SuPi - Supertree Pipeline

Manual Version 1.0 (August 3, 2010)

©by Anne Kupczok, Heiko A Schmidt and Arndt von Haeseler

Anne Kupczok
Center for Integrative Bioinformatics Vienna,
Max F. Perutz Laboratories,
Dr. Bohr-Gasse 9/6, A-1030 Vienna, Austria.
email: `anne.kupczok @ mfpl.ac.at`

Heiko A Schmidt
Center for Integrative Bioinformatics Vienna,
Max F. Perutz Laboratories,
Dr. Bohr-Gasse 9/6, A-1030 Vienna, Austria.
email: `heiko.schmidt @ mfpl.ac.at`

Arndt von Haeseler
Center for Integrative Bioinformatics Vienna,
Max F. Perutz Laboratories,
Dr. Bohr-Gasse 9/6, A-1030 Vienna, Austria.
email: `arndt.von.haeseler @ mfpl.ac.at`

Contents

1 Legal Stuff

©by Anne Kupczok, Heiko A Schmidt and Arndt von Haeseler

This program is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or (at your option) any later version. This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

2 Introduction

SuPi is a simulation pipeline to generate alignment and phylogeny data sets for a subsequent supertree analysis. It does not do any supertree computations, instead it provides the source trees generated with different models (i.e. lengths variation, true gene tree variation). In addition to computing the source trees, it will also compute the superalignment tree.

3 Installation

The command-line program is freely available from http://www.cibiv.at/software/supi. It is written in python and should run on every computer with python 2, version 2.4 or newer. It will not run with python 3. Python can be downloaded from `http://www.python.org/`. First unzip `SuPi-1-0.zip`, then change into the directory `SuPi` and type `python SuPi.py [options]` to run the program.

The program uses several external programs that are assumed to be installed on your system:

- Seq-gen (Rambaut and Grassly, 1997), available from http://tree.bio.ed.ac.uk/software/seqgen/.
- IQPNNI (Vinh and von Haeseler, 2004), available from http://www.cibiv.at/software/iqpnni/.
- Seqboot and Consense from the Phylip package, available from http://evolution.genetics.washington.edu/phylip.html.

Please note that the pipeline was mainly written for our own simulations in Kupczok *et al.* (2010). So, be aware that many errors that may occur in input and processing are not catched. However, all calls to external programs are printed to the screen. If an error occures shortly after an external program was called, please try first whether the command worked correctly on your input data.

4 Command-line options

Run `python SuPi.py -h` to print out a short description of available options:

```
Usage: SuPi.py [options]

Options:
  -h, --help            show this help message and exit

  Input options:
    -s DATA, --supermatrix=DATA
                        location of original supermatrix in phylip or nexus
                        format, for simulation matrix.nex1.phy etc. are
                        assumed to be in same directory or written there,
                        default=./matrix.nex
    -t SEC, --section=SEC
```

```
                        file with lengths of individual genes in alignment

  Simulation options:
    -o START, --offset=START
                        number of first simulation, default=1
    -e END, --end=END   number of last simulation, default=1
    -u, --multsource    use gene trees for simulation, stored in
                        'model_tree[1-source].tree' in data directory,
                        default: off (only 'model_tree')
    -g SEQGEN, --seqgen=SEQGEN
                        file in simulation directory with options for seq-gen,
                        default 'seq-gen.para' (at least substitution model
                        needed)
    -f SCALE, --tree_scale=SCALE
                        file with tree lengths scalings for sections, default
                        no scaling
    -l LSCALE, --length_scale=LSCALE
                        file with sequence lengths scalings for alignments,
                        default 1
    -v THETA, --theta=THETA
                        specify theta if gene trees should be generated from
                        species tree by coalescent model (default: no gene
                        trees)
    -p PHYLO, --phylo=PHYLO
                        type of phylogeny reconstruction: a - all (default), i
                        - input trees only, s - superalignment tree only, n -
                        none
    -i IQPNNI, --iqpnni=IQPNNI
                        command line parameters for iqpnni, default '-w gamma'
    -b BOOT, --bootstrap=BOOT
                        number of bootstrap replicates for ipqnni (0-off,
                        default)

  Output options:
    -d DIR, --dir=DIR   simulation directory, default is current directory
```

4.1 Input options

`-s DATA, --supermatrix=DATA` This filename (potentially including a path) gives the name of the supermatrix according to which new data sets are simulated. The pattern of missing data is read from this supermatrix. Missing data is marked as `?`. Additional gaps that are not missing data may be marked with `-`. For both, phylip and nexus format, a space between the taxaname and the sequence is assumed. See examples `matrix.nex` and `matrix.phy` in the `example`-directory.

`-t SEC, --section=SEC` This filename gives the lengths of the genes in the alignment. Each length should be on one line. See example `matrix.sec` in the `example`-directory.

4.2 Simulation options

`-o START, --offset=START` The number of the first simulation. A directory named `sim_i` will be created in the output directory for each `i` in `start`, ..., `end`

`-e END, --end=END` The number of the last simulation.

`-u, --multsource` If this option is set, each gene tree is defined independently. Thus not `model_tree` is assumed to be in the results directory (see output options), but each `model_tree{i}.tree`. In this case, no taxa are pruned after simulation, since the model trees are assumed to have the correct taxa already. The order of the model trees coincides with the sections given with `-t`. See `results2` for an example with multiple gene trees.

`-g SEQGEN, --seqgen=SEQGEN` Name of the file with simulation parameters for Seg-Gen (Rambaut and Grassly, 1997). The file should be located in the output directory (see output options). The file may contain only one line, then the parameters are equal for each gene, or exactly one line for each section, then each section can have different parameters. Note that you should not end the file with an empty line.

`-f SCALE, --tree_scale=SCALE` File with tree lengths scalings. The file is assumed to contain random numbers, thus each number is used only once and it should contain at least (number of sections)×(number of simulations) numbers.

`-l LSCALE, --length_scale=LSCALE` Sequence lengths scaling for alignments. Each alignment in each simulation can be scaled with the same number.

`-v THETA, --theta=THETA` θ parameter for generating gene trees from the species trees before generating alignments along the species tree. For the definition of θ see Ewing *et al.* (2008). In each simultation directory there will be a file `org.genetrees` that contains the gene trees in the respective order.

`-p PHYLO, --phylo=PHYLO` Determine whether and which phylogeny reconstruction with iqpnni is done. It is possible to reconstruct all trees, only the gene/input trees, only the superalignment tree or no tree at all.

`-i IQPNNI, --iqpnni=IQPNNI` Command line parameters for IQPNNI (Vinh and von Haeseler, 2004). Each gene is reconstructed with the same model given by the parameters, check `iqpnni -h` for possible options.

`-b BOOT, --bootstrap=BOOT` If this option is set, then the source trees are computed via bootstrapping, i.e. IQPNNI is run multiple times on bootstrapped alignments and the source tree is the majority-rule consensus of the bootstrapped trees. With this command, also the supermatrix tree is computed via bootstrapping. This command uses the phylip programs `seqboot` and `consense`.

4.3 Output options

`-v HEADER, --header=HEADER` Directory where the simulation output will be written. A file named `model_tree` is assumed to be in this directory. `model_tree` contains the model tree for simulations in newick format.

5 Output specification

After running SuPi, a directory is created for each simulation replicate where you can find the results. Depending on the parameters, slightly different files will be created (see simulation options). The final results can be found in `alg.source.trees` and `sm.tree`. The first contains the source trees for supertree analysis in the respective order and the last one contains the superalignment (supermatrix, concatenation) tree. Furthermore the simulated superalignment is in `alg.phy` and the gene alignments are in `alg{i}.phy`.

6 Examples

A simulation with coalescent model and no tree reconstruction:

```
SuPi.py -o 1 -e 1 -s example/matrix.phy -t example/matrix.sec -d results/ -v 0.005
-p n
```

A simulation where each gene tree has its own model tree:

```
SuPi.py -o 1 -e 1 -s example/matrix.nex -t example/matrix.sec -d results2/ -u
```

A simulation with doubled alignment length, 10 bootstrap replicates, random tree length scaling and the GTR model for reconstruction:

```
SuPi.py -o 1 -e 1 -s example/matrix.phy -t example/matrix.sec -d results3/ -l 2 -b
10 -f example/random.txt -i '-m GTR -w gamma'
```

7 Acknowledgements

We thank Gregory Ewing for providing his program GenGTree that simulates gene trees under the coalescent model.

References

Ewing, Gregory B, Ingo Ebersberger, Heiko A Schmidt, and Arndt von Haeseler (2008) Rooted triple consensus and anomalous gene trees. *BMC Evol. Biol.*, **8**:118.

Kupczok, Anne, Heiko A Schmidt, and Arndt von Haeseler (2010) Accuracy of phylogeny reconstruction methods combining overlapping gene data sets. *submitted to Alg. Mol. Biol.*

Rambaut, Andrew and Nick C Grassly (1997) Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, **13**(3):235–238.

Vinh, Le Sy and Arndt von Haeseler (2004) IQPNNI: Moving Fast Through Tree Space and Stopping in Time. *Mol. Biol. Evol.*, **21**(8):1565–1571.