

An introduction to the phylotools package

Jinlong Zhang, Xiangcheng Mi, Nancai Pei

Oct 2010

Contents

1	Introduction	1
2	How to install phylotools	2
3	Data Input	2
4	Data Output	2
5	Step by step guide building a supermatrix	3
5.1	Step 1: Copy the phy files to working directory	3
5.2	Step 2 Set working directory	3
5.3	Step 3 Build supermatrix using supermat	4
5.4	Step 4 Save the supermatrix to a Phylip file	4
6	Literature cited	4

1 Introduction

”phylotools” is an R package for construction of ”supermatrix” which will be used in the further analysis of phylogenies from DNA barcoding sequences. It is primary designed for creat supermatrix more easily in R. See Kress et al. 2010 for further information. The slashes at both ends of one sequence in the aligned sequences will be replaced by ”?”, the ones at the internal parts of the sequences will be retained.

2 How to install phylotools

The `phylotools` is available at CRAN mirror <http://cran.r-project.org/web/packages/>. To install the package, just type `"install.packages("phylotools")"`. The R will automatically download and install the package. Once it is installed, users have to type it can be loaded by typing:

```
> library(phylotools)
```

3 Data Input

User have to provide the aligned sequences in phyliip format. Sequences from the same species must have the same name in different phyliip files. For example: the sequences from *Ulmus pumila*, should have have same name in different alinment files. These files maybe generated by the software for sequence alignments, for example: ClustalW, ClustaX, MUSCLE.

4 Data Output

Output is a super matrix with unknown sites represent by "?". See Tabel 1 for detail. In order to build the supermatrix, users have to provide the files listed as follow.. (1) One file for *rbcLa*, in phyliip format (2) One file for *matK*, in phyliip format. (3) Files for *trnH-psbA* aligned sequences. The *trnH-psbA* sequences evolve much more rapidly and often aligned by species in on order or even family. Users for realy data may have tens of *trnH-psbA* aligned files in phyliip format.

Table 1. The format of output supermatrix.

species	rbclA	matK	trnH-psbAI	trnH-psbAII	trnH-psbAIII
species1	ACTTCCGGTCAACTTTTA	CCTAAGGTTCTCTAGACAACA	ACTTAAGGCCTTTACC	?????????????	????????????????????
species2	ACTTCCGGTCAACTTTTA	CCTAAGGTTCTCTAGACAACA	ACTTAAGGCCTTTA??	?????????????	????????????????????
species3	??TTCAGGTCAACTTTTA	CCTAAGGTTCTCTAGACAACA	?CTTAAGGCCTTTACC	?????????????	????????????????????
species4	ACTTCCGGTCAACTTTTA	CCTAAGGTTCTCTAGACAACA	ACTTAAG---TTTACC	?????????????	????????????????????
species5	ACTTCCGGTCAACTTTTA	CCTAAG---CTCTAGACAACA	??TAAGGCCTTTACC	?????????????	????????????????????
species6	ACTTCCGGTCAACTTTTA	CCTAAGGTTCTCTAGACAACA	ACTTAAGGCCTTTACC	?????????????	????????????????????
species7	ACTAOCGGTCAACTTTTA	CCTAAGGTTCTCTAGACA???	????????????????	CTCTCAATTGGCA	????????????????????
species8	ACTTCCGGTCAACTTTTA	CCTAAGGTTCTCTAGACAACA	????????????????	CTCTCAA-TGGCA	????????????????????
species9	ACTTCCGGTCAACTTTTA	??TAAGGTTCTCTAGACAACA	????????????????	CTCTCAATTGTCA	????????????????????
species10	ACTTCCGGTCAACTTTTA	CCTAAGGTTCTC---ACAACA	????????????????	CTCT---TGGCA	????????????????????
species11	ACTTCCGTCTCAACTTTTA	CCTAAGGTTCTCTAGACAACA	????????????????	CT-TCAATCGGCA	????????????????????
species12	ACTTCCGGTCAACTTTTA	CCTAAGGTTCTCTAGACAACA	????????????????	CTCTCAATTAGCA	????????????????????
species13	??TCCGGTCAACTTTTA	CCTAAGGTTCTCTAGACAACA	????????????????	CTCTCACTGGCA	????????????????????
species14	ACTTCCGGTCAACTTTTA	????????????????????	????????????????	????????????????	TT----ATTAACTCTTACA
species15	ACTTCCGGTCAACTTTTA	CCTAAGGTTCTCTAGACAACA	????????????????	????????????????	TTCCGGATTAACTCTTACA
species16	ACTTCCGGTCAACTTTTA	CCTAAGGTTCTCTAGACAACA	????????????????	????????????????	??CCGGATT-----TACA
species17	ACTTCCGGTCAACTTTTA	CCTAAGGTTCTCTAGACAACA	????????????????	????????????????	TTCCGGATTAACTCTTACA
species18	ACTT---GTCAACTTTTA	CCTAAGGTTCTCTAGAC???	????????????????	????????????????	TTCCGGATTAACTCTT???
species19	ACTTCCGGTCCACTTTTA	?????GTCTCT---GACAACA	????????????????	????????????????	????????????????????
species20	ACTTCCGGTCAACTTTTA	CCTAAGGTTCTCTAGACAACA	????????????????	????????????????	????????????????????

5 Step by step guide building a supermatrix

Assume we have the aligned files in phylip format named :

"rbclA.phy", "matK.phy", "trn1.phy", "trn2.phy", "trn3.phy", "trn4.phy"

Users may follow the steps to build the supermatrix and save the results to file:

5.1 Step 1: Copy the phy files to working directory

Copy ALL the .phy files to the same directory. Here we use the data provided in the package in directory "extdata"

5.2 Step 2 Set working directory

type the following command to set the working directory the files exists.

```
> dir <- system.file("extdata", package = "phylotools")
> setwd(dir)
```

5.3 Step 3 Build supermatrix using supermat

Use the function "supermat()" to build a "super" matrix representing the relationships between the sequences. Type:

```
> supermat <- supermat(rbcl = "rbcla.phy", matk = "matK.phy",  
+   trn = c("trn1.phy", "trn2.phy", "trn3.phy", "trn4.phy"))
```

5.4 Step 4 Save the supermatrix to a Phylip file

Save the supermatrix to file

```
> write.mat(supermat, "result.phy")
```

6 Literature cited

- Kress W., Erickson D., Jones F., Swenson N., Perez R., Sanjur O., Bermingham E., Plant DNA barcodes and community phylogeny of a tropical forest dynamics plot in Panama. Proceedings of the National Academy of Sciences of the United States of America. 2009 18621-18626