Shifted stochastic processes evolving on trees: application to models of adaptive evolution on phylogenies

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

[1] UMR 518 MIA - AgroParisTech/INRA, Paris
[2] UR 1404 MaIAGE - INRA, Jouy en Josas

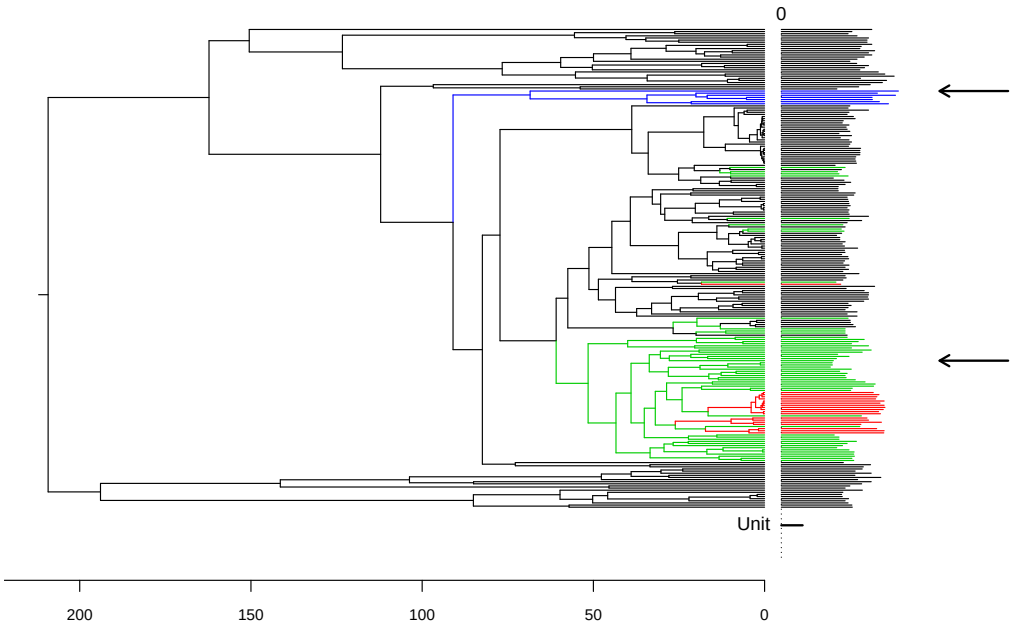
2 June 2015

AgroParisTech

INRA SCIENCE & IMPACT

Mathématiques et Informatique Appliquées MaIAGE du Génome à l'Environnement

Introduction



Dermochelys Coriacea

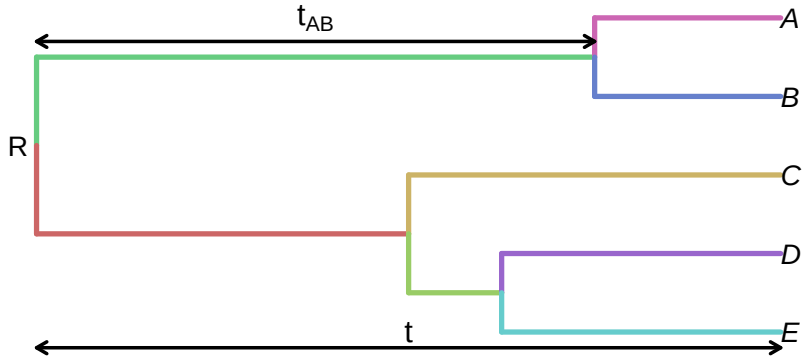
Homopus Areolatus

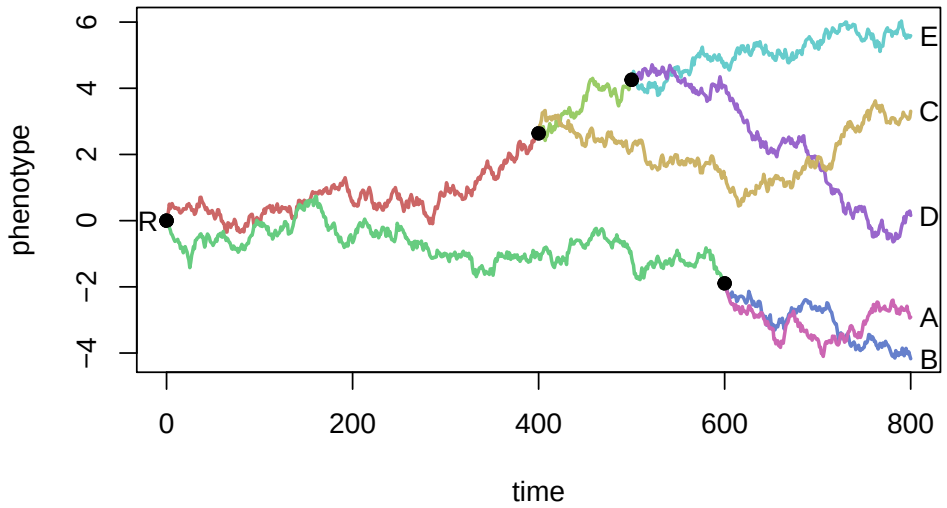
Chelonian phylogenetic tree with habitats. (Jaffe et al., 2011).

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

Stochastic Process on a Tree

(Felsenstein, 1985)





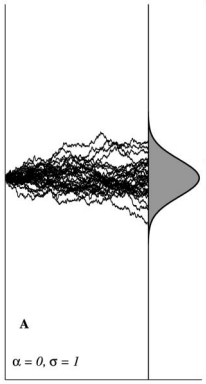
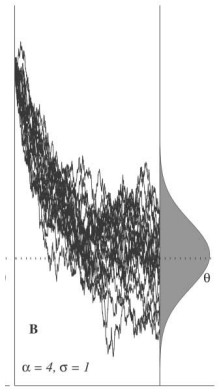
Brownian Motion:

$$\mathbb{V}\mathrm{ar}\left[A \mid R\right] = \sigma^2 t$$

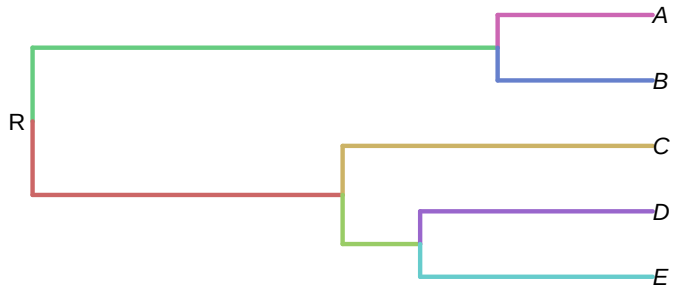
$$\mathbb{C}\mathrm{ov}\left[A; B \mid R\right] = \sigma^2 t_{AB}$$

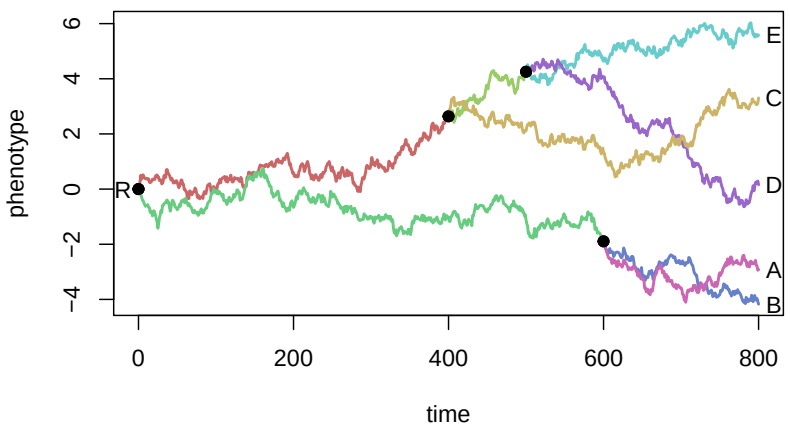
BM vs OU

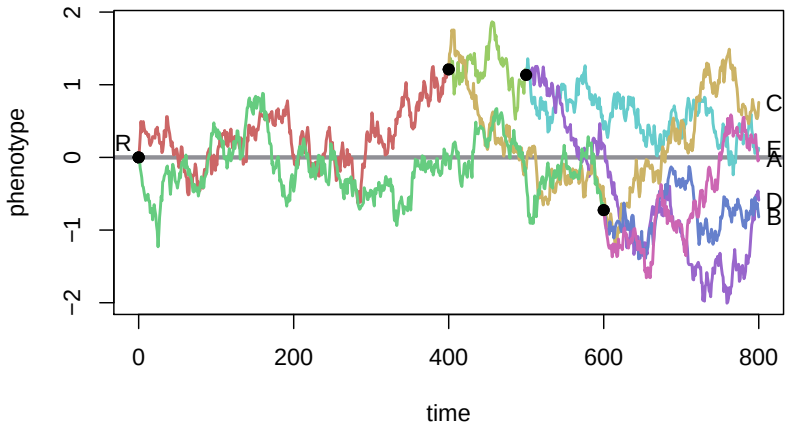
(Hansen, 1997; Butler and King, 2004)

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

Shifts







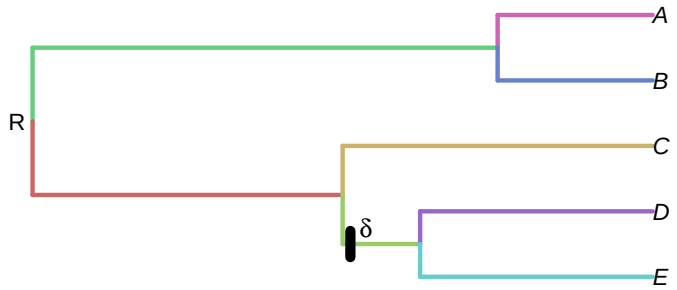
BM Shifts in the **mean**:

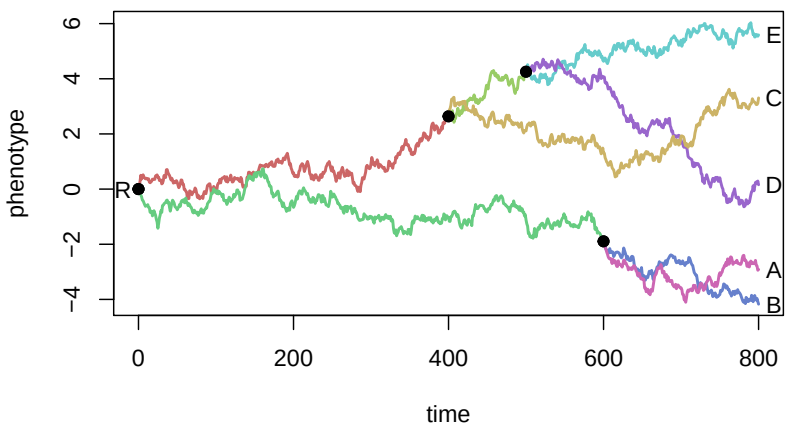
$$m_j = m_{\mathrm{pa}(j)} + \sum_k \mathbb{I}\{\tau_k = b_j\}\delta_k$$

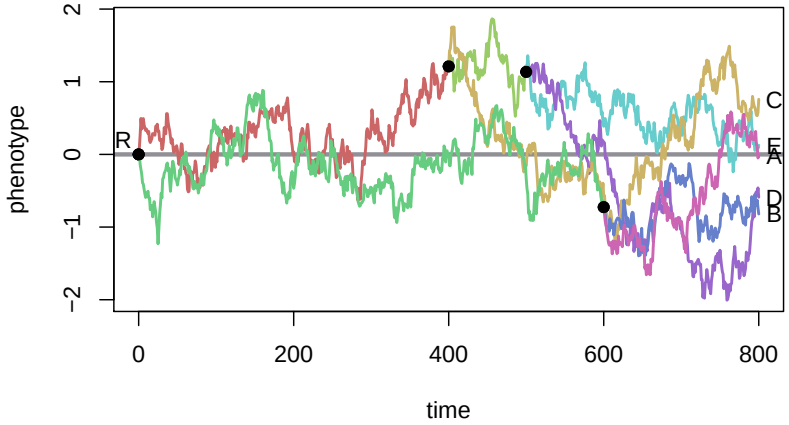
OU Shifts in the **optimal value**:

$$\beta^j = \beta^{\mathrm{pa}(j)} + \sum_k \mathbb{I}\{\tau_k = b_j\}\delta_k$$

Shifts







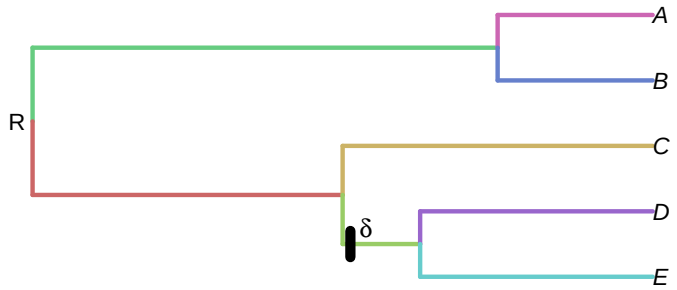
BM Shifts in the **mean**:

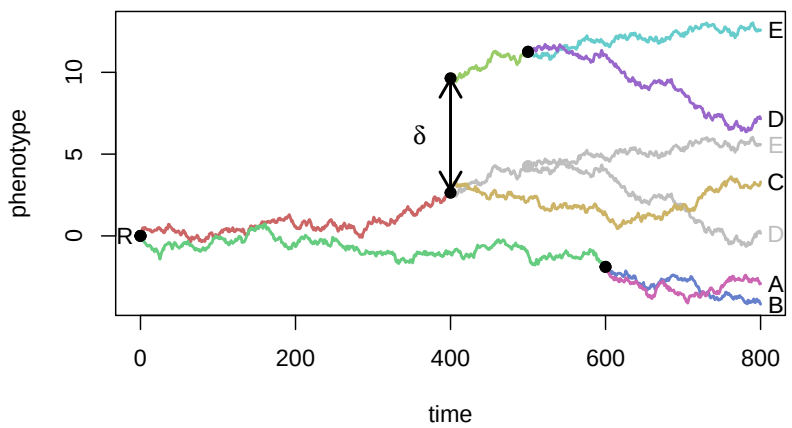
$$m_j = m_{\mathrm{pa}(j)} + \sum_k \mathbb{I}\{\tau_k = b_j\}\delta_k$$

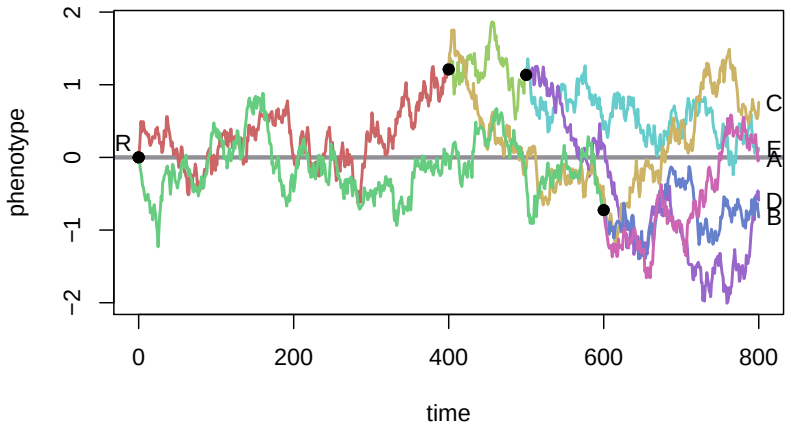
OU Shifts in the **optimal value**:

$$\beta^j = \beta^{\mathrm{pa}(j)} + \sum_k \mathbb{I}\{\tau_k = b_j\}\delta_k$$

Shifts







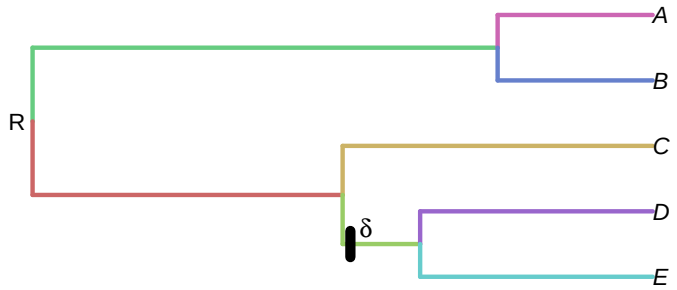
BM Shifts in the **mean**:

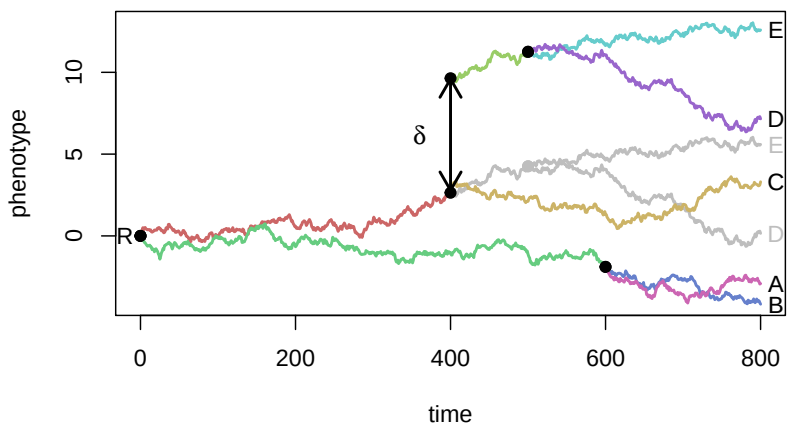
$$m_j = m_{\mathrm{pa}(j)} + \sum_k \mathbb{I}\{\tau_k = b_j\}\delta_k$$

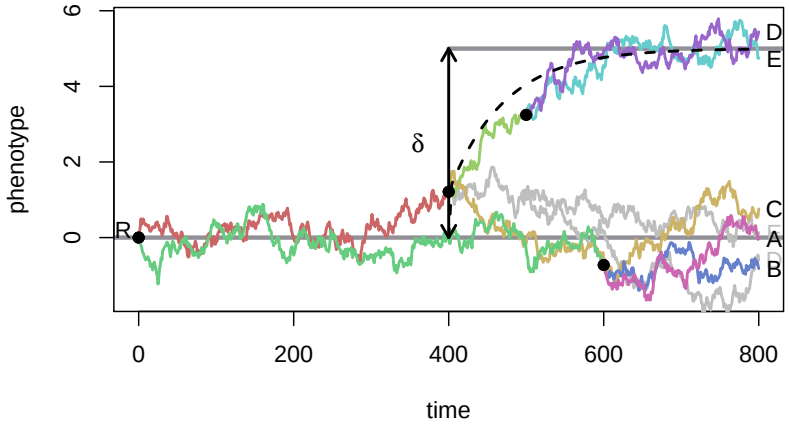
OU Shifts in the **optimal value**:

$$\beta^j = \beta^{\mathrm{pa}(j)} + \sum_k \mathbb{I}\{\tau_k = b_j\}\delta_k$$

Shifts







BM Shifts in the **mean**:

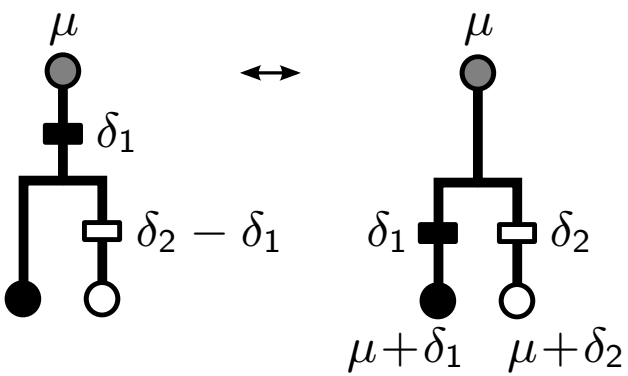
$$m_j = m_{\mathrm{pa}(j)} + \sum_k \mathbb{I}\{\tau_k = b_j\}\delta_k$$

OU Shifts in the **optimal value**:

$$\beta^j = \beta^{\mathrm{pa}(j)} + \sum_k \mathbb{I}\{\tau_k = b_j\}\delta_k$$

Equivalencies

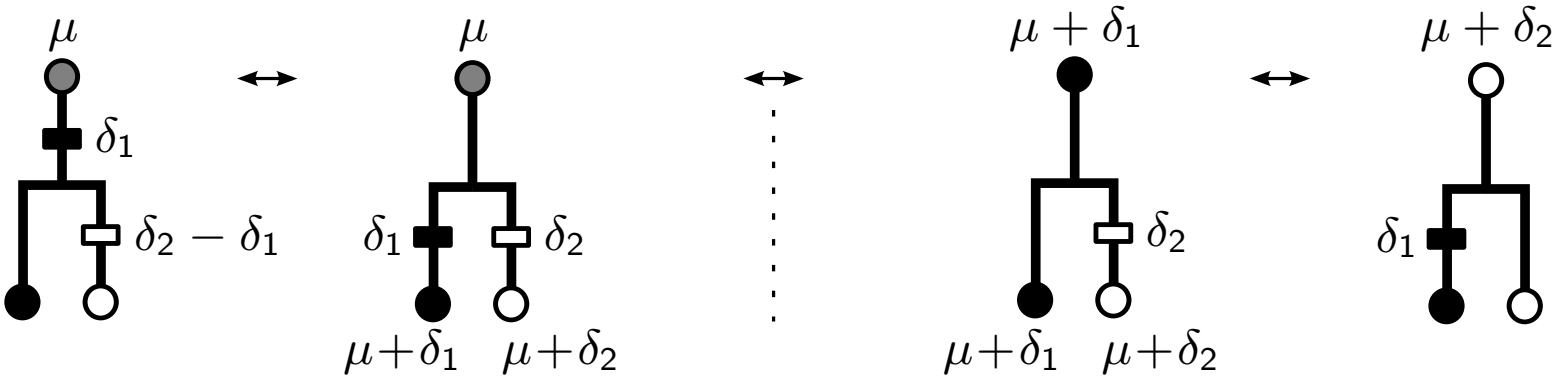
- K fixed, several equivalent solutions.



- Problem of over-parametrization: parsimonious configurations.

Equivalencies

- K fixed, several equivalent solutions.



- Problem of over-parametrization: parsimonious configurations.

Parsimonious Solution : Definition

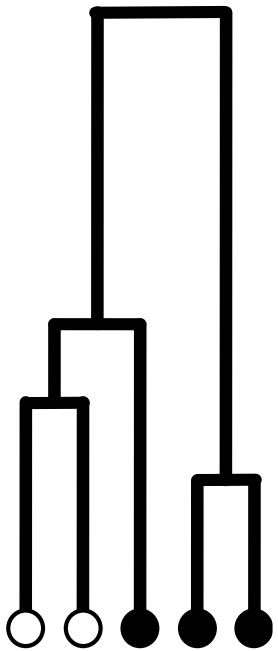
Definition

A coloring of the tips being given, a *parsimonious* allocation of the shifts is such that it has a minimum number of shifts.

Parsimonious Solution : Definition

Definition

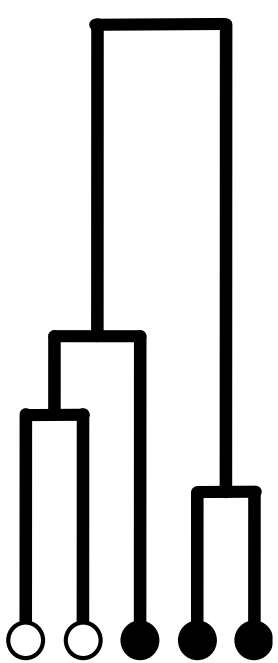
A coloring of the tips being given, a *parsimonious* allocation of the shifts is such that it has a minimum number of shifts.

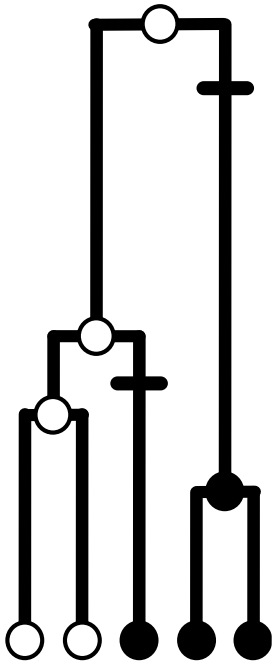


Parsimonious Solution : Definition

Definition

A coloring of the tips being given, a *parsimonious* allocation of the shifts is such that it has a minimum number of shifts.

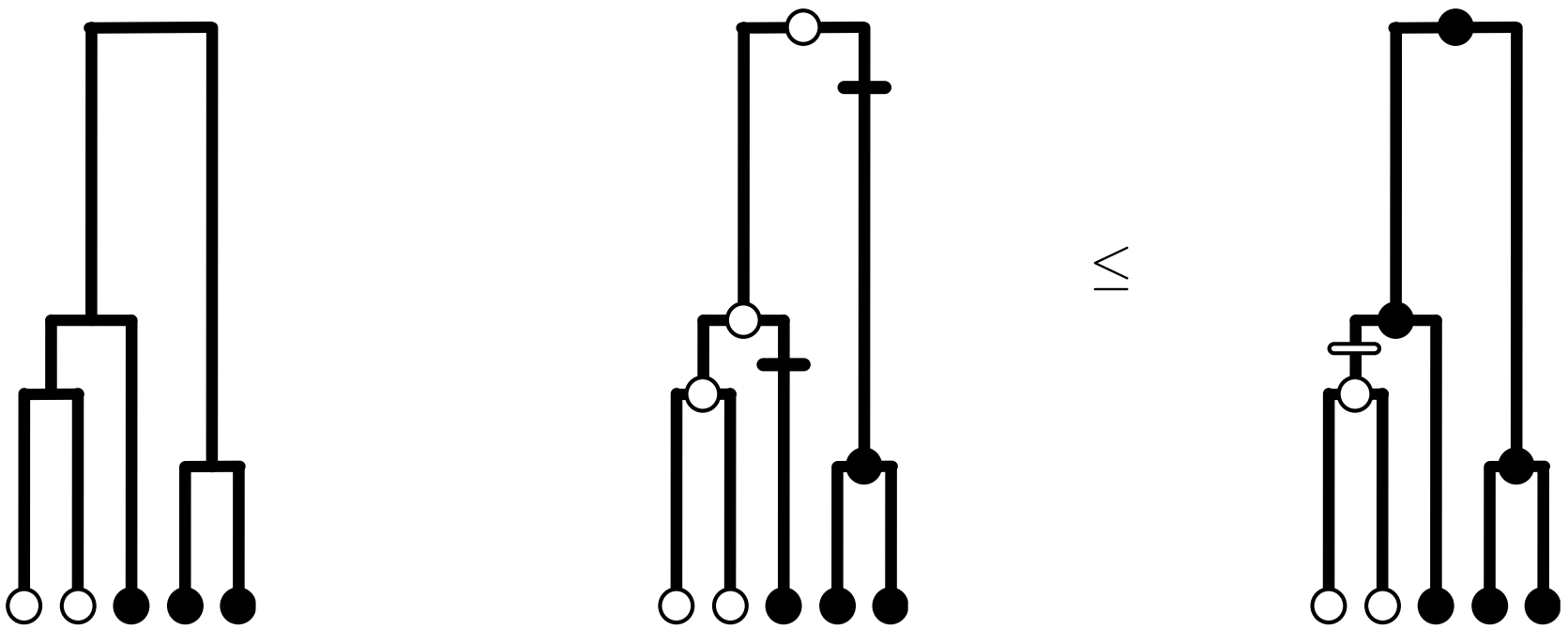




Parsimonious Solution : Definition

Definition

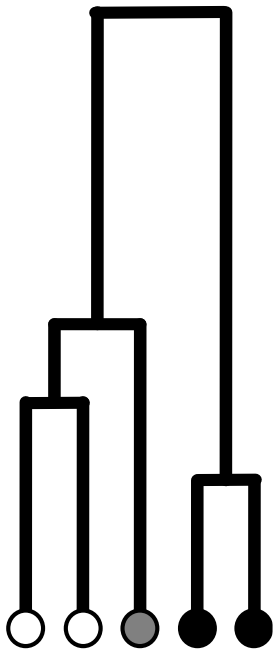
A coloring of the tips being given, a *parsimonious* allocation of the shifts is such that it has a minimum number of shifts.



Parsimonious Solution : Definition

Definition

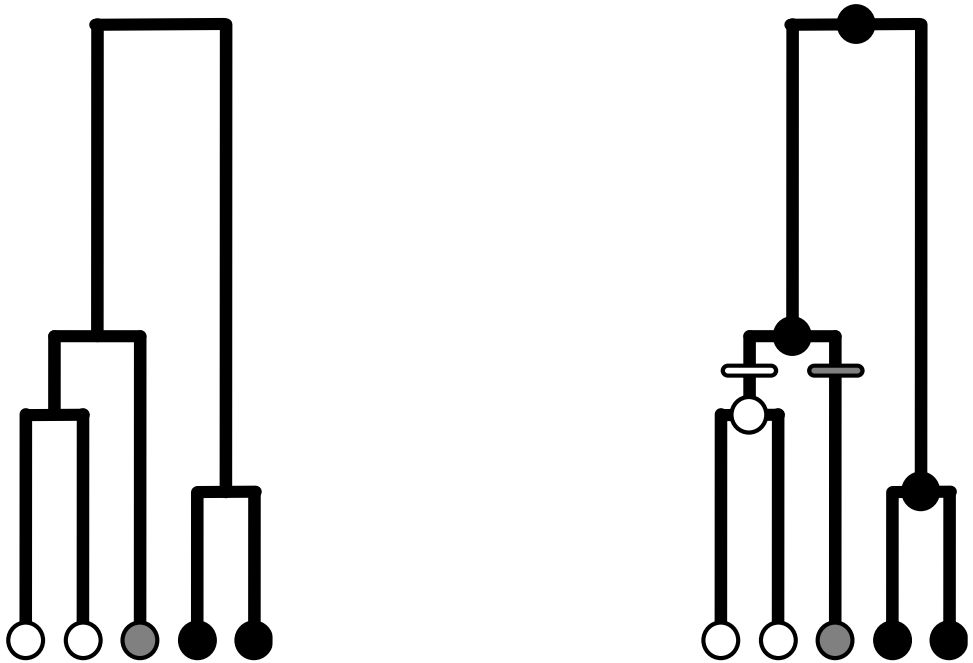
A coloring of the tips being given, a *parsimonious* allocation of the shifts is such that it has a minimum number of shifts.



Parsimonious Solution : Definition

Definition

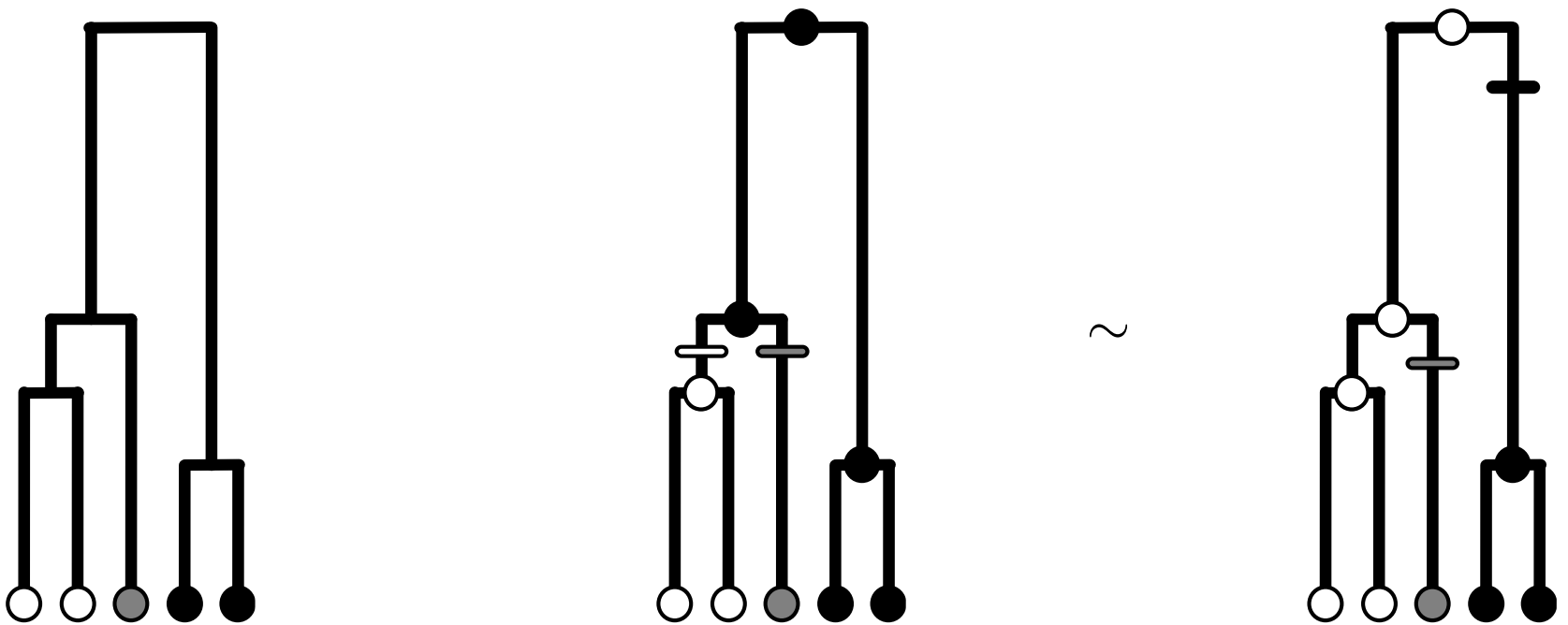
A coloring of the tips being given, a *parsimonious* allocation of the shifts is such that it has a minimum number of shifts.



Parsimonious Solution : Definition

Definition

A coloring of the tips being given, a *parsimonious* allocation of the shifts is such that it has a minimum number of shifts.



Equivalent Parsimonious Allocations

Definition

Two allocations are said to be *equivalent* (noted $\sim$) if they are both parsimonious and give the same colors at the tips.

Find one solution Several existing Dynamic Programming algorithms (see Felsenstein, 2004).

Enumerate all solutions New recursive algorithm, adapted from previous ones (and implemented in R).

Number of Models with K Shifts

Hypothesis "No Homoplasy": 1 shift = 1 new color.

$$\text{“}K \text{ shifts} \iff K+1 \text{ colors”}$$

Bijection

$$\mathcal{S}_K^{PI} = \mathcal{S}_K^P / \sim; \quad \mathcal{S}_K^P = \{\text{Parsimonious allocations of } K \text{ shifts}\}$$

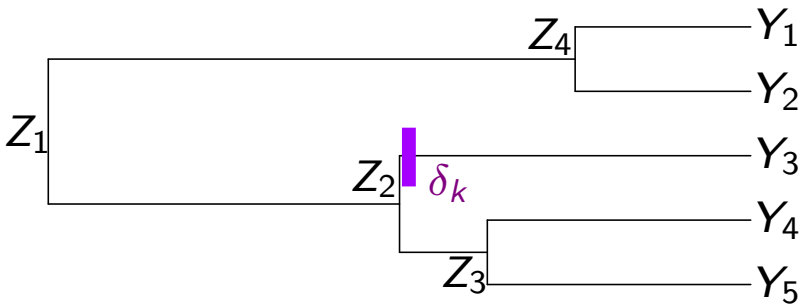
$$\mathcal{S}_K^{PI} \simeq \{\text{Tree compatible coloring of tips in } K+1 \text{ colors}\}$$

Problem Size of $\mathcal{S}_K^{PI}$?

Proposition

- $\left|\mathcal{S}_K^{PI}\right| \leq \binom{m+n-1}{K}$
- $\left|\mathcal{S}_K^{PI}\right|$ *depends on the topology of the tree. It can be computed with a recursive algorithm.*
- *For a binary tree:* $\left|\mathcal{S}_K^{PI}\right| = \binom{2n-2-K}{K}$.

Incomplete Data Model : EM



$$X_j | X_{\mathrm{pa}(j)} \sim \mathcal{N}\left(q_j X_{\mathrm{pa}(j)} + r_j + s_j \sum_k \mathbb{I}\{\tau_k = b_j\}\delta_k, \sigma_j^2 \right)$$

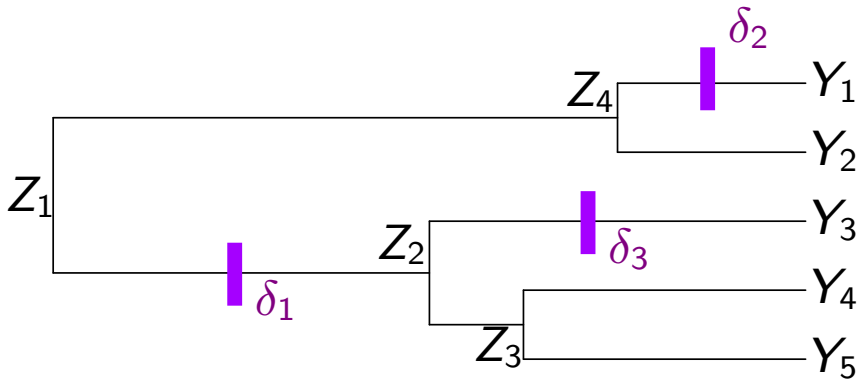
EM Algorithm Maximize $\mathbb{E}_\theta[\log p_\theta(Z, Y) \mid Y]$.

E step "Upward-Downward" Algorithm.

M step OU: increase objective function (GM).

Initialization LASSO regression (see next).

Linear Regression Model



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

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

$$BM : \quad Y = T\Delta + E$$

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

Model Selection on K

Proposition (Form of the Penalty and guaranties (α known))

Under our setting:

$$Y = R\Delta + \gamma E \quad \textit{with} \quad E \sim \mathcal{N}(0, V) \quad \textit{and} \quad \mathcal{S} = \{S_\eta, \eta \in \mathcal{M}\}, \quad \mathcal{M} = \bigcup_{K \geq 0} \mathcal{S}_K^{PI}$$

Define the following penalty:

$$\text{pen}(K) = A \frac{n-K-1}{n-K-2} \text{EDkhi}[K+2, n-K-2, e^{-L_K}], \quad L_K = \log \left| \mathcal{S}_K^{PI} \right| + 2\log(K+2)$$

$$\textit{and the estimator:} \quad \hat{\eta} = \underset{\eta \in \mathcal{M}}{\text{argmin}} \, \| Y - \hat{s}_\eta \|_V^2 \left(1 + \frac{\text{pen}(K_\eta)}{n - K_\eta - 1} \right)$$

Under some reasonable technical hypothesis, we get the non-asymptotic bound:

$$\mathbb{E}\left[\frac{\left\| s - \hat{s}_{\hat{\eta}} \right\|_V^2}{\gamma^2} \right] \leq C(A, \kappa) \left[\inf_{\eta \in \mathcal{M}} \left\{ \frac{\| s - s_\eta \|_V^2}{\gamma^2} + D_\eta (3 + \log(n)) \right\} + 1 + \log(n) \right]$$

Model Selection: Important Points

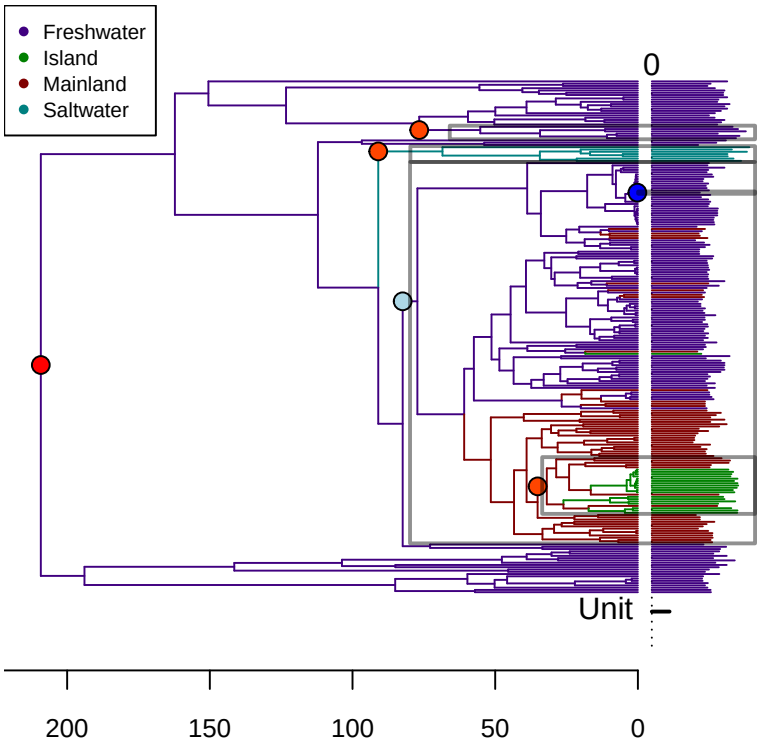
Based on Baraud et al. (2009)

- Non-asymptotic bound.
- Unknown variance.
- No constant to be calibrated.

Novelties

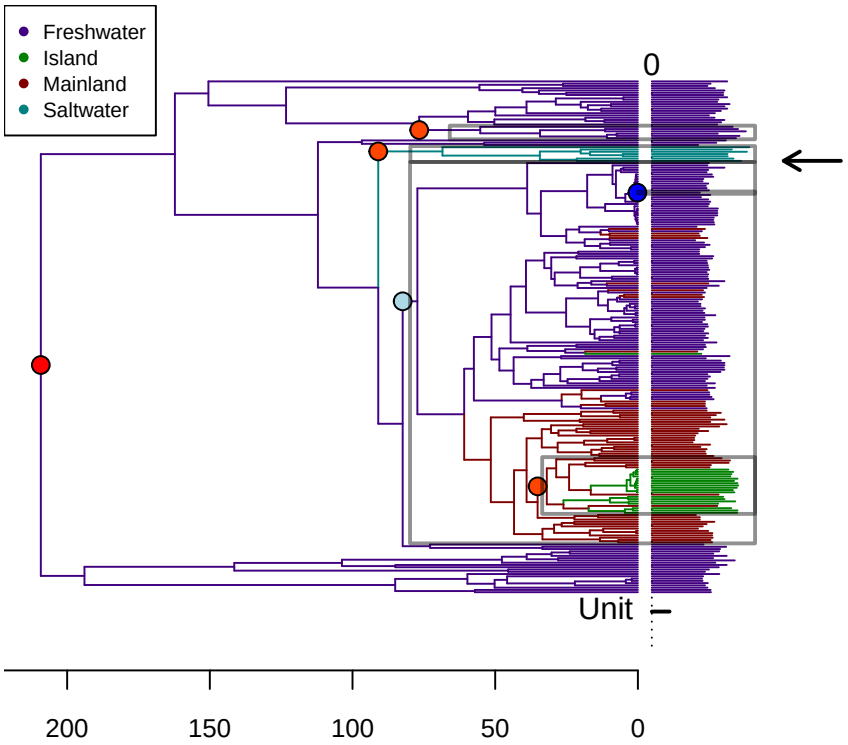
- Non iid variance.
- Penalty depends on the tree topology (through $\left|\mathcal{S}_K^{PI}\right|$).

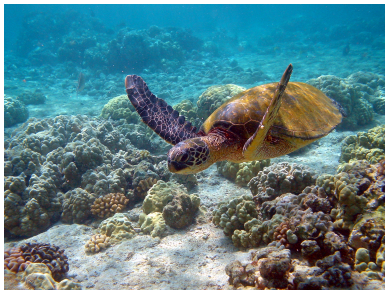
Chelonia Dataset



Colors: habitats.
Boxes: selected EM regimes.

Chelonia Dataset

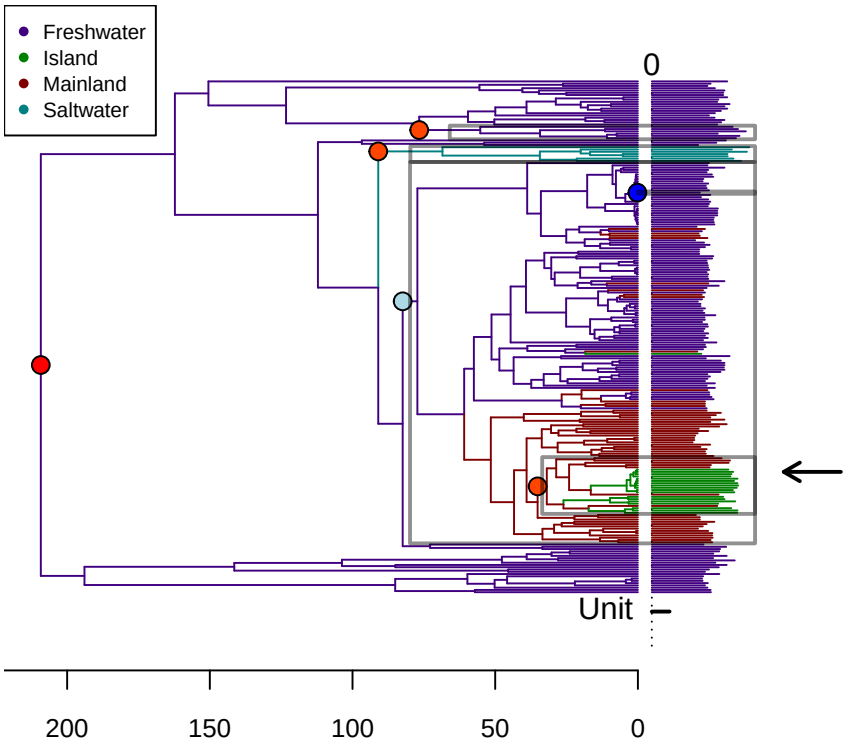




Chelonia mydas

Colors: habitats.
Boxes: selected EM regimes.

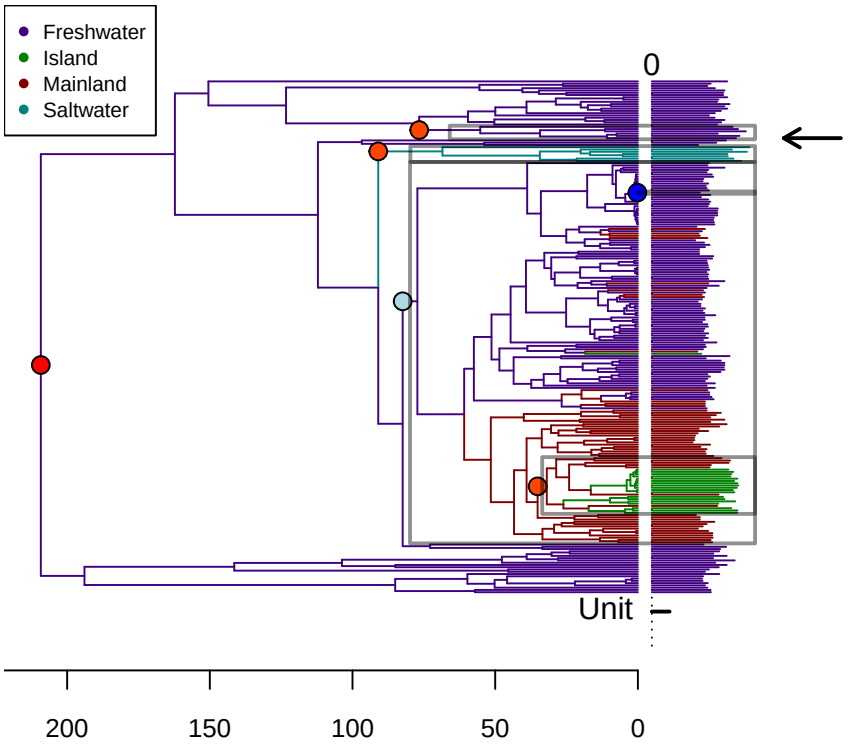
Chelonia Dataset



Colors: habitats.
Boxes: selected EM regimes.

Geochelone nigra abingdoni

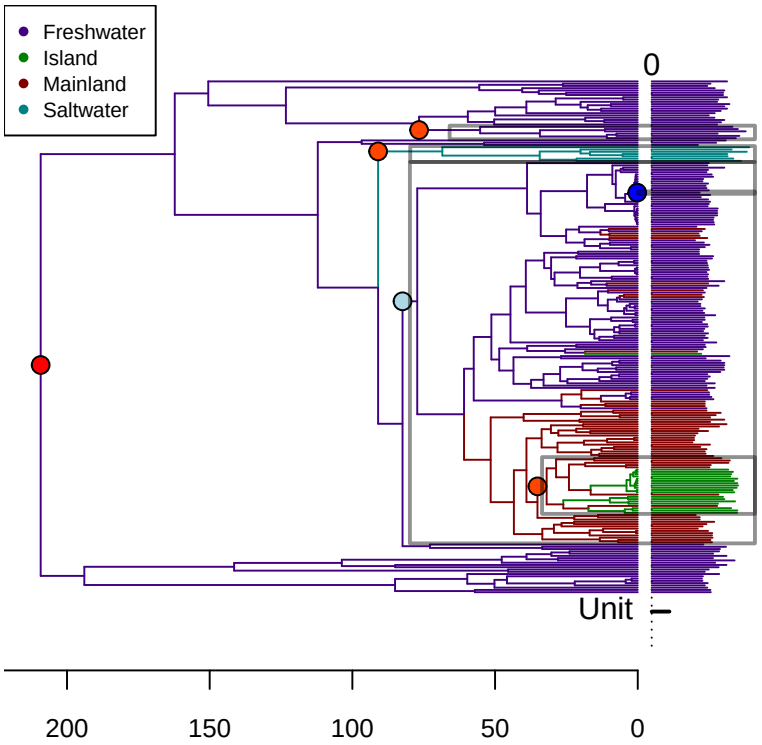
Chelonia Dataset



Chitra indica

Colors: habitats.
Boxes: selected EM regimes.

Chelonia Dataset



	Habitat	EM
No. of shifts	16.00	5.00
No. of regimes	4.00	6.00
lnL	-135.56	-97.59
$\ln 2/\alpha$ (%)	7.83	5.43
γ^2	0.35	0.22
CPU time (min)	1.25	134.49

Colors: habitats.
Boxes: selected EM regimes.

Conclusion and Perspectives

A general inference framework for trait evolution models.

Conclusions

- Some problems of identifiability arise.
- An EM can be written to maximize likelihood.
- Adaptation of model selection results to non-iid framework.

R codes Available on GitHub:
`https://github.com/pbastide/Phylogenetic-EM`

Perspectives

- Multivariate traits.
- Deal with uncertainty (tree, data).
- Use fossil records.

Bibliography

Y. Baraud, C. Giraud, and S. Huet. Gaussian model selection with an unknown variance. *Annals of Statistics*, 37 (2):630–672, Apr. 2009. doi: 10.1214/07-AOS573.

M. A. Butler and A. A. King. Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. *The American Naturalist*, 164(6):pp. 683–695, 2004. ISSN 00030147.

J. Felsenstein. Phylogenies and the Comparative Method. *The American Naturalist*, 125(1):pp. 1–15, Jan. 1985. ISSN 00030147.

J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Suderland, USA, 2004.

T. F. Hansen. Stabilizing selection and the comparative analysis of adaptation. *Evolution*, 51(5):1341–1351, oct 1997.

A. L. Jaffe, G. J. Slater, and M. E. Alfaro. The evolution of island gigantism and body size variation in tortoises and turtles. *Biology letters*, 2011.

Photo Credits :

- "Parrot-beaked Tortoise Homopus areolatus CapeTown 8" by Abu Shawka - Own work. Licensed under CC0 via Wikimedia Commons
- "Leatherback sea turtle Tinglar, USVI (5839996547)" by U.S. Fish and Wildlife Service Southeast Region - Leatherback sea turtle/ Tinglar, USVIUploaded by AlbertHerring. Licensed under CC BY 2.0 via Wikimedia Commons
- "Hawaii turtle 2" by Brocken Inaglory. Licensed under CC BY-SA 3.0 via Wikimedia Commons
- "Dudhwalive chitra" by Krishna Kumar Mishra — Own work. Licensed under CC BY 3.0 via Wikimedia Commons
- "Lonesome George in profile" by Mike Weston - Flickr: Lonesome George 2. Licensed under CC BY 2.0 via Wikimedia Commons

Thank you for listening