Package ‘indelmiss’

September 15, 2023

Type Package
Title Insertion Deletion Analysis While Accounting for Possible Missing Data
Version 1.0.10
Date 2023-09-13
Author Utkarsh J. Dang and G. Brian Golding
Maintainer Utkarsh J. Dang <utkarshdang@cunet.carleton.ca>
Description Genome-wide gene insertion and deletion rates can be modelled in a maximum likelihood framework with the additional flexibility of modelling potential missing data using the models included within. These models simultaneously estimate insertion and deletion (indel) rates of gene families and proportions of “missing” data for (multiple) taxa of interest. The likelihood framework is utilized for parameter estimation. A phylogenetic tree of the taxa and gene presence/absence patterns (with data ordered by the tips of the tree) are required. See Dang et al. (2016) <doi:10.1534/genetics.116.191973> for more details.
License GPL (>= 2)
Imports Rcpp (>= 0.11.2), ape (>= 3.2), numDeriv (>= 2012.9.1), phangorn (>= 1.99.13)
LinkingTo Rcpp
Suggests testthat
Depends R (>= 2.10)
LazyData true
NeedsCompilation yes
Repository CRAN
Date/Publication 2023-09-15 16:52:05 UTC

R topics documented:

- indelmiss-package .................................................. 2
- gardnerelladata ....................................................... 3
- indelrates ............................................................. 4
indelmiss-package

Insertion deletion analysis while accounting for possible missing data.

Description

Genome-wide gene insertion and deletion rates can be modelled in a maximum likelihood framework with the additional flexibility of modelling potential missing data using the models included within. These models simultaneously estimate insertion and deletion (indel) rates of gene families and proportions of “missing” data for (multiple) taxa of interest. The likelihood framework is utilized for parameter estimation. A phylogenetic tree of the taxa and gene presence/absence patterns (with data ordered by the tips of the tree) are required. For more details, see Utkarsh J. Dang, Alison M. Devault, Tatum D. Mortimer, Caitlin S. Pepperell, Hendrik N. Poinar, G. Brian Golding (2016). Gene insertion deletion analysis while accounting for possible missing data. Genetics (accepted).

Details

Package: indelmiss
Type: Package
Version: 1.0.7
Date: 2016-08-21
License: GPL (>=2)

Author(s)

Utkarsh J. Dang and G. Brian Golding
<utkarshdang@cunet.carleton.ca>

References

**gardnerelladata**


Wickham, Hadley (2012). stringr: Make it easier to work with strings. R package version 0.6.2.

---

**gardnerelladata**

_Gardnerella vaginalis phyletic data (gene family memberships) and phylogenetic tree_

---

**Description**

Gene family memberships for thirty five species from the genus _Gardnerella vaginalis_.

**Usage**

data("gardnerelladata")

**Format**

Contains a list that comprises a tree (called "tree") and phyletic gene family membership data (called "phyl") as its components. The tree is in the ape package phylo format. The data component consists of a matrix of 0/1 patterns with the different patterns as the rows and the 35 taxa as the columns. An entry of 1 (0) describe gene presence (absence) in the taxa.

**References**


**Examples**

data(gardnerelladata)
indelrates

Function for estimating indel rates while accounting for missing data.

Description

This is the function used for running the four gene gain/loss rate models. The four models being run are "M1", "M2", "M3", and "M4". The first model only estimates indel rates where both the insertion and deletion rates are the same. The second model tries to account for possible missing data while estimating indel rates. The third model estimates insertion and deletion rates separately. The fourth model tries to account for possible missing data while estimating insertion and deletion rates separately. See modelnames in the arguments below.

Usage

indelrates(verbose = FALSE, usertree = NULL, userphyl = NULL, 
matchtipstodata = FALSE, datasource = "user", seed = 1, taxa = 5, 
brlensh = c(1, 4), mu = 1, nu = 1, phyl = 5000, 
pmiss = 0, toi = 1, bgtype = "listofnodes", bg = NULL, 
zerocorrection = TRUE, rootprob = "stationary", rpvec = NULL, 
optmethod = "nlminb", init = 0.9, lowlim = 0.001, uplim = 100, 
umhessian = TRUE, modelnames = c("M1", "M2", "M3", "M4"), ...)

Arguments

verbose

If TRUE, will print out estimates and standard errors from each model. Note that the order of printing is the deletion rate followed by the insertion rate. Default value is FALSE.

usertree

Used with datasource = "user". Rooted binary tree of class "phylo". Read in Newick tree using read.tree() from package ape before passing this argument. The branch lengths must be in expected substitutions per site. Popular tree estimation programs like MrBayes and BEAST, for instance, yield branch lengths on that scale.

userphyl

Used with datasource = "user". Matrix of gene phyletic patterns. Matrix entries denote presence (1) or absence (0) of a specific gene (rows) for an individual taxon (columns).

matchtipstodata

The default is FALSE, which means that the user must ensure that the ordering of the taxa in the data matrix must match the internal ordering of tip labels of the tree. Set to TRUE, if the column names of the data matrix, i.e., the taxa names, are all present, with the same spelling and notation in the tip labels of the tree provided, in which case, the restriction on ordering is not necessary.

datasource

"simulation" or "user". Default is "user".

seed

seed number for replication (default is 1).

taxa

Used with datasource = "simulation". Total number of taxa (default is 5). Tree is generated using ape package.

zerocorrection

FALSE by default.

rootprob

"stationary" by default.

rpvec

NULL by default.

optmethod

"nlminb" by default.

init

0.9 by default.

lowlim

0.001 by default.

uplim

100 by default.

numhessian

TRUE by default.

modelnames

"M1", "M2", "M3", "M4" by default.
### brlensh
Used with datasource = "simulation". Branch lengths are simulated from a beta distribution. The shape parameters can be provided as a vector (default = c(1, 4)). For simulating closely related sequences (i.e., with smaller branch lengths more important in trees with more taxa), use something like c(1, 8).

### mu
(unstandardized) deletion rate (default is 1). Used with datasource = "simulation". Sequences given the trees are generated after the rate matrix is standardized such that one unit of time is what we expect to see one change per gene site in. (cf. Equation 13.14 in Felsenstein, 2004.)

### nu
(unstandardized) insertion rate (default is 1). Used with datasource = "simulation". An example of the conversion that phangorn implements is simulating with mu = 0.6 and nu = 3 when given mu = 1 and nu = 5. Similarly, phangorn simulates with mu = 0.67 (0.75) and nu = 2 (1.5) when given mu = 1 (1) and nu = 3 (2).

### phyl
Used with datasource = "simulation". Total number of gene phyletic patterns (default is 5000).

### pmiss
Proportion of present genes to remove in the taxa of interest for the purpose of simulating "missingness". Primarily provided for used with datasource = "simulation" but can be used with datasource = "user" as well.

### toi
A vector of tiplabels for taxa of interest as determined by using ape::tiplabels() function while plotting a tree of class "phylo". Models "M2" and "M4" will account for possible missing data for these taxa. Using toi = "all" specifies all tips (although this often leads to over-parameterization). Default value is tip 1.

### bgtype
If clade-specific insertion and deletion rates are required to be estimated, specify as "ancestornodes". If, on the other hand, a group of branches (not in a clade) are hypothesized to follow the same rates, use argument option "listofnodes".

### bg
A vector of nodes can be given here if the "ancestornodes" option was chosen for argument "bgtype". If "listofnodes" was chosen, a list should be provided with each element of the list being a vector of nodes that limit the branches that follow the same rates. See examples.

### zerocorrection
Felsenstein's correction for unobserved data. The results are conditional on observing the gene present in at least one taxa. This accounts for the sampling bias. Default is TRUE.

### rootprob
Four options are available for the prior probability of character state frequencies at the root. The "equal" option makes gene absence and presence contribute equally to the likelihood calculation at the root of the tree. The "stationary" option will weight the contributions by the averaged equilibrium frequencies of all different branch groupings. The "maxlik" option estimates the probability of gene family absence at the root. Default is "stationary". If "user" is supplied here (e.g., for empirical frequencies), a vector of root frequency parameters can be provided to argument "rpvec".

### rpvec
If option "user" is specified for argument "rootprob", supply a vector of length two, representing the root frequency parameters.

### optmethod
"nlminb" (default) or "optim". The "indelrate" model being a one-dimensional problem always uses "nlminb". The "optim" option is provided purely as a fallback; in most cases, "nlminb" will converge to the correct solution more stably and quickly as compared to "optim".
init

Initial value for the rates. The default value is 0.9.

lowlim

For finer control of the boundaries of the optimization problem. The default value is 0.001. This usually suffices if the branch lengths are in expected substitutions per site. However, if branch lengths are in different units, this should be changed accordingly.

uplim

For finer control of the boundaries of the optimization problem. The default value is 100. This usually suffices if the branch lengths are in expected substitutions per site. However, if branch lengths are in different units, this should be changed accordingly.

deltamhessian

Set to FALSE if standard errors are not required (or standard errors are being calculated using bootstrapping). This speeds up the algorithm. Default is TRUE.

modelnames

Default is all models. The four options are "M1", "M2", "M3", and "M4". A vector of these values can be given. The first model only estimates indel rates where both the insertion and deletion rates are the same. The second model tries to account for possible missing data while estimating indel rates. The third model estimates insertion and deletion rates separately. The fourth model tries to account for possible missing data while estimating insertion and deletion rates separately.

... Passing other arguments of the optimization algorithms used. See "nlminb" or "optim" documentation. For example, control = list(trace = 5) with method = "nlminb" will print progress at every 5th iteration.

Details

Gene presence/absence should be coded as 1/0. By default, any datapoints (in the data supplied) greater than 1 are changed to 1 and any rows consisting of zeros only are removed. Gene presence/absence patterns should be ordered by the tips of the tree.

Value

All arguments used while calling the indelrates function are attached in a list. Moreover, the following components are also returned:

call

Function call used.

conv

A vector of convergence indicators for each model run. 0 denotes successful convergence.

time

Time taken in seconds.

tree

The phylogenetic tree used.

bg

List of group of nodes modelled with individual insertion and deletion rates.

results

List of results including modelname, parameter estimates: insertion ("nu") and deletion ("mu") rates and proportion of missing data ("p"), standard errors, number of parameters fit, and AIC and BIC values. For the models with missing data proportions, an estimate (rounded off) of the number of genes missing for the taxa of interest specified is also provided. Furthermore, details from the optimization routine applied are also available.

data_red

Gene phyletic patterns observed.

w

Number of times each gene phyletic pattern was observed.
indelrates

Author(s)
Utkarsh J. Dang and G. Brian Golding
<utkarshdang@cunet.carleton.ca>

See Also
See also print.indelmiss, plot.indelmiss, and plottree.

Examples

###User supplied tree and data###
#Simulate data
library(phangorn)
set.seed(1)
usertree <- rtree(n = 7, br = rbeta(n = 7, shape1 = 1, shape2 = 7))
data <- simSeq(usertree, l = 5000, type = "USER", levels = c(0, 1),
bf = c(1/(1 + 5), 5/(1 + 5)), Q = 1) #1 and 5 correspond to
#unstandardized rates. See item help descriptions on mu and nu.
datab <- matrix(as.numeric(as.character(data)), nrow = 7)
userphyl <- t(datab)
#Run the models.
indel_user <- indelrates(datasource = "user", usertree = usertree,
userphyl = userphyl, toi = 1, zerocorrection = TRUE, rootprob = "stationary",
modelnames = c("M3", "M4"), optmethod = "nlminb",
control = list(trace = 10))
print(indel_user)

#####Simulation#####
#Simulate a dataset with default options and run algorithm.
indel1 <- indelrates(verbose = TRUE, datasource = "simulation",
control = list(trace = 5))
print(indel1)

#Estimate insertion/ deletion rates from gene presence/absence
#data simulated on a simulated five taxon tree.
indel2 <- indelrates(datasource = "simulation", seed = 1, taxa = 5,
brlensh = c(1, 8), mu = 1, nu = 5, phyl = 5000, pmiss = 0, toi = 1,
zerocorrection = TRUE, rootprob = "stationary",
modelnames = c("M1", "M2", "M3", "M4"), optmethod = "nlminb",
control = list(trace = 5))#1 and 5 correspond to unstandardized rates.
#See item help descriptions on mu and nu.
print(indel2)

#With toi="all"
indel3 <- indelrates(datasource = "simulation", seed = 1, taxa = 5,
brlensh = c(1, 8), mu = 1, nu = 5, phyl = 5000, pmiss = c(0, 0.15, 0.25, 0, 0), toi = "all",
zerocorrection = TRUE, rootprob = "maxlik", modelnames = c("M3", "M4"),
optmethod = "nlminb")
print(indel3)

#Compare with
indel3 <- indelrates(datasource = "simulation", seed = 1, taxa = 5,
brlensh = c(1, 8), mu = 1, nu = 5, phyl = 5000, pmiss = c(0.15, 0.25), toi = c(2, 3),
# Here, a vector of ancestor nodes specify the nodes which
# along with all their descendants have unique indel rates.

indel4 <- indelrates(datasource = "simulation", seed = 1, taxa = 10,
brlensh = c(1, 8), mu = 1, nu = 5, phyl = 5000, pmiss = 0, toi = 1,
bgtype = "ancestornodes", bg = c(15), zerocorrection = TRUE, rootprob =
"maxlik", modelnames = c("M3", "M4"), optmethod = "nlminb")
print(indel4)
plot(indel4, model = "M4")

# Above command prints two plots that can be obtained individually.
# These are confidence intervals based on asymptotic normality
# of the maximum likelihood estimators.
# Different confidence interval levels can be specified with the cil option.
plotrates(indel4, model = "M4", ci = TRUE, cil = 95)
plotp(indel4, model = "M4", ci = TRUE, cil = 95)

# This is an alternate (more flexible but potentially less user-friendly)
# way to specify groups of nodes which have unique indel rates.
# A list of nodes is used here.

indel5 <- indelrates(verbose = TRUE, datasource = "simulation", seed = 1,
taxa = 5, brlensh = c(1, 8), mu = 1, nu = 3, phyl = 5000, pmiss = 0,
toi = 1, bgtype = "listofnodes", bg = list(c(7, 1, 2),
c(6, 8, 3, 7, 9, 5, 4, 9)), zerocorrection = TRUE, rootprob = "maxlik",
modelnames = c("M1", "M2", "M3", "M4"), optmethod = "nlminb")

# Mycobacterium data example
data(mycobacteriumdata1)
indel_myco <- indelrates(verbose = TRUE, usertree = mycobacteriumdata1$tree, modelnames = "M4",
usertip = mycobacteriumdata1$tip, matchtipstodata = TRUE,
datasource = "user", toi = c(3, 4, 6:10), bgtype = "listofnodes",
zerocorrection = TRUE, rootprob = "stationary", optmethod = "nlminb",
umhessian = TRUE, control = list(eval.max = 50000, iter.max = 50000))

---

**Description**

Gene family memberships and a phylogenetic tree for ten species from the genus *Mycobacterium*.

**Usage**

```r
data("mycobacteriumdata1")
```
**mycobacteriumdata2**

**Format**

Contains a list that comprises a tree (called "tree") and phyletic gene family membership data (called "phyl") as its components. The tree is in the ape package phylo format. The data component consists of a data frame of 0/1 patterns with the different patterns as the rows and the 10 taxa as the columns. An entry of 1 (0) describe gene presence (absence) in the taxa.

**References**


**Examples**

```r
data(mycobacteriumdata1)
```

---

**mycobacteriumdata2**  
*Alternate Mycobacterium tree.*

**Description**

Gene family memberships and a phylogenetic tree for ten species from the genus *Mycobacterium*.

**Usage**

```r
data("mycobacteriumdata2")
```

**Format**

Contains a list that comprises a tree (called "tree") and phyletic gene family membership data (called "phyl") as its components. The tree is in the ape package phylo format. The data component consists of a data frame of 0/1 patterns with the different patterns as the rows and the 10 taxa as the columns. An entry of 1 (0) describe gene presence (absence) in the taxa.

**Examples**

```r
data(mycobacteriumdata2)
```
Description

Plotting command for use on an object of class "indelmiss". Will draw two plots: one with the estimates for the rates and the other for the "missingness" parameter. plot.indelmiss() calls plotp and plotrates.

Usage

## S3 method for class 'indelmiss'
plot(x, model = NULL, ci = TRUE, cil = 95, ...)

Arguments

x An object of class "indelmiss".
model One of "M1", "M2", "M3", or "M4".
ci TRUE plots confidence intervals around the estimates.
cil Confidence interval level.
... Any further commands to plot.

Value

Draws two plots: one with the estimates for the rates and the other for the "missingness" parameter.

Author(s)

Utkarsh J. Dang and G. Brian Golding
<utkarshdang@cunet.carleton.ca>

See Also

See also indelrates, plotrates and plotp.

Examples

indel <- indelrates(datasource = "simulation", seed = 1, taxa = 5, 
mu = 1, nu = 5, phyl = 5000, pmiss = 0, toi = 1, 
bgtype="ancestornodes", bg = c(7, 9), 
zerocorrection = TRUE, rootprob="maxlik", 
modelnames = c("M1", "M2", "M3", "M4"), 
optmethod = "nlminb")

print(indel)
plot(indel, model="M4")
**plotp**

*Plot estimates for the parameter that accounts for possible missing data*

**Description**

Plotting command for use on an object of class "indelmiss".

**Usage**

```r
plotp(x, model, ci = TRUE, cil = 95, ...)
```

**Arguments**

- `x`: An object of class "indelmiss".
- `model`: One of "M2" or "M4".
- `ci`: TRUE plots confidence intervals around the estimates.
- `cil`: Confidence interval level.
- `...`: Any further commands to plot.

**Value**

Plot estimates for the parameter that accounts for possible missing data. Called by `plot.indelmiss`.

**Author(s)**

Utkarsh J. Dang and G. Brian Golding

<utkarshdang@cunet.carleton.ca>

**See Also**

See also `plot.indelmiss`.

**Examples**

```r
indel <- indelrates(datasource = "simulation", seed = 1, taxa = 5,
 mu = 1, nu = 5, phyl = 5000, pmiss = 0, toi = 1,
 bgtype="ancestornodes", bg = c(7, 9),
 zerocorrection = TRUE, rootprob = "stationary",
 modelnames = c("M1", "M2", "M3", "M4"),
 optmethod = "nlminb")
print(indel)
plotp(indel, model="M4")
```
plotrates  

Plot estimates for insertion and deletion rates

Description

Plotting command for use on an object of class "indelmiss".

Usage

plotrates(x, model, ci = TRUE, cil = 95, ...)

Arguments

x     An object of class "indelmiss".
model One of "M1", "M2", "M3", or "M4".
    ci   TRUE plots confidence intervals around the estimates.
    cil  Confidence interval level.
    ...  Any further commands to plot.

Value

Plot estimates for insertion and deletion rates. Called by plot.indelmiss.

Author(s)

Utkarsh J. Dang and G. Brian Golding
<utkarshdang@cunet.carleton.ca>

See Also

See also plot.indelmiss.

Examples

indel <- indelrates(datasource = "simulation", seed = 1, taxa = 5,
                     mu = 1, nu = 5, phyl = 5000, pmiss = 0, toi = 1,
                     bgtype="ancestornodes", bg = c(7, 9),
                     zerocorrection = TRUE,
                     modelnames = c("M1", "M2", "M3", "M4"),
                     optmethod = "nlminb")

print(indel)
plotrates(indel, model="M4")
plottree

Plot the tree used the branches colored according to the different specified branch groupings (or clades) following unique rates.

Description

Plotting command for use on an object of class "indelmiss".

Usage

plottree(x, toilabel = TRUE, colors = NULL, ...)

Arguments

x
An object of class "indelmiss".

toilabel
If this is TRUE, a plus sign is printed next to the taxa of interest for which a missing data proportion was estimated. Note that the taxa labels being referred to can be seen by using ape::tiplabels().

colors
Vector of colours the same length as length(x$bg). Note that these colours are used to colour the different branch groupings associated with unique insertion and/or deletion rates.

...
Any further commands to ape::plot.phylo.

Value

Plot the tree used the branches colored according to the different specified branch groupings (or clades) following unique rates.

Author(s)

Utkarsh J. Dang and G. Brian Golding
<utkarshdang@cunet.carleton.ca>

See Also

See also plot.indelmiss and plot.phylo.

Examples

indel <- indelrates(datasource = "simulation", seed = 1, taxa = 5,
mu = 1, nu = 5, phyl = 5000, pmiss = 0, toi = 1,
bgtype="ancestornodes", bg = 7,
zerocorrection = TRUE,
modelnames = c("M1", "M2", "M3", "M4"),
optmethod = "nlminb")
print(indel)
plottree(indel,colors=c("blue","red"))
ape::tiplabels()
print.indelmiss

Print summary information from fit

Description

Summary command for use on an object of class "indelmiss". Depending on the model, the rates (mu: deletion; nu: insertion), missing data proportion (p), and prior probability of gene family absence at the root are printed. If branch groupings (or clades) were specified, then the rates (and corresponding standard errors) are displayed in a matrix with the columns representing the different branch groupings (ordered by the subsets of x$bg where x is an object of class "indelmiss"). The rows represent the gene deletion and insertion rate, respectively.

Usage

```r
## S3 method for class 'indelmiss'
print(x, ...)
```

Arguments

- `x`: An object of class "indelmiss".
- `...`: Ignore this.

Value

Output summary information from the fit.

Author(s)

Utkarsh J. Dang and G. Brian Golding
<utkarshdang@cunet.carleton.ca>

See Also

See also `in德尔rates` and `plot.indelmiss`.

Examples

```r
indel <- indelrates(datasource = "simulation", seed = 1, taxa = 5,
                     mu = 1, nu = 3, phyl = 5000, pmiss = c(0.2, 0.5), toi = c(1, 3),
                     zerocorrection = TRUE,
                     modelnames = c("M1", "M2", "M3", "M4"),
                     optmethod = "nlminb")
print(indel)
```
Index

* datasets
  gardnerelladata, 3
  mycobacteriumdata1, 8
  mycobacteriumdata2, 9

gardnerelladata, 3
indelmiss (indelmiss-package), 2
indelmiss-package, 2
indelrates, 4, 10, 14

mycobacteriumdata1, 8
mycobacteriumdata2, 9

plot.indelmiss, 7, 10, 11–14
plot.phylo, 13
plotp, 10, 11
plotrates, 10, 12
plottree, 7, 13
print.indelmiss, 7, 14