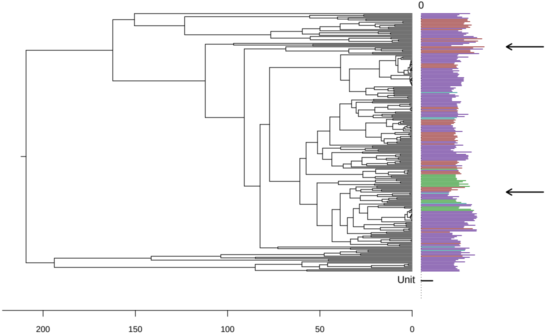
Segmentation sur arbre: quelques applications

C. Ambroise[1], P. Bastide[2,3,4], A. Bichat[1,5], *M. Mariadassou*[4], S. Robin[2]

[1]LaMME, UEVE, Université Paris-Saclay, Evry, France
[2]MIA 518, INRAE, AgroParisTech, Paris, France
[3]MAIAGE, INRAE, Université Paris-Saclay, Jouy-en-Josas, France
[4]IMAG, CNRS, Université de Montpellier, Montpellier, France
[5]Enterome, Paris, France

Journées MMB
12 Mars 2020

Introduction



Dermochelys Coriacea

Homopus Areolatus

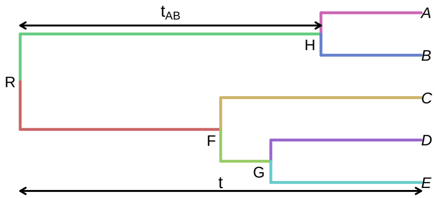
Turtles phylogenetic tree with habitats.
(Jaffe et al., 2011).

- How can we explain the diversity, while accounting for the phylogenetic correlations ?
- Modelling: a shifted stochastic process on the phylogeny.

Outline

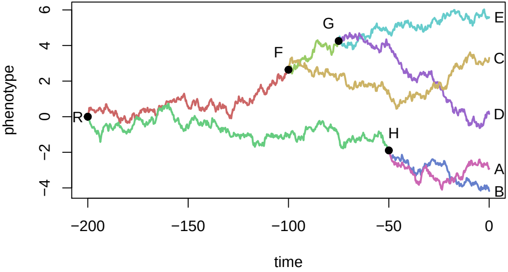
Plan

Stochastic Process on a Tree (Felsenstein, 1985)



The tree is known.

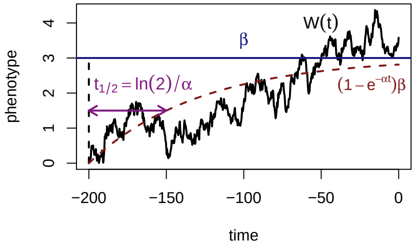
Only *tip* values are observed



Brownian Motion:

$$\mathbb{V}\text{ar}\,[A \mid R] = \sigma^2 t$$

$$\mathbb{C}\text{ov}\,[A; B \mid R] = \sigma^2 t_{AB}$$



$$dW(t) = \alpha[\beta(t) - W(t)]dt + \sigma dB(t)$$

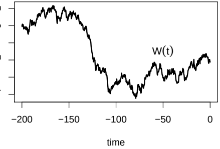
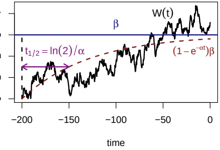
Deterministic part:

- $\beta(t)$: primary optimum, mechanistically defined.
- $\ln(2)/\alpha$: phylogenetic half live.

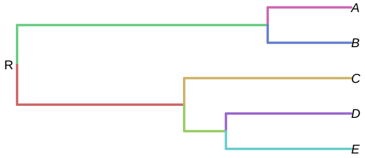
Stochastic part:

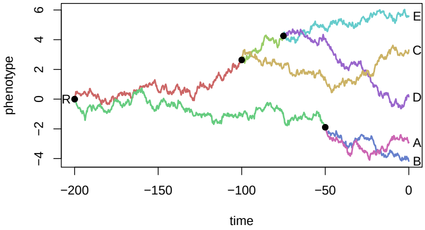
- $W(t)$: actual optimum (trait value).
- $\sigma dB(t)$ Brownian fluctuations.

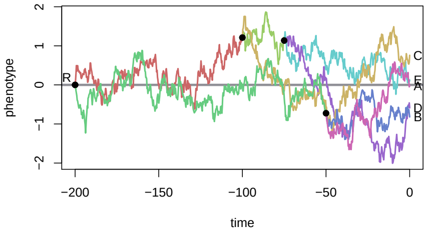
BM vs OU

	Equation	Stationary State	Variance
	$dW(t) = \sigma dB(t)$	None.	$\sigma_{ij} = \sigma^2 t_{ij}$
	$dW(t) = \sigma dB(t) + \alpha[\beta(t) - W(t)]dt$	$\begin{cases} \mu = \beta_0 \\ \gamma^2 = \dfrac{\sigma^2}{2\alpha} \end{cases}$	$\sigma_{ij} = \gamma^2 e^{-\alpha(t_i+t_j)} \times (e^{2\alpha t_{ij}} - 1)$

Shifts





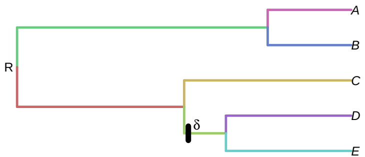


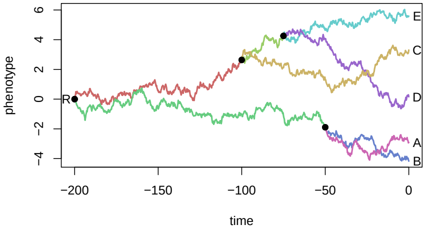
BM Shifts in the **mean**:

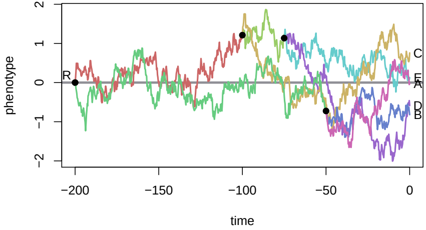
$$m_{\text{child}} = m_{\text{parent}} + \delta$$

OU Shifts in the **optimal value**:

$$\beta_{\text{child}} = \beta_{\text{parent}} + \delta$$







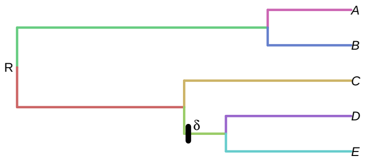
BM Shifts in the **mean**:

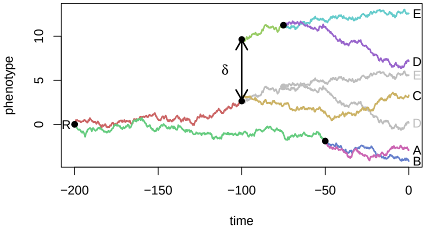
$$m_{\text{child}} = m_{\text{parent}} + \delta$$

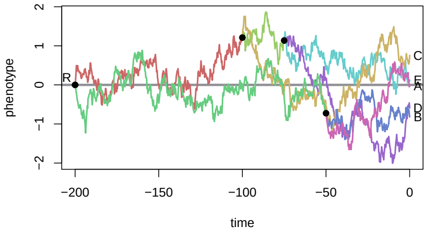
OU Shifts in the **optimal value**:

$$\beta_{\text{child}} = \beta_{\text{parent}} + \delta$$

Shifts





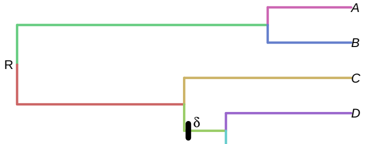


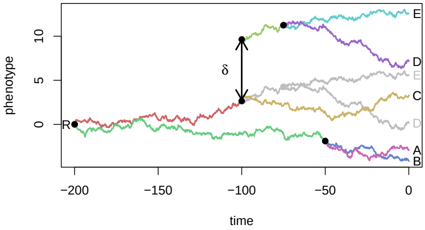
BM Shifts in the **mean**:

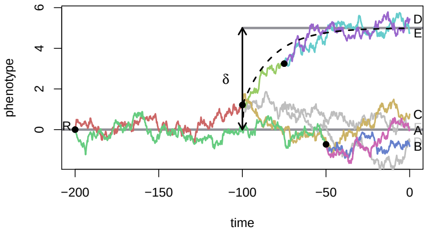
$$m_{\text{child}} = m_{\text{parent}} + \delta$$

OU Shifts in the **optimal value**:

$$\beta_{\text{child}} = \beta_{\text{parent}} + \delta$$







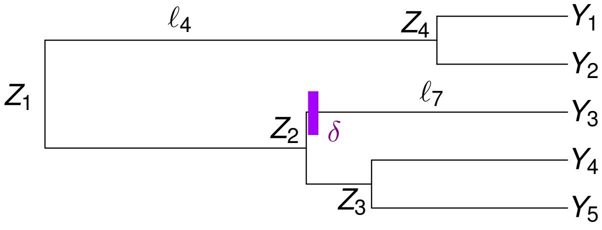
BM Shifts in the **mean**:

$$m_{\text{child}} = m_{\text{parent}} + \delta$$

OU Shifts in the **optimal value**:

$$\beta_{\text{child}} = \beta_{\text{parent}} + \delta$$

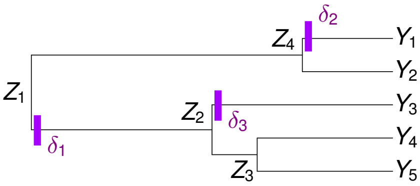
Incomplete Data Model



$$\text{BM} \quad Z_4 | Z_1 \sim \mathcal{N}\left(Z_1 \quad , \sigma^2 \ell_4\right)$$
$$Y_3 | Z_2 \sim \mathcal{N}\left(Z_2 + \delta, \sigma^2 \ell_7\right)$$

$$\text{OU} \quad Y_3 | Z_2 \sim \mathcal{N}\left(Z_2 e^{-\alpha \ell_7} + (1 - e^{-\alpha \ell_7})(\beta_{Z_2} + \delta), \frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha \ell_7})\right)$$

Linear Regression Model

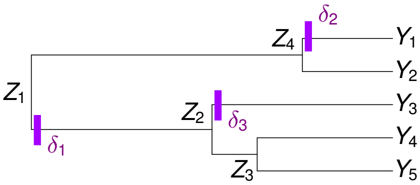


$$\Delta = \begin{pmatrix} \mu \\ \delta_1 \\ 0 \\ 0 \\ \delta_2 \\ 0 \\ \delta_3 \\ 0 \\ 0 \end{pmatrix} \qquad T\Delta = \begin{pmatrix} \mu + \delta_2 \\ \mu \\ \mu + \delta_1 + \delta_3 \\ \mu + \delta_1 \\ \mu + \delta_1 \end{pmatrix}$$

$$T = \begin{array}{c} \\ Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{array} \begin{array}{c} \begin{array}{ccccccccc} Z_1 & Z_2 & Z_3 & Z_4 & Y_1 & Y_2 & Y_3 & Y_4 & Y_5 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

$BM: \quad Y = T\Delta^{BM} + E^{BM}$

Linear Regression Model



$$\Delta = \begin{pmatrix} \lambda \\ \delta_1 \\ 0 \\ 0 \\ \delta_2 \\ 0 \\ \delta_3 \\ 0 \\ 0 \end{pmatrix} \quad TW(\alpha)\Delta = \begin{pmatrix} \lambda + w_5\delta_2 \\ \lambda \\ \lambda + w_2\delta_1 + w_7\delta_3 \\ \lambda + w_2\delta_1 \\ \lambda + w_2\delta_1 \end{pmatrix}$$

$$W(\alpha) = \mathrm{Diag}(1 - e^{-\alpha(h - t_{\mathrm{pa}(i)})}, 1 \leq i \leq m + n)$$

$$\lambda = \mu e^{-\alpha h} + \beta_0(1 - e^{-\alpha h})$$

$$BM: \quad Y = T\Delta^{BM} + E^{BM}$$

$$OU: \quad Y = TW(\alpha)\Delta^{OU} + E^{OU}$$

OU $\iff$ BM

Expectations

$$\mathbb{E}[Y \mid X_1 = \mu] = T \underbrace{W(\alpha)\Delta^{OU}}_{\Delta^{BM}}$$

Remark: $\mu^{BM} = \lambda^{OU} = \mu e^{-\alpha h} + \beta_0(1 - e^{-\alpha h})$

Variance

$$\mathbb{C}\text{ov}\left[Y_i; Y_j \mid X_1 = \mu\right] = \sigma^2 \times \underbrace{\frac{1}{2\alpha} e^{-2\alpha h}(e^{2\alpha t_{ij}-1})}_{t'_{ij}}$$

OU $\iff$ BM on a re-scaled tree with $t' = \frac{1}{2\alpha} e^{-2\alpha h}(e^{2\alpha t} - 1)$

OU $\iff$ BM

$$\text{OU} \iff \text{BM on a re-scaled tree with } t' = \frac{1}{2\alpha} e^{-2\alpha h}(e^{2\alpha t} - 1)$$

Remarks:

- This only works for an *ultrametric* tree.
- The laws of the internal nodes is changed.
- This is *not* the following standard time transformation:

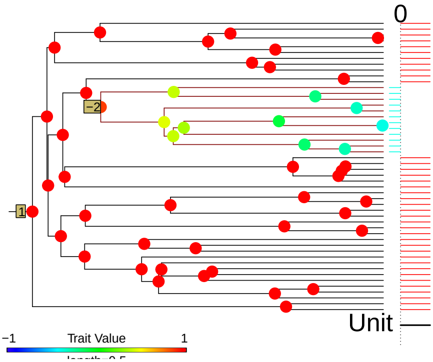
Lemma (Brownian Solution for the OU)

The stochastic process defined by:

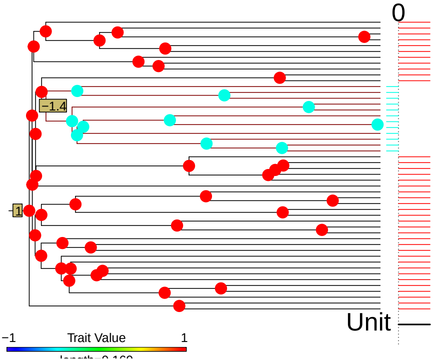
$$X_t = X_0 e^{-\alpha t} + \beta(1 - e^{-\alpha t}) + \frac{\sigma}{\sqrt{2\alpha}} e^{-\alpha t} B_{e^{2\alpha t} - 1}$$

is an OU, solution of the EDS $dX_t = \alpha(\beta - X_t) + \sigma dB_t$.

OU $\iff$ BM on a re-scaled tree with $t' = \frac{1}{2\alpha} e^{-2\alpha h}(e^{2\alpha t} - 1)$



OU: $\lambda = \beta_0 = \mu = 1$ and $t_{1/2} = 0.5$

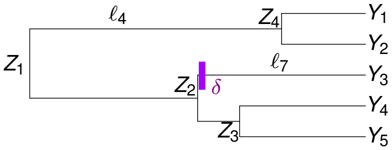


Equivalent BM on a re-scaled tree

Plan

1. Stochastic Processes on Trees
2. Statistical Inference
   - EM Algorithm
   - Model Selection
3. Application in Ecology
4. Application in metagenomics

EM Algorithm: number of shifts K fixed



$$Z_4 | Z_1 \sim \mathcal{N}\left(\lambda_1 Z_1 + (1-\lambda_1)\beta \quad , \frac{\sigma^2}{2\alpha}(1-\lambda_1^2)\right)$$

$$Y_3 | Z_2 \sim \mathcal{N}\left(\lambda_7 Z_2 + (1-\lambda_7)(\beta+\delta), \frac{\sigma^2}{2\alpha}(1-\lambda_7^2)\right)$$

$$\lambda_i = e^{-\alpha \ell_i}$$

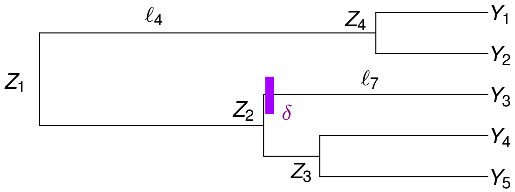
$$p_\theta(Z, Y) = p_\theta(Z_1) \prod_{1<j\leq m} p_\theta(Z_j | Z_{\text{parent}(j)}) \prod_{1\leq i\leq n} p_\theta(Y_i | Z_{\text{parent}(i)})$$

EM Algorithm $\log p_\theta(Y) = \mathbb{E}_\theta[\log p_\theta(Z, Y) \mid Y] - \mathbb{E}_\theta[\log p_\theta(Z) \mid Y]$

E step Given θ^h, compute $p_{\theta^h}(Z \mid Y)$

M step $\theta^{h+1} = \text{argmax}_\theta \, \mathbb{E}_{\theta^h}[\log p_\theta(Z, Y) \mid Y]$

E step



Compute the following quantities:

$$\mathbb{E}^{(h)}[Z_j \mid Y], \ \mathbb{V}\mathrm{ar}^{(h)}\left[Z_j \mid Y\right], \ \mathbb{C}\mathrm{ov}^{(h)}\left[Z_j, Z_{\mathrm{parent}(j)} \mid Y\right]$$

- Using Gaussian properties. Need to invert matrices: complexity in $O(n^3)$.
- Using Gaussian properties **and** the tree structure: "Upward-Downward" algorithm. Complexity in $O(n)$.

M Step

Maximize:

$$\mathbb{E}\left[\log p_\theta(X) \mid Y\right] = -\sum_{j=2}^{m+n} C_j(\alpha, \text{shifts}) + \mathcal{F}^{(h)}\left(\mu, \gamma^2, \sigma^2, \alpha\right)$$

- μ, γ^2, σ^2: simple maximization
- Discrete location of K shifts
  - $\mapsto$ Exact and fast for the BM
  - $\mapsto$ Hill-climbing heuristics for the OU
- α: numerical maximization and/or on a grid
  - $\mapsto$ Generalized EM

Starting point and choice of K

Starting point

Shifts: Fast estimate based on Lasso regression (see next section).
Selection strength α: Initialization using couples of tips and robust estimate of α.

Starting point and choice of K

Starting point

Shifts: Fast estimate based on Lasso regression (see next section).
Selection strength α: Initialization using couples of tips and robust estimate of α.

Choosing K

Assumption α fixed

$$Y = TW(\alpha)\Delta + \gamma E \qquad , \qquad E \sim \mathcal{N}(0, V(\alpha))$$

Models

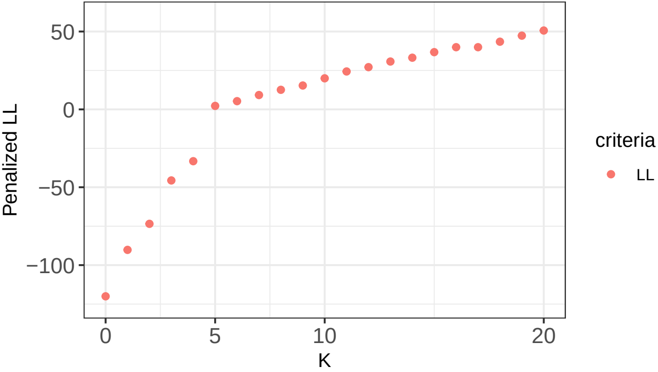
$\eta \in \bigcup_{K=0}^{p-1} \mathcal{S}_K^{PI}$: Identifiable parcimonious allocations of shifts

EM Estimators

$$\hat{Y}_K = \underset{\eta \in \mathcal{S}_K^{PI}}{\operatorname{argmin}} \left\| Y - \hat{Y}_\eta \right\|_V^2$$

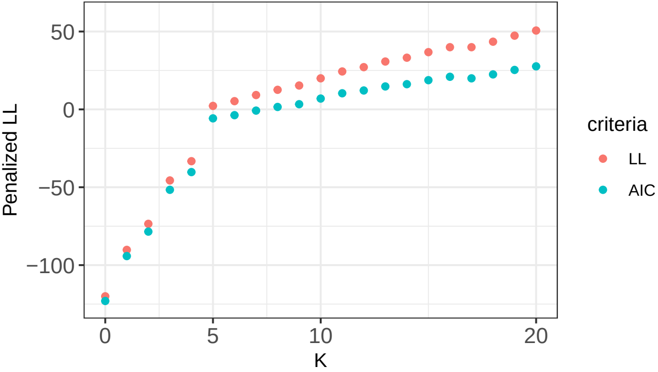
Model Selection: Penalized Likelihood

Idea $$\hat{K} = \underset{0 \leq K \leq p-1}{\text{argmax}} \left\{ \frac{n}{2} \log \left(\frac{1}{n} \left\| Y - \hat{Y}_K \right\|_V^2 \right) - \frac{1}{2} \text{pen}'(K) \right\}$$



Model Selection: Penalized Likelihood

Idea $$\hat{K} = \underset{0 \leq K \leq p-1}{\text{argmax}} \left\{ \frac{n}{2} \log \left(\frac{1}{n} \left\| Y - \hat{Y}_K \right\|_V^2 \right) - \frac{1}{2} \text{pen}'(K) \right\}$$

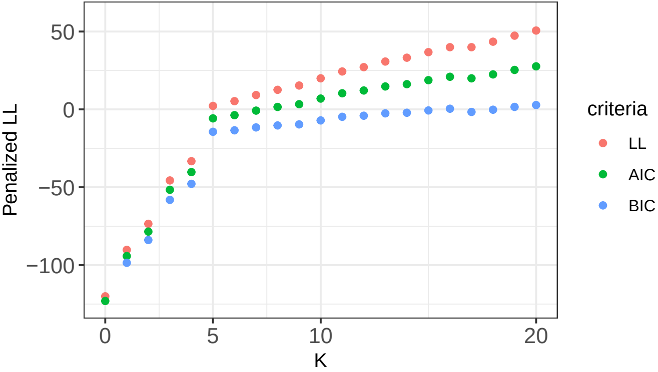


Penalties:

AIC $K + 3$

Model Selection: Penalized Likelihood

Idea $\hat{K} = \underset{0 \leq K \leq p-1}{\operatorname{argmax}} \left\{ \frac{n}{2} \log \left(\frac{1}{n} \left\| Y - \hat{Y}_K \right\|_V^2 \right) - \frac{1}{2} \operatorname{pen}'(K) \right\}$



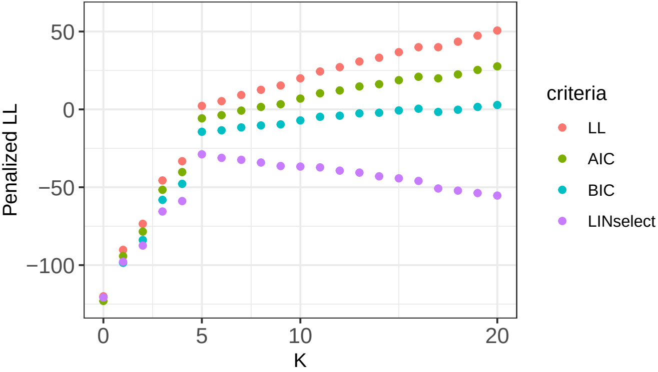
Penalties:

AIC $K + 3$

BIC $\frac{1}{2}(K+3)\log(n)$

Model Selection: Penalized Likelihood

Idea $\hat{K} = \underset{0 \leq K \leq p-1}{\operatorname{argmax}} \left\{ \frac{n}{2} \log\left(\frac{1}{n} \left\| Y - \hat{Y}_K \right\|_V^2 \right) - \frac{1}{2} \operatorname{pen}'(K) \right\}$



Penalties:

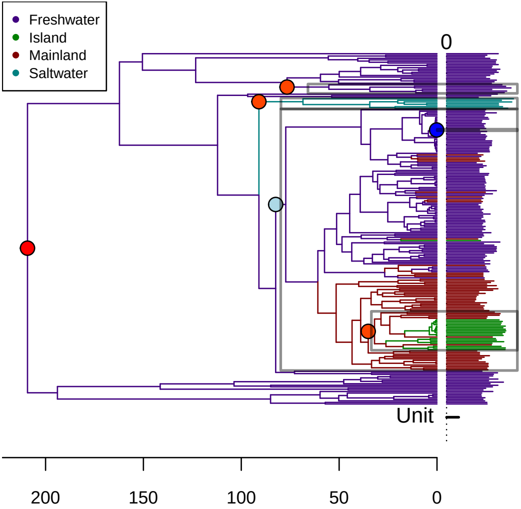
AIC $K + 3$

BIC $\frac{1}{2}(K+3)\log(n)$

LINselect $\operatorname{pen}(n, K, |\mathcal{S}_K^{PI}|)$

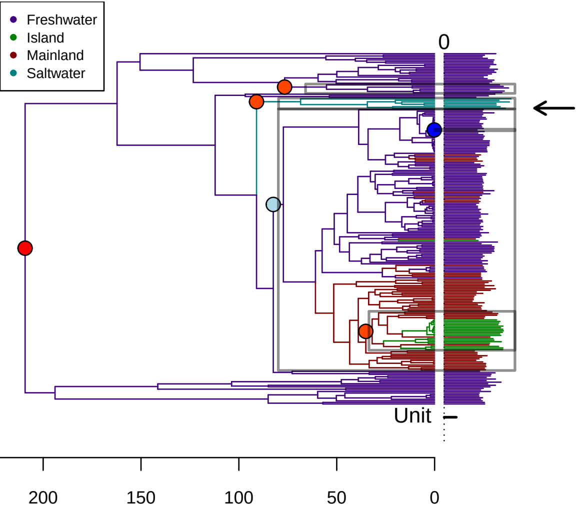
Plan

Turtles Dataset



Colors: habitats.
Boxes: selected EM regimes.

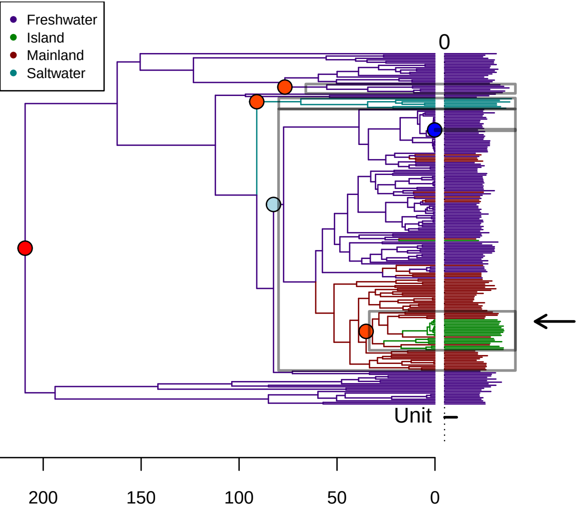
Turtles Dataset



Colors: habitats.
Boxes: selected EM regimes.

Chelonia mydas

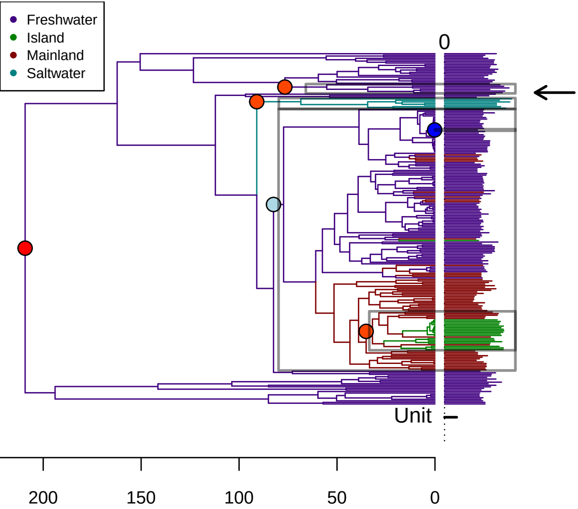
Turtles Dataset



Colors: habitats.
Boxes: selected EM regimes.

Geochelone nigra abingdo

Turtles Dataset



Chitra indica

Colors: habitats.
Boxes: selected EM regimes.

References

Univariate framework

P. Bastide, M. Mariadassou, S. Robin (2016), Detection of adaptive shifts on phylogenies by using shifted stochastic processes on a tree. *JRSS-B*. doi:10.1111/rssb.12206

Extension to multivariate framework

P. Bastide, C. Ané, S. Robin, M. Mariadassou (2018), Inference of Adaptive Shifts for Multivariate Correlated Traits. *Syst. Biol.*. doi:10.1093/sysbio/syy005

Package `PhylogeneticEM`: available on GitHub, on the CRAN.

Plan

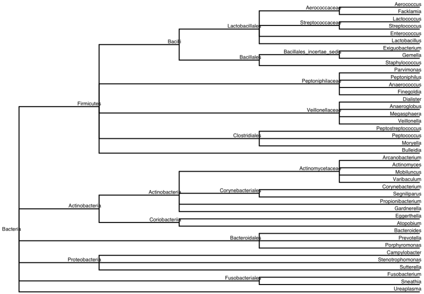
Metabarcoding data (Ravel et al., 2011)

A species × sample count table

	Taxa	A1	A2	A3	B1	B2	B3
1	Lactobacillus	2318	1388	1361	2256	88	1770
2	Prevotella	0	1	1	0	525	7
3	Megasphaera	0	1	0	0	402	0
4	Sneathia	0	0	0	0	302	0
5	Atopobium	0	1	0	0	84	0
6	Streptococcus	0	0	3	0	0	0
7	Dialister	0	1	0	0	152	4
8	Anaerococcus	0	1	3	2	0	9
9	Peptoniphilus	0	1	0	0	7	2
10	Eggerthella	0	0	0	0	2	0

Taxonomic / phylogenetic tree

	Phylum	Class	Order	Family	Genus
1	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Actinobaculum
2	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Actinomyces
3	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Arcanobacterium
4	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Mobiluncus
5	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Varibaculum
6	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium
7	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Gardnerella

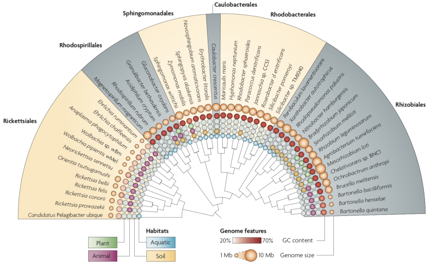


Differential abundance analysis

- For each taxa i (in $\{1, \ldots, n\}$), test
  - H_{0i}: Abundances are equal in groups A and B
  - H_{1i}: Abundances are not equal in groups A and B
- Hundred of univariate tests and p-values
- Need for a multiple testing correction procedure

Differential abundance analysis

- For each taxa i (in $\{1, \dots, n\}$), test
  - H_{0i}: Abundances are equal in groups A and B
  - H_{1i}: Abundances are not equal in groups A and B
- Hundred of univariate tests and p-values
- Need for a multiple testing correction procedure



- Taxa / group associations may show a **phylogenetic signal**
- Similar taxa $\Rightarrow$ similar levels of association
- Can we leverage the tree when correcting the tests?

Mathematical Model

Standard assumptions on *p*-values

- Under $H_{0i}, p_i \sim \mathcal{U}(0, 1)$
- Under $H_{1i}, p_i \preccurlyeq \mathcal{U}(0, 1)$

Mathematical Model

Standard assumptions on *p*-values

- Under $H_{0i}, p_i \sim \mathcal{U}(0,1)$
- Under $H_{1i}, p_i \preccurlyeq \mathcal{U}(0,1)$

Standard assumptions on *z*-scores

- Under H_{0i}, $z_i = \Phi^{-1}(p_i) \sim \mathcal{N}(0,1)$
- Under H_{1i}, $z_i = \Phi^{-1}(p_i) \sim \mathcal{N}(m_i,1)$ with $m_i < 0$

Mathematical Model

Standard assumptions on *p*-values

- Under $H_{0i}, p_i \sim \mathcal{U}(0,1)$
- Under $H_{1i}, p_i \preccurlyeq \mathcal{U}(0,1)$

Standard assumptions on *z*-scores

- Under H_{0i}, $z_i = \Phi^{-1}(p_i) \sim \mathcal{N}(0,1)$
- Under H_{1i}, $z_i = \Phi^{-1}(p_i) \sim \mathcal{N}(m_i,1)$ with $m_i < 0$

Tractable assumptions on *z*-scores vector

- $Z = (z_1, \dots, z_n) \sim \mathcal{N}(M, V(\alpha))$ where
  - $M = (m_1, \dots, m_n) \in \mathbb{R}^n_-$
  - $V(\alpha)$ is the variance matrix of an OU on a tree.

Mathematical Model

Standard assumptions on p-values

- Under $H_{0i}, p_i \sim \mathcal{U}(0,1)$
- Under $H_{1i}, p_i \preccurlyeq \mathcal{U}(0,1)$

Standard assumptions on z-scores

- Under H_{0i}, $z_i = \Phi^{-1}(p_i) \sim \mathcal{N}(0,1)$
- Under H_{1i}, $z_i = \Phi^{-1}(p_i) \sim \mathcal{N}(m_i,1)$ with $m_i < 0$

Tractable assumptions on z-scores vector

- $Z = (z_1, \ldots, z_n) \sim \mathcal{N}(M, V(\alpha))$ where
  - $M = (m_1, \ldots, m_n) \in \mathbb{R}_-^n$
  - $V(\alpha)$ is the variance matrix of an OU on a tree.

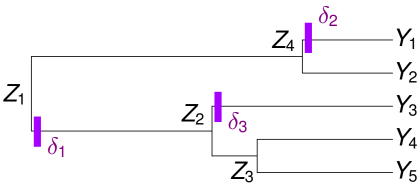
Assume that z-scores evolve as an OU on the tree with a sign constraint on the mean.

Linear regression model

Tree-structure enforced by decomposition $M = TW(\alpha)\Delta$.

Linear regression model

Tree-structure enforced by decomposition $M = TW(\alpha)\Delta$.



$$\Delta = \begin{pmatrix} 0 \\ \delta_1 \\ 0 \\ 0 \\ \delta_2 \\ 0 \\ \delta_3 \\ 0 \\ 0 \end{pmatrix} \qquad TW(\alpha)\Delta = \begin{pmatrix} w_5\delta_2 \\ 0 \\ w_2\delta_1 + w_7\delta_3 \\ w_2\delta_1 \\ w_2\delta_1 \end{pmatrix}$$

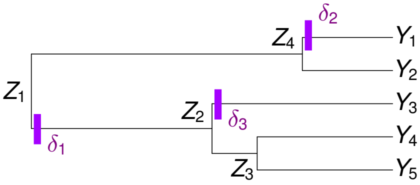
$$T = \begin{array}{c} \\ Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{array} \begin{array}{c} \begin{array}{ccccccccc} Z_1 & Z_2 & Z_3 & Z_4 & Y_1 & Y_2 & Y_3 & Y_4 & Y_5 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

OU : $\quad Z = TW(\alpha)\Delta + E$

$$W(\alpha) = \mathrm{Diag}(1 - e^{-\alpha(h - t_{\mathrm{pa}(i)})}, 1 \leq i \leq m + n)$$

Linear regression model

Tree-structure enforced by decomposition $M = TW(\alpha)\Delta$.



$$\Delta = \begin{pmatrix} 0 \\ \delta_1 \\ 0 \\ 0 \\ \delta_2 \\ 0 \\ \delta_3 \\ 0 \\ 0 \end{pmatrix} \qquad TW(\alpha)\Delta = \begin{pmatrix} w_5\delta_2 \\ 0 \\ w_2\delta_1 + w_7\delta_3 \\ w_2\delta_1 \\ w_2\delta_1 \end{pmatrix}$$

$$T = \begin{array}{c} \\ Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{array} \begin{array}{c} \begin{array}{ccccccccc} Z_1 & Z_2 & Z_3 & Z_4 & Y_1 & Y_2 & Y_3 & Y_4 & Y_5 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

$$OU: \quad Z = TW(\alpha)\Delta + E$$

$$W(\alpha) = \mathrm{Diag}(1 - e^{-\alpha(h - t_{\mathrm{pa}(i)})}, 1 \leq i \leq m + n)$$

Goal: Find $\{i : m_i < 0\}$

Estimating M

The MLE of Δ (and in turn M) is solution to

$$\underset{\Delta \text{ s.t. } TW(\alpha)\Delta \leq 0}{\text{argmax}} \|Z - TW(\alpha)\Delta\|^2_{2,V(\alpha)^{-1}}$$

Equivalent to[1]:

$$\underset{\Delta \text{ s.t. } C\Delta \leq 0}{\text{argmax}} \|Y - X\Delta\|^2_2$$

[1]with C, Y and X some simple transforms of Z and $TW(\alpha)$, $V(\alpha)$

[2]Using a variant of the LASSO shooting algorithm

Estimating M

The MLE of Δ (and in turn M) is solution to

$$\underset{\Delta \text{ s.t. } TW(\alpha)\Delta \leq 0}{\text{argmax}} \|Z - TW(\alpha)\Delta\|^2_{2, V(\alpha)^{-1}}$$

Equivalent to[1]:

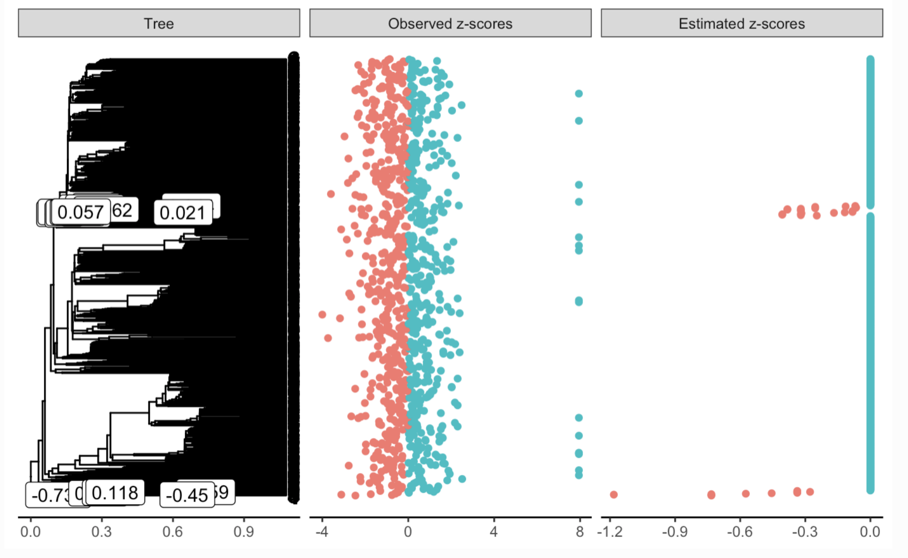
$$\underset{\Delta \text{ s.t. } C\Delta \leq 0}{\text{argmax}} \|Y - X\Delta\|^2_2$$

Add a ℓ_1-penalty to sparsify the solution and solve[2]

$$\hat{\Delta} = \underset{\Delta \text{ s.t. } C\Delta \leq 0}{\text{argmax}} \|Y - X\Delta\|^2_2 + \lambda\|\Delta\|_1 \quad (1)$$

using penalized likelihood for selection of α and λ

[1] with C, Y and X some simple transforms of Z and $TW(\alpha)$, $V(\alpha)$
[2] Using a variant of the LASSO shooting algorithm


Tree
Observed z-scores
Estimated z-scores
0.057
62
0.021
-0.73
0.118
-0.45
0.0
0.3
0.6
0.9
-4
0
4
8
-1.2
-0.9
-0.6
-0.3
0.0

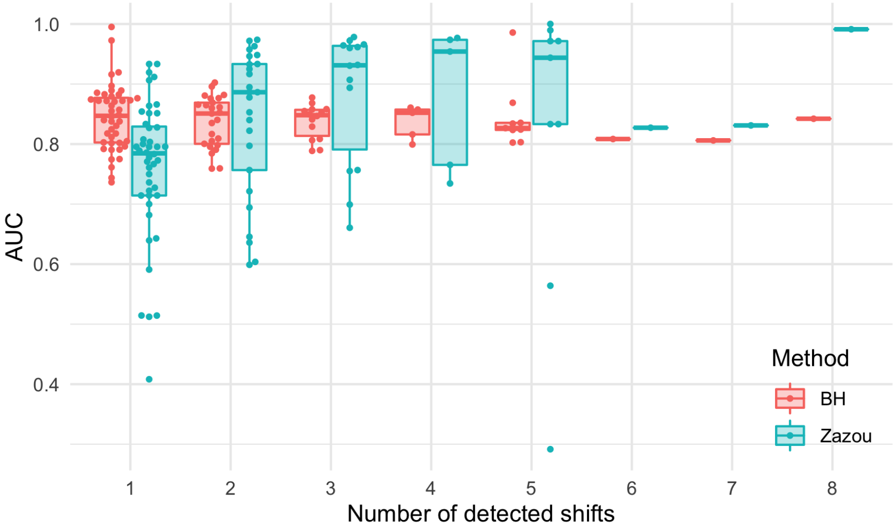
Testing null components of M

Decision rule

- $\hat{m}_i \neq 0 \Rightarrow$ reject H_{0i}
- $\hat{m}_i = 0 \Rightarrow$ do not reject H_{0i}

Comparison with non-hierarchical procedures

AUC on simulated data (higher is better)



Perspectives

- `zazou` R package: under active development on GitHub.

Perspectives

- `zazou` R package: under active development on GitHub.

Good performance when the number of shifts is not too small

- Selection model adapted to *tests* rather than *prediction*

Perspectives

- `zazou` R package: under active development on GitHub.

Good performance when the number of shifts is not too small

- Selection model adapted to *tests* rather than *prediction*

$\hat{M} = TW(\alpha)\hat{\Delta}$ is bias

- Unbias $\hat{M}$ (using desparsified lasso)
- Build consistent confidence intervals for m_i

Bibliography

J. Felsenstein. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39(4):783–791, July 1985. doi: 10.2307/2408678. URL http://links.jstor.org/sici?sici=0014-3820(198507)39:4%3C783:CLOPAA%3E2.0.CO;2-L.

T. F. Hansen. Stabilizing selection and the comparative analysis of adaptation. *Evolution*, 51(5):1341–1351, Oct. 1997. URL http://www.jstor.org/stable/2411186.

A. L. Jaffe, G. J. Slater, and M. E. Alfaro. The evolution of island gigantism and body size variation in tortoises and turtles. *Biol Lett*, 7(4):558–561, Aug 2011. doi: 10.1098/rsbl.2010.1084. URL http://dx.doi.org/10.1098/rsbl.2010.1084.

J. Ravel, P. Gajer, Z. Abdo, G. M. Schneider, S. S. K. Koenig, S. L. McCulle, S. Karlebach, R. Gorle, J. Russell, C. O. Tacket, and et al. Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences*, 108(Suppl. 1): 4680–4687, March 2011. ISSN 1091-6490. doi: 10.1073/pnas.1002611107. URL http://dx.doi.org/10.1073/pnas.1002611107.