

# ScaFoS

**Selection, Concatenation, and Fusion  
of Sequences**

**Version 1.25**

---

<b>1.</b>	<b><u>RATIONALE</u></b>	<b>3</b>
<b>2.</b>	<b><u>PURPOSE</u></b>	<b>4</b>
<b>3.</b>	<b><u>HOW-TO</u></b>	<b>5</b>
3.1	MAIN USAGES	5
3.2	COMPLEMENTARY USAGES	7
<b>4.</b>	<b><u>PRINCIPLE OF SEQUENCE SELECTION</u></b>	<b>10</b>
<b>5.</b>	<b><u>MAKING OF CHIMERA</u></b>	<b>13</b>
<b>6.</b>	<b><u>INPUT FILES FORMAT</u></b>	<b>14</b>
6.1	ALIGNED SEQUENCES FILES	14
6.2	OTHER INPUT FILES	16
<b>7.</b>	<b><u>GRAPHICAL MODE</u></b>	<b>19</b>
7.1	TO RUN SCAFOS	19
7.2	TO CHOOSE THE USAGE	19
7.3	TO MAKE AN OTUS FILE	20
7.4	TO SELECT FILES WITH CHOSEN SPECIES	20
7.5	TO ASSEMBLE DATASETS	21
7.6	TO OBTAIN THE HELP	24
<b>8.</b>	<b><u>COMMAND LINE MODE</u></b>	<b>25</b>
8.1	TO LOAD THE PROGRAM	25
8.2	TO MAKE FILE.OTU	25
8.3	TO SELECT FILES WITH CHOSEN SPECIES	26
8.4	TO CONCATENATE FILES	26
8.5	BATCH MODE	27
<b>9.</b>	<b><u>OUTPUT FILES</u></b>	<b>29</b>
9.1	SPECIES PRESENCE	29
9.2	DATASET ASSEMBLING	29
9.3	FILE SELECTION	31
<b>10.</b>	<b><u>SCAFOS INSTALLATION</u></b>	<b>32</b>
10.1	TECHNICAL REQUIEREMENTS	32
10.2	INSTALLATION UNDER LINUX	32
10.3	INSTALLATION UNDER WINDOWS	34
10.4	INSTALLATION UNDER MAC OSX	35

*If you use this software, please cite:*

Béatrice Roure, Naiara Rodriguez-Ezpeleta and Hervé Philippe. **SCaFoS: a tool for Selection, Concatenation and Fusion of Sequences for phylogenomics**. *BMC Evolutionary Biology* 2007, 7(Suppl 1):S2

## 1. RATIONALE

Phylogenetic inference based on large amounts of sequence data (phylogenomics) is becoming an alternative approach to single gene<sup>1</sup> phylogenies; which are often insufficient to resolve most phylogenetic questions. In this context, handling large amounts of data implies to deal with species and gene sampling, missing data, partial sequences, unequal distribution of species and genes, and with the presence multiple sequences per species (usually due to paralogous genes).

Despite the huge size of molecular databases, large amounts of data are only available for a limited number of species, and a choice has to be made between using a large number of genes or of species. An alternative is to make a compromise between a maximum number of genes and species and a reduced amount of missing data. The different ways to minimize missing data include the use of the most broadly sequenced genes and species and the combination of sequences from closely related species into a single one. Once species and genes have been selected, many approaches exist to infer phylogenies, but the most common ones can be defined as (i) super-matrix approaches or (ii) super-tree approaches. SCaFoS is a tool that allows the easy selection of sequences, species and genes and the construction of datasets suitable for these approaches. In particular, it helps in maximizing the amount of data usable for phylogenomic analyses.

---

<sup>1</sup> Generally, one file corresponds to one gene; in this user guide, the two terms will be used equally

## 2. PURPOSE

The construction of a phylogenomic dataset is often confronted to a logical choice of sequences according to various constraints:

- existence of partial sequences (e.g. EST or low-cover genome sequencing),
- absence of genes for some species (e.g. gene loss or incomplete sequencing not finished),
- existence of multiple sequences for one species (e.g. paralogous or xenologous genes).

SCaFoS is an useful software in this phylogenomic context allowing the easy handling of multiple aligned files. Starting from alignments of proteins or nucleic acids, SCaFoS is able to select genes, species and sequences according to the needs of the user.

The various options of SCaFoS include:

- the concatenation of multiple aligned files into a single super-matrix,
- the selection of genes for super-tree generation,
- the selection of species according to their frequency of presence or other user defined criteria,
- the selection among different paralogous sequences from the same species,
- the creation of chimerical genes from closely related species,

All these abilities taken into account, SCaFoS helps in minimizing the amount of missing data.

SCaFoS can be used in an intuitive easy-to-use graphical mode for which the options' choice is highly facilitated, or in command line mode that can be implemented in a workflow.

### 3. HOW-TO

The main goal of SCaFoS is to select the *best sequence* for an operational taxonomic unit (OTU) for each gene included in the phylogenomic analysis. To determine the best sequence, in general the longest slowest sequence between orthologous sequence, three usages are available in SCaFoS and are designed to be used sequentially:

1. SPECIES PRESENCE: defines the potential organisms (species or strains) to be selected as OTUs for following phylogenomic analyses,
2. FILE SELECTION: creates files of aligned sequences which contain only the species or strains as previously defined by the user,
3. DATASET ASSEMBLING: creates a concatenation of the selected aligned sequences files with the defined OTUs.

Figure 1 shows the current steps of a phylogenomic analysis, including the options of SCaFoS involved in each step. For a detailed description of various options, see the [graphical mode](#) or the [command line mode](#). First, the three main usages are described; next, some particular uses of SCaFoS with the needed options are described at the end of this section: to minimize the quantity of missing data and to eliminate potential paralogous sequences.

---

#### 3.1 MAIN USAGES

##### 3.1.1 SPECIES PRESENCE

Starting from a directory containing all files of aligned sequences, the SPECIES PRESENCE usage allows the creation of an OTU file that includes the name of the species<sup>2</sup> present in at least one file and the percentage of genes in which this species is present. If a file containing information about the taxonomic affiliation of each species is provided, the species will be classified according to their taxonomic group; otherwise the species will be displayed in alphabetic order or in decreasing order of frequency.

This initial OTU file is the starting point to create the correct OTU files for the subsequent steps; the user can use his/her preferred text editor to do it. In fact, this important step tends to define OTUs of closely related species that can be combined into chimeral sequences. This is a classical approach to maximize the amount of molecular information. The creation of these groups takes into account the frequency of species in different aligned files to minimize missing data; of course, an OTU may contain one only species when the complete

---

<sup>2</sup> To simplify this guide, the term species will often be used to cover all levels of organism as species, strains or so on. To differentiate various strains of a same species, by convention, the identifier of the sequence is composed in the identifier of the species and the identifier of the strain separated by the at character (@)

genome is available by example. An OTU is simply the grouping of some species names according to a global name that characterizes the OTU. See a description of these files in [input files format](#) section. These first steps correspond to the points ① and ② in Figure 1.

### 3.1.2 FILE SELECTION

It is the simplest way to create new aligned files containing only sequences according to a list indicated in a given OTU file. This usage needs a directory with the files of aligned sequences and the OTU file; neither selection between several sequences from the same OTU nor chimera is made during this usage, so, it is not necessary to create groups in the OTU file. According to the choice of the user, three options are available:

- Extract all aligned files but with only the species specified in the OTU file: this option is used to realize quickly phylogenetic analysis only with the species of interest,
- Extract all aligned files that contain at least one species from those specified in a OTU file (all other species are also conserved): it is a way to determine aligned files which contain some species that have a special importance,
- Extract only aligned files containing at least one species for each OTU are selected: this option is particularly interesting to produce a dataset containing only files with at least a partial sequence for each OTU.

This usage corresponds to the point ③ in Figure 1. This step does not obligatory realize before the concatenation in the next usage, but it is easier to determine unambiguous positions separately for each gene and only with species of interest, especially when the fast evolving species are removed. This step is also useful to do preliminary phylogenetic trees that can be used to interactively determine the best sequence during the next step.

### 3.1.3 DATASET ASSEMBLING

This is the most complex feature of SCaFoS with a lot of options and potential output files, but it is also the most useful and innovative. Starting from a directory containing the files of aligned sequences and the OTU file, it allows a rapid selection and concatenation from multiple aligned files in a single step. The main potentialities of SCaFoS are displayed in step ④ of the Figure 1 and allow super-matrix and super-tree approaches:

- The more or less automatic selection among multiple sequences from an OTU (see Principle of sequence selection for more details):
  - automatically depending on the length of the sequences, i.e. without any phylogenetic criterion,
  - automatically depending on the evolutionary distance calculated by TREE-PUZZLE,
  - by the user, interactively or using default sequences files (see description below);
- The creation of chimerical sequences within an OTU: if any complete sequence can be selected for an OTU in a given gene, partial sequences of OTU's species are merged,
- The concatenation of sequences within an OTU among multiple genes: creation of a single sequence constituted with the best sequence determine for each gene.

For a better selection of sequences according to phylogenetic criteria, inference of phylogenetic trees for each gene is recommended (see step ⑤, tree inference, in Figure 1); this inference is not implemented in SCaFoS and the user could use its/her favorite methods to do it.

To easily reconstruct the dataset, for example with OTU files being slightly different or after adding few new sequences, it is recommended to create files of default sequences, i.e. text files containing the user selected sequence when SCaFoS can't automatically determine the only one best sequence. The first time the user chooses the default option, SCaFoS asks for the correct sequence within OTU for which ambiguity exists and save the user choice in the default file to be reused after (see a description of these files in [input files format](#) section).

---

## 3.2 COMPLEMENTARY USAGES

---

### 3.2.1 Minimizing of missing data

To eliminate genes with too few OTUs:

- Choose the **DATASET ASSEMBLING** usage,
- Create subdirectories according to the number of missing OTUs: in each subdirectory, only files with less or equal than the number of missing genes specified by the subdirectory name are copied.

To eliminate genes with too missing data:

- Choose the **DATASET ASSEMBLING** usage,
- Create subdirectories according to the percent of missing positions: in each subdirectory, only files with less or equal than the percent of missing positions specified by the subdirectory name are copied.

A concatenated file will be done for each subdirectory. For each OTU, the number of real positions (i.e. characters for which the state is known and is not a deletion) in the concatenated sequence is added to the name of the OTU, so it is easy to know which concatenated sequence has a lot of missing data.

### 3.2.2 Elimination of too divergent sequences

To avoid inherent risks in phylogenetic inference, the goal of this action is dual: (i) remove high evolutionary sequences that might be involved in long branch attraction artifact (LBA), (ii) avoid incorporation of paralogous genes. This elimination could be done in an automatic or a semi-automatic mode, as the former is more drastic, the latter is recommended to take into account the user expertise to conserve sequence of interest.

#### Automatic mode:

- Choose the **DATASET ASSEMBLING** usage, and let the program determine the evolutionary distance with the **Minimal evolutionary distance** option
- Select a small **Threshold**: all OTUs for which the inner phylogenetic distance average is greater than the average of phylogenetic distance for all sequences with a value higher than the chosen threshold are automatically eliminated from the concatenation.

#### Semi-automatic mode:

- 
- Make a phylogenetic tree for each file,
  - Choose the **DATASET ASSEMBLING** usage,
  - Select **Using default sequence for an OTU** option without checking **With automatic choice by SCaFoS**,
  - Choose a not too small **Threshold**,
  - When SCaFoS finds multiple sequences within an OTU, choose the correct one by analysing the corresponding phylogenetic tree in the light of the evolutionary distance, the number of missing positions and the variation in composition.



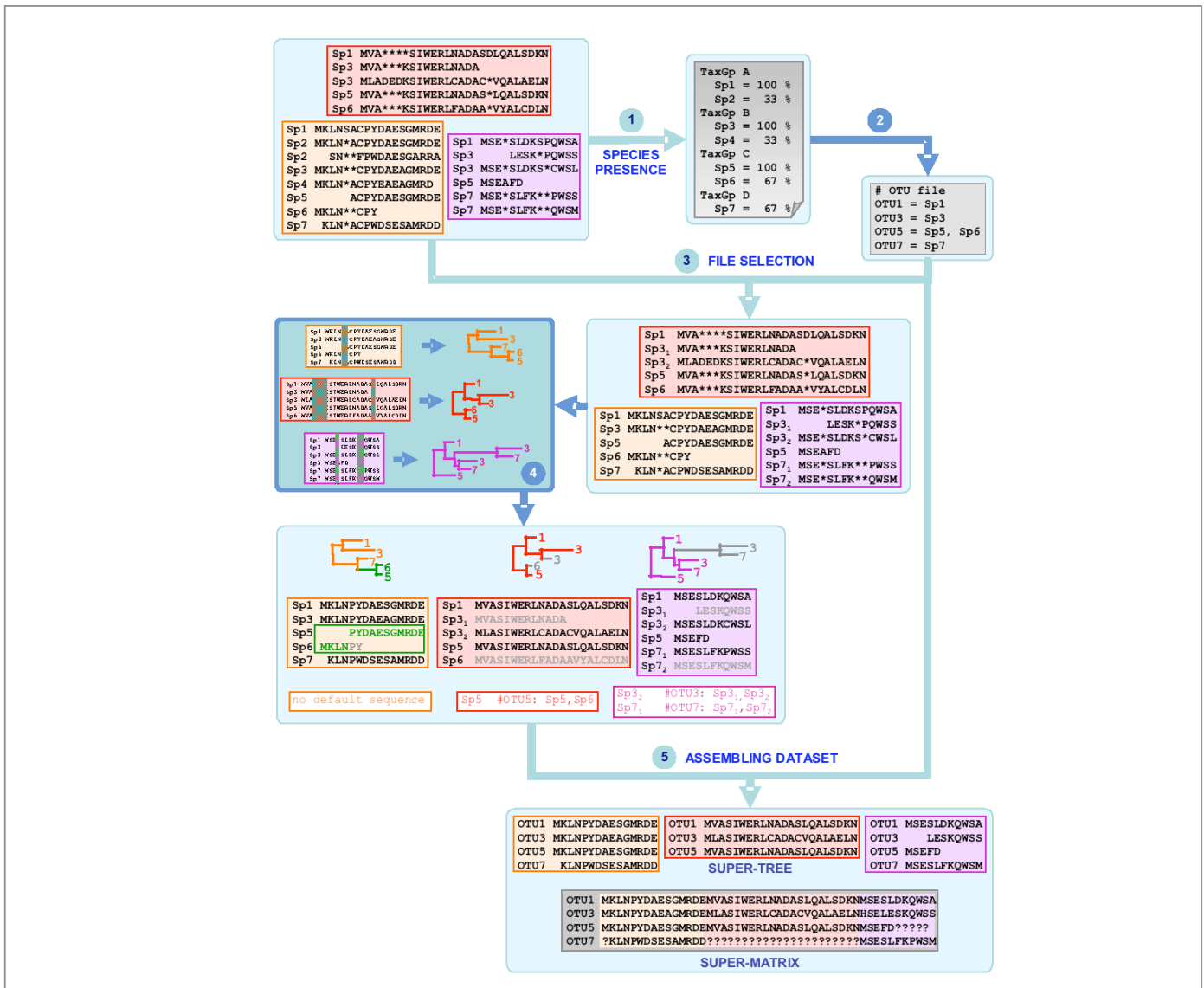


Figure 1: **Main steps to use SCaFoS** (steps 1, 3 and 5 are done by SCaFoS)

1. **SPECIES PRESENCE:** listing of all species present in the files of aligned sequences followed by their frequency of presence and, if desired, classified into taxonomic groups (specified by TaxGp in the figure).
2. Definition by the user of the species to be selected and their respective OTUs
3. **FILE SELECTION:** creation of files containing only the selected species
4. Discarding ambiguously aligned positions (displayed in dark colour) with a tool such as GBlocks; making phylogenetic trees (using PHYML or PAUP for example)
5. **DATASET ASSEMBLING:** selection of sequences and chimera construction according to an OTU file and default sequence files: creation of single gene files including chimeras and selected sequences and creation of concatenated files for super-tree and super-matrix approaches respectively.

In the last step, three typical cases are represented: (i) construction of a chimera (OTU5) in the orange file, (ii) selection of the less divergent sequence within an OTU (Sp6 in OTU5) and elimination of a short sequence (Sp3<sub>1</sub>) in the red file and (iii) elimination of potential paralogous sequences by the user (Sp3<sub>1</sub> and Sp7<sub>1</sub>) in the purple file. Eliminated sequences are drawn in grey. The corresponding default sequences files are displayed under their respective sequence files.

## 4. PRINCIPLE OF SEQUENCE SELECTION

Combining closely related species in an OTU is an effective way to minimize the amount of missing data. SCaFoS allows its ability to realize the selection of one sequence per group, being a list of one or more species (see description of OTU file in the [Input files format](#) section). During the concatenation step, SCaFoS searches for the best sequence<sup>3</sup> within all sequences of all species contained in the OTU and this for each gene; so in the concatenated file, each sequence is the result of the juxtaposition of the best sequence within an OTU for a particular gene, and its name is the OTU name.

The definition of the best sequence depends on the options chosen by the user and on the algorithm of SCaFoS (see Figure 2); several criteria alone or all together can be combined for the sequence selection:

- **Sequence length:** the longest sequences in the OTU (i.e. except gaps and missing characters) will be chosen by SCaFoS. This criterion is useful when a fast but rough selection criterion is preferred.
- **Evolutionary distance:** the less divergent sequences in the OTU will be chosen by SCaFoS. Evolutionary distances among each pair of sequences are calculated by TREE-PUZZLE (ref) and, for each OTU, the sequence presenting the lowest average distances to the rest of the sequences is chosen. The model of substitution is this chosen by TREE-PUZZLE<sup>4</sup>, according to the data, with a unique substitution rate. This criterion is applied on the complete sequences only according to the value defined by the user to determine a complete sequence.
- **Maximum percent to consider a sequence as complete:** when the making of chimera is activated, the evolutionary distance is only estimated for complete sequences, and then partial sequences are used to create a chimerical sequence. In some cases, it will be useful to consider large partial sequences such as complete sequences, this option defined this threshold.
- **Minimum sequence length:** if any chimera is made, even partial sequences are taken into account to estimate evolutionary distances except for the sequences smaller than this minimum length; created chimera must also be longer than this value.
- **Default sequences:** it is the sequence that will be systematically kept by SCaFoS for the OTU and the given gene. By default, SCaFoS selects the sequence with the smallest divergence distance within an OTU according to all sequences. But for various reasons (partial sequence, paralogous sequence, OTUs with too divergent representative species, and so on) the user could prefer to indicate another one. At the first use of this option, one default file is created for each aligned sequences file with the user selected sequence in each OTU for which sequence assignment is ambiguous. So information about selected sequences is conserved to be reused in other steps. Depending on the existence of previous default files,

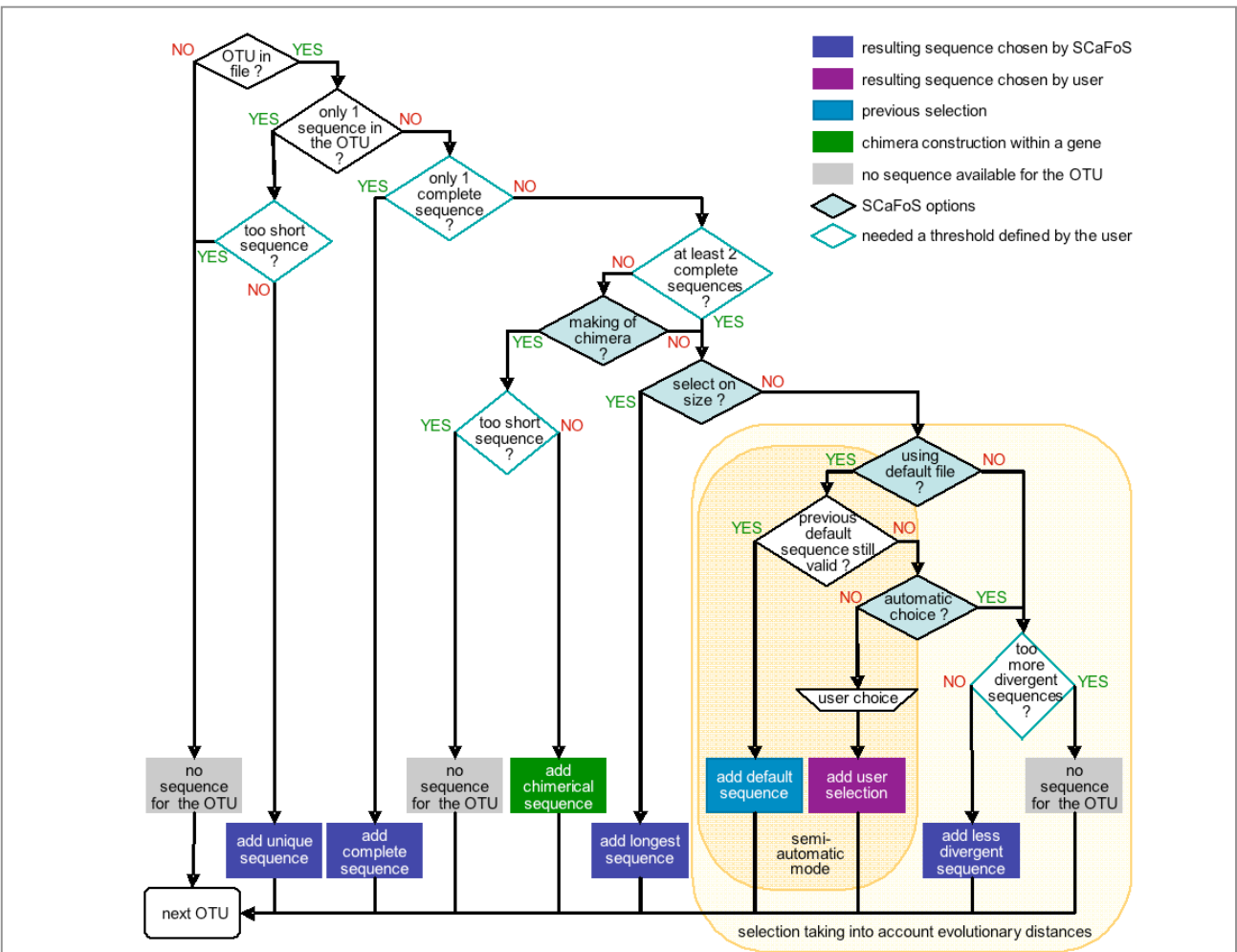
---

<sup>3</sup> *The best sequence is the longest slowest sequence evolving orthologous sequence*

<sup>4</sup> *TREE-PUZZLE 5.1 can not compute with more than 256 sequences*

if no default sequence is defined or if selection criterion is ambiguous, the user must specify the default sequence. In ambiguous cases, you could force SCaFoS to select the less divergent sequence. If the list of potential sequences within an OTU is modified (i.e. adding or removing sequences in the aligned file for species of this OTU, or modifying list of species in the OTU file), the user must redefine the default sequence

- **Divergence threshold:** OTUs that overcome the threshold are systematically removed for the given gene. For each OTU, the average of pairwise phylogenetic distances within the OTU is calculated and compared with the average pairwise distances for all other sequences; if the in-group average distance overcomes the out-group average with a value greater than the divergence threshold, the OTU is not taken into account for the given gene and an empty sequence is used. It is an efficient way to detect and eliminate paralogous sequences and to identify highly divergent sequences that could generate long branch attraction artifact. This criterion is not used if a default sequence is defined; higher is the threshold, less drastic is the filter.
- **Order in OTU file:** if all previous criteria have not determined a unique sequence, SCaFoS keeps the first sequence in the list of species of the group according to the OTU file between the less divergent sequences.



**Figure 2: Flowchart of sequences selection and construction of chimera for an OTU in a given gene**

For each OTU of each gene, SCaFoS selects the sequence that best represents the OTU. See text for a detailed description of the process. Three thresholds (empty blue rhombus) with default or user specific values are important: (i) the maximal percentage of characters present with respect to the longest sequence to keep a sequence, (ii) the minimal percentage of characters present with respect to the longest sequence to consider a sequence as complete and (iii) the maximum in-OTU/out-OTU distances ratio (see text) to keep an OTU. The user should select if he/she desires to create or not chimerical sequences and chose among the different sequence selection criteria (filled blue rhombus). If the selection criterion is the sequence size, no other options should be checked. If the selection criterion is the evolutionary rate of the sequences, the user must chose between a fully automatic or a semi-automatic choice of sequences and specify if he/she desires to use a previously defined selection.

## 5. MAKING OF CHIMERA

Another way to minimize the lack of data is the creation of a chimerical sequence from several partial sequences from species belonging to a monophyletic group. If no complete sequence is available for an OTU in an aligned file, SCaFoS makes a chimera with portions of different sequences according to the specified OTU. Sequences used to make chimera are taken into account in decreasing order of number amino acids or nucleotides (i.e. except gaps and missing characters). This chimerical sequence will be used in the concatenation file. You could force SCaFoS to take into account a partial sequence as a full sequence if you specify the percent of missing characters authorized in this case. If all partial sequences lengths are smaller than the minimal length authorized to take into account a sequence during the selection, they are used to create a chimera. The information used to create chimera, like parts of sequences involved in the chimerical construct, as described is the out.chim file (see [Output files format](#)).

**CAUTION:** *This option applies only to create chimerical sequences within a gene. But, even if it is not chosen by the user, the concatenation of sequences may provide chimera: i.e. concatenated sequences from various species.*

```

Uredinales :
P.graminis@ti.715                               LGPERFRATEILFNPELIGEEFPGIHQDLPERK ST
P.pachyrhizi@ti.710                             REKAGRRTTGIVSGDGVTHSV                RFRATEILFNP
P.pachyrhizi@ti.712 LTEAPLNPKKDREKA
P.graminis@ti.717                               PERKYST
chimera      LTEAPLNPKKDREKAGRRTTGIVSGDGVTHSV???????LGPERFRATEILFNPELIGEEFPGIHQDLPERKYST

```

**Figure 3: Example of chimera making**

Sequence fragments are combined from the longest sequence to the shortest; the length is calculated according to the number of characters: selected parts are displayed in black; the chimerical sequence is the

## 6. INPUT FILES FORMAT

SCaFoS needs aligned sequences files stored in a unique directory. These files can contain more than one sequence per species, but it is not necessary that all species be present in each file. SCaFoS can automatically recognize files in FASTA, MUST, NEXUS or PHYLIP format with the most common extensions. All files must have the same format and contain the same type of sequences (protein or nucleic acid). Other input files are created by SCaFoS itself or by MUST and are described below. You could use programs such as [readseq](#) or [PAUP](#) to convert aligned sequences files in FASTA, PHYLIP or TREE-PUZZLE format.

Aligned sequence files are recognized according to the extension names allowed in the table 1.

Table 1: **Allowed extension names for input files**

FASTA	fasta, fast, fa, fsa, seq, nt, aa
PHYLIP	phylip, phlp, phy, ph
NEXUS	nexus, nex
MUST	ali

### 6.1 ALIGNED SEQUENCES FILES

#### 6.1.1 Specific sequence name format

To take into account paralogous and xenologous genes as well as pseudogenes, one of the specificities of SCaFoS is its ability to deal with several copies from the same gene within one organism (species or strain); therefore the sequence identifier is composed as follow:

**Organism@Gene**

where **Organism** is the identifier used to described the organism (see below a detailed format),

**Gene** is a string of characters (like the accession number by example) that is specific to the corresponding sequence,

at symbol (@) is reserved to separate the organism identifier and the gene identifier and it must not be used in the organism name

It is important that the sequence names respect this format otherwise SCaFoS will consider that the various sequences correspond to different organisms and it will be unable to select the best sequence among all the sequences of this organism.

SCaFoS has the ability to distinguish between species name and strain name if the sequence name format respects the second important specification: nor genus name nor species name should contain an underscore (\_) which is the usual character used to separate genus name/species name from strain name. This restriction is useful to deal with several strains of the same species and so to do phylogenetic analysis on a small scale,

otherwise, the entire sentence preceding the gene identifier, if available, will be completely kept as the organism name.

To accommodate all these name constraints, SCaFoS is able to manage sequence name with the following format:

**GenusName SpeciesName\_StrainName@AccessionNumber**

Not all parts of the sequence name are necessary. SCaFoS looks for the first space character to determine genus name and eventually the unique underscore and the unique at symbol.

The other authorized characters in the sequence name are:

- all alphabetic characters,
- all numeric characters,
- point (.), dash (-) and space.

***CAUTION:** For obvious reasons of sequence differentiation, SCaFoS needs that all sequences in one file have a unique identifier. In consequence, when several sequences have an identical name, an incremental number is added to the name to distinct them.*

### 6.1.2 Gaps and missing characters

For all available input formats, star (\*) or dash (-) on the one hand and question mark (?) or character X (x or x) on the other hand indicate gaps and unknown characters respectively. The MUST format also accepts space as unknown character.

### 6.1.3 Comments

All characters following a hash sign (#) are considered comments.

***CAUTION:** These characters are removed during the processing.*

### 6.1.4 FASTA format

This format is defined by the sequence name on the first line with a 'greater than' character (>) at the beginning of the line, and the sequence on the next lines (sequence may be written on several lines) and eventually with space between character blocks. The spacers between characters blocks (space or end of line) are removed during the concatenation.

```
>Homo
ATGCAACGTTGACCTAGCATGAGA
>Mus
ATCCAACGTTGACCTAGCATAAGA
```

### 6.1.5 PHYLIP format

The number of sequences and the sequence length must be present on the first line.

The PHYLIP (ref) format accepts two different formats: interleaved or sequential.

The letter is the simplest with the sequence name on 10 characters and all the sequence on the same line or cut on multiple lines before the following sequence.

In the interleaved format sequences are cut in short fragments: a block of fragments contains the homologous sequence part for all sequences; the first block of fragments is preceded by the sequence names, the next block displays the second part for each sequence, and so on for all the fragments.

Interleaved and sequential formats could separate sequence in blocks with spaces.

```

2 24
Homo ATGCAACGTT GACCTAGCAT
Mus  ATCCAACGTT GACCTAGCAT
      GAGA
      AAGA

```

### 6.1.6 NEXUS format

SCaFoS ignores all NEXUS options, only **taxa** and **data** blocks are treated. Only a data block with **ntax** and **nchar** values, and the **matrix** is needed by SCaFoS. A NEXUS file must begin by **#NEXUS** keyword and all comments and empty lines are ignored. Interleaved or sequential formats and placeholder character are also supported.

```

#NEXUS
begin data;
  dimensions ntax=2 nchar=24;
  format datatype=RNA gap='-';
  matrix
    Homo ATGCAACGTT GACCTAGCAT GAGA
    Mus  ATCCAACGTT GACCTAGCAT AAGA
  ;
end;

```

### 6.1.7 MUST format

This is the format of the ED program from the MUST package, a tool to manage aligned sequences. It is very similar to the FASTA format, except that the file must begin with at least two lines beginning with a hash sign (#), and that the sequence for a given species should be on a single line.

---

## 6.2 OTHER INPUT FILES

According to usage and relative options, some other input files may be needed. These various files will be described below.



### 6.2.1 OTU file (<out>.otu)

It is created during the first use of SCaFoS and used in the next steps. The line format of OTU file is:

```
OTU name : species1 name (comment1), species2 name(comment2), ...
```

where **OTU name** is used to characterize the selected sequences from this OTU;

**species1 name...** are all the name of related species that can be grouped in the OTU with a possible comment that is not be considered.

Because some programs of phylogenetic inference don't support sequence names longer than 10 characters, the sequence names are truncated; make sure that OTU names are all different for the first ten characters to avoid subsequent problems.

In the SPECIES PRESENCE usage, SCaFoS creates an OTU file where each OTU contains one species and its name also is the species name; one OTU is created for each species present at least once in the aligned files:

```
species1 name : species1 name (percentage of genes present)
species2 name : species2 name (percentage of genes present)
...
```

Depending on the needs, the user creates his/her own OTUs by juxtaposition of all the species names that constitute the new group:

```
new group name : species1 name, species2 name, ...
```

### 6.2.2 Systematic file (<taxa>.nom)

It is a text file in which the taxonomic group is indicated for each species. This file could be used to create the OTU file ordered by taxonomic group to facilitate species selection and grouping. The line format of the systematic file is:

```
species name, taxa name
```

### 6.2.3 Default sequence files (<prefix of aligned sequence file>.def)

The automatic selection of a sequence requires that SCaFoS has been able to select only one sequence for an OTU in a given gene. If this unique selection is not possible, the user could interactively specify the correct one. The previously specified sequence chosen for representing an OTU is conserved in the default files; each aligned sequences file has its own default file. These files are created into the output directory the first time the `default files` option is used and they must be moved in the input directory containing the aligned sequences files to be reused later in the subsequent steps. The line format of default file is:

```
selected sequence id # OTU : sequence1 id, sequence2 id, sequence3 id
```

It is possible to modify a default sequences file, but the user must take care to respect the line format, particularly the list of all the sequence identifiers following the OTU name in which the selected sequence identifier must be present.

#### 6.2.4 Tree file (<prefix of aligned sequence file>.ps)

This kind of files is necessary to have a funny visualization of the selection made by SCaFoS for each gene, but the choice of this option does not influence the selection itself.

A tree file is a phylogenetic tree in postscript format generated by TREEPLOT, another program from the MUST package. If a tree file exists in the input directory, SCaFoS will change the color of sequence names according to the sequences selection realized by SCaFoS:

**green** chimera

**blue** selected sequence

**grey** discarded sequence

## 7. GRAPHICAL MODE

### 7.1 TO RUN SCAFoS

Under **Linux** system, the program is loaded by typing

```
scafes
```

or the following command in interpreted mode:

```
perl scafos.pl
```

In **Windows** environment, open a `command prompt` window and type one of the previous commands should.

In a standard installation of Windows, to open `command prompt`:

click **start**,  
point to **All Programs**,  
point to **Accessories**,  
and then click **Command Prompt**.

Or shorter:

chose **Run** in **start menu**  
and type `cmd` (XP or 2000) or `command` (98 or Me).

In **Mac OS X** environment, open a `x11 window` and type one of the previous commands run SCAFoS. The X11 windows are accessible in the `Applications>Utilities` folder.

In all cases, be sure that the `scafes` environment variable is defined previously to run SCAFoS

Input and output files are described in [Input files format](#) section and [OutPut Files format](#) section respectively.

### 7.2 TO CHOOSE THE USAGE

The three previously described usages (see the [How-to](#) section) are available according to the chronological use of SCAFoS (SPECIES PRESENCE, SELECT FILES and DATASET ASSEMBLING).

- **SPECIES PRESENCE**: to list all species in aligned sequence files and create an OTU file,
- **FILE SELECTION**: to extract sequences according to species included in the OTU file,
- **DATASET ASSEMBLING**: to create datasets for super-matrix or super-tree approaches.

According to his/her need, the user must click on the corresponding radio-button:

The screenshot shows a graphical user interface with three radio buttons at the top: "SPECIES PRESENCE" (selected), "FILE SELECTION", and "DATASET ASSEMBLING". Below these buttons are two input fields labeled "Input directory" and "Output directory". To the right of these fields is a "clear all" button.

Last options selected are automatically reloaded. By clicking on the **clear all** button, all options are removed and default values are reloaded.

Input and output directories are always needed. If the output directory already exists, you must confirm to replace the old one or type a new name.

**CAUTION:** *It is forbidden to use root directory or user home directory as output directory because a previous output directory could be automatically removed from the computer. To prevent any involuntary action, such output names will be systematically refused by SCAFoS*

For each usage, some parameters are needed or optional and only potential parameters will be activated; the description of these options will follow. Once the parameters are chosen, a simple click on **RUN** button executes the program. A new window will appear while the application is running. Close the result window, by clicking on the **CLOSE** button; the user can make another analysis by changing options or quit by clicking on the **QUIT** button.

With the **VERBOSE** option, a detailed description of the run is displayed.

In all following explanations, `out` will be used to refer to the output directory.

**CAUTION:** *As files may be modified (see [ALIGNED SEQUENCE FILES](#) in [Input files format](#) section), original files are saved in a `bak` subdirectory.*

### 7.3 TO MAKE AN OTUS FILE

To create OTU file, first choose the **SPECIES PRESENCE** radiobutton. The following options are available:

- To obtain an OTUs file in taxonomic order, you must specify the **systematic file** where the correspondence between species and taxa are defined; the button **browser** open a window where it is possible to select the systematic file
- You could minimize species number by a threshold corresponding to the **Minimum frequency of species** throughout aligned files, so too rare species are automatically eliminated from the OTUs file.

The image shows a screenshot of a software interface with two input fields. The first field is labeled "Systematic file" and has a "browser" button to its right. The second field is labeled "Minimum frequency of species [0, 100]".

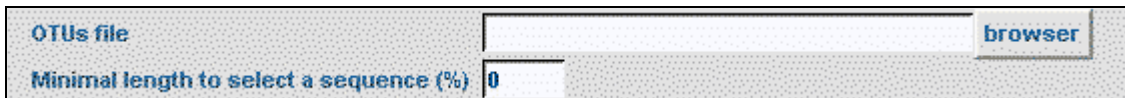
It is possible to continue with a concatenation according to the length of sequences, the program asks the user to this option: a concatenated sequence will be made for all species found at least once in the aligned sequences files. The concatenated file resulting of this rough concatenation will give a first draft of phylogeny because this dataset contains many missing data and lacks good choice of paralogous sequences.

### 7.4 TO SELECT FILES WITH CHOSEN SPECIES

Choose radiobutton **FILE SELECTION**.

Since SCaFoS will perform sequences selection, it's necessary to indicate the **OTUs file**. To easily select OTUs file, click on the **browser** button to open a file selection window.

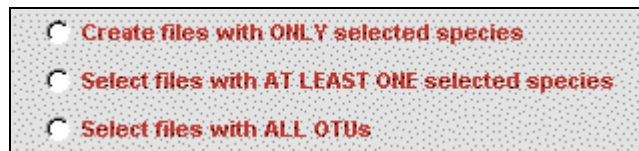
You could automatically eliminate sequences that are too short by typing the **Minimal length to choose a sequence**. By default sequences with a length shorter than 10 percent of the longest sequence are removed from the analysis and are not taken into account to calculate the evolutionary distance.



The screenshot shows a form with two input fields. The first field is labeled 'OTUs file' and has a 'browser' button to its right. The second field is labeled 'Minimal length to select a sequence (%)' and contains the value '0'.

According to the checked option, selected files of aligned sequences are written in output directory; three options are available:

- writing new files **ONLY** with the sequences of species defined in the OTUs file
- copying aligned files that contain **AT LEAST ONE SPECIES** defined in the OTUs file
- copying files that contain at least one species for **EACH GROUP** defined in the OTUs file



The screenshot shows three radio button options for file selection criteria:

- Create files with **ONLY** selected species
- Select files with **AT LEAST ONE** selected species
- Select files with **ALL** OTUs

Except for the first option for which unselected sequences are removed from the file, the selected aligned sequences files are identical to the initial files.

## 7.5 TO ASSEMBLE DATASETS

This step is the central point of SCaFoS because it generates a lot of information that could be used to choose genes and species, so several options are associated with this usage for which the user needs to click the radiobutton **DATASET ASSEMBLING**.

For details to deal with the **OTUs file**, see chapter 7.3

### 7.5.1 Selection criteria

When **MAKING OF CHIMERA** is chosen, if no complete sequence exists for an OTU, a new sequence is created with fragments of different species within the OTU for the current gene (see the [Making Chimera](#) section above). But even if this option is not checked, the concatenated sequences may be the result of a chimera: i.e. the concatenation of sequences from various species within the given OTU.

Several criteria are used to determine the best sequence within an OTU (see [Principle of sequence selection](#) for more details):

- **Minimal evolutionary distance:** SCaFoS selects the sequence with the smallest evolutionary distance to all other sequences within an OTU (default option of selection)
  - **With gamma:** the evolutionary distance is calculated according to a gamma distribution with 4 categories
  - The **Threshold** is the value used to eliminate too more divergent OTUs in order to reduce cause of LBA and the risk of including paralogous sequences (default value: 25)
- Select Using default sequence for an OTU allows the user to select its/her preferred sequence in ambiguous selection case except if the corresponding default sequences file already exists
  - **With automatic choice by SCaFoS:** when no previous choice is defined in the default file, let SCaFoS choose the best sequence, even in ambiguous case
- **Remove file when at least one OTU is too divergent:** this option is only available in the automatic mode
- **Longest sequence:** a fast but rough selection

Making of Chimera

Criteria to choose sequence within an OTU

Minimal evolutionary distance (Tree-Puzzle)     With Gamma    Threshold [ $\geq 0\%$ ]

Using DEFAULT sequence for an OTU     With automatic choice by SCaFoS

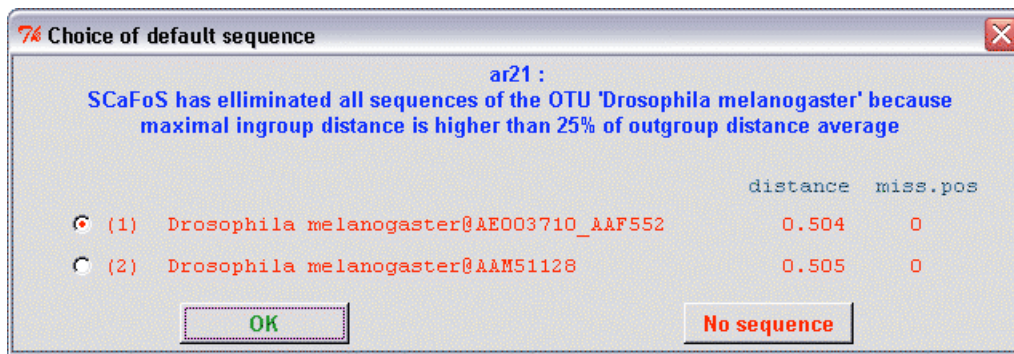
Remove file when at least one OTU is too divergent

Longer sequence

### 7.5.2 Selection of default sequences

In the semi-automatic mode, when SCaFoS is unable to select a unique sequence for the OTU (i.e. when **Using DEFAULT sequence for an OTU** option is selected), the user must validate the choice of the sequence guided by the evolutionary distance, the amount of missing positions, and eventually the deviation in composition. Remember that the list of sequences is written in red when at least one sequence is too divergent; otherwise the list is written in blue. Three options are available:

- accept the first sequence which has the smallest evolutionary distance (checked sequence) by clicking on the **OK** button
- select an other sequence and click **OK**
- remove the OTU for this aligned sequence file by clicking on the **No sequence** button



If all sequences are identical (i.e. same evolutionary distance and same length), SCaFoS automatically selects the first sequence, according to the order defined in the OTUs file, without asking the user.

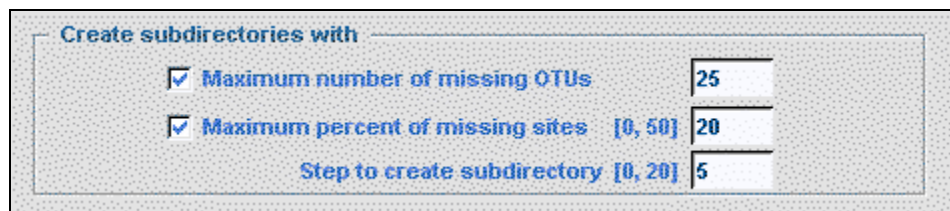
Sequence selection in semi-automatic mode could be very time consuming, but selection of default sequences is a real gain when the same dataset is use for various purposes with minor variations in definition of OTUs. After closing SCaFoS, copy the `.def` files newly created from the output directory to the input directory; on the next run, the previously selected sequences will be used and no question will be asked to the user for these OTUs.

**CAUTION:** One exception in use of default files. Even the default file exists, a question could be asked to the user if the list of sequences within an OTU has changed since the creation of the default file..

### 7.5.3 Creation of subdirectories

To minimize frequency of missing characters in final data set, user can create subdirectories with selected files according to:

- **Maximum number of missing OTUs:** create directories with files containing between 0 and <value> missing OTUs
- **Maximum number of missing sites:** create directories with files containing between 0 and <value> percent of missing sites with a <value> step

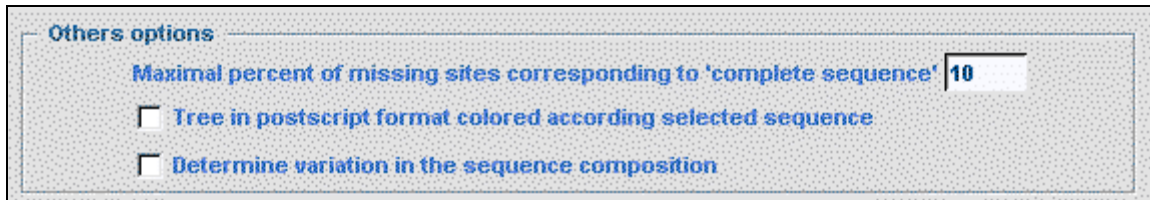


### 7.5.4 Global options

The value given for the option **Maximal percent of missing sites corresponding to complete sequence** determines the threshold for which a sequence is treated as complete and not included in the chimerical sequence. This option is used only if the **MAKING of CHIMERA** option is checked; otherwise all sequences are taken into account to calculate the evolutionary distance independently of their completeness.

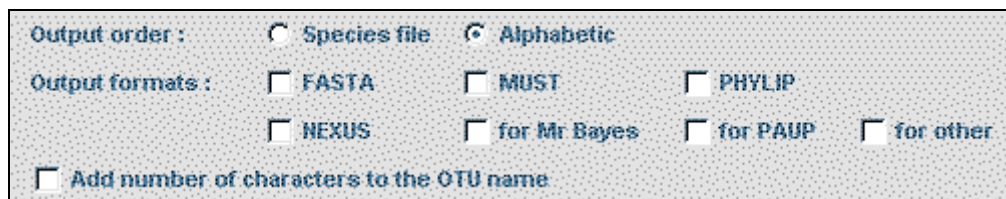
To obtain a phylogenetic tree where sequence names are written in different colors according to their selection by SCaFoS, check the corresponding option: **Tree in postscript format colored according to selected sequence**.

In complementary of computing of evolutionary distances, it is possible to display the variation in sequence composition towards the average composition of all sequences in the file. By checking the option: **Determine variation in sequence composition**, this information is displayed during the interactive choice by the user and in various output files.



The concatenated file can be output in several formats: FASTA, PHYLIP in sequential format, MUST, NEXUS. Just check the desired options. Commands can be automatically added to the nexus file to obtain output files suitable for PAUP or MrBayes. The user can also modify these optional commands or add his/her own commands in the NEXUS file with the **for other** option: a text editor will be displayed to type appropriate commands. Be careful because no verification for the validity of these commands is made by SCaFoS. For all format it is possible to choose the order of the concatenated sequences with the **output order** option. These options are also used for the concatenated files in subdirectories if needed.

Checking the **Add number of characters to the OTU name** adds the length of the concatenated sequence at the end of each OTU name: i.e. the number of real characters.



## 7.6 TO OBTAIN THE HELP

A short help about each option is displayed when the mouse stay on the corresponding option.

A global help is displayed by click on the **HELP** button. From the newly window, it is possible to display the full online help available directly on our web site.



## 8. COMMAND LINE MODE

This possibility may be useful to call SCaFoS from other programs as in a pipeline. But even if you don't plan to use the graphical interface, Perl-Tk has to be installed on your computer. In case of scripting use, you must handle interactive questions:

- in all usages, when the output directory already exists
- in **DATASET ASSEMBLING** usage, in cases of ambiguous selections, because the user must decide if default sequence option is chosen. To avoid this problem use **auto=yes** option

Synopses of various usages are given below.

### 8.1 TO LOAD THE PROGRAM

Type `scafos` and the needed options or the following command in interpreted mode:

```
scafos <options>
```

On **Windows** environment, you must open a **command prompt** window and type the previous commands in the new opened window. On **Mac OS X**, you must open an **x11** window. For details see [section 7.1](#).

The available options are the same that the ones described in the [graphical usage](#) section. Please refer to this section for more details. In the following sections, only the synopsis of the command line and the option codification for each usage will be explained:

- words in angular brackets (<>) are general terms that must be replaced by specific values
- brackets ([]) indicate optional features, if omitted, the default values are used

### 8.2 TO MAKE FILE.OTU

```
SCaFoS in=<dirIn> out=<dirOut>
      [freq=<freq>] [name=<file.nom>] [l=<minlen>]
      [cat=yes]
```

File	Description
dirIn	directory of input files of aligned sequences
dirOut	output directory
freq	minimum frequency of species in files
file.nom	list of species with corresponding taxa
minlen	minimal percent of sequence length to keep a sequence (default=10%); i.e. minimal length needed to calculate the evolutionary distance or to conserve the chimerical sequence
cat=yes	done an automatic concatenation after creation of OTU file (default=no)

### 8.3 TO SELECT FILES WITH CHOSEN SPECIES

```
SCaFoS in=<dirIn> out=<dirOut> otu=<file.otu>
      o=[ofav] [l=<minlen>]
```

File	Description
dirIn	directory of input files of aligned sequences
dirOut	output directory
file.otu	list of OTUs used for assembling datasets
minlen	minimal percent of sequence length to keep a sequence (default=10%); i.e. minimal length needed to calculate the evolutionary distance or to conserve the chimerical sequence
o	write files with ONLY SELECTED SPECIES in dirOut <i>! not compatible with f and a options</i>
f	copy files with AT LEAST ONE SELECTED SPECIES in dirOut <i>! not compatible with o and a options</i>
a	copy files with ALL OTUs in dirOut <i>! not compatible with o and f options</i>
v	verbose

### 8.4 TO CONCATENATE FILES

```
SCaFoS in=<dirIn> out=<dirOut> otu=<file.otu>
      [gamma=yes] [t=<threshold>] [puz=no]
      o=[gsclrv g=<number> s=<percent> p=<step>]
      [def=yes auto=yes] [m=<missing>] [l=<minlen>]
      [format=fpmnba] [color=yes] [cmp=yes]
```

File	Description
dirIn	directory of input files of aligned sequences
dirOur	output directory
file.otu	list of OTUs used to assemble the datasets
gamma=yes	use of 4 gamma categories to calculate evolutionary distances
threshold	threshold used to compare distances from puzzle and eliminate OTUs containing too divergent sequences (default=25%)
puz=no	doesn't use PUZZLE to determine best sequence (default=use of PUZZLE) <i>! no compatible with def option</i>
g	make subdirectories with files containing less than 'number' missing OTUs (default = number of OTUs)
s	make subdirectories of files containing less than 'number' % of missing sites (default = 20%)

c	make chimera with partial sequences in an OTU
l	add character number in concatenated sequence after sequence name
r	remove completely a file when at least one OTU does not respect the <b>threshold</b> default=no removal action <i>! not compatible with def option</i>
v	verbose
step	step of percent of missing sites to create sub-directories (default=5) <i>! available only with s option</i>
def=yes	use of input files *.def for selected sequences (default=sequences selected by SCaFoS computation)
auto=yes	use SCaFoS computation and don't ask user if more than one sequences are possible (default=no)
missing	percent of missing sites to use sequence as a complete sequence [0:50]
minlen	minimal percent of sequence length to keep a sequence; i.e. minimal length needed to calculate the evolutionary distance or to conserve the chimerical sequence (default=10%)
f	output files in FASTA format
p	output files in PHYLIP format
m	output files in MUST format
n	output files in NEXUS format
b	output files in NEXUS format with commands for MrBayes
a	output files in NEXUS format with commands for PAUP
color=yes	modify colors in tree.ps files (default=no) : <b>green</b> -> chimera <b>blue</b> -> selected sequence <b>grey</b> -> discarded sequence (default=no)
cmp=yes	display the variation of the composition in amino acids or nucleotides

## 8.5 BATCH MODE

To automate some analyses, it may be useful to run SCaFoS without human interaction. The easy way is to create a file containing parameters corresponding to the needed answers in a chronological order. In the example following, SCaFoS is loaded to create OTU file; the file `file.par` contains only the answer to confirm the overload of output directory:

**y**

where **y** authorizes the software to overload the output directory

All parameters in parameter file must be on a new line, without other characters (neither space nor comments).

To run SCaFoS, type the line below on a command line:

```
SCaFoS in=<dir1> out=<dir2> < file.par
```

**CAUTION:** *To run SCaFoS in the semi-automatic mode with a parameter file, it is necessary to use default files without change neither in the OTU file nor in the sequences existing in the input files of aligned sequences. This condition is needed to avoid questions to the user.*

## 9. OUTPUT FILES

In the output directory, and eventually its sub-directories, you obtain various files according to the usage of SCaFoS; all output files are not always available, depending on the chosen options. In the next tables, `out` means the name of output directory described in the previous chapter.

**CAUTION:** *It is forbidden to use root directory or user home directory as output directory because a previous output directory could be automatically removed from the computer. To prevent any involuntary action, such output names will be systematically refused by SCaFoS*

### 9.1 SPECIES PRESENCE

File	Description
<code>out-freq.otu</code>	List of species that appear at least in one aligned file; species are displayed in decreasing order of frequency
<code>out-name.otu</code>	List of species that appear in at least one aligned file; species are displayed in alphabetic order
<code>out-taxa.otu</code>	List of species that appear in at least one aligned file; species are displayed in alphabetic order according to a systematic classification provided by the user
<code>out.log</code>	Information about the genes and species content of each aligned file

### 9.2 DATASET ASSEMBLING

The output directory has a hierarchical structure according to the optional output files expected by the user. If the output directory is called `out`, this structure is described at the right with `N` as the higher number of missing OTUs, and `M` the higher percent of missing sites.

**CAUTION:** *according to the chosen options, all files described below are not always present.*

```

out
├── misgen
│   ├── misgen_1
│   ├── . . .
│   └── misgen_N
└── missit
    ├── missit_1
    ├── . . .
    └── missit_M

```

#### 9.2.1 The main output directory

This directory contains the concatenated aligned sequence files for all separated aligned sequence files and all sequences selected by SCaFoS, and- or the user in semi-automatic mode, according to the OTUs defined in the OTU file. The concatenated file appears in various formats according to the choices of the user. Some

informational files are also displayed to describe actions performed by SCaFoS to complete its sequence selection. The state file is particularly useful to determine which files or which groups must be eliminated according to bulk of missing data.

File	Description	Directory
out.fasta	Concatenation in FASTA format according to the OTUs defined in the OTU file	out
out.phylip	Concatenation in PHYLIP format according to the OTUs defined in the OTU file	out
out.ali	Concatenated in MUST format according to the OTUs defined in the OTU file	out
out.nex	Concatenation in NEXUS format according to the OTUs defined in the OTU file, eventually with commands add by the user	out
out_PAUP.nex	Concatenation in NEXUS format according to the OTUs defined in the OTU file; commands for PAUP are added in this nexus file	out
out_MB.nex	Concatenation in NEXUS format according to the OTUs defined in the OTU file; commands for MrBayes are added in this nexus file	out
out.len	Maximal sequence length for each aligned sequences file	out
out.dist out.outdist	Display of evolutionary distances of sequence within each OTU and each file. Those files could be useful to verify which OTUs are removed depending on their too divergent distances. The composition of the sequences is also displayed.	out
out.stat	State file that contains a table of percent of missing positions for each OTU within each file, and the global percent of missing positions, missing genes and chimera by OTU. The global percent of missing positions, missing OTUs and chimera are also displayed by file.	out
out.cmp	State file that contains a table of composition in amino acids or nucleotides for each sequence within each file	out
out.seq4otu	Within each file, list of selected sequence for each OTU. Creation of chimera and existence of any sequence within an OTU are indicated.	out
out.chim	Description of each sequence fragment used to create chimera	out
out.misotu	List of files according number of missing OTUs	out

**CAUTION:** the commands manually added in NEXUS file are not checked by the program, be cautious when you added such commands

### 9.2.2 Partial selection directories

The global analysis could be completed by creating the subdirectories `misgen` and `missit` where files are put according to the maximum number of missing OTUs and the maximum percent of missing sites respectively within each aligned file. These separated aligned sequence files are reconstructed only with the OTUs responding to the selection criteria currently used.

For each `misgen_X` subdirectory, concatenated files are created with the corresponding state file and the corresponding OTU file.

### In the `misgen` directories

File	Description	Directory
<code>misgen.stat</code>	General information on subsequent directories	<code>misgen</code>
<code>misgen_X.fasta</code> <code>misgen_X.phylip</code> <code>misgen_X.ali</code> <code>misgen_X.nex</code>	Concatenation in various formats containing at most X missing OTUs according to the OTU list in the file.otu	<code>misgen_X</code>
<code>misgen_X.otu</code>	OTUs file with only the OTUs present in the concatenation X	<code>misgen_X</code>
<code>misgen_X.len</code>	General information about the concatenation in <code>misgen_X</code> directory	<code>misgen_X</code>
<code>misgen_X.stat</code>	State file that contains a table of percent of missing positions for each OTU within each file, and the global percent of missing positions, missing genes and chimera by OTU. The global percent of missing positions, missing OTUs and chimera are also displayed by file.	<code>misgen_X</code>
aligned sequence files	Separated aligned sequence files containing only the retained OTUs used for the concatenation	<code>misgen_X</code>

### In the `missit` directories

File	Description	Directory
<code>missit.stat</code>	General information on subsequent directories	<code>missit</code>
<code>missit_X.stat</code>	General information about the file selection in <code>missit_Y</code> directory	<code>missit_Y</code>
aligned sequence files	Separated aligned sequence files containing only the retained OTUs selected according to the Y percent of missing sites	<code>missit_Y</code>

## 9.3 FILE SELECTION

According to the option chooses by the user, selected files of aligned sequences are put directly in output directory. The three options are:

- writing new files `ONLY` with the sequences of species defined in the OTUs file
- copying aligned files that contain `AT LEAST ONE SPECIES` defined in the OTUs file
- copying files that contain at least one species for `EACH GROUP` defined in the OTUs file

For more details see the [File selection section](#) in the main usages description.

## 10. SCAFOS INSTALLATION

### 10.1 TECHNICAL REQUIEREMENTS

The current version runs under Linux, Mac OSX and Windows. SCaFoS has been tested on the following operating systems: Windows XP Home and Professional editions, Red Hat and Fedora, Mac OSX. Theoretically, SCaFoS is able to run under all Unix-like systems and Windows Vista, but it is provided as is.

SCaFoS requires Perl version 5.8.0 or later, Tk version 8.4.5 or later and Tree-puzzle version 5.1 or later.

You can download the latest version from their respective web site:

- Perl: <http://www.perl.org/> or <http://www.activestate.com/Products/languages.plex?tn=1> (Windows XP or Max OS X)
- Perl/Tk: [http://sourceforge.net/project/showfiles.php?group\\_id=10894&package\\_id=10452](http://sourceforge.net/project/showfiles.php?group_id=10894&package_id=10452)
- Tree-Puzzle: <http://www.tree-puzzle.de/>

To install these modules it's better to have administrator permissions.

As SCaFoS has to be independent of the TREE-PUZZLE version, it is needed to rename the executable file with the standard name `puzzle`.

Before installing Tk, be sure to have the X11 development library on your system.

The easiest way to install Tk is to use CPAN by typing the command:

```
perl -MCPAN -e'install Tk'
```

*REMARK: In the following commands, XXX will refered to the number version of SCaFoS and must be changed to the current version.*

### 10.2 INSTALLATION UNDER LINUX

#### 10.2.1 Required tools

On most linux systems (Debian, Fedora, Gentoo, Mandrake, RedHat, slackware, SuSE, ...), Perl will be probably already installed and you only have to verify the version.

On linux you can verify if Perl and Perl/Tk are already installed with the following commands:

```
perl -v  
perl -e 'use Tk'
```

The first command lists the perl version if installed. The second one displays an error message if the Tk module is not installed.

Before installing Tk, be sure to have the X11 development library on your system.



To install these modules it's better to have administrator permissions. The purpose of this part is not to explain in detail how to install these pre-required softwares, but only to guidelines to typical user. For more information, please, refer to the online sites of the products that take into account the specificities of the various operating systems.

3 major ways are available to install packages on linux:

- from the sourcefile, the different usual steps to install a package located in the current directory are (in all cases, read the README file which includes installation specificities):

- uncompress and unpack the sourcefile:

```
gzip -d SourceFile.tar.gz
tar -xvf SourceFile.tar
```

- built the package from the newly created directory:

```
cd 'NewDirectory'
./configure
perl MakeFile.PL
make
```

- install the package:

```
make install
```

- from a RedHat Package Manager file (RPM) :

- for a new installation:

```
rpm -ivh PackageFile.rpm
```

- for updating a package:

```
rpm -Uvh PackageFile.rpm
```

- from the Comprehensive Perl Archive Network (CPAN), you need to be connected as root and to have access to internet; you have to type the following commands and answer to the questions (a lot are predefined and you have just to type ENTER) :

```
perl -MCPAN -e'install Perl'
perl -MCPAN -e'install Tk'
```

*If errors appear during test steps of the Tk Installation, you can force the installation by the following command in a CPAN session:*

```
force install Tk
```

### 10.2.2 SCaFoS

Create a SCaFoS directory and change to this directory:

```
mkdir scafos
cd scafos
```

To install the Perl script, copy the source file in the SCaFoS directory and uncompress it by typing the following commands (xxx means the version number):

```
gzip -d scafos_scr_linux.xxx.tgz
tar -xvf scafos_src_linux.xxx.tar
```

To be able to run SCaFoS from all directories, it is necessary to add a new environment variable in your system.

To simplify software loading, an executable file (not compiled) is provided: its file name is `scafos`

You must add the following line in the file `.bash_profile` or type it before each running session of SCaFoS:

```
export SCAFOS=scafos
let scafos be the full directory name where SCaFoS is installed ; for example
export SCAFOS=/home/toto/scafos if SCaFoS is installed in the directory scafos of the user
toto
```

To enable the environment variable, you must close your session or use the `source` UNIX command to force the shell to read the `.bash_profile` file:

```
source .bash_profile
```

---

## 10.3 INSTALLATION UNDER WINDOWS

---

SCaFoS installation needs permission to create directories and system variables. To install the pre-required tools under Windows, the easy way is to use the MSI installer with the files provided by the [ActiveState web site](#) and to follow the explanations.

To install SCaFoS, create a SCaFoS directory, copy the source file in this directory and uncompress it: a right click on the `scafos_src_linux.xxx.zip` file opens a menu where you chose the **Extract All** function. You could also use your favorite uncompress tool like Winzip or another one.

The easy way to create the environment variable under Windows:

- click on the **start menu**,
- point to **My Computer**, open a new menu by a right click,
- click on **Properties**,
- in the new opened window choose the **Advanced tab**,
- click on **Environment Variables** button,
- select **New** in the user variables,
- in the new window, type SCAFOS in the **variable name zone** and the full path of the directory where SCaFoS is installed in the **variable value zone**
- close all windows by click on the **OK** buttons

---

## 10.4 INSTALLATION UNDER MAC OSX

---

Mac OS X is compatible with UNIX systems. So to run SCaFoS, not only Perl must be installed, but also X11 program and perl-tk. The following procedure has been tested for Intel-based Macs, but it should work for PowerPC-based Macs also.

### 10.4.1 Required tools

Generally, **Perl** is preinstalled on your system: you can test it in a terminal window by the following command that displays the perl version if installed:

```
perl -v
```

As UNIX application, ScaFoS runs in a terminal window and some installation steps require this type of window also. You access the UNIX operating system in Mac OS X by using the **terminal application**. Terminal is in the **Applications>Utilities** folder.

If you did not install **X11** during Mac OS X installation:

- insert the **Mac OS X Install** disc,
- double-click the **Optional Installs package**
- follow the on-screen instructions until you see a list of software packages
- open **Applications**
- select **x11**

To test if Tk is already installed, type the following command in a terminal window:

```
perl -e 'use Tk'
```

If the Tk module is not installed, an error message is displayed, nothing otherwise.

To install **Perl-tk**, you need to be a superuser to perform the following steps (you may need to type an administrator password to make these changes):

- open **NetInfo Manager**, located in **Applications>Utilities**.
- in the top menu, choose **Security>Enable Root User**
- if needed, type a new password for the root account.
- choose **Apple menu>Log Out** to quit the superuser session
- choose **Other** in the **Login window**,
- type **root** in the **Name field**, and the **root password** in the **Password field**
- in the terminal, type the following command to launch the Perl-tk installation :

```
perl -MCPAN -e'install Perl'
```

*If errors appear during test steps, you can force the installation by the following command in a CPAN session:*

---

```
force install Tk
```

- Exit from the root session if you want install SCaFoS in your own environment.

### 10.4.2 SCaFoS

To install **SCaFoS**, create a SCaFoS directory. Copy and uncompress the source file in this directory. To enable the SCaFoS script, you must add the following line in the `.profile` file (if the file does not exist, create it in your Home directory):

```
export SCAFOS=scafos
```

let `scafos` be the full directory name where SCaFoS is installed; for example

```
export SCAFOS=/Users/toto/scafos if SCaFoS is installed in the directory scafos of the user  
toto
```

To enable the environment variable, you must close the terminal window or use the `source` UNIX command to force the shell to read the `.profile` file:

```
source .profile
```