M2 Écologie,Évolution,Biométrie 
UE Description Statistique des Structures Biologiques

Signification statistique

D. Chessel

Notes de cours cssb3

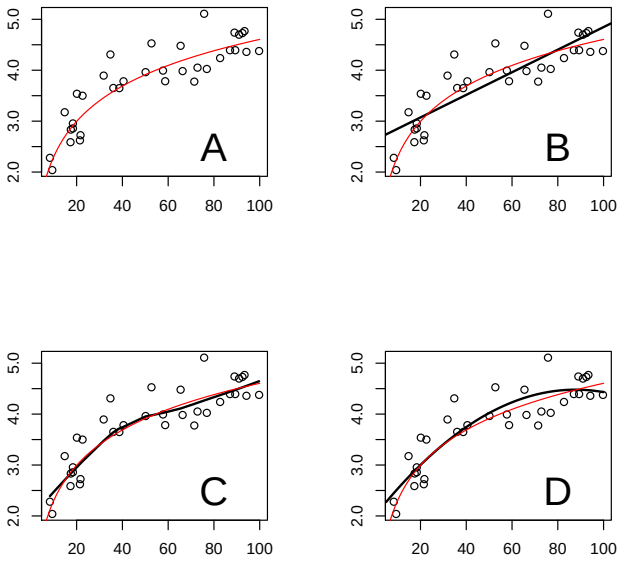
Pour répondre à la demande d'une probabilité critique, quelques illustrations des tests de randomisation. Elles portent sur des structures de données fondamentales en biologie évolutive : associations interspécifiques, collections de cartes et phylogénies compromis.

Table des matières

1 Introduction

La demande d'une probabilité critique (p-value) est générale en biologie évolutive. Il est recommandé de lire l'article de G. Yoccoz [5].

```
set.seed(27082006)
n <- 35
x = runif(n, 5, 100)
x2 <- x^2
a <- 5:100
a2 <- a^2
y <- log(x) + rnorm(n, sd = 0.35)
par(mfrow = c(2, 2))
plot(x, y, xlab = "", ylab = "")
lines(a, log(a), lwd = 1, col = "red")
text(80, 2.5, "A", cex = 2.5)
plot(x, y, xlab = "", ylab = "")
abline(lm(y ~ x), lwd = 2)
lines(a, log(a), lwd = 1, col = "red")
text(80, 2.5, "B", cex = 2.5)
plot(x, y, xlab = "", ylab = "")
lines(lowess(x, y, f = 0.6), lwd = 2)
lines(a, log(a), lwd = 1, col = "red")
text(80, 2.5, "C", cex = 2.5)
plot(x, y, xlab = "", ylab = "")
lines(a, predict(lm(y ~ x + x2), new = list(x = a, x2 = a2)), lwd = 2)
lines(a, log(a), lwd = 1, col = "red")
text(80, 2.5, "D", cex = 2.5)
```



La ligne rouge indique la réalité vraie mais inconnue, les points sont l'échantillon et les lignes noires sont les modèles. Si vous dites qu'on voit sur la figure (A) que la quantité mesurée est croissante avec l'explicative, le referee demandera une preuve statistique. Le referee a tort. Si vous dites qu'on voit sur la figure (B) que la quantité mesurée est linéairement croissante avec l'explicative (p=4e-10, ***) vous avez tort. Vous auriez le droit de dire que la variable mesurée n'est pas constante (p=4e-10, ***) mais ça crève les yeux. Si vous dites qu'on

voit sur la figure (C) que la quantité mesurée est croissante avec l'explicative et que la croissance tend à ralentir, le referee demandera une preuve statistique. Il a tort, mais c'est déjà moins évident. Si vous dites qu'on voit sur la figure (D) que la quantité mesurée est croissante avec l'explicative et tend vers un optimum (p=2e-3, **) vous avez tort mais vous aurez la paix. Il est bien connu que c'est un langage de menteurs.

En fait, tout ceci n'a aucun intérêt parce qu'il n'y a pas de signification biologique derrière cette simulation.

Ici, on parle de description de structures, mais on peut se poser la question de l'existence d'une structure à décrire. Si le besoin est on peut construire des tests de randomisation bien adaptés à la situation. On donne des exemples.

2 Les associations inter spécifiques

Reprendre l'exemple `cortes`.

```
load(url("http://pbil.univ-lyon1.fr/R/donnees/cortes.rda"))
```

On a 25 îles (lignes) et 20 espèces (colonnes). La fréquence des espèces est très inégale (c'est une généralité !). La richesse des sites l'est tout autant.

```
dim(liz)
```

```
[1] 20 25
```

```
liz <- as.data.frame(t(liz))
srel = apply(liz, 1, sum)
srel
```

```
 A  B  C  D  E  F  G  H  I  J  K  L  M  N  O  P  Q  R  S  T  U  V  W  X  Y
13  4  9  4  3  2  3  4  4  5  2  5  2 10 10 10  7  6  6  3  3 11  8 11  6
```

```
sesp = apply(liz, 2, sum)
sesp
```

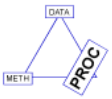
```
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
 7 23  6  6  2 18  8  8  1 22  2  2  9  2  1 18  9  3  3  1
```

Il y a 151 présences au total. Le nombre de *cooccurrences* est une bonne statistique. Il y a structure si ce nombre est trop grand.

```
wtw = t(liz) %*% as.matrix(liz)
diag(wtw) = 0
sum(wtw)/2
```

```
[1] 512
```

Si certaines espèces ont tendance à se retrouver ensemble cet effectif sera trop grand. Si certaines espèces s'évitent il aura tendance à être trop faible. On pourrait avoir les deux phénomènes mais ne compliquons pas tout de suite. Examiner la fonction `outer`. Que fait ce calcul ?.




```
(1:4) %o% (1:3)
proba <- (srel/151) %o% (sesp/151)
sum(proba)
sample(1:500, 151, prob = proba, rep = F)
```

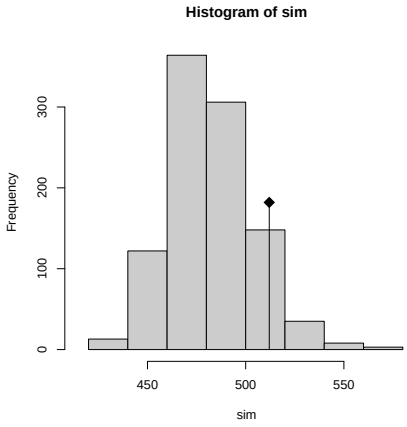
Écrire alors une fonction d'échantillonnage :

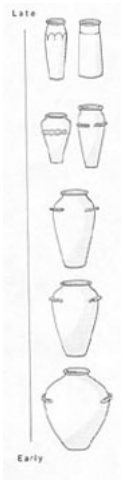
```
sim1 <- function(k) {
    vec = rep(0, 500)
    tir = sample(1:500, 151, prob = proba, rep = F)
    vec[tir] = 1
    vec = matrix(vec, 25, 20)
    wtw = t(vec) %*% as.matrix(vec)
    diag(wtw) = 0
    sum(wtw)/2
}
```

L'utiliser et conclure. A noter la classe des `randtest`, très simple à utiliser.

```
library(ade4)
sim = unlist(lapply(1:999, sim1))
test = as.randtest(sim, 512)
plot(test)
test
```

```
Monte-Carlo test
Call: as.randtest(sim = sim, obs = 512)
Observation: 512

Based on 999 replicates
Simulated p-value: 0.073
Alternative hypothesis: greater

    Std.Obs Expectation    Variance
  1.528205  480.835836  415.860799
```

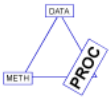




La question centrale est : y a-t-il une structure dans l'assemblage des espèces ? Le résultat n'est pas claire. Les espèces sont-elles pourtant liées ? Pour voir la structure utiliser une méthode d'ordination. En sciences humaines, on dit sériation. Mais une ordination est-elle vraiment légitime ? Le modèle nul est peu vraisemblable, mais on ne peut le rejeter. A cause de la statistique ? A cause du modèle de l'indépendance ? A cause de la stratégie non paramétrique ? On pourrait permuter les présences-absences en gardant les sommes marginales. C'est l'objet du débat [6].

3 Distances et tests de Mantel

Un procédé très général est celui du test de Mantel. Il s'applique à l'origine pour tester le lien entre deux matrices de distances. Une structure est un ensemble de différences entre une série d'objets. Ces distances peuvent être observées (appréciation directe) ou calculées (après évaluation pour chaque objet de divers caractères). C'est particulièrement utile quand on utilise des marqueurs (des variables qui varient) et que seule une typologie induite entre objets est en cause [1]. Cette stratégie est particulièrement pratiquée en données sensorielles, génétique et écologie des communautés. Voir les fonctions :



- `dist.binary` (dissimilarités sur données binaires)
- `dist.prop` (distances entre profils)

- `dist.dudi` (distances euclidiennes dérivées des schémas de dualité)
- `dist.neig` (distances dérivées des graphes de voisinages)
- `dist.genet` (distances génétiques multi-loci)
- `dist.quant` (distances sur variables quantitatives, morphométrie)

Reprenons les 25 îles de l'exemple `cortes`. Calculons entre sites les distances spatiales :

```
dspat <- dist(xy)
```

Calculer entre sites les dissimilarités entre listes faunistiques (indice de Jaccard) :

```
dfau <- dist.binary(liz, 1)
```

On trouve une presentation détaillée du test de Mantel dans [2][p.70-75]. L'espace est connu par une matrice $\mathbf{S}$ de distances spatiales. Les données forment un tableau duquel on déduit une distance entre les individus consignée dans une matrice de distances $\mathbf{D}$. La correlation entre les deux est mesurée directement par $\sum_{i=1}^{n}\sum_{i=1}^{n} s_{ij}d_{ij}$

Les couples ii ne jouent aucun role puisque les distances sont nulles. Peu importe également que l'on compte une fois ou deux fois les couple ij et ji. Seul importe le type de permutations utilisées. Une des matrices est laissée en place et dans l'autre lignes et colonnes sont permutées a l'identique, par exemple :

$$25134 \Rightarrow \begin{bmatrix} 11 & 12 & 13 & 14 & 15 \\ 21 & 22 & 23 & 24 & 25 \\ 31 & 32 & 33 & 34 & 35 \\ 41 & 42 & 43 & 44 & 45 \\ 51 & 52 & 53 & 54 & 55 \end{bmatrix} \rightarrow \begin{bmatrix} 22 & 25 & 21 & 23 & 24 \\ 52 & 55 & 51 & 53 & 54 \\ 12 & 15 & 11 & 13 & 14 \\ 32 & 35 & 31 & 33 & 34 \\ 42 & 45 & 41 & 43 & 44 \end{bmatrix}$$

Pour chacune de m permutations de ce type, on calcule la statistique

$$\sum_{i=1}^{n}\sum_{i=1}^{n} s_{ij}d_{ij}$$

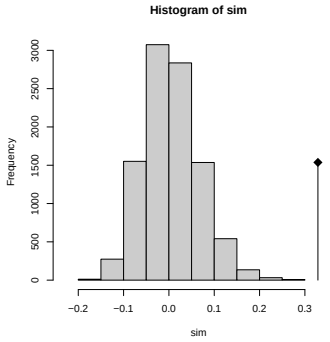
et on compare la valeur observée a l'ensemble des permutations. L'habitude veut que l'on corrige par les moyennes et les écarts-types pour faire apparaître exactement la corrélation entre les deux statistiques :

$$\begin{bmatrix} d_{21} & d_{31} & d_{32} & d_{41} & d_{42} & d_{43} & \ldots & d_{n1} & d_{n2} & \ldots & d_{n(n-1)} \\ s_{21} & s_{31} & s_{32} & s_{41} & s_{42} & s_{43} & \ldots & s_{n1} & s_{n2} & \ldots & s_{n(n-1)} \end{bmatrix}$$

```
w <- mantel.randtest(dspat, dfau, 9999)
w
```

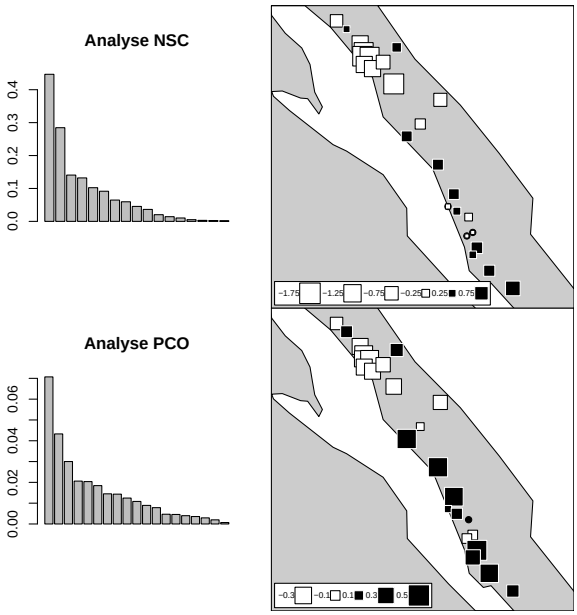
```
Monte-Carlo test
Call: mantel.randtest(m1 = dspat, m2 = dfau, nrepet = 9999)
Observation: 0.3287427

Based on 9999 replicates
Simulated p-value: 1e-04
Alternative hypothesis: greater

    Std.Obs Expectation    Variance
5.221511381 0.005758681 0.003826217
```

```
plot(w)
```



Sur une logique radicalement différente, l'information est maintenant claire. Il y a une structure statiale de la liste faunistique. Donc l'ordination est légitime :

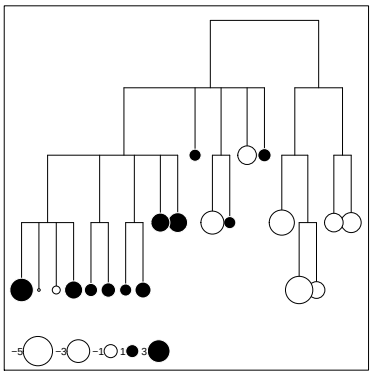
```
par(mfrow = c(2, 2))
w <- dudi.nsc(liz, scann = F)
barplot(w$eig)
title(main = "Analyse NSC")
bkgnd()
s.value(xy, w$li[, 1], add.p = T)
w <- dudi.pco(dfau, scann = F)
par(mar = c(5.1, 4.1, 4.1, 2.1))
barplot(w$eig)
title(main = "Analyse PCO")
bkgnd()
s.value(xy, w$li[, 1], add.p = T)
```



4 Orthogrammes

Plus particulier, mais avec des concepts voisins, la notion de test sur l'inertie phylogénétique. Retourner à l'example `galabiose`. Refaire la normalisation de la réponse :

```
phy1 <- newick2phylog(galabiose$tre, add = T)
n <- galabiose$rep$oui + galabiose$rep$non
trait <- galabiose$rep$oui/n
p0 <- sum(galabiose$rep$oui)/sum(galabiose$rep)
trait <- (trait - p0)/sqrt(p0 * (1 - p0)/n)
symbols.phylog(phy1, circ = trait, csi = 1.5)
```

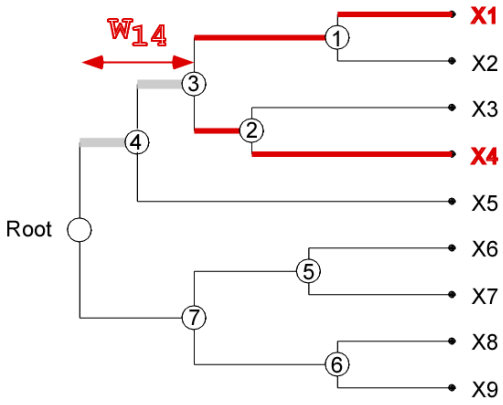


Etudier l'objet :

```
phy1
```

```
Phylogenetic tree with 20 leaves and 11 nodes
$class: phylog
$call: newick2phylog(x.tre = galabiose$tre, add.tools = T)
$tre: (((a,b)I1,((c,d)I2,e)I3)I...I7,(q,r,s,t)I8)I9)I10)Root;
        class   length content
$leaves numeric 20     length of the first preceeding adjacent edge
$nodes  numeric 11     length of the first preceeding adjacent edge
$parts  list    11     subsets of descendant nodes
$paths  list    31     path from root to node or leave
$droot  numeric 31     distance to root

         class      dim   content
$Wmat    matrix     20-20 W matrix : root to the closest ancestor
$Wdist   dist       190   Nodal distances
$Wvalues numeric    19    Eigen values of QWQ/sum(Q)
$Wscores data.frame 20-19 Eigen vectors of QWQ '1/n' normed
$Amat    matrix     20-20 Topological proximity matrix A
$Avalues numeric    19    Eigen values of QAQ matrix
$Adim    integer    1     number of positive eigen values of QAQ
$Ascores data.frame 20-19 Eigen vectors of QAQ '1/n' normed
$Aparam  data.frame 11-3  Topological indices for nodes
$Bindica data.frame 20-19 class indicator from nodes
$Bscores data.frame 20-19 Topological orthonormal basis '1/n' normed
$Blabels character  11    Nodes labelling from orthonormal basis
```

Vérifier la définition de la matrice W :



Donner la signification de la matrice de distance (distance phylogénétique) `Wdist` définie par :

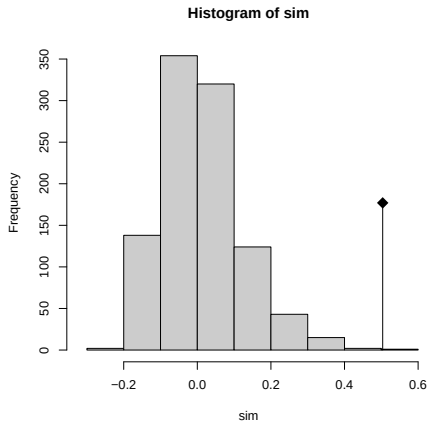
$$d_{ij} = \sqrt{w_{ii} + w_{jj} - 2w_{ij}}$$

Ceci renvoie à :

$$d_{ij} = \sqrt{\|f_i - f_j\|^2_{\mathbf{W}}} = \sqrt{\|f_i\|^2_{\mathbf{W}} + \|f_j\|^2_{\mathbf{W}} - 2\langle f_i \mid f\rangle_{\mathbf{W}}}$$

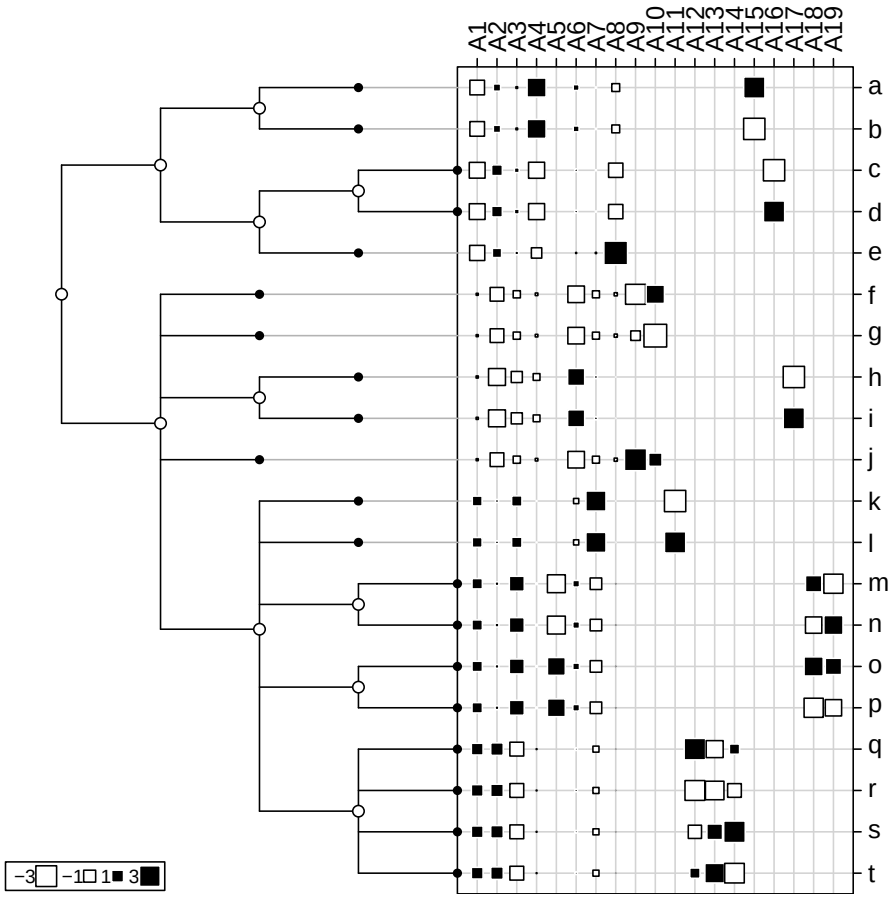
La distance phylogénétique est euclidienne. Le lien entre distances phylogénétiques et distances observées est très significatif :

```
plot(mantel.randtest(phy1$Wdist, dist(trait)))
```



Examiner la notion de scores canoniques :

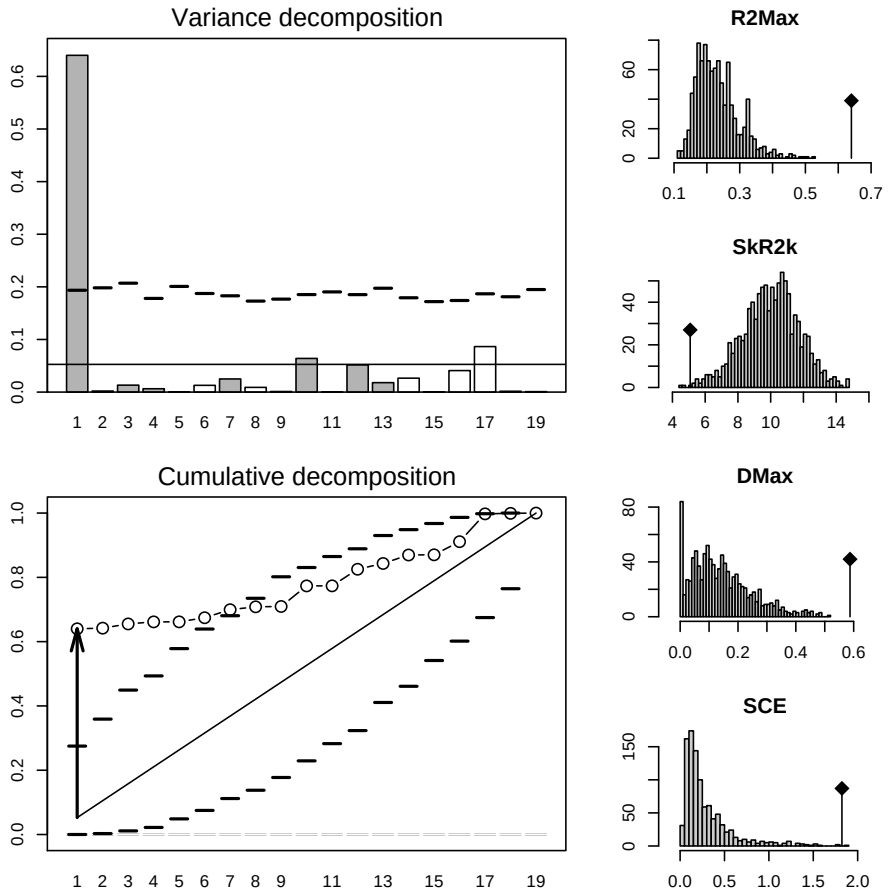
```
table.phylog(phy1$Ascores, phy1)
```



et son usage dans la fonction `orthogram` [3]

```
 orthogram(trait, phy1$Ascore)
class: krandtest
Monte-Carlo tests
Call: orthogram(x = trait, orthobas = phy1$Ascore)
Test number:   4
Permutation number:   999
   Test       Obs   Std.Obs     Alter Pvalue
1 R2Max 0.6400153  6.302764   greater  0.001
2 SkR2k 5.0848195 -2.844721      less  0.003
3  Dmax 0.5873837  4.264868 two-sided  0.001
4   SCE 1.8228286  5.447539   greater  0.002

other elements: NULL
```

Un orthogramme est la représentation de la décomposition de la variance d'une variable en composantes définies par une base orthonormée de dimension $n-1$ (la droite des constantes définit la dimension restante). Ceci se comprend comme extension du modèle linéaire. L'ordre des vecteurs de la base a un sens (ici l'ancienneté et l'importance de division dans l'arbre). L'orthogramme est une suite de carrés de corrélation qui indique où se définit l'inertie phylogénétique. L'accumulation de la variance expliquée, qui croit régulièrement quand la phylogénie n'entretient aucune relation avec le trait, se fait soit en peu d'endroits marqués (mutation conservée) soit de façon continue (modèle autorégressif). On trouvera des exemples reproductibles dans l'article cité.



On est passé dans la description de structure avec l'assurance, si besoin est, qu'une structure est à décrire. On pourra encore s'intéresser à la matrice de proximité topologique (on n'utilise que l'existence des ancêtres communs) dans la composante `Amat`. Pour comprendre cette dernière illustration, faire l'exercice suivant.

```
a <- "((a,b)A,(c,d,(e,f)B)C)D;"
wa <- newick2phylog(a)
wx <- enum.phylog(wa)
dim(wx)
```

```
[1] 48  6
```

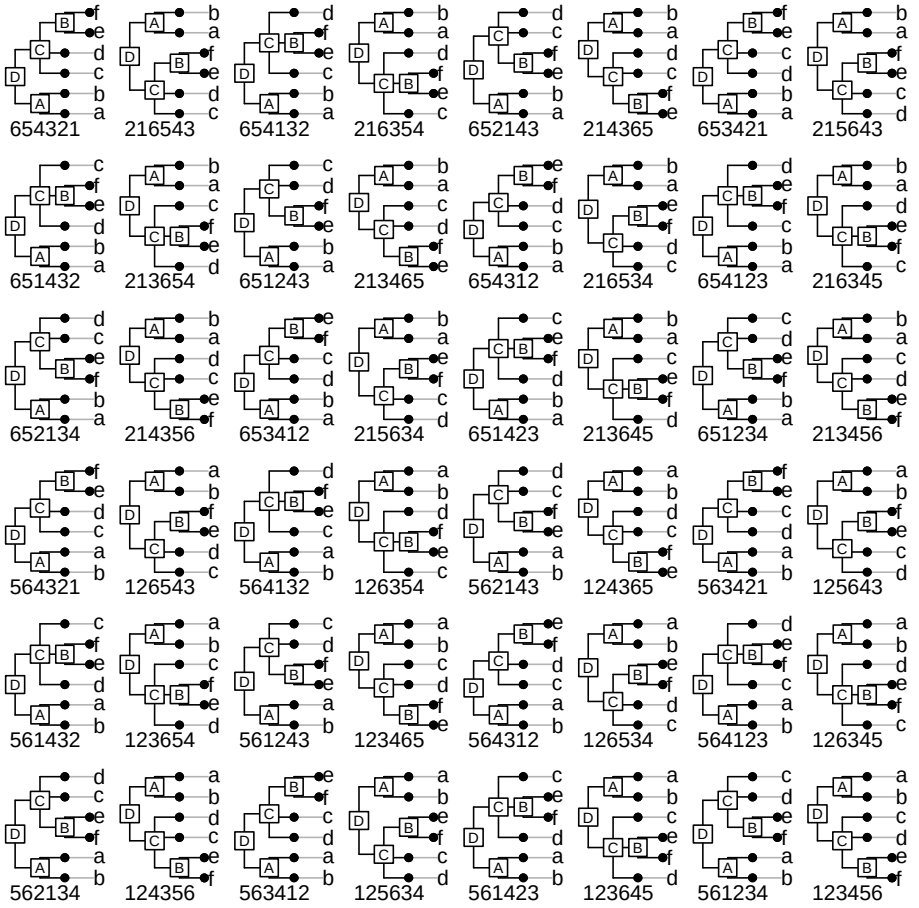
```
par(mfrow = c(6, 8))
fun <- function(x) {
    w <- NULL
    lapply(x, function(y) w <<- paste(w, as.character(y), sep = ""))
    plot(wa, x, clabel.n = 1.25, f = 0.75, clabel.l = 2, box = FALSE,
        cle = 1.5, sub = w, csub = 2)
    invisible()
}
invisible(apply(wx, 1, fun))
par(mfrow = c(1, 1))
wa$Adist
```

```
NULL
```

On a énuméré toutes les permutations des feuilles compatibles avec la phylogénie. Il y en a ici 48. Ce nombre grandit très vite.

```
prod(factorial(unlist(lapply(wa$parts, length))))
```
```
[1] 48
```
```
prod(factorial(unlist(lapply(phy1$parts, length))))
```
```
[1] 88473600
```

Quand on représente une phylogénie, on utilise une des très nombreuses possibilités de le faire. On peut représenter dans l'exemple toutes ces possibilités :



Ceci induit la possibilité de caractériser la proximité de deux feuilles par la proportion des possibilités qui place une des feuilles juste au dessus de l'autre. Vérifier :

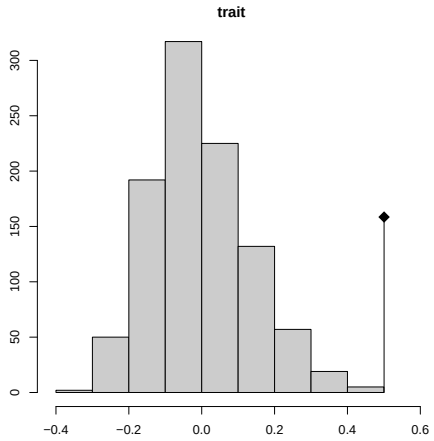
```
wa$Amat
```
```
          1         2         3         4         5         6
a 0.2500000 0.5000000 0.0833333 0.0833333 0.0416667 0.0416667
b 0.5000000 0.2500000 0.0833333 0.0833333 0.0416667 0.0416667
c 0.0833333 0.0833333 0.1666667 0.3333333 0.1666667 0.1666667
d 0.0833333 0.0833333 0.3333333 0.1666667 0.1666667 0.1666667
e 0.0416667 0.0416667 0.1666667 0.1666667 0.0833333 0.5000000
f 0.0416667 0.0416667 0.1666667 0.1666667 0.5000000 0.0833333
```

Cette matrice a une propriété fondamentale. Laquelle ? Utilisée comme matrice de proximité, elle permet les tests de Geary et de Moran. On trouvera tous les détails p. 30-37 dans :

http://pbil.univ-lyon1.fr/R/stage/stage8.pdf

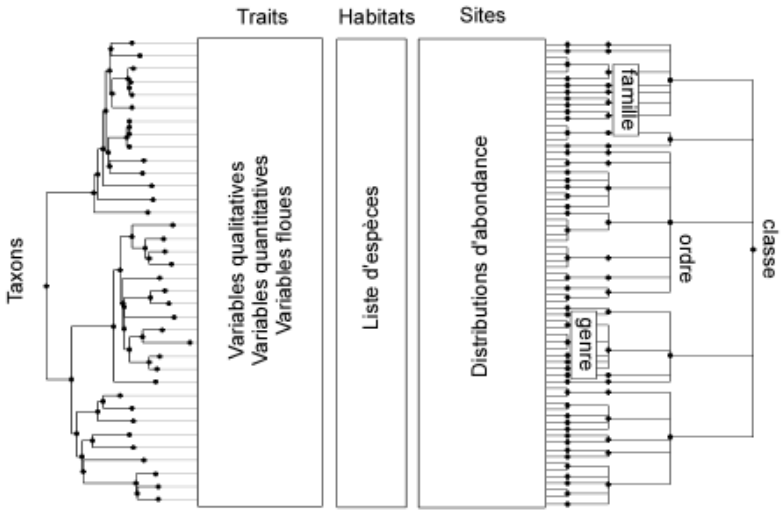
Les deux stratégies sont strictement identiques dans le cas des matrices bistochastiques. C'est donc un cas idéal :

```
plot(gearymoran(phy1$Amat, as.data.frame(trait)))
```



Il y a très souvent plusieurs manières d'obtenir un même résultat. Mieux vaut être prudent.

Les phylogénies font maintenant partie des structures de données utilisables en biologie évolutive à l'intérieur de l'ensemble des outils statistiques. Au centre de ces progrès : la librairie `ape` et l'ouvrage d'E. Paradis [4]. `ape` explore les modèles phylogénétique tandis qu'`ade4` est orienté vers les structures de données multivariées du type :



Références

[1] P. Legendre and L. Legendre. *Numerical ecology.* Elsevier Science BV, Amsterdam, 2nd english edition edition, 1998.

[2] B.F. Manly. *Multivariate Statistical Methods. A primer. Second edition.* Chapman & Hall, London, 1994.

[3] S. Ollier, P. Couteron, and D. Chessel. Orthonormal transform to detect and characterize phylogenetic signal. *Biometrics*, 62(2) :471–477, 2006.

[4] E. Paradis. *Analysis of Phylogenetics and Evolution with R.* Springer, New York, 2006.

[5] N. G. Yoccoz. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America*, 72 :106–111, 1991.

[6] A. Zaman and D. Simberloff. Random binary matrices in biogeographical ecology - instituting a good neighbor policy. *Environmental and Ecological Statistics*, 9 :405–421, 2002.