SIMIL version 1.0 - User's Manual

Supplementary material of:

SIMIL: an R (CRAN) scripts collection for computing genetic structure similarities based on STRUCTURE 2 outputs.

Alvarez Nadir[1], Arrigo Nils[2], IntraBioDiv Consortium[3].

[1] Laboratoire d'Entomologie Evolutive, Institut de Biologie, Université de Neuchâtel, 11 rue Emile-Argand, CH-2000 Neuchâtel, Suisse.

[2] Laboratoire de Botanique Evolutive, Institut de Biologie, Université de Neuchâtel, 11 rue Emile-Argand, CH-2000 Neuchâtel, Suisse.

[3] members of the consortium: www.intrabiodiv.eu/IMG/pdf/IntraBioDiv_Consortium_v10.pdf

Index

1 Installing SIMIL

1.1 Installing R (CRAN)

Download and install R (CRAN) from http://lib.stat.cmu.edu/R/CRAN/.

Check that your version of R (CRAN) includes the following libraries: "tcltk" and "combinat". A simple way to check it is to copy-paste the following code lines (the code lines are indicated in **red** in the present manual) in the console of R (CRAN):

```
require(tcltk)
require(combinat)
```

R(CRAN) will output "TRUE" for each library if it is already installed on your computer (Fig.1).

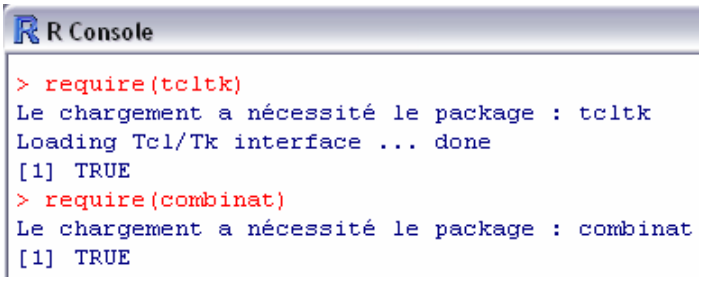

Figure 1 Screenshot of the R Console. This is the windows where commands must be copy-pasted. In red: the commands that were previously typed by the user (the ">" is added automatically and is not part of the command), in blue: the answers of R(CRAN).

If R(CRAN) outputs "FALSE" to your request, you need to install one or both libraries. To perform this, you may use the installing menu of R(CRAN): "Packages / Install Packages" and select the needed library in the proposed list (Fig.2).

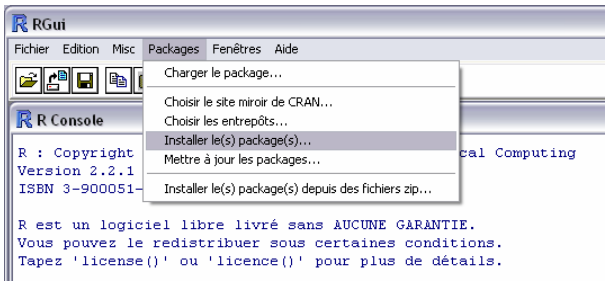

Figure 2 Installing a new package in R(CRAN).

1.2 Starting SIMIL

Download SIMIL at http://www2.unine.ch/webdav/site/ebolab/shared/Programs/SIMIL.zip and unzip it. The zip file contains several files:

- The present Manual file.
- "SIMIL.collection.txt" – is the code of SIMIL.
- Structure outputs: "Cerastium.3_f", "Cirsium.3_f" and "Luzula.3_f".
- ".pop" files (one file per species): "Cer.pop", "Cir.pop" and "Luz.pop".

Open the "SIMIL.collection.txt" file, and copy-paste the whole code that it contains in the R Console. Use SIMIL via its Graphical User Interface (GUI).

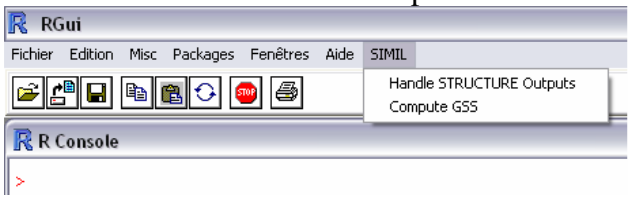

Figure 3 The SIMIL Graphical User Interface.

2 Using SIMIL

SIMIL is a program written in the R(CRAN) language that use input files produced by the program STRUCTURE (versions 2.0 to 2.2, Pritchard et al. http://pritch.bsd.uchicago.edu/structure.html).

Two main features are implemented:

- Exporting probability tables stored in the files produced by STRUCTURE.
- Computing the Genetic Structure Similarity index (abbreviated "GSS index" in the present manual) between two species as described in Alvarez and Arrigo 2007.

2.1 Project organisation and requirements

The analysis uses STRUCTURE outputs that satisfy the three following conditions:

- The whole dataset must be organised in a way that allows comparisons between the tested species (i.e., the matrix lines should represent elements that belong to a common reference system for the whole dataset, e.g. lines may represent a population, a spatial site, or a cell within a common geographical grid). This implies that both species respect a same nomenclature.
- Both tested species must share at least two common elements.
- The tested number of groups or genetic pools (i.e. the K value in STRUCTURE) must be the same for both species.

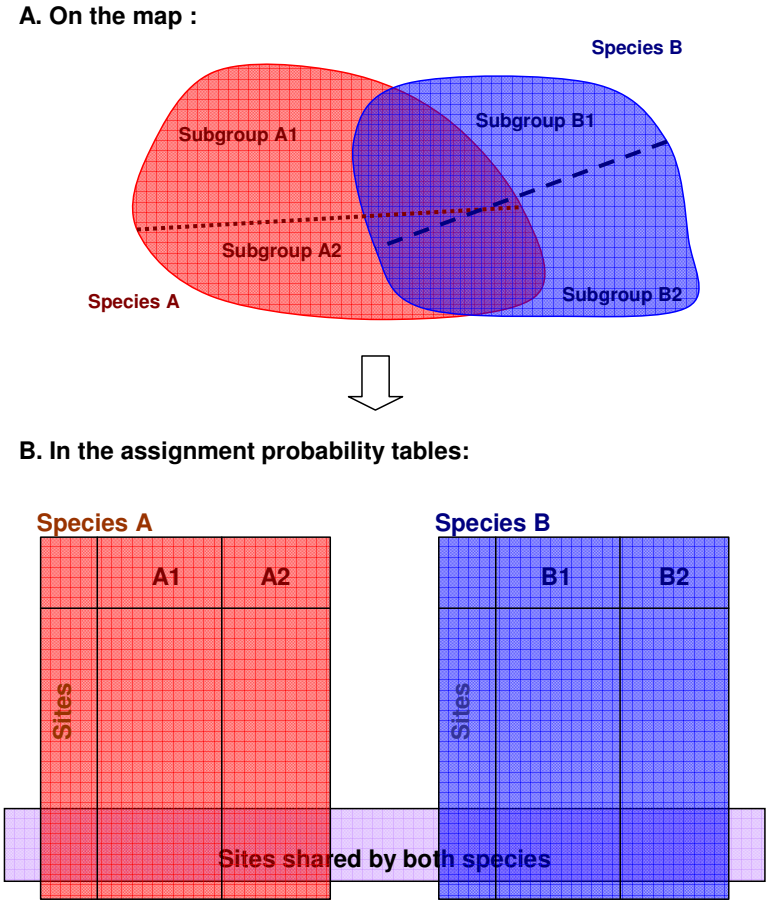

Figure 4 Example of dataset. This case study involves two species (namely A and B), each of them being split into two phylogeographical regions (K value = 2, species A divided into subgroups A1 and A2, species B idem, B1 and B2). Species A and B must have a common nomenclature system as this is the only way to identify their shared spatial sites.

2.2 Populations nomenclature

A common way to analyse genetic variation is to sample several individuals per population, and several populations per region. Therefore, a structuration analysis can be performed at two levels: the individual and / or the population level. If the population variable is provided, STRUCTURE calculates assignment probabilities for both levels and the two resulting assignment probability tables are stored in a same file.
SIMIL is designed to compute the GSS with the population probabilities table. The user has to set STRUCTURE in order to obtain that information.

STRUCTURE provides outputs including the population nomenclature and the analysis can be performed properly as long as the population nomenclature is the same for the different species.

However, SIMIL includes an option to rename the probability tables, directly from the STRUCTURE outputs. This option is especially convenient when STRUCTURE outputs were performed with different nomenclature from one analysed species to the other.

The user has to provide a ".pop" file that contains the list of sites (or populations) names. **A single file per species must be saved in the STRUCTURE outputs' folder.**
This file can be edited with a spread-sheet program and only contains one column, without column name. The site names are stored in that column, in the same order as they are arranged in the STRUCTURE output (consult example file).
The ".pop" file is edited with a spread-sheet program and saved in a tab-delimited format. It has to be renamed as follows:

- The suffix of the file name should begin with the **first three letters** of the names of the related STRUCTURE outputs.
- Its extension must be changed from "*.txt" to "*.pop".

Example:
STRUCTURE outputs : "Cerastium.3_f1", "Cerastium.3_f2" … to "Cerastium.3_f10"
Possible names for the "*.pop" file: **"Cer.pop"** or **"CerastiumAlpinum.pop"**.

The renaming option takes place as soon as a ".pop" file is present in the folder of the STRUCTURE outputs. The first lines of the SIMIL code attempts to detect this file; the option is not applied if the ".pop" file is absent.

2.3 Handling STRUCTURE outputs

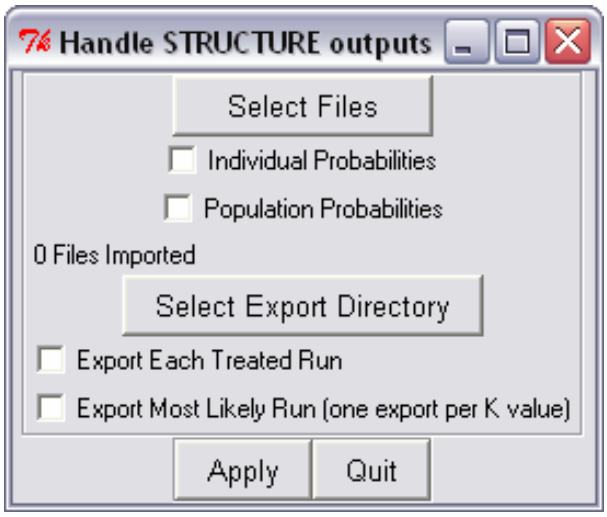

Figure 5 Screenshot of the exporting options.

1. Choose the STRUCTURE outputs to export with the "Select files" button. The user can either select one or several files at a same time.
2. Choose to export either individual and / or populations probabilities with the "Individual" or "Population Probabilities" check-buttons.
3. Select the exporting directory. R(CRAN) will create a new subdirectory (named "EXPORTS.IND" or "EXPORTS.POP" according to which probability table was selected) within the chosen directory.
4. Choose to export each selected file or only the most likely one (the run that maximizes the maximum likelihood criterion).

The exporting procedure builds a matrix containing: the number of the individual / population, the likelihood value of the run, the label given to the individual and / or to the population and the assignment probabilities (Table 1).

Table 1 Example of exported file (Population probability table).

	CodePop	EstLn	lab	P	P	P	NbIndivs
1	E27	-3221.1	01:00	0.986	0.009	0.004	3
2	F26	-3221.1	02:00	0.967	0.028	0.005	3
3	F28	-3221.1	03:00	0.983	0.014	0.004	3

The exported files are saved in a tab-delimited format.

Computing the GSS index

The SIMIL algorithm works as follows:

I. It selects the cells of species A and B that belong to the overlapping range of both species, by comparing the sites nomenclature.

II. It shuffles columns of the STRUCTURE output-matrices in order to maximize the level of homology between species. Indeed, STRUCTURE affiliates samples to groups in a Bayesian framework, in which related samples should always belong to the same group, whereas the ordering of groups (*i.e.* the columns of the output-matrices) might change from one run to the other. As a consequence, the two selected subsets must be reorganised in a way that allows probabilities comparisons. This homology research is done iteratively and can be time consuming, especially for large values of K. One subset (say species B) has its columns systematically shuffled, the mean absolute difference (Eq. 1) is then applied on the shuffled matrix and the other subset (species A). The columns order that minimizes this difference is considered to reflect the best homology between both subsets and kept for the following calculations.

III. It calculates the unweighted GSS index (Eq. 1). The unweighted GSS index ranges from zero (absolute difference between species A and B) to one (absolute similarity of patterns between species A and B). Three derivatives of the original index, demonstrating different kinds of weightings are proposed: A. weighting by the area of the species with the smaller sampling distribution (Eq. 2. - GSS.smallest), B. weighting by the area of the whole additive sampling distribution of both species (Eq. 3. - GSS.overall) and C. applying the arithmetic mean of GSS.smallest and GSS.overall (Eq. 4. - GSS.mean).

Eq. 1. $GSS = 1 - \frac{1}{n} \sum (\left| subsetA - subsetB_{shuffled} \right|)$

where n = number of cells of subsetA = number of cells of subsetB, subsetA = subset of species A, $subsetB_{shuffled}$ = subset of species B shuffled with the best column order.

Eq. 2. $GSS.smallest = GSS * \frac{n.overlap}{n.min}$

n.overlap = number of overlapping cells and n.min = number of cells of the species that shows the smallest distribution

Eq. 3. $GSS.overall = GSS * \frac{n.overlap}{n.tot}$

where n.overlap = number overlapping cells and n.tot = total number of cells in the dataset (n.tot = cells of species A + cells of species B – overlapping cells).

Eq. 4. $GSS.mean = \frac{GSS * n.overlap}{2} * \frac{n.tot + n.min}{n.tot * n.min}$

3 Using command lines

3.1 Opening a single STRUCTURE output

```
IMPSTRUCT.P(file)
IMPSTRUCT.I(file)
```

Used to import the individual probabilities (**IMPORTMULTI.I**) or the population probabilities (**IMPORTMULTI.P**) into R(CRAN) of a single STRUCTURE output

file = path of file to import. Typically generated with the **choose.files()** function.

Example:

```
path=choose.files(default = "", caption = "Select Structure Run Species A",multi
= TRUE, index = nrow(Filters))
table=IMPSTRUCT.P(path[1])
table
```

3.2 Computing the GSS

```
SIMIL()
```

This is an "all in one" function that only can compute the GSS index for a pair of species. This function always asks the user to select manually the files to compare and does not allow routines programming.

```
SIMIL.R(pathA,pathB)
```

This function allows routines programming, as pathA and pathB are the paths of files to compare. The function produces five elements: GSS.unweighted, GSS.smallest, GSS.overall, GSS.mean and PercentOverlap.

Example:

```
pathA=choose.files()
pathB=choose.files()
SIMIL.R(pathA,pathB)
```

Both **SIMIL** and **SIMIL.R** produce a list named "GSS.out" in the R(CRAN) working environment. "GSS.out" contains: the GSS calculations results (as presented above) AND the best shuffled matrix. This object is not showed during the procedure, but it can be visualised and exploited by calling it via the R Console: **GSS.out**

- Calling the GSS calculations: **GSS.out$ GSS**
- Calling the best shuffled matrix (subset$B_{shuffled}$ in Eq.1): **GSS.out$ SpB.shuffled**

3.3 Handling STRUCTURE outputs:

```
IMPORTMULTI.I(listfiles,path=getwd(),best=T)
IMPORTMULTI.P(listfiles,path=getwd(),best=T)
```

Used to import and export the individual probabilities (**IMPORTMULTI.I**) or the population probabilities (**IMPORTMULTI.P**) of several STRUCTURE outputs.

listfiles = list of files to import. Typically generated with the **choose.files()** function.
path = path used to save exported txt files
best = T (TRUE) or F (FALSE), exports only the most likely if best=T (default option).

3.4 Programming pairwise comparisons

This code is not included in the SIMIL scripts collection. It is inspired from the functions "**pairwise.prop.test**" and "**pairwise.table**" of the package "stats".

Choose files to compare (use multiple selections if needed, "Ctrl" + click)

```
listfiles=choose.files(default = "", caption = "Select STRUCTURE Runs to
confront",multi = TRUE, index = nrow(Filters))

compare.levels=function(i,j) {
  SIMIL.R(listfiles[i], listfiles[j])[1]
  }
```

Note that **SIMIL.R(listfiles[i], listfiles[j])[1]** will compute the GSS.unweighted. Replace **[1]** by **[2]** to obtain GSS.smallest, by **[3]** to obtain GSS.overall, by **[4]** to obtain GSS.mean or by **[5]** to obtain PercentOverlap.

```
level.names <- seq(along = listfiles)
ix <- seq(along = level.names)
names(ix) <- level.names
pp <- outer(ix[-1], ix[-length(ix)], function(ivec, jvec) sapply(seq(along =
ivec),
        function(k) {
            i <- ivec[k]
            j <- jvec[k]
            if (i > j)
                compare.levels(i, j)
            else NA
        }))
pp
```

4 Further versions and improvements

1. R(CRAN) gives the opportunity to code functions in the C language, such programming fastens considerably the algorithm.
2. Other approaches to compute genetic structuration comparisons can be implemented.
3. Since the code involves a small number of functions, SIMIL will stay as a script collection. As soon as needed, it will be implemented in a package.

5 How to cite

Alvarez N, Arrigo N and IntraBioDiv Consortium, 2007. **SIMIL: an R (CRAN) scripts collection for computing genetic structure similarities based on STRUCTURE 2 outputs.** Mol Ecol Notes In revision (will be updated).