEN et al. 2019](#) la figure 1 présente l'arbre phylogénétique :

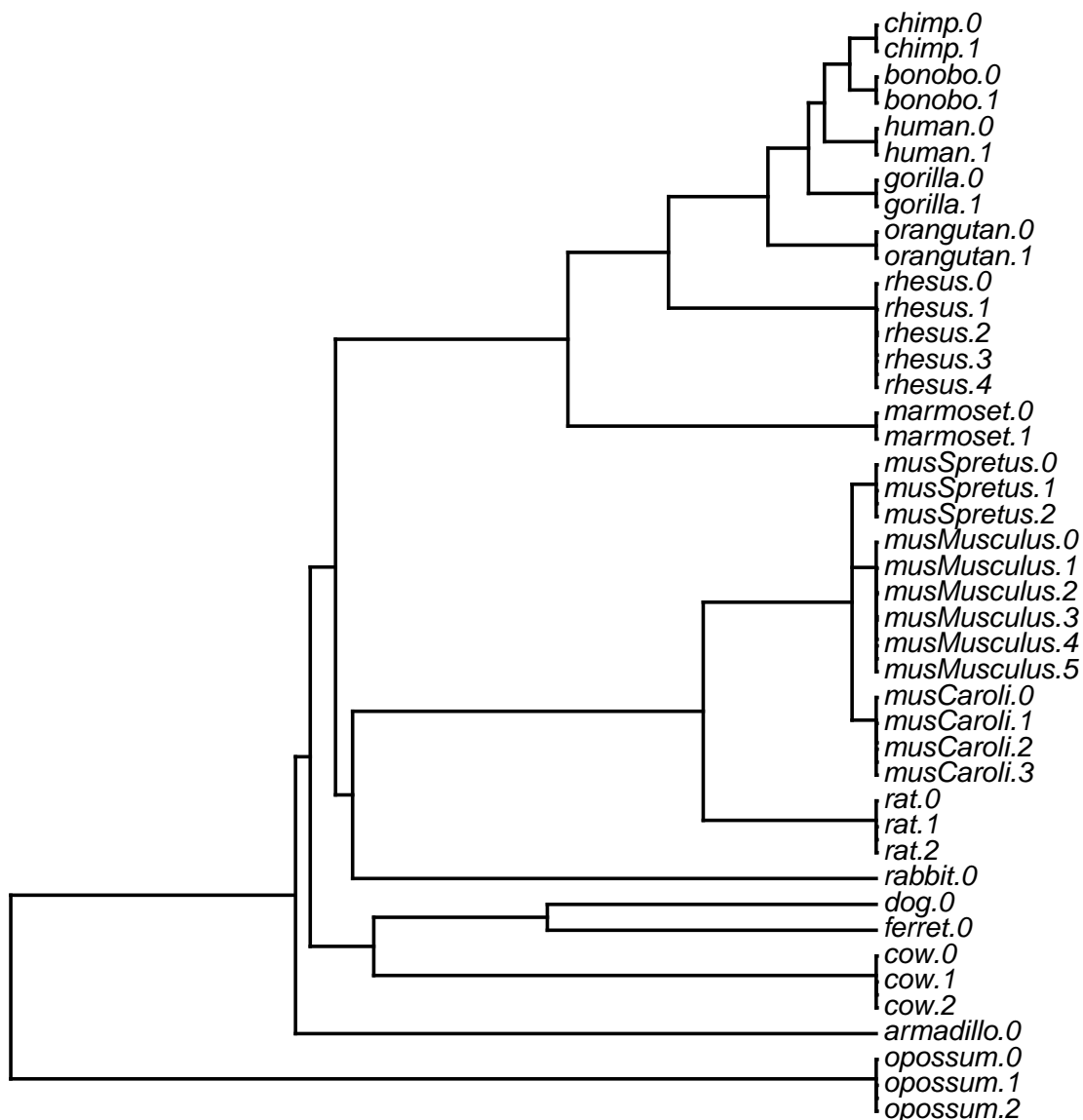


FIGURE 1 – Arbre phylogénétique de [CHEN et al. 2019](#)

Transition, c'est pourquoi on va tester la méthode d'ANOVA phylogénétique avec cette forme de données. But ? Etudier cette méthode et les résultats

Un gène, comparer les moyennes d'expression d'un gène On connaît les groupes exemple individus malade/sain

Comparaison non pas sur individus malades/pas malades mais sur espèces différentes. Pas possible de supposer iid, existe relations entre les individus et les groupes que l'on compare donc besoin de les prendre en compte.

Modele mixte la matrice des temps de divergences, BM simple sans erreurs, avec erreur (ajustement du ratio) avec OU...

2 Méthodes

Ici les rappels sur l'ANOVA, l'explication de l'ANOVA phylogénétique. La démonstration des limites de l'ANOVA phylogénétique par des simulations Méthode : la partie maths anova, anova phylo, satterthwaite,

2.1 L'ANOVA

L'ANOVA est un cas classique du modèle linéaire, nous utilisons ici les notations et le formalisme de [BEL et al. s. d.](#)

Le principe de l'ANOVA est d'expliciter le lien entre une variable quantitative et une ou plusieurs variables qualitatives.

La forme usuelle de l'ANOVA à 1 facteur est la suivante :

$$Y_{ik} = \mu_i + E_{ik}, \quad i = 1, \dots, I, k = 1, \dots, n_i, E_{ik} \sim \mathcal{N}(0, \sigma^2) \quad (1)$$

où dans cette équation, reprise du livre ([BEL et al. s. d.](#)), i représente le niveau du facteur et k indique le numéro de l'observation dans ce niveau. I est le nombre total de niveaux du facteur, n_i le nombre d'observation du niveau i .

L'ANOVA se généralise à deux facteurs, plus facilement compréhensible avec cette forme, non identifiable :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + E_{ijk}, \quad i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, n_{ij}, E_{ijk} \sim \mathcal{N}(0, \sigma^2) \quad (2)$$

où μ représente un effet moyen de la population (*intercept*), α_i l'effet du premier facteur de niveau i , β_j l'effet du second facteur de niveau j .

Les paramètres de l'ANOVA sont estimables, grâce par exemple à la méthode du maximum de vraisemblance et ont des formules bien connues.

2.2 L'ANOVA phylogénétique

parler du BM? PUIS de la matrice V ou K qui donne la structure phylogénétique

Etre assez concis sur l'histoire de la projection et le modèle et les différences avec l'ANOVA.

2.3 Approximation de Satterthwaite

Pourquoi vouloir l'utiliser? Réduire nbre de degrés de liberté utilisés dans la stat de test. Le but est d'approximé le nbre de degré de Liberté. On se basera sur la documentation du package lmer [KUZNETSOVA, BROCKHOFF et CHRISTENSEN 2017](#) pour ensuite

implémenter une approximation de Satterthwaite.

$$Y = X\beta + u + \epsilon \quad (3)$$

$$\text{où } \mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \quad u \sim \mathcal{N}_n(0, \sigma_{phy}^2 K), \quad \epsilon \sim \mathcal{N}_n(0, \sigma_{err}^2 I_n)$$

$$\text{Alors } Y \sim \mathcal{N}_n(X\beta, \sigma_{phy}^2 K + \sigma_{err}^2 I_n) \quad \text{et} \quad Var_\theta(Y) = V(\theta) = \sigma_{phy}^2 K + \sigma_{err}^2 I_n$$

De là on obtient :

$$C(\theta) = (Cov(\beta_i, \beta_j))_{i,j} = (X^T V(\theta)^{-1} X)^{-1} = (X^T (\sigma_{phy}^2 K + \sigma_{err}^2 I_n)^{-1} X)^{-1} \quad (4)$$

Toujours en suivant la documentation [KUZNETSOVA, BROCKHOFF et CHRISTENSEN 2017](#) on obtient une expression pour les degrés de liberté df ainsi qu'une approximation. Ce qui nous donne :

$$df = \frac{2(l^T \hat{C} l)^2}{[Var(l^T \hat{C} l)]} = \frac{2(f(\hat{\theta}))^2}{[Var(f(\hat{\theta}))]} \approx \frac{2(f(\hat{\theta}))^2}{[\nabla f(\hat{\theta})]^T A [\nabla f(\hat{\theta})]} \quad (5)$$

$$\text{où } \hat{C} = C(\hat{\theta}) \quad \text{et} \quad f(\theta) = l^T C(\theta) l$$

On va donc dans la suite calculer $\nabla f(\theta)$ puis l'appliquer en $\hat{\theta}$ et A la matrice de variance-covariance de $\hat{\theta} = (\hat{\sigma}_{phy}^2, \hat{\sigma}_{err}^2)$

Calcul du gradient. Nous voulons calculer les dérivées partielles $\partial_{\sigma_{phy}^2} f(\theta)$ et $\partial_{\sigma_{err}^2} f(\theta)$. Pour les premières étapes de calculs, on écrira seulement ∂ sans distinction car ce sont les mêmes expressions pour les 2 dérivées. On utilisera dans la suite les formules de [PETERSEN et PEDERSEN 2012](#) pour les dérivées de matrice

$$\partial f(\theta) = l^T \partial C(\theta) l$$

$$\partial C(\theta) = \partial (X^T V(\theta)^{-1} X)^{-1} = -C(\theta) \partial (X^T V(\theta)^{-1} X) C(\theta)$$

$$\partial (X^T V(\theta)^{-1} X) = \partial (X^T V(\theta)^{-1}) X + \cancel{X^T V(\theta)^{-1} \partial (X)} \quad (\partial_{\sigma_{phy}^2} (X) \text{ et } \partial_{\sigma_{err}^2} (X) \text{ sont nulles})$$

□

3 Méthodologie

lrt ANOVA normale VANILLA = ANOVA phylo sans correction des degrés de liberté
 $df1 = K - 1, df2 = n - K$ ANOVA phylo (avec REML)

test sur arbre quelconque puis sur arbre avec petites branches ?

Ou faire une partie à part entière avec 1) ANOVA vs ANOVA phylo sans correction des degrés de liberté b) avec une sous partie sur le REML

2) ANOVA phylo avec approximation de Satterthwaite a) prez a) simulation et résultats b) instabilités numériques -> correction avec la Hessienne ? c) La hessienne analytique ? A voir si besoin d'une partie supplémentaire

3 parties : - théo - méthodo par simu - appli aux données réelles

3.1 Simulations

Simu : Plusieurs design, tailles etc On sait la vérité, on peut connaître les vrais positifs etc Qu'est ce qu'on prend en entrées qu'est ce qu'on veut en sortie

Bien insister sur l'arbre d'entrée et l'objectif de la simu : quelle approche pour mieux détecter les gènes différentiellement exprimés.

Simulations :

- soit selon l'arbre des données
- soit partir sur regarder l'impact de la taille de l'arbre etc.

4 Données

Revenir sur explication de gènes différentiellement exprimées etc.

Applications aux données réelles de Chen mais ne pas perdre de temps à expliquer en détails EVEmodel (dire que c'est State of the art).

5 Résultats

6 Discussion et conclusion

Intro

Application/Résultats : décrire les données, vite fait normalisation avec vrai arbre, on ne connaît pas Discussion/Conclusion ? Interprétation des résultats sinon la mettre dans les

Références

BASTIDE, Paul et Julien CLAVEL (s. d.). “Continuous Trait Evolution”.

BASTIDE, Paul, Mahendra MARIADASSOU et Stéphane ROBIN (juill. 2022). “Modèles d'évolution de caractères continus”. In : DIDIER, Gilles et Stéphane GUINDON. *Modèles et méthodes pour l'évolution biologique*. ISTE Group, p. 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> (visité le 14/11/2023).

BASTIDE, Paul, Charlotte SONESON et al. (1^{er} jan. 2023). “A Phylogenetic Framework to Simulate Synthetic Interspecies RNA-Seq Data”. In : *Molecular Biology and Evolution* 40.1, msac269. ISSN : 1537-1719. DOI : 10.1093/molbev/msac269. URL : <https://doi.org/10.1093/molbev/msac269> (visité le 20/11/2023).

BEL, L et al. (s. d.). *Le Modèle Linéaire et ses Extensions*.

Bgee (2023). *Bgee : Gene Expression Data in Animals*. URL : <https://www.bgee.org/> (visité le 20/11/2023).

CHEN, Jenny et al. (jan. 2019). “A Quantitative Framework for Characterizing the Evolutionary History of Mammalian Gene Expression”. In : *Genome Res* 29.1, p. 53-63. ISSN : 1549-5469. DOI : 10.1101/gr.237636.118. pmid : 30552105.

- GOMEZ-MESTRE, Ivan, Robert Alexander PYRON et John J. WIENS (2012). “Phylogenetic Analyses Reveal Unexpected Patterns in the Evolution of Reproductive Modes in Frogs”. In : *Evolution* 66.12, p. 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> (visité le 13/11/2023).
- KUZNETSOVA, Alexandra, Per B. BROCKHOFF et Rune H. B. CHRISTENSEN (2017). “**lmerTest** Package : Tests in Linear Mixed Effects Models”. In : *J. Stat. Soft.* 82.13. ISSN : 1548-7660. DOI : 10.18637/jss.v082.i13. URL : <http://www.jstatsoft.org/v82/i13/> (visité le 01/03/2024).
- PETERSEN, Kaare Brandt et Michael Syskind PEDERSEN (2012). *The Matrix Cookbook*. Version 20121115. URL : <http://matrixcookbook.com>.
- SATTERTHWAITE, F. E. (déc. 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> (visité le 08/01/2024).
- Wide Cross-species RNA-Seq Comparison Reveals Convergent Molecular Mechanisms Involved in Nickel Hyperaccumulation across Dicotyledons - García de La Torre - 2021 - New Phytologist - Wiley Online Library* (2023). URL : <https://nph.onlinelibrary.wiley.com/doi/full/10.1111/nph.16775> (visité le 20/11/2023).