

# Package ‘P2C2M’

October 23, 2014

**Type** Package

**Title** Posterior Predictive Checks of Coalescent Models

**Version** 0.5

**Date** 2014-10-23

**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 and \*BEAST, respectively. 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, installation of the Python libraries NumPy (>= 1.9.0) and DendroPy (= 3.12.0) is necessary.

**License** GPL (>= 2)

**OS\_type** unix

**NeedsCompilation** yes (automatic)

**SystemRequirements** Python (= 2.7)

## R topics documented:

P2C2M-package	2
p2c2m.complete	3
sim.E.003.small	5
viz_example_1	5
viz_example_2	5

Index	6
-------	---

## Description

**P2C2M** provides functions to read default output from BEAST (Drummond and Rambaut 2007) and \*BEAST (Heled and Drummond 2010) and to conduct posterior predictive checks of coalescent models (Reid et al. 2014) via several descriptive statistics and 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 this requirement 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 dependencies manually. Prior to 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, and to <http://docs.python.org/2/using/mac.html> for the manual installation of Python libraries.

### Study Design Requirements

Every species should be represented by at least two alleles in a user's data set. While species that are represented by only a single allele do not terminate an analysis in **P2C2M**, they cannot be included in the calculation of some descriptive statistics. The statistic 'GSI', for example, requires that all species are represented by at least two alleles; otherwise it prints 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 folder with three different types of input files: (a) a file with species trees, (b) a file with gene trees for each gene under study, and (c) a 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 usually 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., in prep. 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 = "GTP,NDC", beast.vers = "1.8", single.allele = c("0"),
  num.reps = 100, use.sorted = FALSE, use.mpi = FALSE, verbose = FALSE,
  dbg = FALSE)
```

### Arguments

path	an absolute file path to an input folder, specified as a double-quoted string; if "/home/user/Desktop/" (the default), then the desktop itself is considered the input folder.
xml.file	the name of the file in .xml-format that was generated by BEAUTi in order to perform a species tree inference in *BEAST, specified as a double-quoted string; if "beast.xml" (the default), then the BEAUTi file in .xml-format is expected to be named "beast.xml".

<code>descr.stats</code>	the name(s) of the descriptive statistic(s) the user wishes to utilize, specified as a double-quoted string, whereby the individual statistics are separated by commas. Four descriptive statistics are currently available: "GSI", "GTP", "NDC", and "RAY". The default is "GTP,NDC".
<code>beast.vers</code>	the version of BEAST/*BEAST used to perform the species tree inference, specified as a double-quoted string. Appropriate parsers must be located in the R subdirectory <code>exec/</code> to be recognized. Currently, the possible choices are: "1.7" and "1.8". By default, the version of BEAST/*BEAST is set to "1.8".
<code>single.allele</code>	the name of those species that are represented by only a single allele, specified as a variable of mode vector. Species so defined do not contribute towards the calculation of descriptive statistic 'GSI'. By default, the name is set to <code>c("O")</code> .
<code>num.reps</code>	the number of simulation replicates to be conducted, specified as an integer. By default, the number of simulation replicates is set to 100.
<code>use.sorted</code>	a logical specifying if the descriptive statistics from the posterior distribution and the posterior predictive distributions are ranked by magnitude prior to the calculation of differences. By default, this parameter is set to FALSE.
<code>use.mpi</code>	a logical specifying if <b>P2C2M</b> takes advantage of multiple CPUs present on the system (if such exist) in order to speed up the calculation of descriptive statistic. By default, this parameter is set to FALSE.
<code>verbose</code>	a logical specifying if <b>P2C2M</b> prints status information to the screen. Default is FALSE.
<code>dbg</code>	a logical specifying if <b>P2C2M</b> is to be run in its 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 during the code execution. Argument <code>dbg = TRUE</code> must be set in combination with argument <code>verbose = TRUE</code> . Default is FALSE.

## Details

The default setting of parameter `single.allele` expects that the data set contains an outgroup that consists of only a single allele.

## Value

Description of the output here.

## 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.

## Examples

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

# The absolute path to the input folder 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	<i>Default output of input data set "sim.E.003.small"</i>
-----------------	---

---

### Description

A data set for illustrating the default output of the co-supplied input data set "sim.E.003.small".

---

viz_example_1	<i>Data set of visualization example 1</i>
---------------	--

---

### Description

A data set for illustrating the differences between ranked and unranked descriptive statistics. See package vignette "Result Visualization for P2C2M" for details.

---

viz_example_2	<i>Data set of visualization example 2</i>
---------------	--

---

### Description

A data set for illustrating the distribution of false positive results. See package vignette "Result Visualization for P2C2M" for details.

# Index

P2C2M (P2C2M-package), [2](#)

P2C2M-package, [2](#)

p2c2m.complete, [3](#)

sim.E.003.small, [5](#)

viz\_example\_1, [5](#)

viz\_example\_2, [5](#)