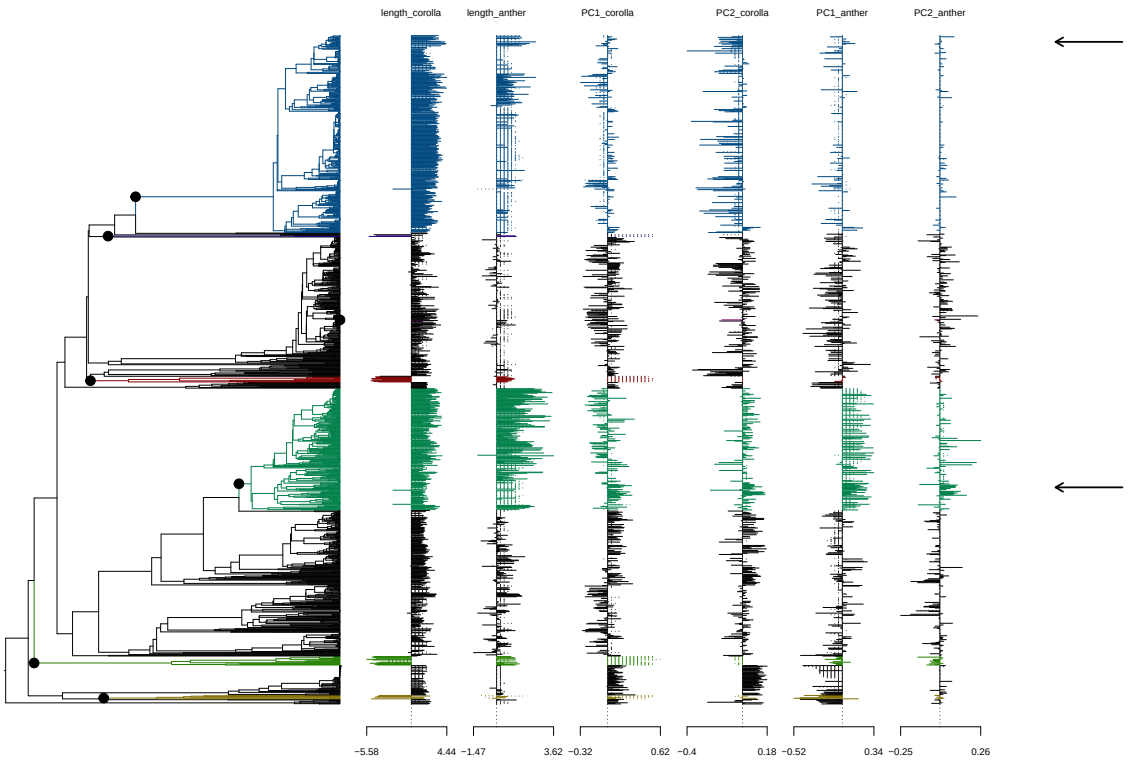
PhylogeneticEM: An R Package for Change-point Detection on Phylogenetic Trees

Cécile Ané[1,2], Paul Bastide[3,4], Mahendra Mariadassou[4], Stéphane Robin[3]

[1] Department of Statistics, University of Wisconsin–Madison, WI, 53706, USA
[2] Department of Botany, University of Wisconsin–Madison, WI, 53706, USA
[3] INRA - AgroParisTech, UMR518 MIA-Paris, F-75231 Paris Cedex 05, France
[4] INRA, UR1404 Unité MaIAGE, F78352 Jouy-en-Josas, France.

23 February 2017

Introduction



Rhododendron dalhousieae

Vaccinium myrtilloides

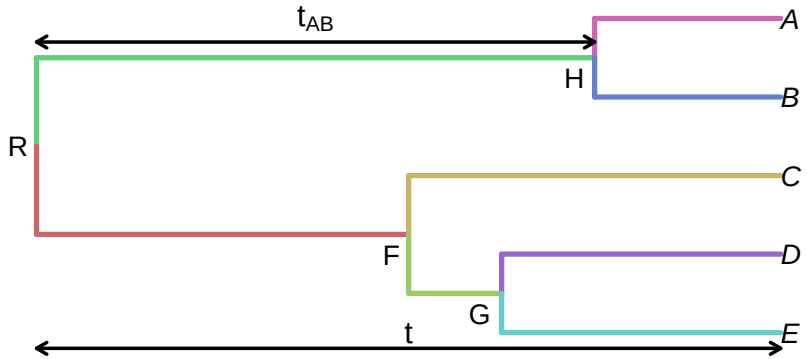
Ericaceae dataset.

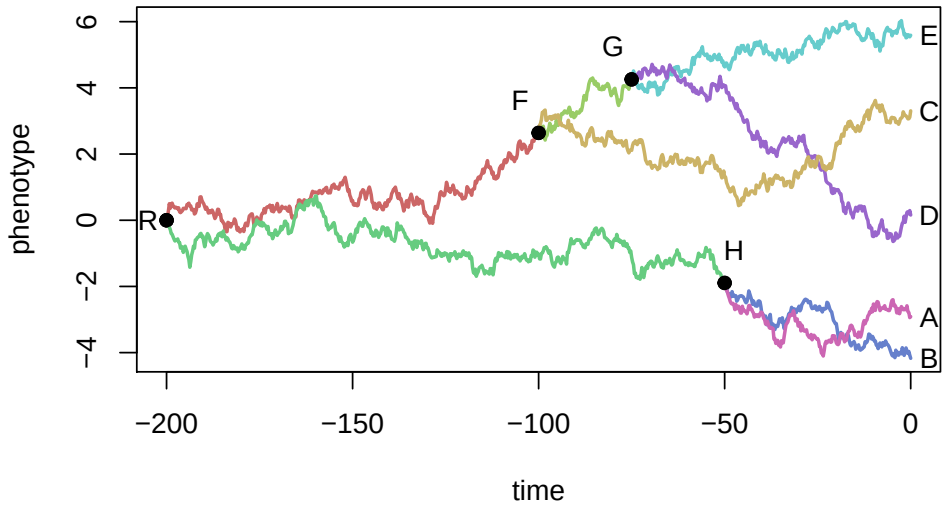
- How can we explain the diversity, while accounting for the phylogenetic correlations ?

Outline

Stochastic Process on a Tree

(Felsenstein, 1985)





The tree is known.

Only *tip* values are observed

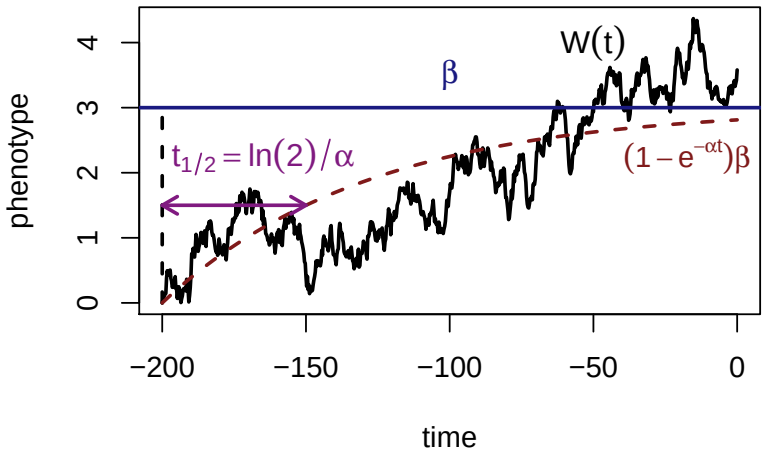
Brownian Motion:

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

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

OU Modeling

(Hansen, 1997)



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

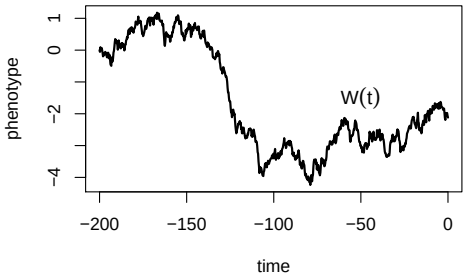
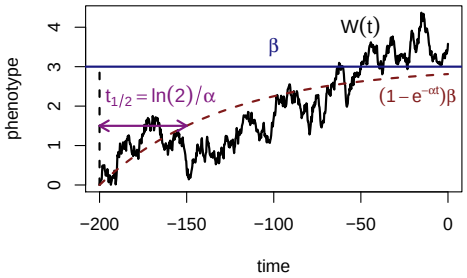
Deterministic part:

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

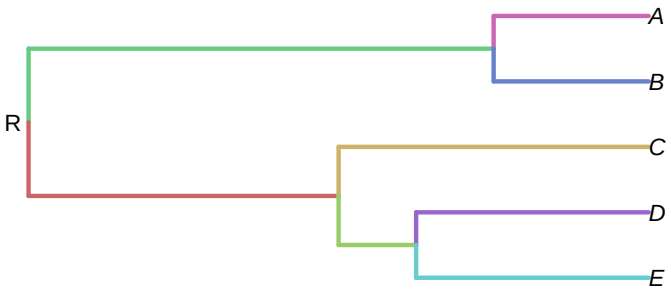
Stochastic part:

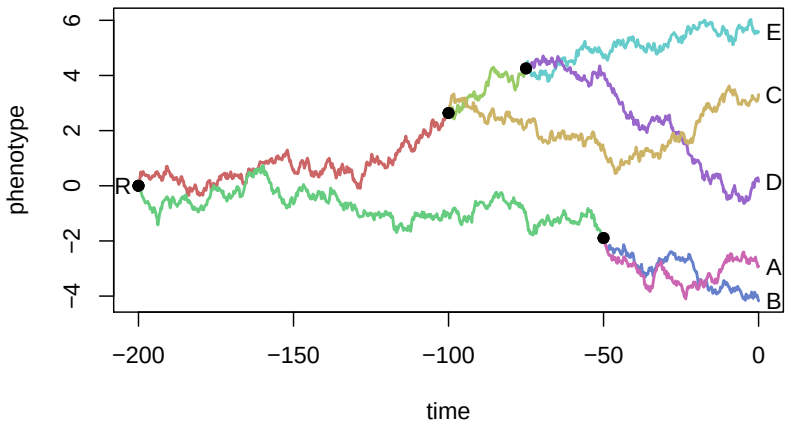
- $W(t)$: actual optimum (trait value).
- $\sigma dB(t)$ Brownian fluctuations.

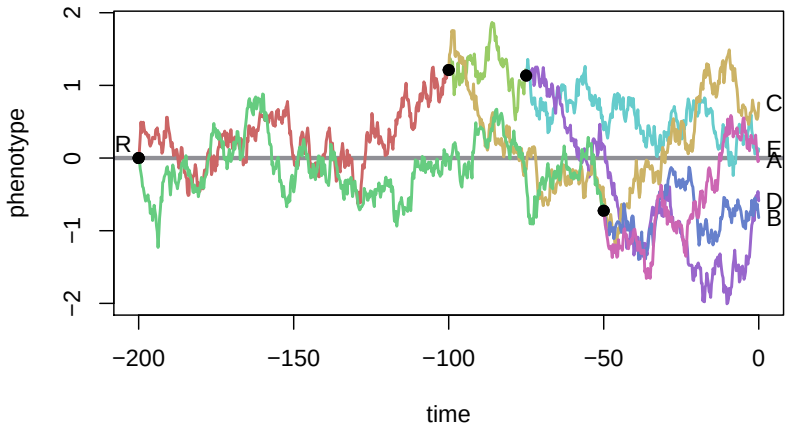
BM vs OU

	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







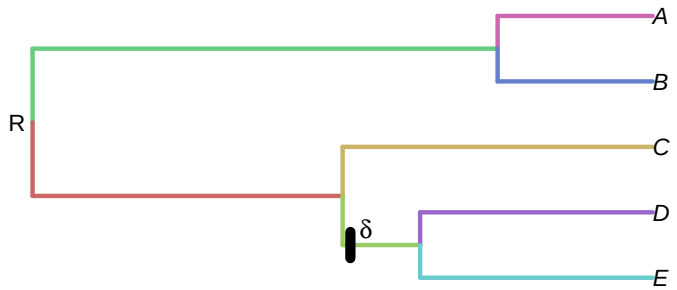
BM Shifts in the **mean**:

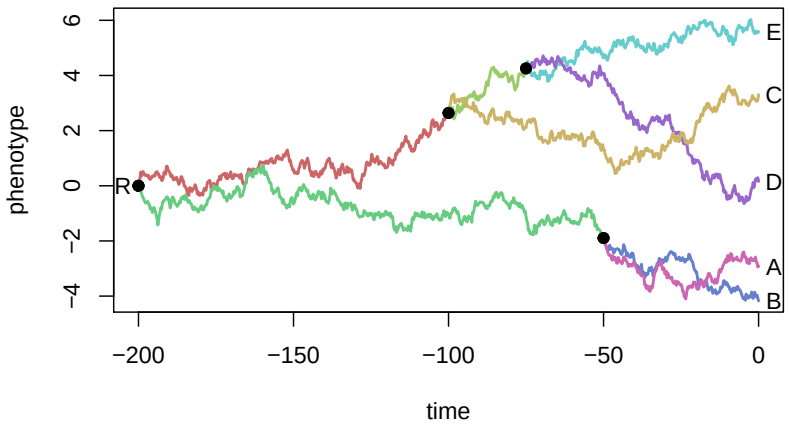
$$m_{\text{child}} = m_{\text{parent}} + \delta$$

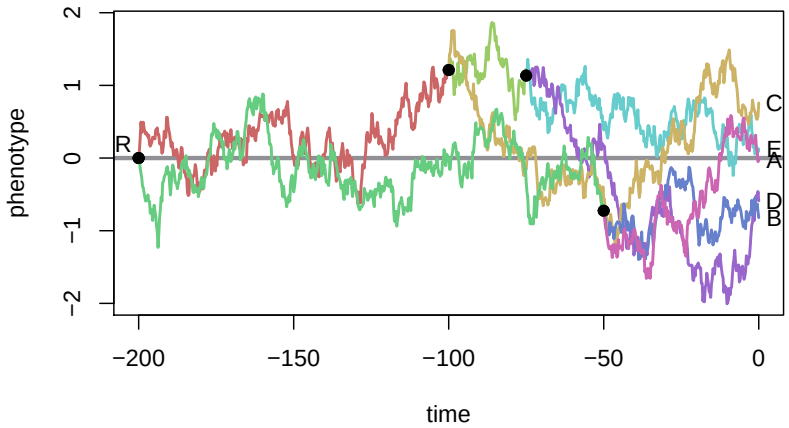
OU Shifts in the **optimal value**:

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

Shifts







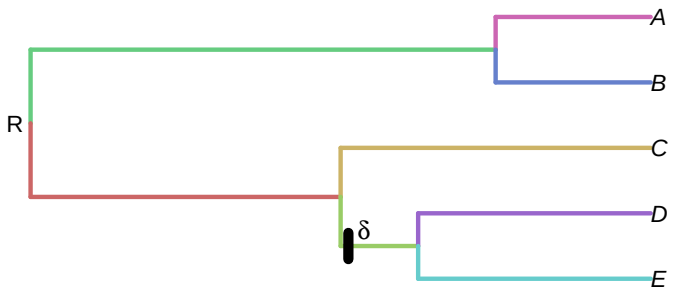
BM Shifts in the **mean**:

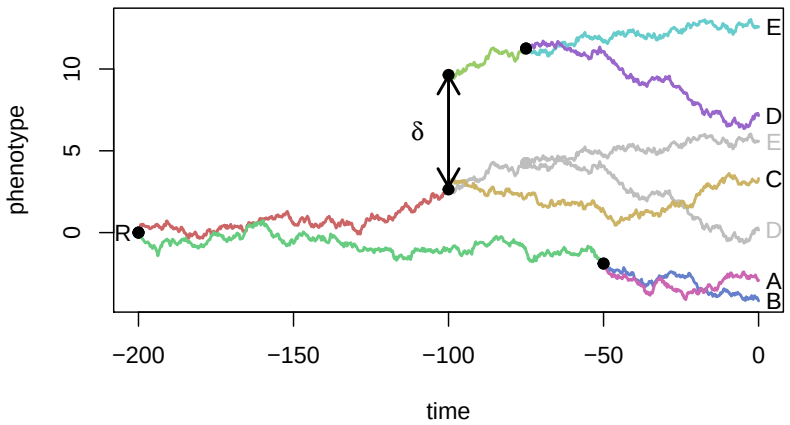
$$m_{\text{child}} = m_{\text{parent}} + \delta$$

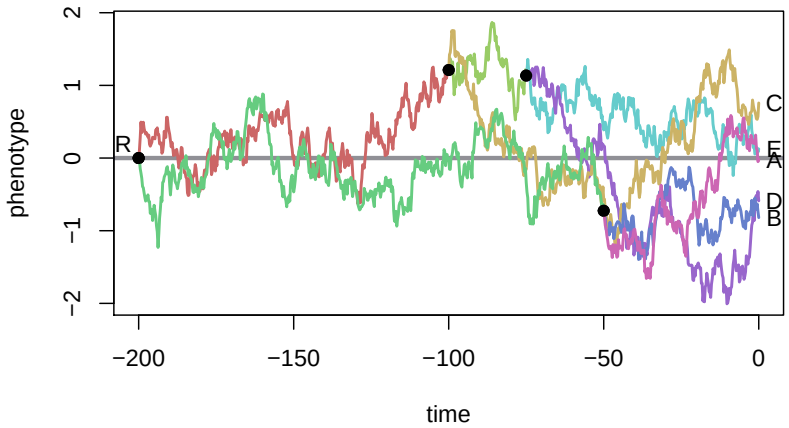
OU Shifts in the **optimal value**:

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

Shifts







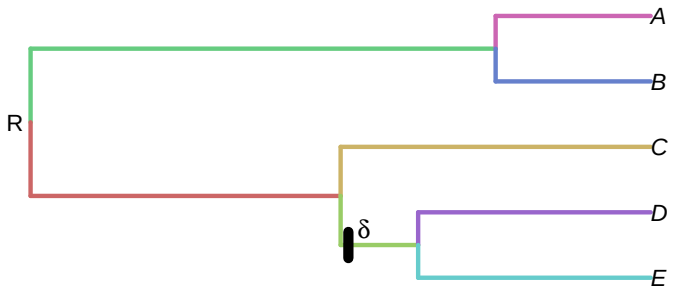
BM Shifts in the **mean**:

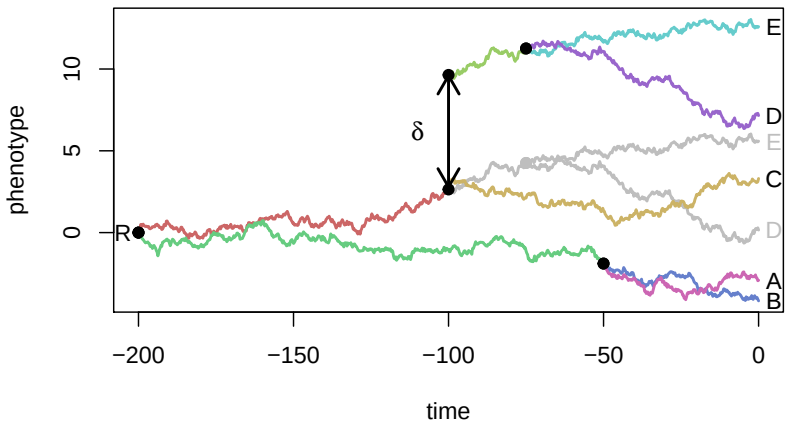
$$m_{\text{child}} = m_{\text{parent}} + \delta$$

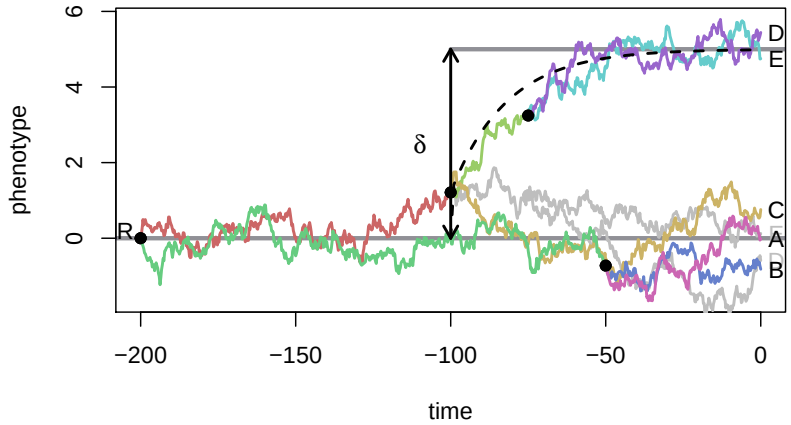
OU Shifts in the **optimal value**:

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

Shifts







BM Shifts in the **mean**:

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

OU Shifts in the **optimal value**:

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

Multivariate BM vs OU

Equation	$\mathbb{C}\text{ov}\left[\mathbf{Y}_i; \mathbf{Y}_j\right]$
$d\mathbf{W}(t) = \mathbf{\Sigma} d\mathbf{B}_t$	$t_{ij}\mathbf{R}, \text{ with } \mathbf{R} = \mathbf{\Sigma}\mathbf{\Sigma}^T$
$d\mathbf{W}(t) = -\mathbf{A}(\mathbf{W}(t) - \beta(t))dt + \mathbf{\Sigma} d\mathbf{B}_t$	$e^{-\mathbf{A}t_i}\mathbf{\Gamma} e^{-\mathbf{A}^T t_j} + e^{-\mathbf{A}(t_i - t_{ij})}\left(\int_0^{t_{ij}} e^{-\mathbf{A}v}\mathbf{R}e^{-\mathbf{A}^T v} dv\right) e^{-\mathbf{A}^T(t_j - t_{ij})}$

$\mapsto$ All the characters shift at the same time

Multivariate BM vs scOU

We use the scalar OU (scOU).

Equation	$\mathbb{C}\text{ov}\left[\mathbf{Y}_i; \mathbf{Y}_j\right]$
$d\mathbf{W}(t) = \mathbf{\Sigma} d\mathbf{B}_t$	$t_{ij}\mathbf{R}$, with $\mathbf{R} = \mathbf{\Sigma}\mathbf{\Sigma}^T$
$d\mathbf{W}(t) = -\alpha(\mathbf{W}(t) - \beta(t))dt + \mathbf{\Sigma} d\mathbf{B}_t$	$\frac{1}{2\alpha} e^{-\alpha(t_i+t_j)} \left(e^{2\alpha t_{ij}} - 1\right) \mathbf{R}$

$\mapsto$ All the traits shift at the same time.

$\mapsto$ All the traits shift have the same selection strength.

Outline

Simulation: Tree

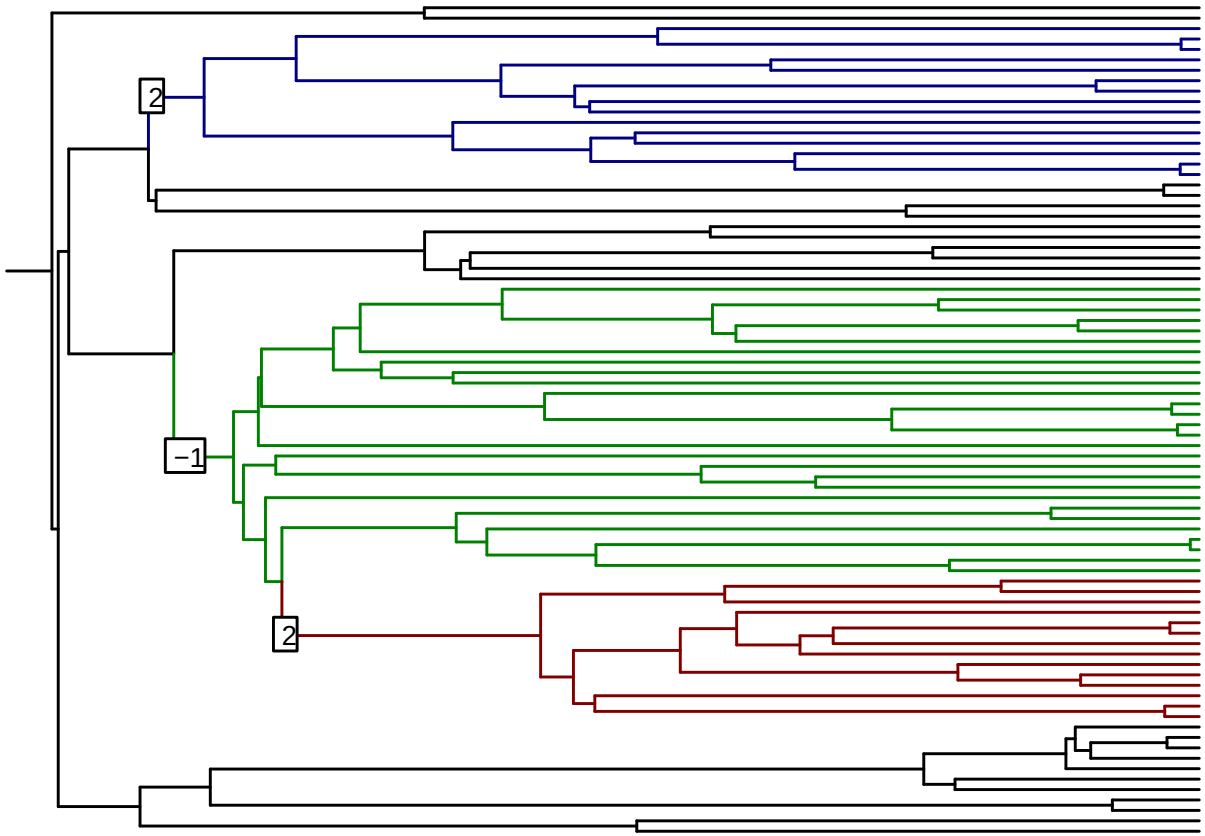
```
library(PhylogeneticEM)
```

```
set.seed(17920902)
ntaxa = 80
tree <- TreeSim::sim.bd.taxa.age(n = ntaxa, numbsim = 1, lambda = 0.1, mu = 0,
                                 age = 1, mrca = TRUE)[[1]]
```

```
params <- params_process("OU",                                  ## Process
                         p = 2,                                 ## Dimension
                         variance = diag(0.5, 2, 2) + 0.5,      ## Rate matrix
                         selection.strength = 3,                ## Selection Strength
                         random = TRUE,                         ## Root is random
                         stationary.root = TRUE,                ## Root is stationary
                         edges = c(29, 25, 127),                ## Positions of the shifts
                         values = cbind(c( 2,  1),              ## Values of the shifts
                                        c(-1,  2),
                                        c( 2, -1)))
```

Simulation: Parameters

```
plot(params, phylo = tree, traits = 1, value_in_box = TRUE, shifts_bg = "white")
```



Simulation: Data

```
sim <- simul_process(params, tree)
```

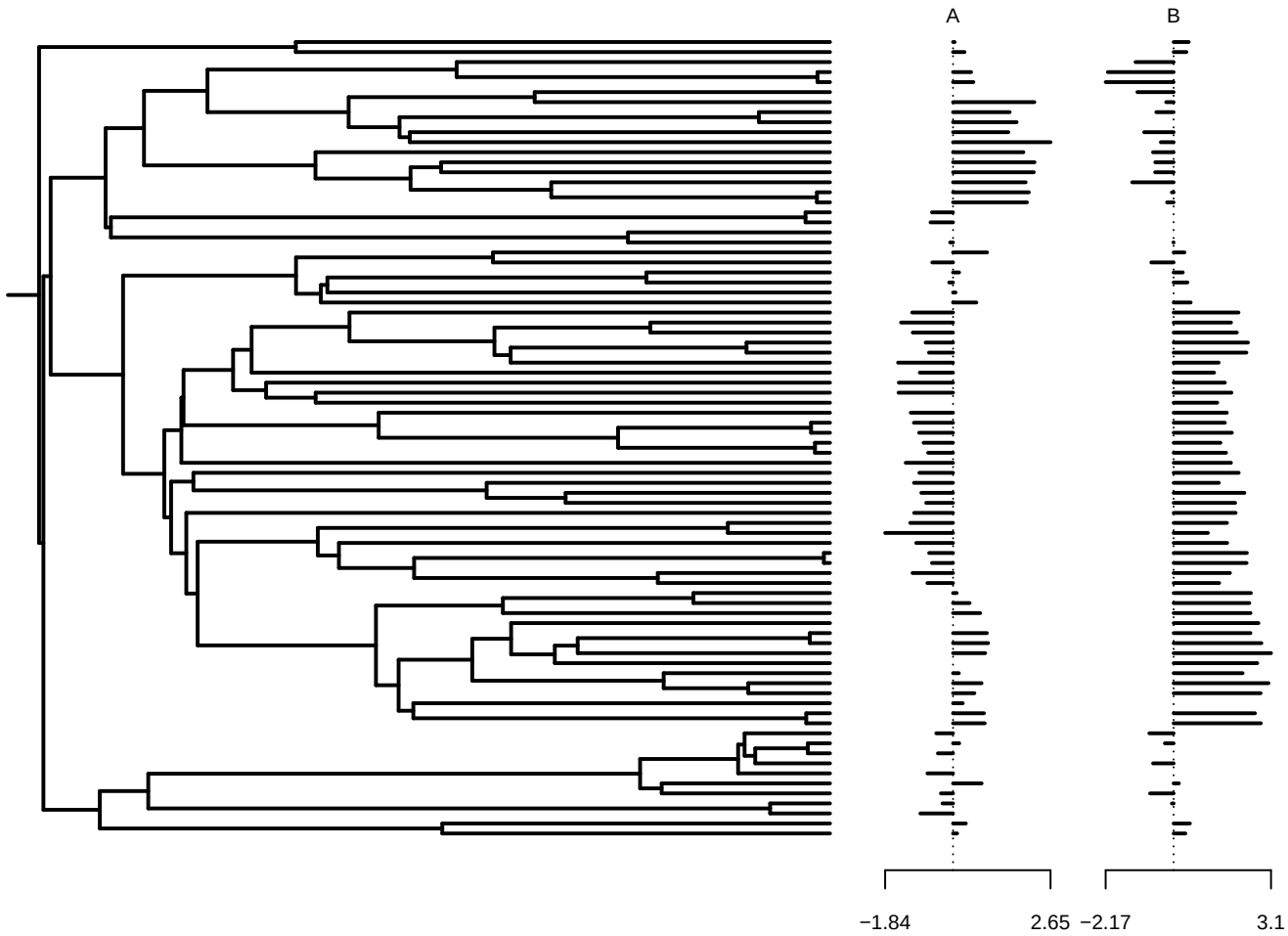
```
data <- extract(sim,              ## The simul_process object
                what = "states", ## We want the actual values
                where = "tips")  ## Only at the tips of the tree
```

```
rownames(data) <- c("A", "B")
```

```
nMiss <- floor(ntaxa * 2 * 0.1)                          ## 10% of missing data
miss <- sample(1:(2 * ntaxa), nMiss, replace = FALSE) ## sample missing randomly
chars <- (miss - 1) %% 2 + 1                             ## Trace back rows and columns
tips <- (miss - 1) %/% 2 + 1
for (i in 1:nMiss){
 data[chars[i], tips[i]] <- NA                           ## Forget some values
}
```

Simulation: Data

```
plot(params_BM(p=2), phylo = tree, data = data, edge.width=2)
```



Inference

```
## Grid on alpha
alpha_grid <- c(2, 2.5, 3, 3.5)

## Run algorithm
res <- PhyloEM(phylo = tree,
               Y_data = data,
               process = "scOU",                  ## scalar OU model
               random.root = TRUE,                ## Root is stationary (true model)
               stationary.root = TRUE,
               alpha = alpha_grid,                ## On a grid of alpha
               K_max = 10,                        ## Maximal number of shifts
               parallel_alpha = TRUE,             ## This can be set to TRUE for
               Ncores = 2)                        ## parallel computations
```

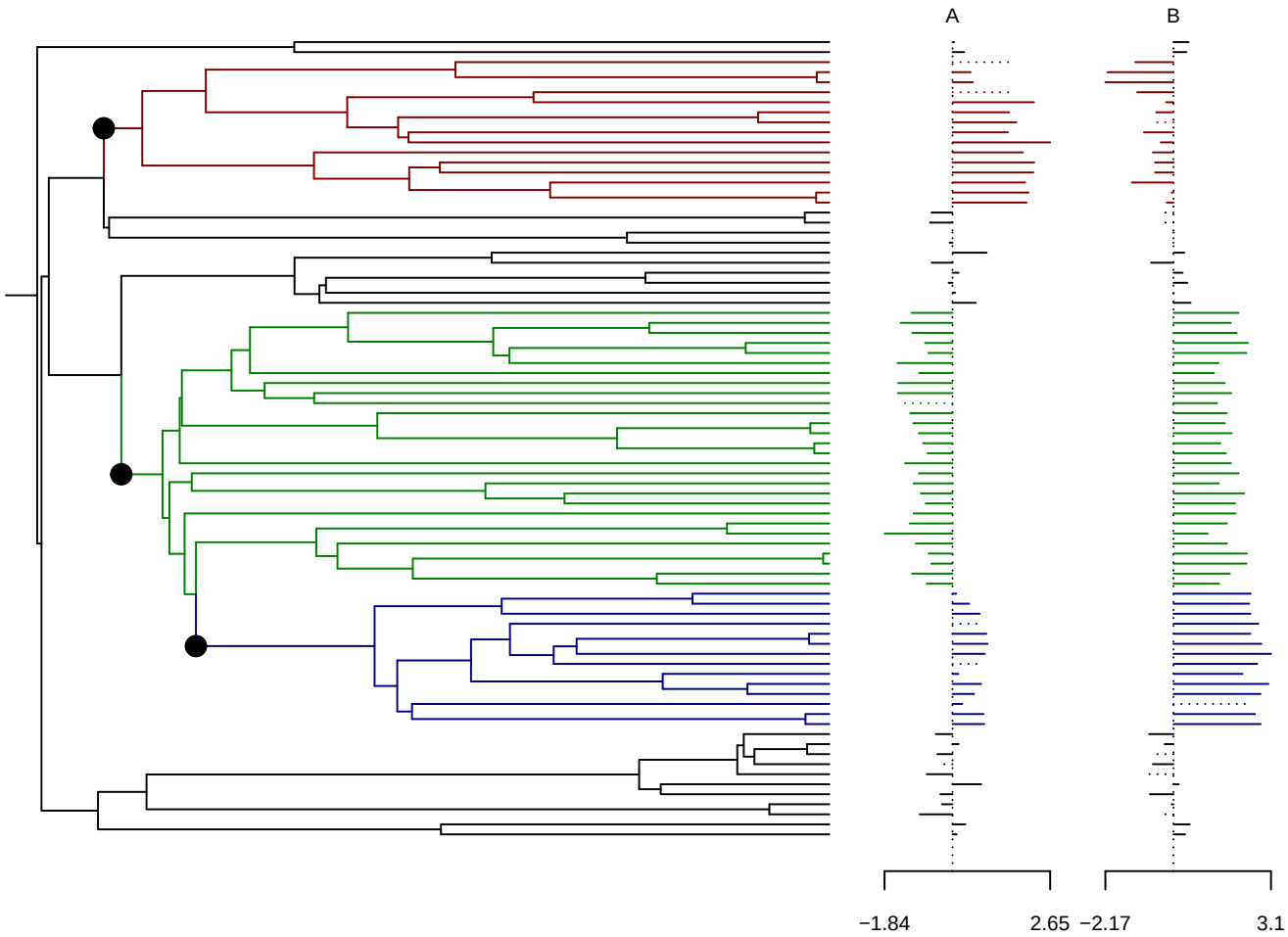
Inference

```
params_process(res)


##
## 2 dimensional scOU process with a random stationary root.
##
## Root expectations:
## [1] 0.02048531 0.14878724
##
## Root variance:
##           [,1]      [,2]
## [1,] 0.1704274 0.1178765
## [2,] 0.1178765 0.1694839
##
## Process variance:
##          [,1]     [,2]
## [1,] 1.022564 0.707259
## [2,] 0.707259 1.016904
##
## Process selection strength:
##      [,1] [,2]
## [1,]    3    0
## [2,]    0    3
##
## Process root optimal values:
## [1] 0.02048531 0.14878724
##
## Shifts positions on branches: 127, 25, 29
## Shifts values:
```

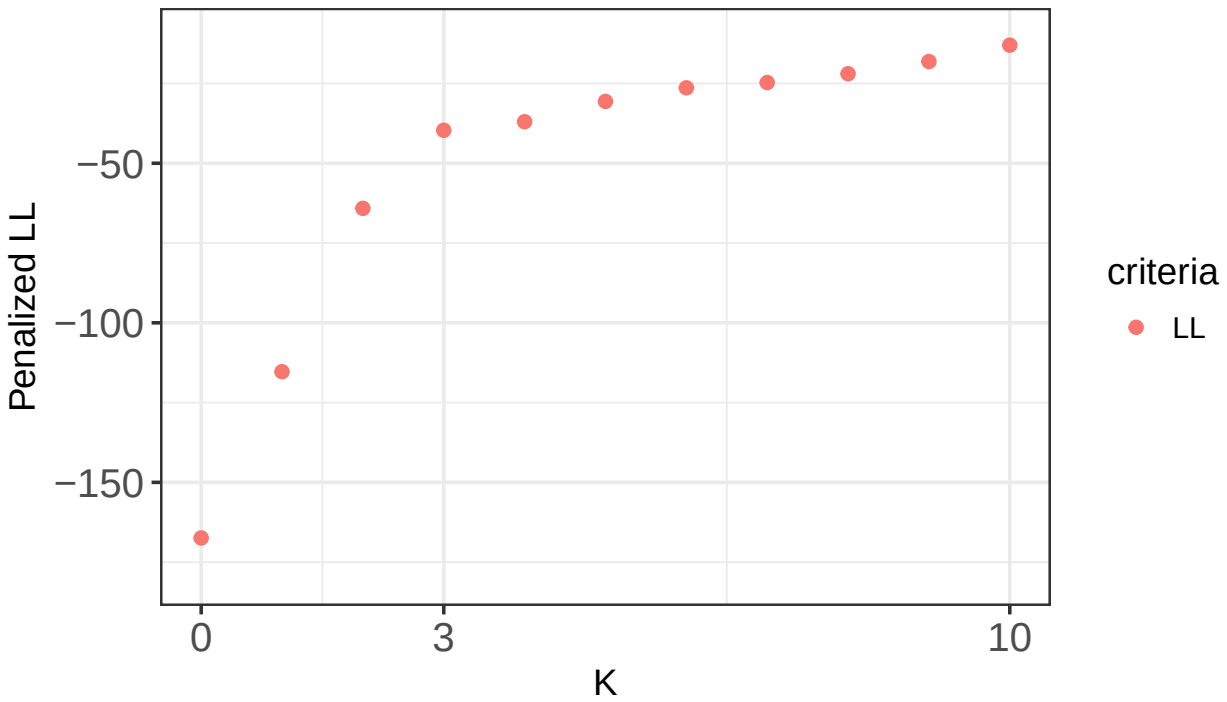
Analysis

```
plot(res)
```



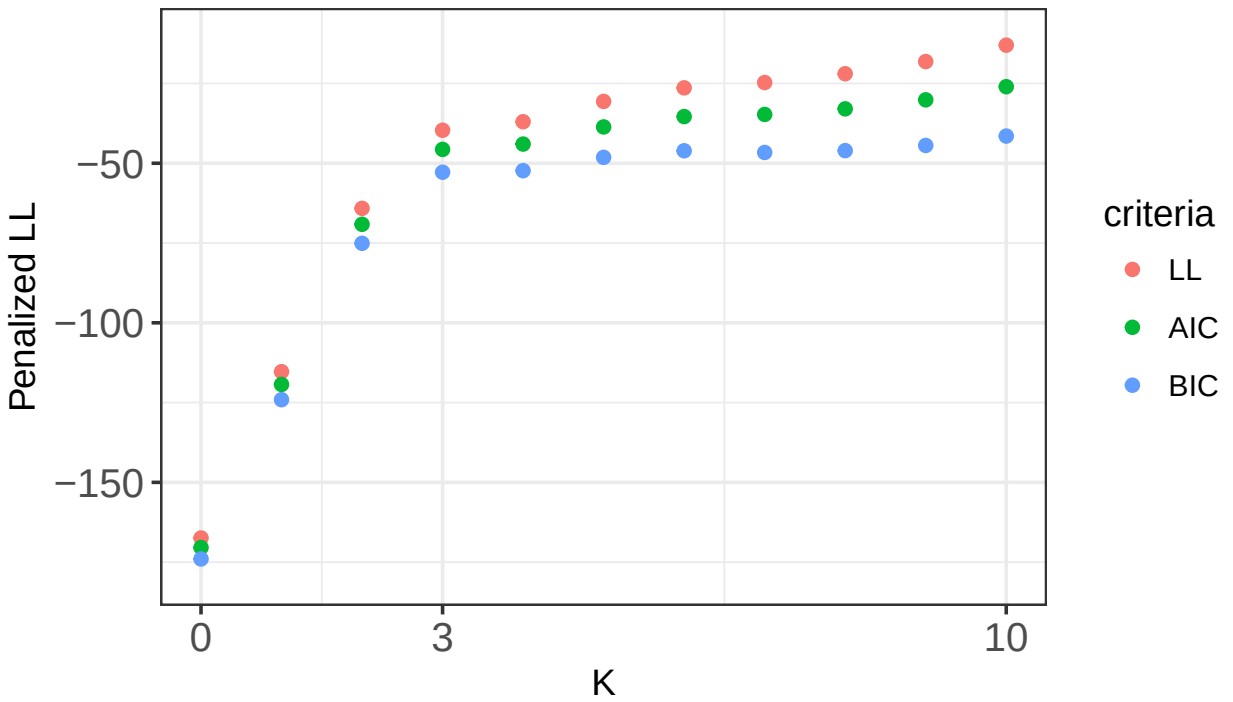
Model Selection: Penalized Likelihood

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



Model Selection: Penalized Likelihood

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



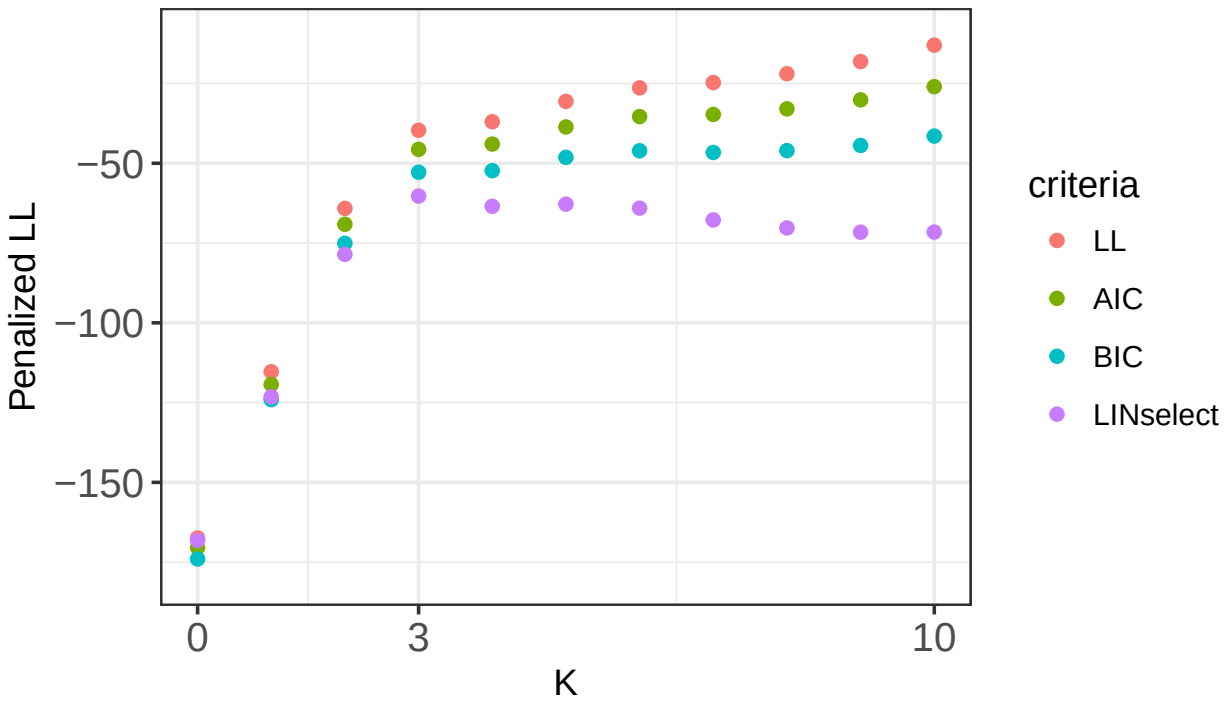
Penalties:

AIC $K + 3$

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

Model Selection: Penalized Likelihood

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



Penalties:

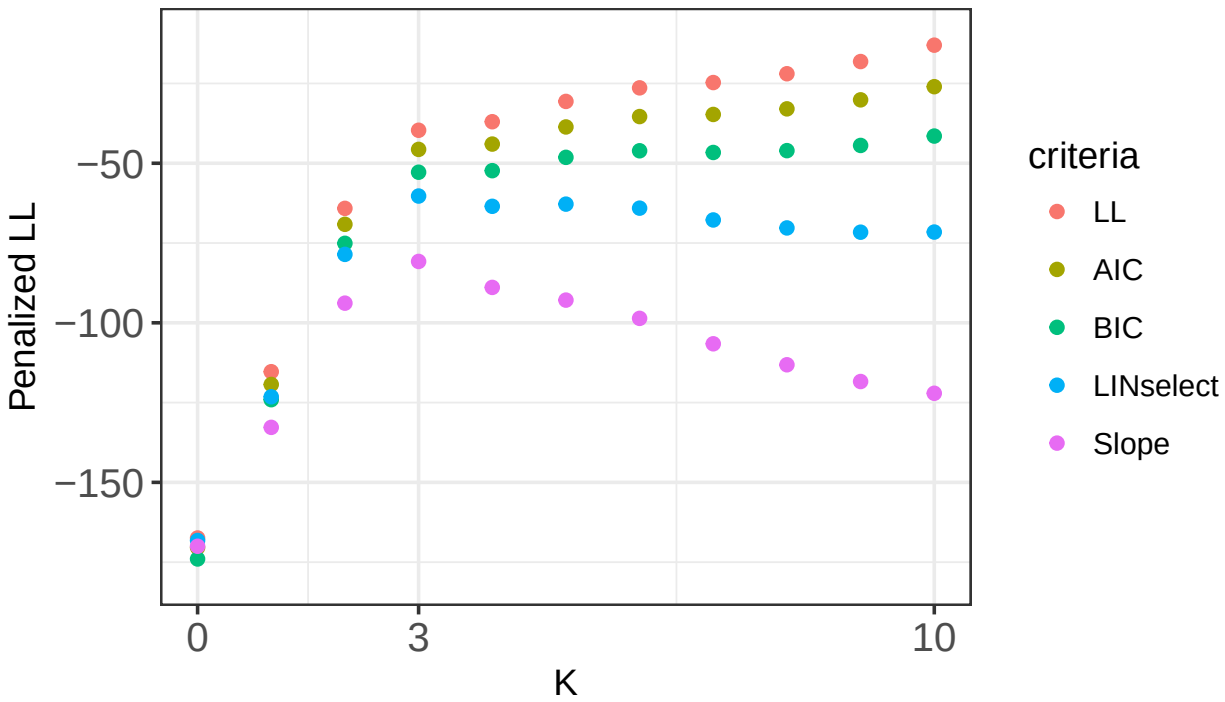
AIC $K + 3$

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

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

Model Selection: Penalized Likelihood

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



Penalties:

AIC $K + 3$

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

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

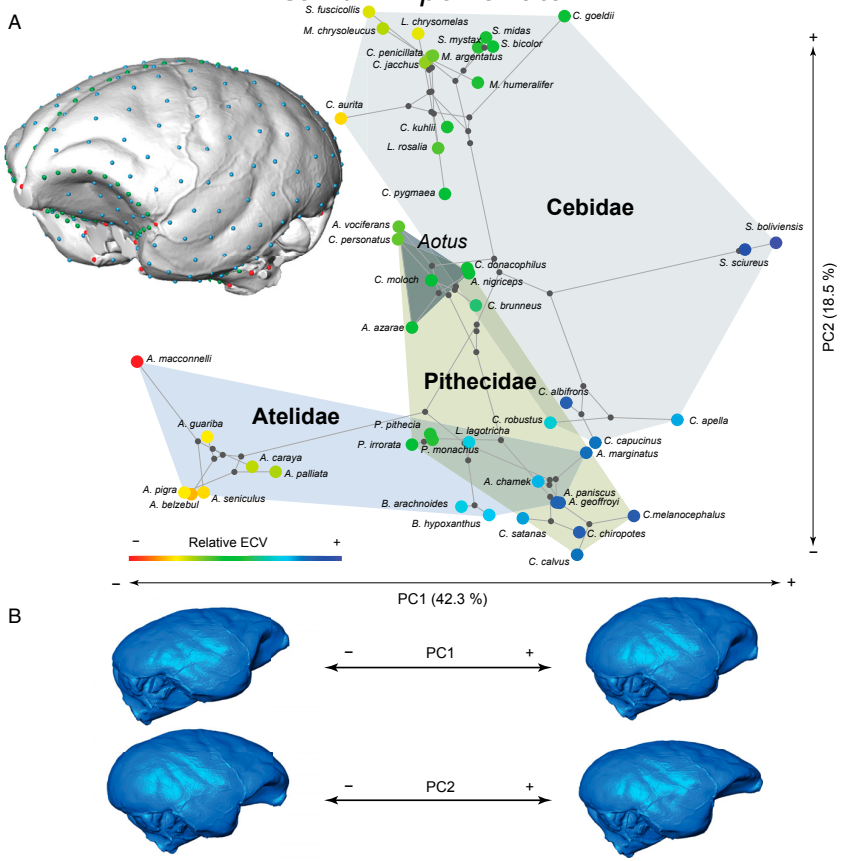
Slope $A\,\text{pen}_s(n, K, |\mathcal{S}_K^{PI}|)$

New World Monkeys

(Aristide et al., 2016)

Callithrix penicillata

Alouatta palliata

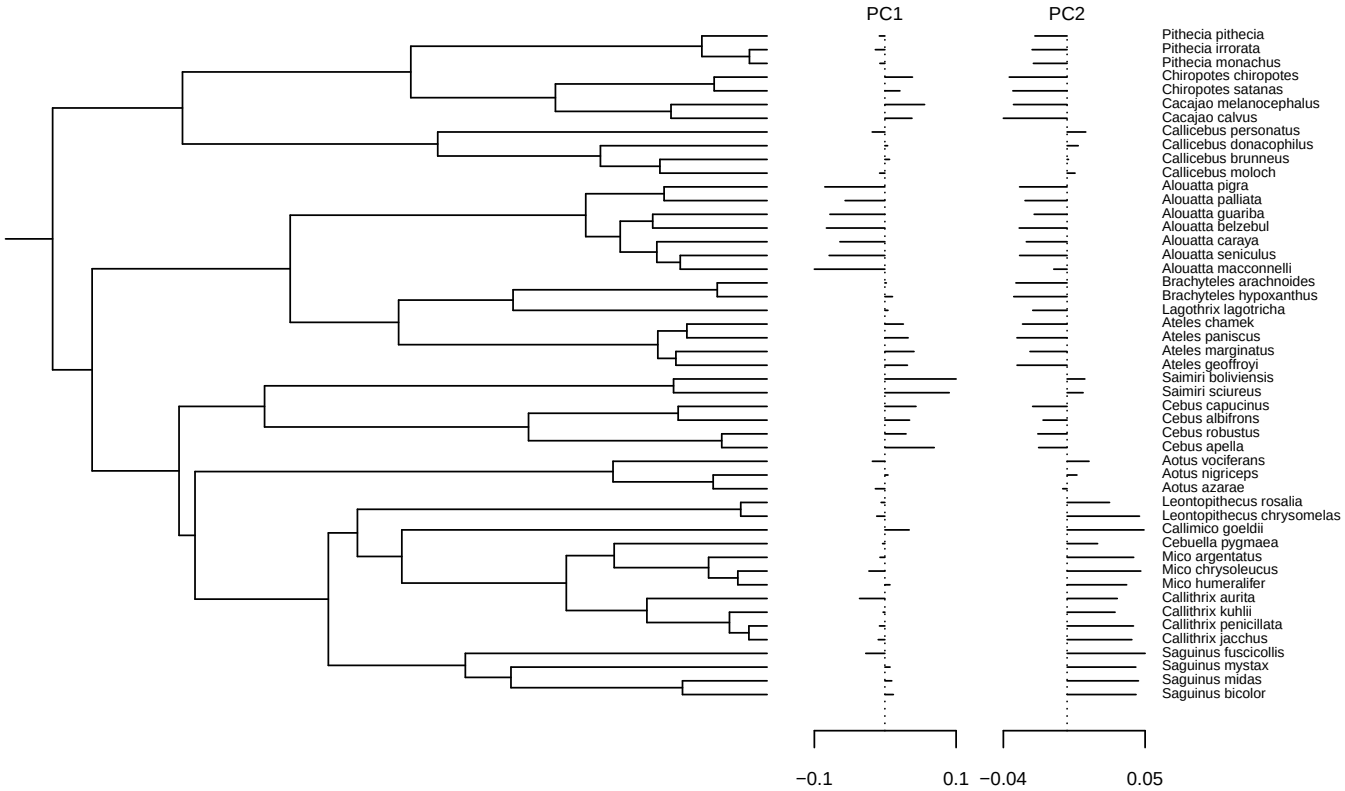


Saimiri sciureus

Monkey Dataset (Aristide et al., 2016)

```
data(monkeys)

plot(params_BM(p=2), data = monkeys$dat, phylo = monkeys$phy, show.tip.label = TRUE)
```



Analysis

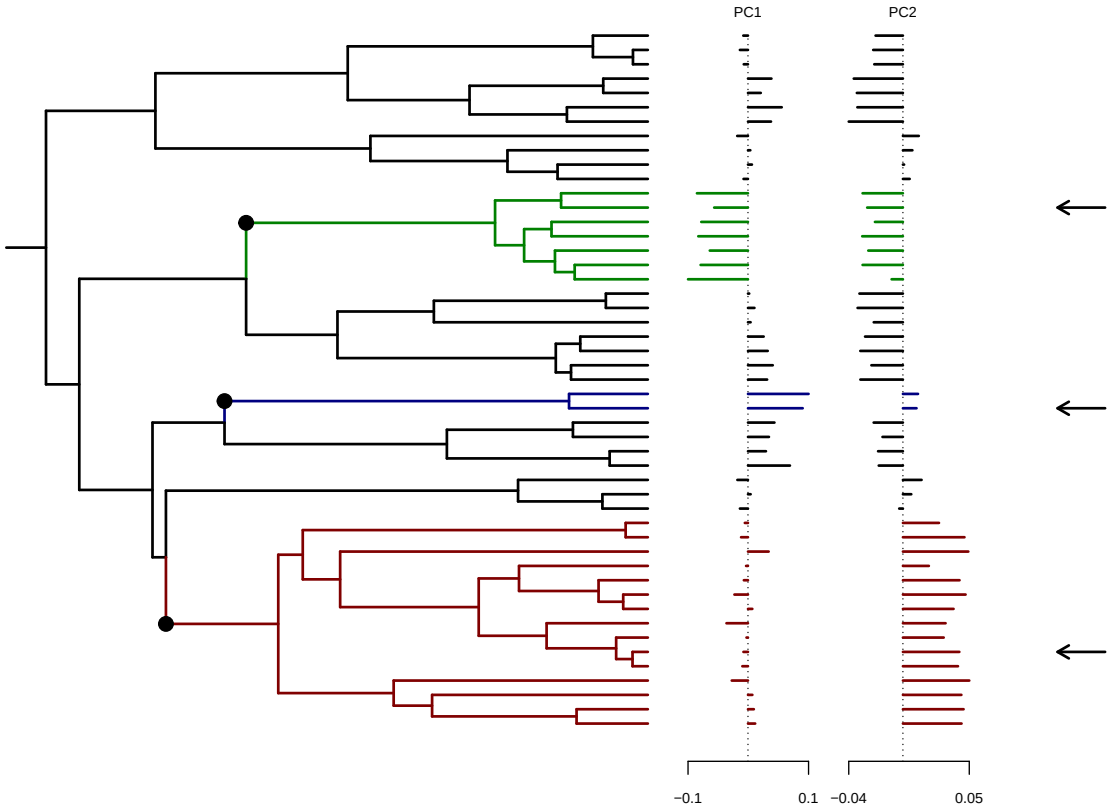
We use function `PhyloEM`:

```
res <- PhyloEM(Y_data = monkeys$dat,          ## data
               phylo = monkeys$phy,           ## phylogeny
               process = "scOU",              ## scalar OU
               K_max = 10,                    ## maximal number of shifts
               nbr_alpha = 4,                 ## number of alpha values
               parallel_alpha = TRUE,         ## parallelize on alpha values
               Ncores = 2)
```

Then plot the solution selected by the default method:

```
plot(res, edge.width = 2)
```

Result



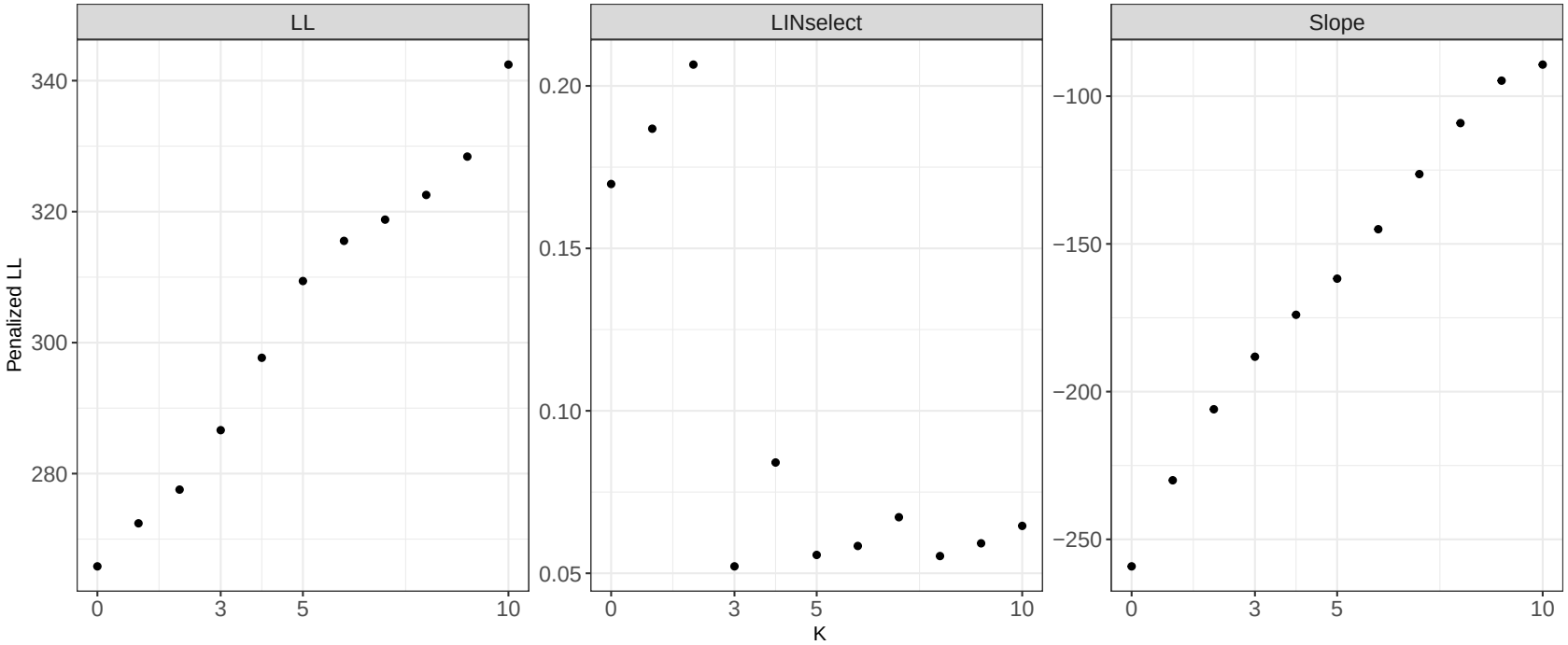
Alouatta palliata

Saimiri sciureus

Callithrix penicillata

Model Selection

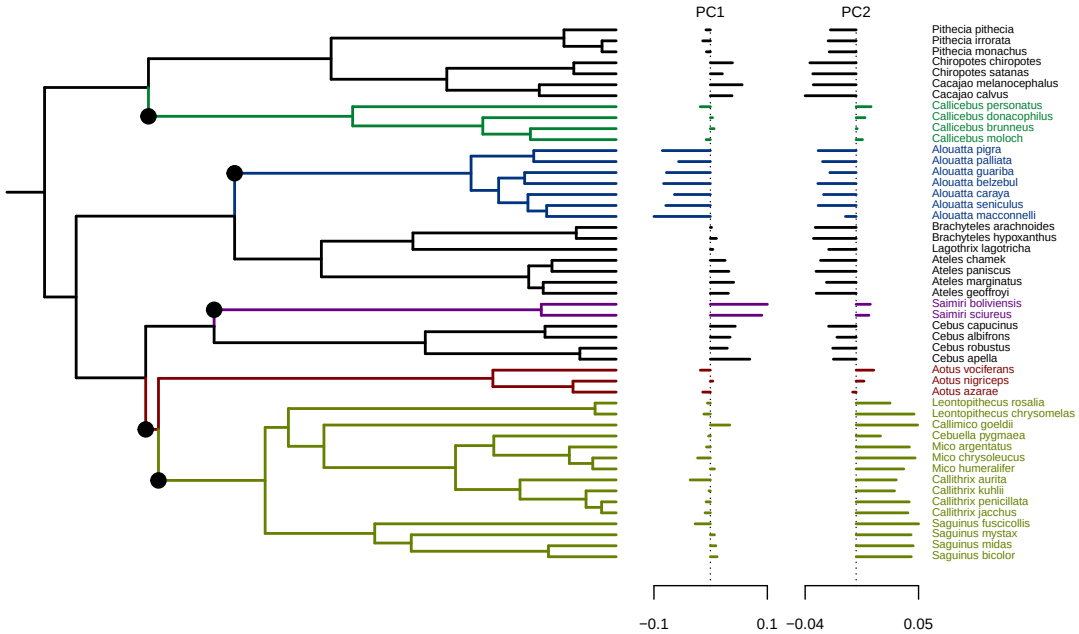
Solution with $K = 5$ seems to be a good solution too.



Solution for $K = 5$

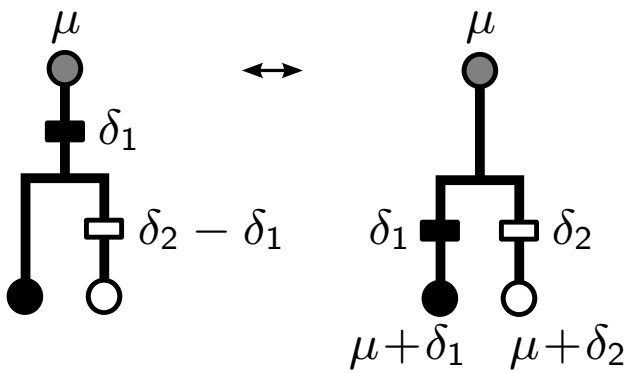
```
plot(res, params = params_process(res, K = 5), edge.width = 2, show.tip.label = TRUE)
```

```
## Warning in params_process.PhyloEM(res, K = 5):  There are several equivalent solutions for
this shift position.
```



Equivalencies

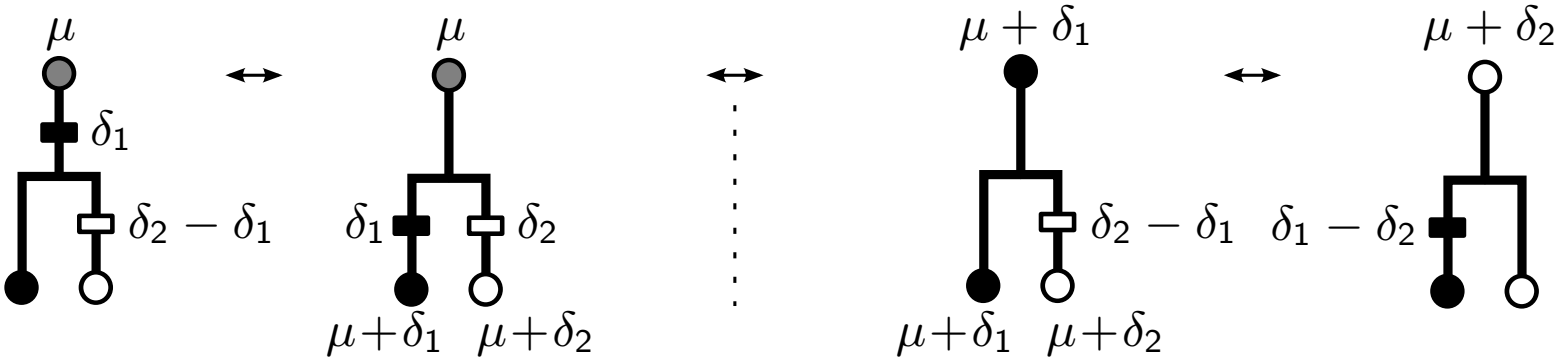
- Number of shifts K fixed, several equivalent solutions.



- Problem of over-parametrization: parsimonious configurations.

Equivalencies

- Number of shifts K fixed, several equivalent solutions.



- Problem of over-parametrization: parsimonious configurations.

Parsimonious Solution: Definition

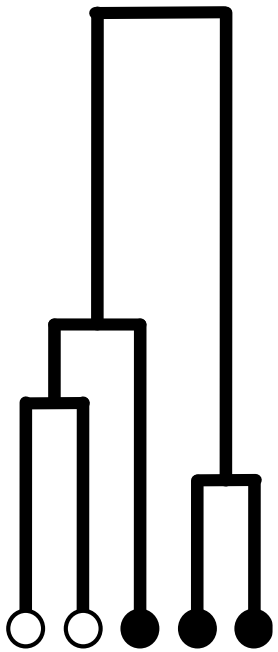
Definition (Parsimonious Allocation)

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 (Parsimonious Allocation)

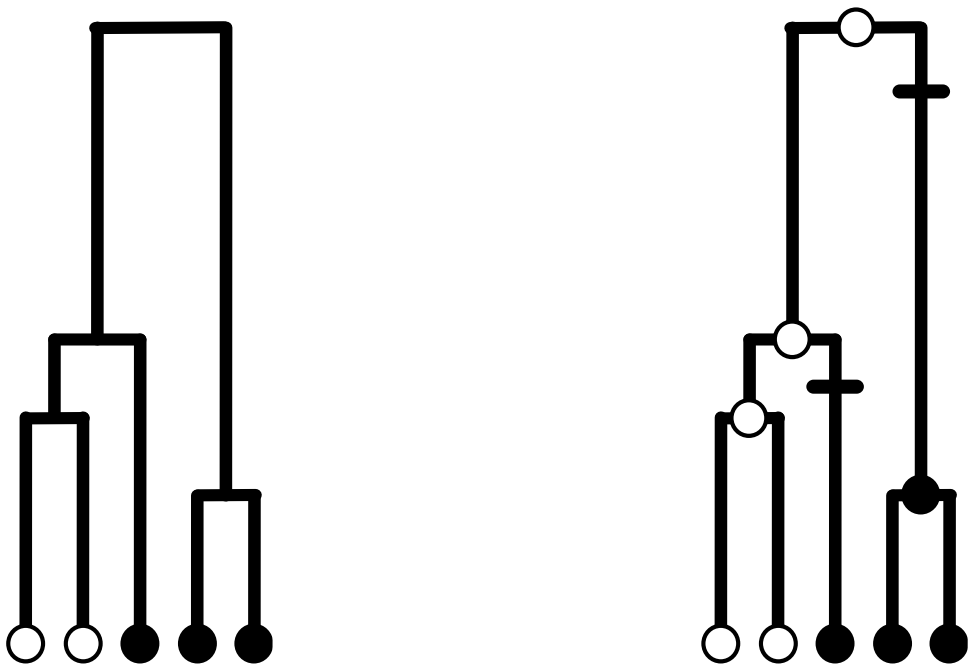
A coloring of the tips being given, a *parsimonious* allocation of the shifts is such that it has a minimum number of shifts.



Parsimonious Solution: Definition

Definition (Parsimonious Allocation)

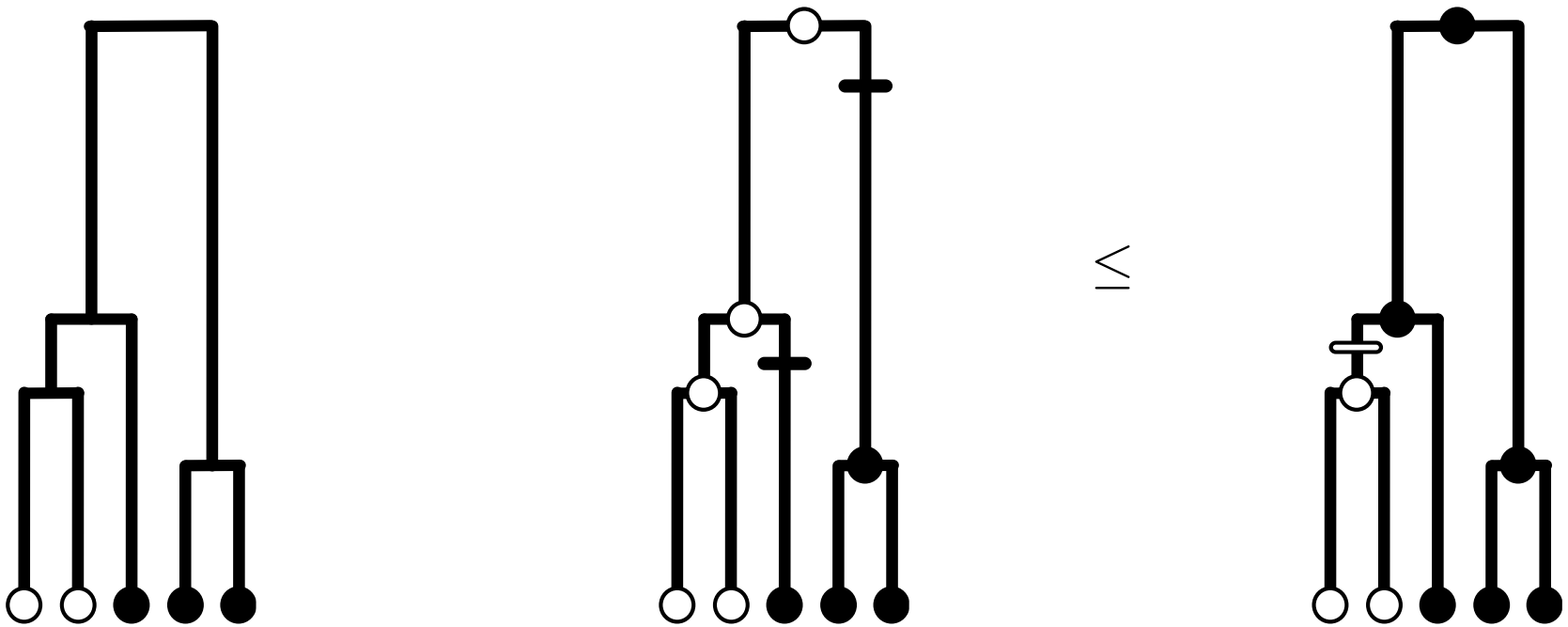
A coloring of the tips being given, a *parsimonious* allocation of the shifts is such that it has a minimum number of shifts.



Parsimonious Solution: Definition

Definition (Parsimonious Allocation)

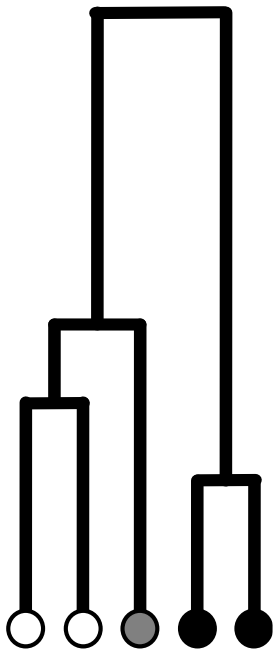
A coloring of the tips being given, a *parsimonious* allocation of the shifts is such that it has a minimum number of shifts.



Parsimonious Solution: Definition

Definition (Parsimonious Allocation)

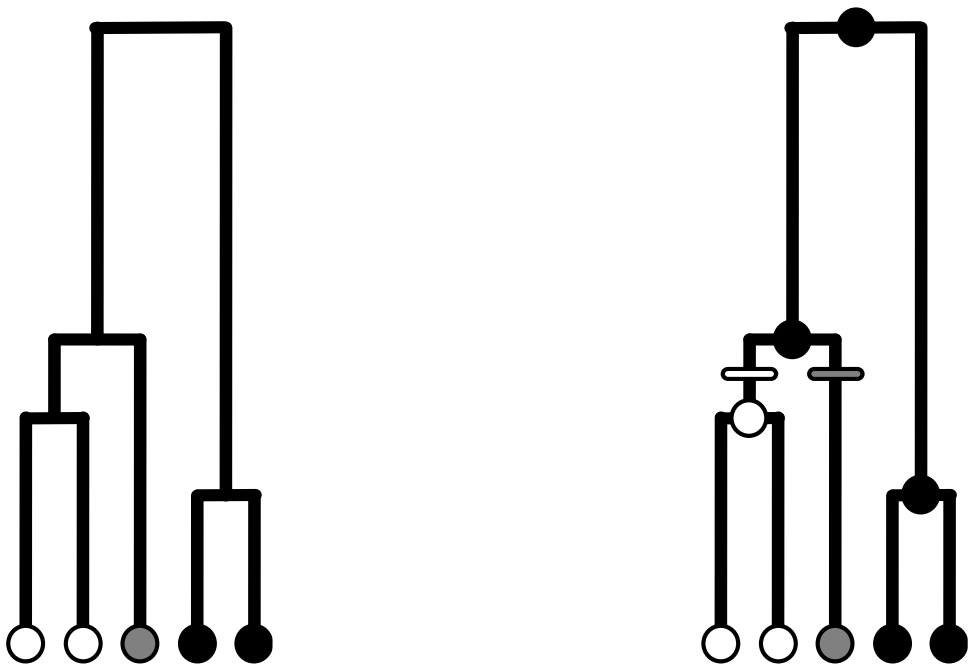
A coloring of the tips being given, a *parsimonious* allocation of the shifts is such that it has a minimum number of shifts.



Parsimonious Solution: Definition

Definition (Parsimonious Allocation)

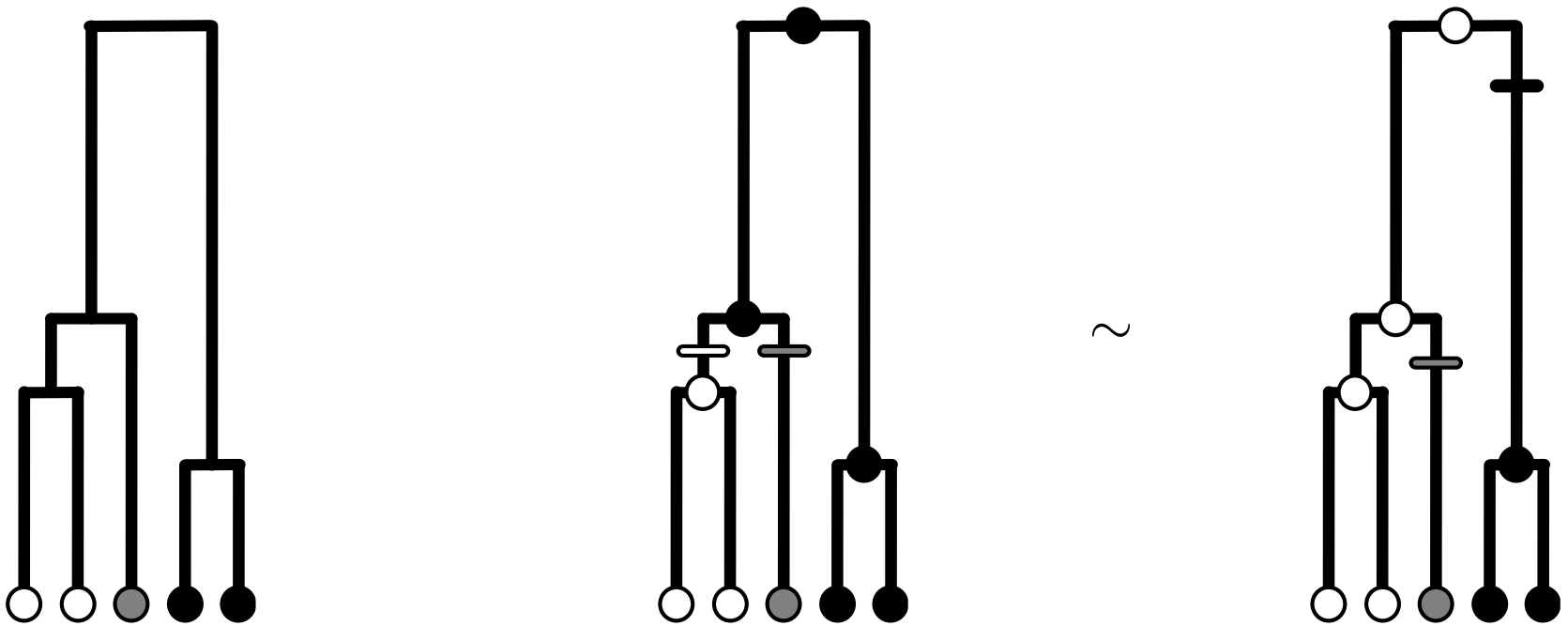
A coloring of the tips being given, a *parsimonious* allocation of the shifts is such that it has a minimum number of shifts.



Parsimonious Solution: Definition

Definition (Parsimonious Allocation)

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 (Equivalency)

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

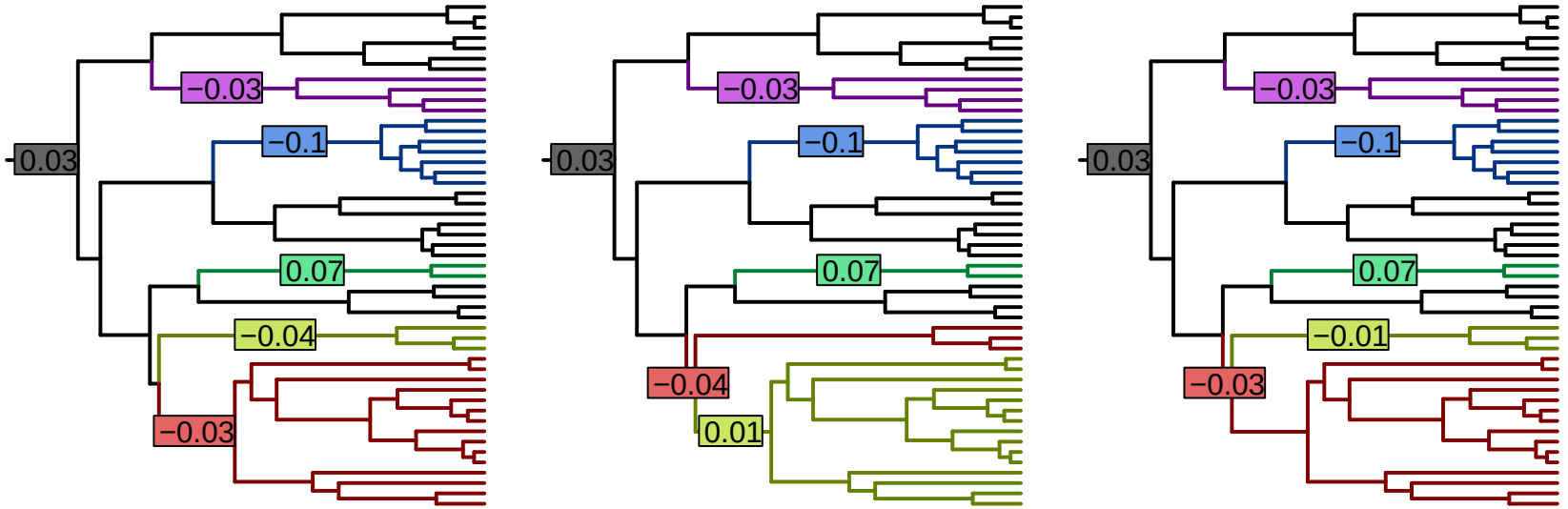
Find one solution Several existing Dynamic Programming algorithms (Fitch, Sankoff, see Felsenstein, 2004).

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

Solution for $K = 5$

```
params_5 <- params_process(res, K = 5)
eq_shifts <- equivalent_shifts(monkeys$phy, params_5)
```

```
plot(eq_shifts)
```

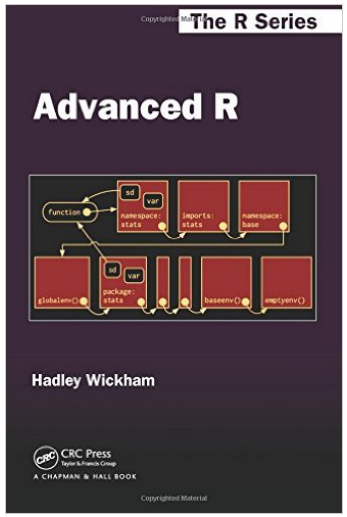


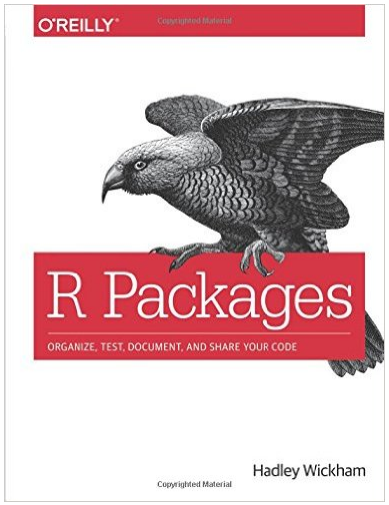
Outline

References

I used mainly these two books by Hadley Wickham:

- Advanced R: `http://adv-r.had.co.nz/`
- R Packages: `http://r-pkgs.had.co.nz/`





- Intensive use of `devtools` and `Rstudio`.

Transparency

All the code is available on `GitHub`:
`https://github.com/pbastide/PhylogeneticEM`

- Package code with verision control.
- Works with the CRAN: tag versions.
- Build status with `Travis CI`.
- Simulations and test cases (reproducibility).
- Automatic documentation with `pkgdown`.

Efficiency

- Profiling with `lineprof`.
- Botelneck: big matrix allocations.
- Solution: use `RcppArmadillo` to code an efficient algorithm (upward-downward).
- Avoid memory leaks with `valgrind`.
- Life saver: the `Armadillo` library.

- Parrallel computations with `doParallel` and `foreach`.

Robustness

- Unitary tests with `testthat`.
- Coverage with `covr` and `codecov`.
- Automated tests with `Travis CI`.

Conclusion and Perspectives

A general inference framework for trait evolution models.

Conclusions
- A complete maximum likelihood procedure
- Taking identifiability problems into account
- With model selection

R Package
- Availble on the CRAN and on GitHub
- Can scale up to big datasets ($\sim$ 1200 species)

Perspectives
- Deal with uncertainty (data).
- Phylogenetic networks: See `julia` package `PhyloNetworks`.

Bibliography

L. Aristide, S. F. dos Reis, A. C. Machado, I. Lima, R. T. Lopes, and S. I. Perez. Brain shape convergence in the adaptive radiation of New World monkeys. *Proceedings of the National Academy of Sciences*, 113(8): 2158–2163, feb 2016. ISSN 0027-8424. doi: 10.1073/pnas.1514473113. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1514473113.

J. Felsenstein. Phylogenies and the Comparative Method. *The American Naturalist*, 125(1):1–15, jan 1985. ISSN 0003-0147. doi: 10.1086/284325. URL http://www.journals.uchicago.edu/doi/10.1086/284325.

J. Felsenstein. Inferring Phylogenies. *American journal of human genetics*, 74(5):1074, 2004. ISSN 00029297. doi: 10.1086/383584.

T. F. Hansen. Stabilizing Selection and the Comparative Analysis of Adaptation. *Evolution*, 51(5):1341, oct 1997. ISSN 00143820. doi: 10.2307/2411186. URL http://www.jstor.org/stable/2411186{%}5Cnhttp://www.jstor.org/stable/pdfplus/10.2307/2411186.pdf?acceptTC=truehttp://www.jstor.org/stable/2411186?origin=crossref.

Photo Credits:
- "Black-tufted marmoset (Callithrix penicillata) in Belo Horizonte Zoo, Brazil." Miguelrangeljr - Own work. Licensed under CC BY-SA 3.0
- "Mantled howler in a wildlife sanctuary, Gulf of Dulce, Costa Rica." Steven G. Johnson - Own work. Licensed under CC BY-SA 3.0
- "Squirrel monkey at The Phoenix Zoo. 2.13.06 Phoenix, Arizona." Braboowi - Own work. Licensed under CC BY-SA 3.0

Thank you for listening

pbastide.github.io

Appendices