Package 'P2C2M'

January 29, 2015

Type Package

Title Posterior Predictive Checks of Coalescent Models

Version 0.6

Date 2015-01-28

Author Michael Gruenstaeudl, Noah Reid

Maintainer Michael Gruenstaeudl <gruenstaeudl.1@osu.edu>

Depends R (>= 3.0.0)

Imports
ape (>= 3.1-4), apTreeshape (>= 1.4-5), ggplot2 (>= 1.0.0), rPython (>= 0.0-5), stringr (>= 0.6.2)

Suggests genealogicalSorting (>= 0.92), phybase (>= 1.3.1), Rmpi (>= 0.6-5), xtermStyle (>= 2.2-4)

Description P2C2M is an R package to conduct posterior predictive checks of coalescent models using gene and species trees generated by BEAST or *BEAST. The functionality of P2C2M can be extended via two third-party R packages that are available from the author websites only: genealogicalSorting (http://www.genealogicalsorting.org) and phybase (http://odyssey.bioinformatics.uga.edu/~lliu/phybase/). To use these optional packages, the installation of the Python libraries NumPy (>= 1.9.0) and DendroPy (= 3.12.0) is required.

License GPL (>= 2)

OS_type unix

NeedsCompilation yes

SystemRequirements gcc (>= 4.9), Python (= 2.7)

R topics documented:

P2C2M-package *Posterior Predictive Checks of Coalescent Models*

Description

P2C2M provides functions to read default output from BEAST (Drummond and Rambaut 2007) and *BEAST (Heled and Drummond 2010) and conduct posterior predictive checks of coalescent models (Reid et al. 2014) with the help of data simulation and summary statistics under various settings.

Note

Installation Instructions

To use **P2C2M**, the default version of Python must be set to Python 2.7. Users of unix-like operating systems can insure that this requirement is fulfilled by setting the following alias:

```
echo 'alias python=python2.7' >> ~/.bashrc
```

Mandatory and optional dependencies of **P2C2M** can be installed automatically via two installation scripts that are co-supplied with the package. These scripts were designed for unix-like operating systems and are located in folder /exec. To use these installation scripts, a correct configuration of python2-setuptools is required. Users of unix-like operating systems can insure a correct configuration by setting the following alias:

```
echo 'alias python-config=python2-config' >> ~/.bashrc
```

To execute the R installer, please run the following commands in R:

```
source('/path_to_P2C2M/exec/P2C2M.installRlibs.R'); p2c2m.install()
```

To execute the Python installer, please run the following command in a terminal:

```
python /path_to_P2C2M/exec/P2C2M.installPylibs.py
```

Special Note for MacOS

Users of the MacOS operating system need to install the dependencies manually. Prior to their installation, please confirm that file '/usr/bin/python2-config' exists in your file system and that it points to the Python 2.7 executable. Please refer to http://cran.r-project.org/bin/macosx/RMacOSX-FAQ.html on how to install R packages manually. For the manual installation of Python libraries, please refer to http://docs.python.org/2/using/mac.html

Study Design Requirements

In the user-supplied data set, every species should be represented by at least two alleles. Species that are represented by only a single allele, by contrast, must be specified via option `"single.allele"` and thereby are not included in the calculation of the summary statistic 'GSI'; misspecifications causes **P2C2M** to print the error message *'Error: given group represents one or fewer taxa. Cannot compute index.'*).

Input File Requirements

In order to execute **P2C2M**, a user must provide a directory with three different types of input files: (a) a file that contains species trees, (b) a file that contains gene trees for each gene under study, and (c) an XML-formatted file generated by BEAUTi, the input generator of BEAST (Drummond and Rambaut 2007). A species tree file contains a draw of s generations from the posterior distribution of species trees. Each gene tree file contains an equally large draw from the respective posterior distribution of ultrametric genealogies. Please note that the generations recorded in the species tree file must match those in the gene tree files exactly. The input file generated by BEAUTi is

formatted in XML markup language and represents the starting point for a species tree inference in *BEAST. Here, it provides information on allele and species names, the association between alleles and species, and ploidy levels to **P2C2M**.

File Name Requirements

The following requirements for input file names are in place: The species tree file must be named 'species.trees'. Each gene tree file must be named 'g.trees', where the letter g is substituted with the actual name of the gene. The name of the xml-formatted input file is not constrained and at the discretion of the user. Please be aware that **P2C2M** uses the name of the xml-formatted input file name to label all subsequent output of the package.

Author(s)

Michael Gruenstaeudl, Noah Reid

Maintainer: Michael Gruenstaeudl <gruenstaeudl.1@osu.edu>

References

Drummond, A.J. and Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, **7**, 214.

Gruenstaeudl, M., Reid, N.M., Wheeler, G.R. and Carstens, B.C., submitted. Posterior Predictive Checks of Coalescent Models: P2C2M, an R package.

Heled, J. and Drummond, A.J. (2010) Bayesian inference of species trees from multilocus data. *Molecular Biology And Evolution*, **27**, 570–580.

Reid, N.M., Brown, J.M., Satler, J.D., Pelletier, T.A., McVay, J.D., Hird, S.M. and Carstens, B.C. (2014) Poor fit to the multi-species coalescent model is widely detectable in empirical data. *Systematic Biology*, **63**, 322–333.

`p2c2m.complete` *Execute the complete P2C2M pipeline via a single command*

Description

This function executes the complete **P2C2M** pipeline from beginning to end.

Usage

```
p2c2m.complete(path = "/home/user/Desktop/", xml.file = "beast.xml",
    descr.stats = "COAL_REID,NDC", beast.vers = "1.8", single.allele = c("O"),
    num.reps = 100, use.sorted = FALSE, use.mpi = FALSE, verbose = FALSE,
    dbg = FALSE)
```

Arguments

`path`	the absolute file path to the input directory, specified as a double-quoted string; if `"/home/user/Desktop/"` (the default), then the desktop itself is considered the input directory.
`xml.file`	the name of the BEAUTi-generated and XML-formatted input file, specified as a double-quoted string. The default is `"beast.xml"`.

`descr.stats` the name(s) of the summary statistic(s) to be applied, specified as a double-quoted string. If multiple statistics are specified, they must be separated by commas. A total of four summary statistics is currently available: `"COAL_LIU"` and `"COAL_REID"` (both Rannala & Yang 2013), `"GSI"` (Cummings et al. 2008), `"NDC"` (Maddison 1997). The default is `"COAL_REID,NDC"`.

`beast.vers` the version of *BEAST (Heled and Drummond 2010) used to perform the species tree inference, specified as a double-quoted string. Data parsers are located in the subdirectory exec/. Currently, the following parsers are available: `"1.7"` and `"1.8"`. The default is `"1.8"`.

`single.allele` the name of a species that is represented by only a single allele, specified as a variable of mode vector. This setting is useful when defining an outgroup, because the species so defined does not contribute towards the calculation of the summary statistic 'GSI'. The default is `c("O")`.

`num.reps` the number of simulation replicates to be conducted, specified as an integer. The default is `100`.

`use.sorted` a logical specifying if the summary statistics generated from the posterior and from the posterior predictive distribution are to be ranked by magnitude prior to the calculation of the differences and the formation of the test distribution. The default is `FALSE`. This argument is only EXPERIMENTAL and should not be selected by regular users.

`use.mpi` a logical specifying if **P2C2M** utilizes multiple computer CPUs (if such exist on the system) in order to speed up the calculations. Computations are then executed as parallel processes. The default is `FALSE`.

`verbose` a logical specifying if **P2C2M** prints status information to the screen. The default is `FALSE`.

`dbg` a logical specifying if **P2C2M** is to be run in a debug mode. If `TRUE`, then (a) only the first 5 percent of input trees are analyzed, (b) intermittent results are saved to file, and (c) information useful for debugging is printed to the screen. Argument `dbg = TRUE` must be set in combination with argument `verbose = TRUE`. The default is `FALSE`. This argument is intended for developers and should not be selected by regular users.

Value

The results of a **P2C2M** run comprise test statistics, measures of data dispersion and deviations marked at several quantile levels (analogous to P-values under different alpha-levels in a parametric simulation) for each gene under study and of the sum of all genes.

Author(s)

Michael Gruenstaeudl, Noah Reid

Maintainer: Michael Gruenstaeudl <gruenstaeudl.1@osu.edu>

References

Cummings, M.P., Neel, M.C. and Shaw, K.L. (2008) A genealogical approach to quantifying lineage divergence. *Evolution*, **62**, 2411–2422.

Gruenstaeudl, M., Reid, N.M., Wheeler, G.R. and Carstens, B.C., submitted. Posterior Predictive Checks of Coalescent Models: P2C2M, an R package.

Heled, J. and Drummond, A.J. (2010) Bayesian inference of species trees from multilocus data. *Molecular Biology And Evolution*, **27**, 570–580.

Maddison, W.P. (1997) Gene trees in species trees. *Systematic Biology*, **46**, 523–536.

Rannala, B. and Yang, Z. (2003) Bayes Estimation of Species Divergence Times and Ancestral Population Sizes Using DNA Sequences From Multiple Loci. *Genetics*, **164**, 1645–1656.

Examples

```
## Example of the minimal data requirements to run P2C2M

# The absolute path to the input directory is set
inPath <- system.file("extdata", "sim.E.003.small/", package="P2C2M")

# The name of the xml-file generated by BEAUTi and located in
# "inPath" is set
inFile <- "sim.E.003.small.xml"

# Posterior predictive simulations with a setting of 2 simulation
# replicates are preformed
sim.E.003.small <- p2c2m.complete(inPath, inFile, num.reps=2)
```

---

sim.E.003.small *Output of the data set 'sim.E.003.small'*

Description

The results that are generated when data set 'sim.E.003.small' is analyzed under the default settings of P2C2M.

---

viz_example_1 *Data set of visualization example 1*

Description

A data set selected to illustrate the differences between calculations performed with and without simulation replication. See the package vignette for details.

---

viz_example_2 *Data set of visualization example 2*

Description

A data set selected to illustrate cases of poor fit to the multispecies coalescent model as detected by P2C2M. See the package vignette for details.

Index