User Level Documentation

Application of a test for reversibility of Markov chains in molecular evolution

Pipeline for the application of the tests Bowker, Stuart, Internal Symmetry and Quasisymmetry on simulated or biological data. The software is executable on the command line and should be able to run on Windows/MacOS/Linux. It comprises two bash scripts, which serve as a link between a C++ program and an R Script for the analysis of the provided data. The program uses IQ-Tree to simulate an alignment (AliSim) provided a tree topology, or to find the maximum likelihood tree (ModelFinder) provided an alignment file, which it then uses for the analysis.

Structure

softproj
│ `analysis_simulation.sh` - Bash script for the analysis of simulated data
│ `analysis_biological.sh` - Bash script for the analysis of real data
└─────bin
-------- │ `analysis.R` - the R script
-------- │ `comp.cpp` - the C++ script
-------- │ `comp` - the compiled C++ script
-------- └─────lib - all libraries and header files needed for the C++ script
---------- │ headerfile1.h
---------- │ headerfile2.h
---------- │ headerfile3.h\

Installation & Requirements

R Script - built under R 4.2.2 (list whichh packages are needed?)
C++ Script - compiled with g++ 9.4.0

There is no installation needed, simply run the desired bash script in the command line, providing the needed input files.

Usage

To start, open a terminal (cmd/Powershell in Windows) and navigate to the path of the program.

`analysis_simulation.sh`

The analysis of simulated data requires a tree file and the parameters for the simulation.

```
bash analysis_simulation.sh -t treefile.tree -m JC -n 1000 -k 100 -s true
```

- -t - the tree file in standard Newick format
- -m - specifies a substitution model to use for the simulation (default: JC)

- -n - specifies the length of the root sequence (default: 500)
- -k - specifies how many simulations to be ran (default: 1)
- -s (optional) - can be true or false, if true it will also compute the saturation tests (default: false)

`analysis_biological.sh`

The analysis of real data requires the multiple sequence alignment file and optionally a tree file. If there is no tree file provided to program will ask to run IQ-Tree for the ML Tree and use that for further computations.

```
bash analysis_biological.sh -a alignment.phy -t treefile.tree -s true
```

(the order of the parameters is important)

- -a - specifies the sequence alignment file in PHYLIP or NEX format
- -t (optional) - the tree file in standard Newick format
- -s (optional) - can be true or false, if true it will also compute the saturation tests (default: false)

Output

Both scripts produce up to 4 `.csv` files.

- `results_raw_<treename>.csv` - contains the raw values of the test statistics for each pair
- `results_raw_<treename>.csv` - contains the p-values against the null hypotheses with significance 0.05
- `results_rev_test.csv` - contains the results of the decision for each pair and test (whether the null hypothesis is retained/rejected)
- `results_sat_test.csv` - contains the results of the decision for each pair and each saturation test (if chosen)

Additionally one PDF file for each of the tests computed:

- `plot_Bowker_test.pdf`
- `plot_Stuart_test.pdf`
- `plot_IS_test.pdf`
- `plot_QS_test.pdf`
- `plot_Sat_Cassius1_test.pdf`
- `plot_Sat_Cassius2_test.pdf`
- `plot_Sat_Chi_test.pdf`

And lastly the result tree with added labels for each test in NEXUS format.

- `<treename>.tree`