Consultation statistique avec le logiciel 

Comment saisir une phylogénie au format Newick ?

D. Chessel

11 juillet 2006

Les phylogénies, compromis de connaissances sur les relations de parentés entre espèces d'un groupe donné, font maintenant partie des corpus de données écologiques. On fournit ici quelques indications pour saisir ce type d'information.

Table des matières

1 Introduction

La fiche a été écrite à l'intention des membres d'un groupe de travail animé par Jorge Rabinovitch (**C**entro de **E**studios **P**arasitológicos y de **Ve**ctores , Universidad Nacional de La Plata, Argentina). Elle peut servir à d'autres. Parmi les données analysées en écologie, on trouve maintenant des phylogénies, modèles plus ou moins précis des relations de parenté dans l'arbre de l'évolution.

On se propose ici de montrer comment on saisit une phylogénie au format Newick. C'est un travail assez long et, comme il faut souvent s'y remettre plusieurs fois, il est utile d'en indiquer la méthode. Supposons que la phylogénie à implanter soit disponible dans un article comme celui de V. Hypsa *et al.* (2002) [2, p. 453]. On part donc de la figure 1.

Fig. 4. A preferred phylogeny of the Triatominae (see Discussion). Newly proposed taxonomic combinations are boldfaced (for discussion on the generic names used in the tree see Systematic implications). SA, SA *Triatoma* clade; 1, *T. infestans* complex; 2, *T. circummaculata* complex; 3, *T. sordida* complex.

FIG. 1 – Source bibliographique initiale de la phylogénie à implanter.

2 Etablir un code espèce

Avant tout autre opération, il convient d'établir un code simple pour l'ensemble des espèces dont on voudra parler. Il vaut mieux ne pas négliger ce point de départ tant la manipulation des noms d'espèce est source de fautes de frappe et de difficultés de toutes sortes. Il vaut mieux faire ces codes dans R. Supposons par exemple que le travail en cours utilise la liste des espèces suivante :

```
binoms <- scan("http://pbil.univ-lyon1.fr/R/donnees/qri.txt", character(),
    sep = "\n")
length(binoms)
```

```
[1] 123
```

```
head(binoms)
```

```
[1] "Alberprosenia goyovargasi" "Alberprosenia malheiroi"
[3] "Belminus costaricensis"    "Belminus herreri"
[5] "Belminus laportei"         "Belminus peruvianus"
```

Pour coder cette liste, dans le but de saisir la phylogénie, on peut faire simplement un code à 4 caractères :

```
cod4 <- paste("e", gsub(" ", "0", formatC(1:123, wi = 3)), sep = "")
head(cod4)
```

```
[1] "e001" "e002" "e003" "e004" "e005" "e006"
```

Pour faire un code plus habituel en écologie, dans le but d'illustrer une carte factorielle, on utilisera une abréviation du nom de genre avec une abréviation du noms d'espèce :

```
genre <- unlist(lapply(binoms, function(x) strsplit(x, " ")[[1]][1]))
espece <- unlist(lapply(binoms, function(x) strsplit(x, " ")[[1]][2]))
codlit <- paste(abbreviate(genre, 3), abbreviate(espece, 3), sep = "")
cbind.data.frame(binoms, cod4, codlit)[c(1:4, 120:123), ]
```

```
                      binoms cod4  codlit
1   Alberprosenia goyovargasi e001  Albgyv
2     Alberprosenia malheiroi e002  Albmlh
3      Belminus costaricensis e003 Blmcstr
4            Belminus herreri e004  Blmhrr
120           Triatoma venosa e120  Trtvns
121        Triatoma vitticeps e121  Trtvtt
122         Triatoma williami e122  Trtwll
123     Triatoma wygodzinskyi e123  Trtwyg
```

On pourra toujours faire des échanges de codes à la demande.

3 Préciser l'arbre à enregistrer

Reporter ensuite sur la figure de départ les numéros des espèces retenues. On notera sur la figure 2 quelques-uns des problèmes rencontrés.

1. Les cinq espèces du haut ne figurent pas dans la liste de référence : la racine est déplacée.
2. Il y a des espèces qui sont dans la zone de saisie mais qui ne figurent pas dans le code ; elles sont éliminées.

3. Un doute pour *Rhodnius tertius*. Dans le code, il y a *Psammolestes tertius*. Problème de synonymes ou rien à voir ? Elle est éliminée.
4. Un doute pour *Triatoma circummaculata* sur la figure. On a *Triatoma circunmaculata* dans le code de référence. On penche pour la faute de frappe et on remplace.
5. Un doute pour *Linshcosteus sp.*. On a 4 espèces dans la liste. On mettra un râteau avec les 4 espèces *Linshcosteus carnifex*, *Linshcosteus chota*, *Linshcosteus confumus* et *Linshcosteus costalis*.

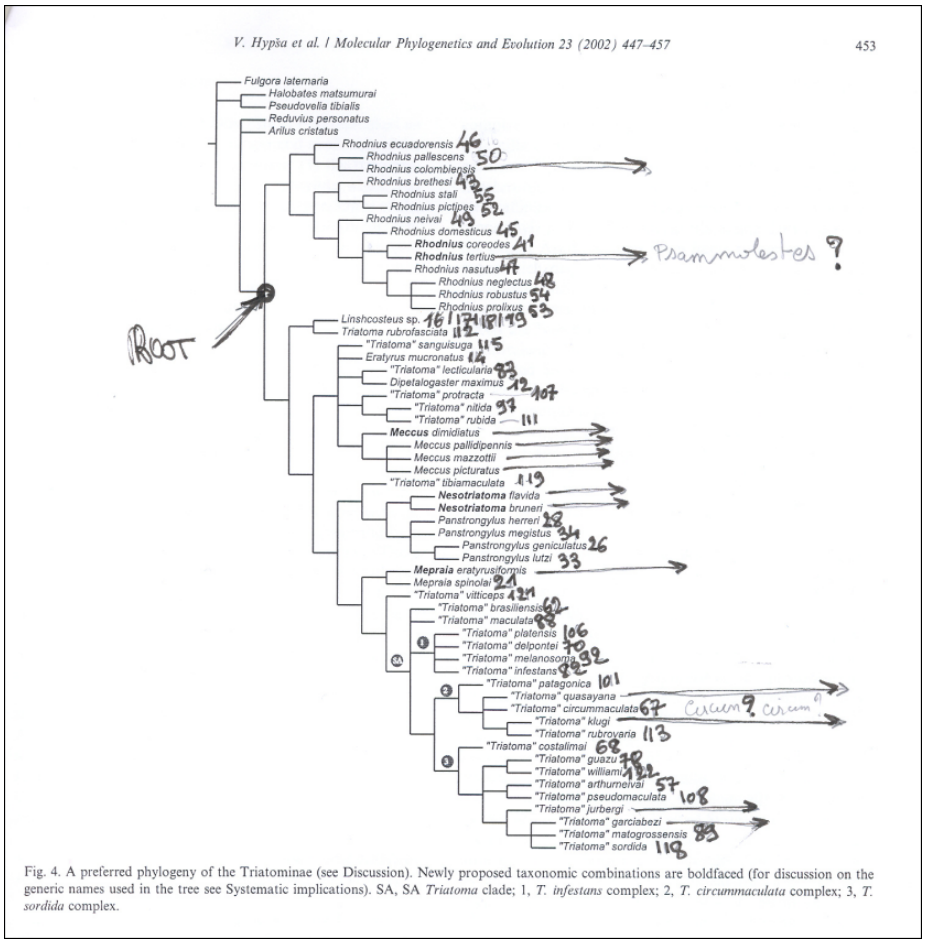


Fig. 4. A preferred phylogeny of the Triatominae (see Discussion). Newly proposed taxonomic combinations are boldfaced (for discussion on the generic names used in the tree see Systematic implications). SA, SA *Triatoma* clade; 1, *T. infestans* complex; 2, *T. circummaculata* complex; 3, *T. sordida* complex.

FIG. 2 – Phylogénie à saisir.

On doit donc enregistrer une phylogénie à 47 espèces.

4 Utiliser un éditeur de phylogénies

Télécharger ensuite un programme qui trace des phylogénies comme `TreeView` [3]. Toute information à :

`http://taxonomy.zoology.gla.ac.uk/rod/treeview/treeview_manual.html`

Dézipper `treev32.zip` et exécuter le *setup*.

Pour les débutants, faire ce petit exercice. Copier le texte ci-dessous entre la première parenthèse et le point-virgule de la fin, le mettre dans un fichier texte et sauvegarder en `example.tre`. Ouvrir ce fichier avec `Treeview.exe`.

```
(((((((Prorocentrum:0.0536, Tetrahymena:0.0721)45:0.0038,((Oryza:0.0135,
(Lycopersicon:0.0046,Citrus:0.0060)84:0.0025)100:0.0414,Saccharomyces:
0.0494)55:0.0051)87:0.0102,(((Homo:0.0009,Mus:0.0019)100:0.0065,
Xenopus:0.0093)100:0.0448, (Drosophila:0.0863,Caenorhabditis:0.0843)
51:0.0029)66:0.0070)73:0.0062,Dictyostelium:0.1226)92:0.0132,
Crithidia:0.1403)65:0.0095,Physarum:0.1347):0.1191, ((((Rhodobacter:0.1032,
(Pseudomonas:0.0468, (Escherichia:0.0399,Ruminobacter:0.0553)81:0.0103)100:0.0325)
100:0.0301,(Leptospira:0.0932, (((Micrococcus:0.0380,Streptomyces:0.0478)
100:0.0654,Pirellula:0.1269)27:0.0014, ((Anacystis:0.0490,EuglenaCP:0.1184)
100:0.0341,(BacillusSubt:0.0262, BacillusStea:0.0184)100:0.0504)86:0.0148)
33:0.0013)70:0.0073)93:0.0161,Thermus:0.1043)100:0.1052, (((Desulfurococcus
:0.0412,Sulfolobus:0.0710)97:0.0145,Thermoproteus:0.0880)100:0.0521,
((Methanococcus:0.0927,Methanobacterium:0.1006)99:0.0282,(Thermoplasma:0.1526,
(HalobacteriumH:0.0515,(Halococcus:0.0543,HalobacteriumM:0.0532)50:0.0030)
100:0.1324)39:0.0027)72:0.0118)96:0.0130)100:0.0634);
```

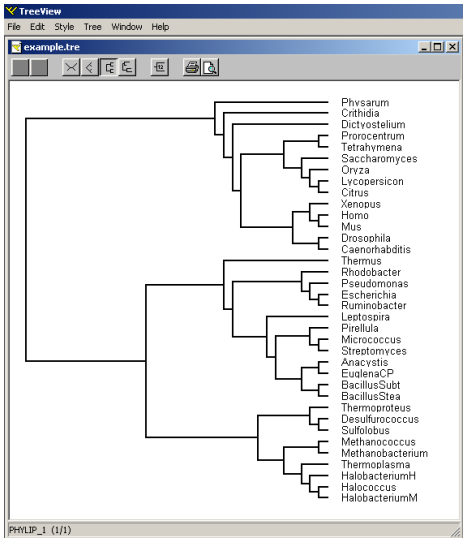


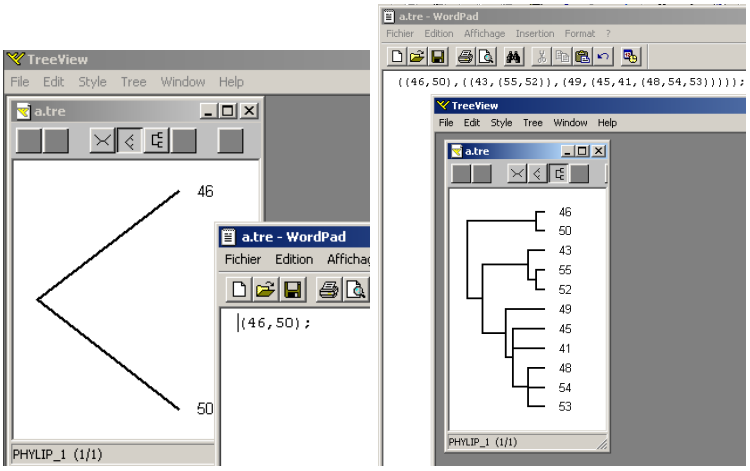
FIG. 3 – `Treeview.exe` dessine des phylogénies au format `Newick`.

5 Ecriture du format Newick

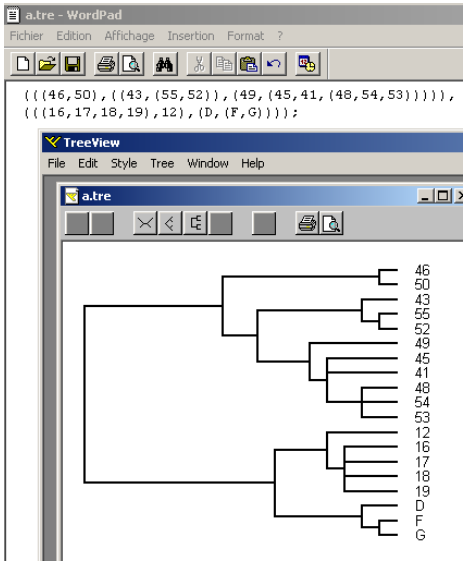
Ce format est une convention d'écriture des arbres avec des virgules et des parenthèses. Toute information à :

`http://evolution.genetics.washington.edu/phylip/newicktree.html`

Ouvrir simultanément le programme graphique et un éditeur de texte (fichier `a.tre`) pour tracer l'arbre au fur et à mesure de sa construction :



Introduire des lettres pour désigner provisoirement des branches qui seront remplacées par des sous-arbres :



...
On peut introduire des retours à la ligne comme on veut. Compléter le code.

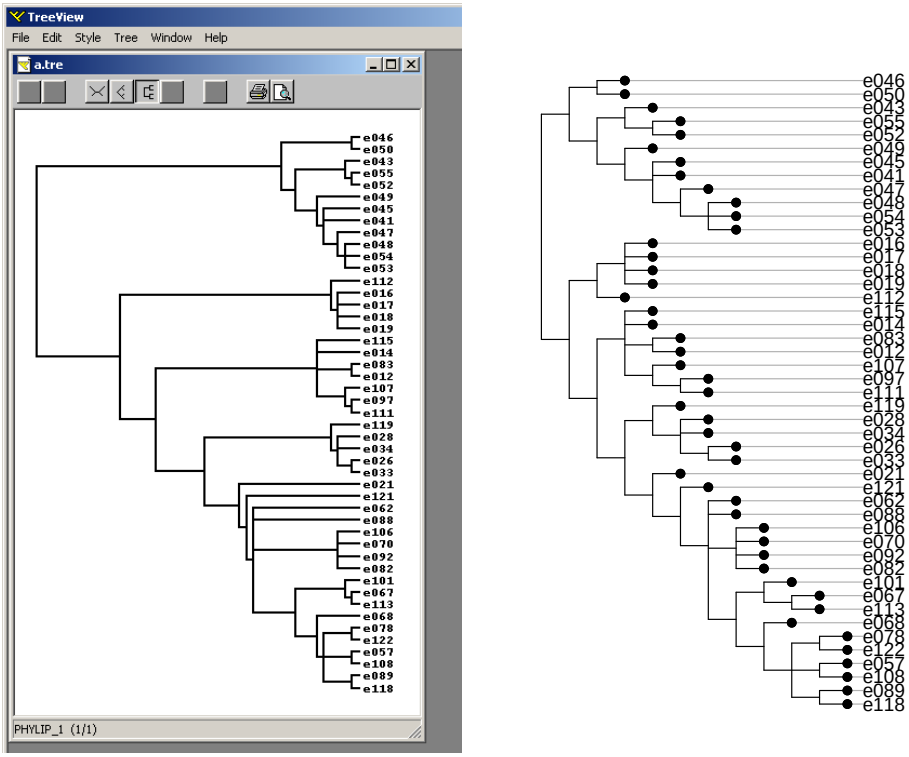


FIG. 4 – Représentation graphique de la phylogénie. A droite dans `treeview.exe`, à gauche dans `ade4`. Noter la différence d'interprétation en cas d'absence de longueurs de branches.

On arrive à la figure 4 à gauche. Le fichier texte contient :

```
(((e046,e050),((e043,(e055,e052)),(e049,(e045,e041,(e047,(e048,e054,e053))))))),
(((e016,e017,e018,e019),e112),((e115,e014,(e083,e012),(e107,(e097,e111))),
((e119,(e028,e034,(e026,e033))),(e021,(e121,(e062,e088,(e106,e070,e092,e082),
((e101,(e067,e113)),(e068,((e078,e122),(e057,e108),(e089,e118))))))))))));
```

Lire alors ce fichier texte dans R.

```
a.tre <- readLines("a.tre")
```

Vérifier et éditer (figure 5 à droite) :

```
a.tre
```

```
[1] "(((e046,e050),((e043,(e055,e052)),(e049,(e045,e041,(e047,(e048,e054,e053))))))),"
[2] "(((e016,e017,e018,e019),e112),((e115,e014,(e083,e012),(e107,(e097,e111))),"
[3] "((e119,(e028,e034,(e026,e033))),(e021,(e121,(e062,e088,(e106,e070,e092,e082),"
[4] "((e101,(e067,e113)),(e068,((e078,e122),(e057,e108),(e089,e118))))))))));"
```

```
library(ade4)
phy1 <- newick2phylog(a.tre, add = F)
plot(phy1, f = 0.75)
```

Changer les étiquettes des feuilles :

```
w <- a.tre
for (k in 1:123) w <- gsub(cod4[k], codlit[k], w)
phy2 <- newick2phylog(w, add = F)
plot(phy2, f = 0.75)
```

```
w <- a.tre
provi <- gsub(" ", ".", binoms)
for (k in 1:123) w <- gsub(cod4[k], provi[k], w)
phy3 <- newick2phylog(w, add = F)
plot(phy3, f = 0.5)
```

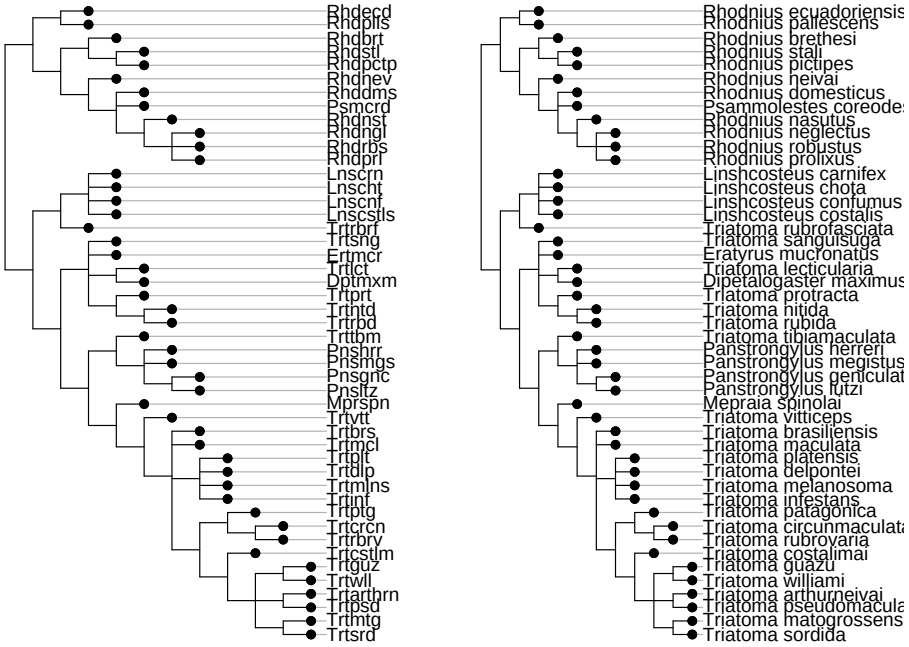


FIG. 5 – Représentation graphique de la phylogénie : changements d'étiquettes

6 Longueurs de branches

Quand elles sont disponibles, il est aisé de les reproduire. Par exemple, l'article de Bried et al. [1] contient un arbre valué sur 36 millions d'années. Compter et étiqueter avec soin les nœuds, indiquer les longueurs de branches en vérifiant la cohérence (figure 6).

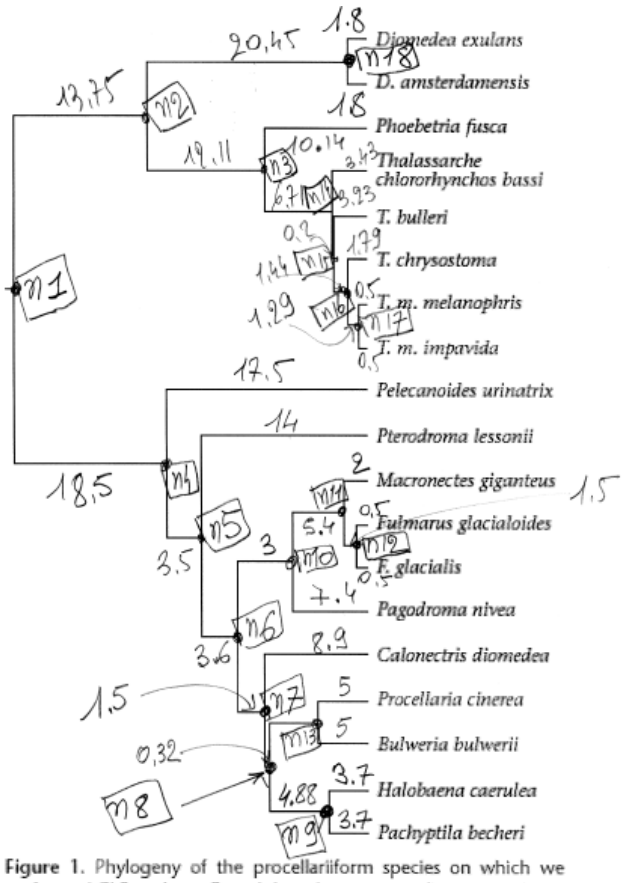


Figure 1. Phylogeny of the procellariiform species on which we performed GLS analyses. Branch lengths represent divergence times, assuming that the divergence between petrels and albatrosses occurred between 36 and 50 million years ago (Viot et al. 1993; Warham 1996). The tree represented here and the results shown in this paper assume a divergence time of 36 million years. We obtained similar results when considering a divergence time of 50 million years.

FIG. 6 – Préparation de la saisie d'une phylogénie avec longueur de branches.

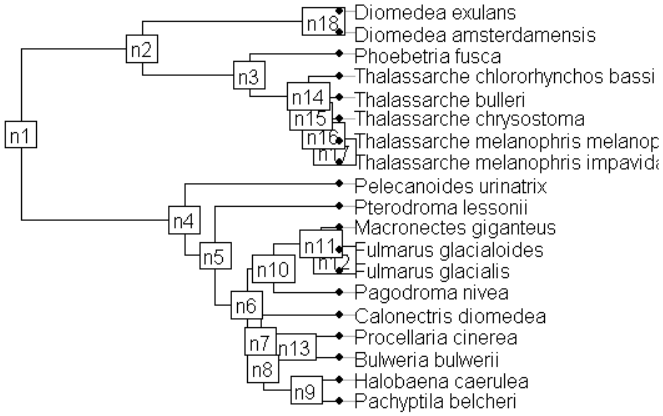
Utiliser ensuite la chaîne de caractères `XXX` suivie par `:xxx` pour associer la longueur `xxx` à la branche qui mène à l'objet (feuille ou nœud intérieur) `XXX`. Ainsi, le nœud `n9` est relié au nœud `n8` par une distance de `4.88` millions d'années ; le nœud `n8` est relié au nœud `n7` par une distance de `0.32` millions d'années ...

Le résultat est consigné dans `procella$tre` :

```
data(procella)
procella$tre[5]
```

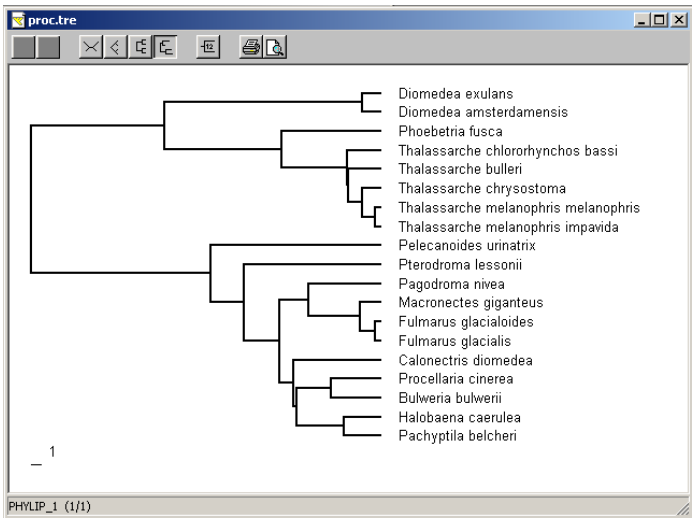
```
[1] "n9:4.88)n8:0.32)n7:1.5)n6:3.6)n5:3.5)n4:18.5)n1:0;"
```

```
procphy <- newick2phylog(procella$tre, add = F)
plot(procphy, clabel.n = 1)
```



Le format Newick permettra les échanges entre logiciels utilisant ce format. Remarquer que la convention qui remplace le blanc du binom latin par un underscore mais qui édite correctement les noms d'espèces est utilisé dans `ade4`, `TreeView` et `ape` [4].

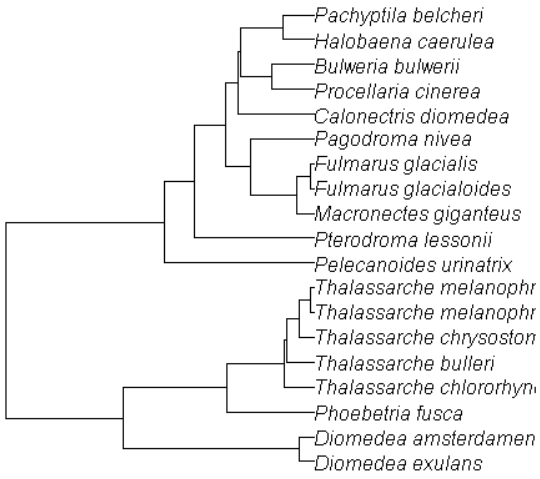
```
write(procella$tre, file = "proc.tre")
```



```
library(ape)
```

```
Le chargement a nécessité le package : gee
Le chargement a nécessité le package : nlme
Le chargement a nécessité le package : lattice
```

```
plot(read.tree("proc.tre"))
```



Références

[1] J. Bried, D. Pontier, and P. Jouventin. Mate fidelity in monogamous birds : a re-examination of the procellariiformes. *Animal Behaviour*, 65 :235–246, 2002.

[2] V. Hypsa, D.F. Tietz, J. Zrzavy, R.O. Rego, C. Galvao, and J. Jurberg. Phylogeny and biogeography of triatominae (hemiptera : Reduviidae) : molecular evidence of a new world origin of the asiatic clade. *Molecular Phylogenetics and Evolution*, 23 :447–457, 2002.

[3] R. D. M. Page. Treeview : An application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences*, 12 :357–358, 1996.

[4] E. Paradis, K. Strimmer, J. Claude, G. Jobb, R. Opgen-Rhein, J. Dutheil, Y. Noel, and B. Bolker. ape : Analyses of phylogenetics and evolution. r package version 1.8-2, 2006.