# Package ‘WGDgc’

October 23, 2015

**Version** 1.2

**Date** 2015-10-22

**Title** Whole genome duplication detection using gene counts

**Depends** R (>= 3.0.1), phylobase, phyext, ape

**Description** Detection of whole genome duplications and triplications on phylogenies using gene count data, with estimation of background rates of gene duplication and loss and estimation of gene retention rates following whole genome duplications/triplications

**Encoding** UTF-8

**License** GPL (>= 3) | file LICENSE

**URL** http://www.stat.wisc.edu/~ane/wgd/

**NeedsCompilation** no

**Author** Tram Ta [aut],Charles-Elie Rabier Rabier [aut],Cécile Ané [aut, cre]

**Maintainer** Cécile Ané <cecile.ane@wisc.edu>

## R topics documented:

`getEdgeOrder` *list the tree nodes in a post-order traversal*

## Description

Preprocessing to list the edges in a post-order traversal, for future use in likelihood calculation. The output includes information on which edges the birth-death process applies to, and which edges represent a whole genome duplication or triplication event.

## Usage

```
getEdgeOrder(phyloMat,nLeaf,wgdTab)
```

## Arguments

| | |
|---|---|
| `phyloMat` | Matrix representation of the species tree and WGD events |
| `nLeaf` | Number of present-day species (i.e. number of leaves) |
| `wgdTab` | Table representation of WGD events with retention rates |

## Details

This function assumes that speciation nodes in phyloMat are given lower indices than singleton nodes when the tree is read in by phyext, that speciation nodes are in pre-order in `phyloMat`, and that 2 singleton nodes are used to represent each WGD.

## Value

Data frame listing the edges in a post-order traversal, with the following components

| | |
|---|---|
| `child` | index of the edge's child node |
| `edge` | index of the edge, i.e. its row in phyloMat |
| `type` | "BD" if birth-death edge, "WGD" or "WGT" if the edge is modelling a WGD/T event, or "rootPrior" if the edge is parent to the root node |
| `scdsib` | `TRUE` if the edge is listed after a sibling edge, `FALSE` otherwise |

## Author(s)

Cécile Ané

## See Also

`processInput`.

## `getLikGeneCount` *Negative log-likelihood of gene count data*

### Description

Calculates the overall negative log-likelihood of gene count data on a phylogenetic tree under a birth-and-death process and whole genome duplication events.

### Usage

```
getLikGeneCount(para, input, geneCountData, mMax=NULL,
                geomProb=NULL, dirac=NULL, useRootStateMLE=FALSE,
                conditioning=c("oneOrMore", "twoOrMore",
                "oneInBothClades", "none"),
                equalBDrates=FALSE, fixedRetentionRates=TRUE)
```

### Arguments

`para` vector of parameters (see Details)

`input` object output by function `processInput`

`geneCountData` data frame with one column for each species and one row for each family, containing the number of gene copies in each species for each gene family. The column names must match the species names in the tree.

`mMax` maximum number of surviving lineages at the root, at which the likelihood will be evaluated.

`geomProb` inverse of the prior mean number of gene lineages at the root.

`dirac` value for the number of genes at the root, when this is assumed to have a fixed value (according to a dirac prior distribution).

`useRootStateMLE` if `TRUE`, the most likely number of genes at the root is determined for each family separately and is used to evaluate the likelihood function.

`conditioning` type of conditioning for the likelihood calculation. The default is to calculate conditional probabilities on observing families with at least 1 gene copy (see Details in `MLEGeneCount`).

`equalBDrates` if `TRUE`, the duplication and loss rates are equal.

`fixedRetentionRates` if `TRUE`, it uses retention rates present in `input$wgdTab`. If `FALSE`, it uses retention rates in `para`.

### Details

The vector `para` for the parameters to be used is of size 1+number of WGD/Ts if the birth and death rates are assumed equal, or 2+number of WGD/Ts otherwise. It starts with `log(StartingBDrates[1])` if `equalBDrates` is `TRUE`, with `log(StartingBDrates)` otherwise. The remaining components correspond to retention rates.

### Value

negative log-likelihood value

### References

Csuros M and Miklos I (2009). Streamlining and large ancestral genomes in archaea inferred with a phylogenetic birth-and-death model. Molecular Biology and Evolution. **26**:2087-2095.

Charles-Elie Rabier, Tram Ta and Cécile Ané (2013). Detecting and locating whole genome duplications on a phylogeny: a probabilistic approach. Molecular Biology and Evolution. **31**(3):750-762.

### See Also

`MLEGeneCount`, `logLik_CsurosMiklos`.

### Examples

```
tre.string = "(D:{0,18.03},(C:{0,12.06},(B:{0,7.06},
              A:{0,7.06}):{0,2.49:wgd,0:0,2.50}):{0, 5.97});"
tre.phylo4d = read.simmap(text=tre.string)
dat = data.frame(A=c(2,2,3,1), B=c(3,0,2,1), C=c(1,0,2,2), D=c(2,1,1,1));
a = processInput(tre.phylo4d, startingQ=0.9)
getLikGeneCount(log(c(.01,.02)),a,dat,mMax=8,geomProb=1/1.5,
                conditioning="oneOrMore")
```

## logLik_CsurosMiklos

*Log-likelihood of count data on a phylogenetic tree*

### Description

Calculates the probability of gene count data on a phylogenetic tree under a birth-and-death process and whole genome duplication (or triplication) events, conditional on n surviving gene lineages at the root. Also computes the probability of a family going extinct.

### Usage

```
logLik_CsurosMiklos(logLamlogMu, nLeaf, nFamily, phyloMat,
                    geneCountData, mMax, wgdTab, edgeOrder)
```

### Arguments

`logLamlogMu` vector of size 1 or 2, for the log of the duplication and loss rates. When a single rate is provided, the duplication and loss rates are assumed to be equal.

`nLeaf` number of present-day species.

`nFamily` number of gene families.

`phyloMat` a phylogenetic matrix with 4 columns: parent (ancestor node), child (descendant node), time (branch length), and species names. The number of rows is the number of nodes in the tree.

`geneCountData` data frame with one column for each species and one row for each family, containing the number of gene copies in each species for each gene family. The column names must match the species names in the tree.

`mMax` maximum number of surviving lineages at the root, at which the likelihood will be computed.

`wgdTab` data frame with 5 columns: node before event, event type (WGD or WGT) and retention rates of 1, 2 and 3 gene copies. The number of rows is the number of WGD events.

`edgeOrder` a data frame listing the tree edges in post-order traversal with information on which are birth-death and WGD/T edges.

## Value

`loglikRoot` matrix of size `nMax+1` by `nFamily` giving the log likelihood of each gene family given that there are n surviving gene lineages at the root in row n+1. Column k corresponds to family k.

`doomedRoot` probability that a single gene lineage present at the root goes extinct.

`doomedRootLeft`
probability that a single gene lineage at the root goes extinct in the clade on the left side of the root.

`doomedRootRight`
probability that a single gene lineage at the root goes extinct in the clade on the right side of the root.

## Author(s)

Cécile Ané

## References

Csuros M and Miklos I (2009). Streamlining and large ancestral genomes in archaea inferred with a phylogenetic birth-and-death model. *Molecular Biology and Evolution*. **26**:2087-2095.

Charles-Elie Rabier, Tram Ta and Cécile Ané (2013). Detecting and locating whole genome duplications on a phylogeny: a probabilistic approach. *Molecular Biology and Evolution*. **31**(3):750-762.

## See Also

`processInput`, `getEdgeOrder`.

## Examples

```
tre.string = "(D:{0,18.03},(C:{0,12.06},(B:{0,7.06},
              A:{0,7.06}):{0,2.49:wgd,0:0,2.50}):{0, 5.97});"
tre.phylo4d = read.simmap(text=tre.string)
dat = data.frame(A=c(2,2,3,1), B=c(3,0,2,1), C=c(1,0,2,2), D=c(2,1,1,1));
a = processInput(tre.phylo4d, startingQ=0.9)
logLik_CsurosMiklos(log(c(.01,.02)), nLeaf=4, nFamily=4,
                    a$phyloMat,dat,mMax=8,a$wgdTab, a$edgeOrder)
```

## `MLEGeneCount` *Maximum likelihood estimation of gene turnover rates with WGD*

### Description

Uses gene count data to estimates rates of gene duplication and gene loss along a phylogeny with zero, one or more whole genome duplication (WGD) or triplication (WGT) events. Also estimates the gene retention rate after each WGD/WGT event.

### Usage

```
MLEGeneCount(tr, geneCountData, mMax=NULL, geomMean=NULL,
             dirac=NULL, useRootStateMLE=FALSE,
             conditioning=c("oneOrMore", "twoOrMore",
             "oneInBothClades", "none"),
             equalBDrates=FALSE, fixedRetentionRates=FALSE,
             startingBDrates=c(0.01, 0.02),startingQ=NULL)
```

### Arguments

`tr` a species tree in SIMMAP format (see Details).

`geneCountData`
data frame with one column for each species and one row for each family, containing the number of gene copies in each species for each gene family. The column names must match the species names in the tree.

`mMax` maximum number of surviving lineages at the root, at which the likelihood will be computed.

`geomMean` the mean of the prior geometric distribution for the number of genes at the root.

`dirac` value for the number of genes at the root, when this is assumed to have a fixed value (according to a dirac prior distribution).

`useRootStateMLE`
if `TRUE`, the most likely number of surviving genes at the root is determined for each family separately, and is used to calculate the overall likelihood of the data. This value at the root may vary with the parameter values during likelihood optimization.

`conditioning` type of conditioning for the likelihood calculation. The default is to calculate conditional probabilities on observing families with at least 1 gene copy (see Details).

`equalBDrates` if `TRUE`, the duplication and loss rates are constrained to be equal.

`fixedRetentionRates`
if `TRUE`, retention rates from the user-defined tree are fixed and used as provided. If `FALSE`, retention rates are considered as parameters and are estimated by maximum likelihood.

`startingBDrates`
Vector of size 2, for the starting values of the duplication and loss rates. When `equalBDrates=TRUE`, only the first component is used.

`startingQ` Vector of starting values for the retention rates at the WGD and WGT events.

### Details

The tree needs to be in simmap format (version 1.1). This format is similar to the newick parenthetical format, except that branch lengths are given inside brackets where states are indicated at specific times along each branch. Along a given branch, the token "0,18" indicates state 0 for a duration of 18 time units. Tokens are separated with ":". State 0 is used to indicate branch segments where only the birth/death process applies for gene duplications and losses. Labels "wgd" or "WGT" are used for branch segments at WGD events, and "wgt" or "WGT" for segments at WGT events. Such segments need to have a length of 0.

For WGT events, the 2 extra copies are assumed to be retained independently. With retention rate $q$, the probability to retain all 3 gene copies is then $q^2$, the probability to retain 2 gene copies is $2 * q * (1 - q)$, and the probability to retain the original gene only is $(1 - q)^2$.

Four types of conditional likelihoods are implemented. The option `conditioning` should match the data filtering process: use `conditioning="oneOrMore"` if all families with one or more gene copies are included in the data, use `"twoOrMore"` to condition on families having two of more genes, `"oneInBothClades"` if the data set was filtered to include only families with at least one gene copy in each of the two main clades stemming from the root. Unconditional likelihoods are used with `conditioning="none"`.

The `geomMean`, `dirac` and `useRootStateMLE` options are incompatible.

By default, `mMax` is set to the maximum family size for an exact likelihood calculation. For data sets with one or more very large families, this can cause `mMax` to be very large and calculation to be very slow. In such cases, the user can set `mMax` to a lower value to speed up calculations, at the cost of an approximation to the likelihood of families with a larger family size.

### Value

`birthrate` birth or duplication rate

`deathrate` death or loss rate

`loglikelihood` log of the likelihood

`WGDtable` a WGD table with 5 columns: node before WGD/WGT, event type, and probabilities that 1, 2 or 3 gene copies are retained. The number of rows is the number of WGD/WGT events.

`phyloMat` data frame with 5 columns to describe the phylogeny: parent (ancestor node), child (descendant node), time (branch length), species names and edge type (e.g. BD or WGD). The number of rows is the number of nodes in the tree.

`call` initial call to the function

`convergence` optimization convergence flag from the `optim` call. 0 means successful convergence.

`mMax` mMax value used for the likelihood calculations

### Author(s)

Tram Ta, Charles-Elie Rabier

### References

Bailey, N. (1964) The Elements of Stochastic Processes. New York: John Wiley \& Sons

Bollback J. P. (2006) SIMMAP: Stochastic character mapping of discrete traits on phylogenies. *Bioinformatics*. **7**:88

De Bie, T. and Cristianini, N. and Demuth, J.P. and Hahn, M.W. (2006) CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*. **22**:1269–1271

Hahn, M.W. and De Bie, T. and Stajich, J.E. and Nguyen, C. and Cristianini, N. (2005) Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res.*. **15**:1153–1160

Crawford, F., Suchard, M. (2012) Transition probabilities for general birth-death processes with applications in ecology, genetics, and evolution. *J Math Biol*. **65**:553-580

Rabier, C., Ta, T. and Ané, C. (2013) Detecting and Locating Whole Genome Duplications on a phylogeny: a probabilistic approach. *Molecular Biology and Evolution*. **31**(3):750-762.

### See Also

`sampleData1`, `sampleData2` for more examples.

### Examples

```
tre.string = "(D:{0,18.03},(C:{0,12.06},(B:{0,7.06},
              A:{0,7.06}):{0,2.49:wgd,0:0,2.50}):{0, 5.97});"
# tree with a single hypothesized WGD event, along the
# internal edge leading to the MRCA of species A and B
tre.phylo4d = read.simmap(text=tre.string)
tre.phylo   = as(tre.phylo4d, "phylo")
## Not run: plot(collapse.singles(tre.phylo))
dat = data.frame(A=c(2,2,3,1), B=c(3,0,2,1), C=c(1,0,2,2), D=c(2,1,1,1))
MLEGeneCount(tre.phylo4d, dat, geomMean=1.5,
             conditioning="oneOrMore", fixedRetentionRates=TRUE)
```

## processInput *Preprocessing function*

### Description

Checking arguments and preparing data for future optimization

### Usage

```
processInput(tr, equalBDrates=FALSE, fixedRetentionRates=TRUE,
               startingBDrates=c(0.01, 0.02),startingQ=NULL)
```

### Arguments

`tr` a species tree in SIMMAP format (see Details of function MLEGeneCount).

`equalBDrates` if `TRUE`, the duplication and loss rates are equal.

`fixedRetentionRates`
if `TRUE`, retention rates will be fixed to `startingQ` during the future optimization. If `FALSE`, retention rates will be considered as parameters and will be estimated by maximum likelihood.

`startingBDrates`
Vector of size 2 as starting values for the duplication and loss rates. When `equalBDrates=TRUE` only the first component is used.

`startingQ` Vector of starting values for retention rates. Default is 0.5 for all WGD events.

### Details

The vector `para` of starting values for the parameters to be optimized is of size 1+number of WGDs if the birth and death rates are assumed equal, or 2+number of WGDs otherwise. It starts with `log(StartingBDrates[1])` if `equalBDrates` is `TRUE`, with `log(StartingBDrates)` otherwise, and the remaining components (corresponding to the retention rates) are `startingQ` if `startingQ` is provided, 0.5 otherwise.

For WGT events, the 2 extra copies are assumed to be retained independently. With retention rate $q$, the probability to retain all 3 gene copies is then $q^2$, the probability to retain 2 gene copies is $2 * q * (1 - q)$, and the probability to retain the original gene only is $(1 - q)^2$.

`lower` and `upper` are vectors whose sizes correspond to the number of parameters for the lower and upper bounds of the different parameters in a subsequent optimization search. The log of the duplication and loss rates are unconstrained, while duplicate retention rates are constrained in [0,1].

### Value

| | |
|---|---|
| `phyloMat` | data frame to represent the phylogeny. The number of rows is the number of nodes in the species tree. There are 5 columns (Parent, Child, Time, Species, type). |
| `nLeaf` | number of present-day species (i.e. number of leaves) |
| `nNode` | number of nodes in the species tree |
| `wgdTab` | data frame with 5 columns. Each row corresponds to a WGD(s) or WGT(s). The first column gives the node just before the WGD/T. The second column 'type' says if the event is a WGD or WGT. The remaining columns contain the probabilities that only the original gene is retained, or if 2 (or 3) gene copies are retained. |
| `para` | Vector of parameters to be optimized. see Details |
| `lower` | Lower bounds for later optimization. see Details |
| `upper` | Upper bounds for later optimization. see Details |

### Examples

```
tre.string = "(D:{0,18.03},(C:{0,12.06},(B:{0,7.06},
              A:{0,7.06}):{0,2.49:wgd,0:0,2.50}):{0, 5.97});"
tre.phylo4d = read.simmap(text=tre.string)
processInput(tre.phylo4d)
```

## `rgenecount` *Random generation of family sizes*

### Description

Generates gene count data for multiple families along a phylogeny, using background rates of duplication and loss and possible whole genome duplication (WGD) or triplication (WGT) event(s), each with its own retention rate.

### Usage

```
rgenecount(nfam, tre, lambdamu, retention, geomMean=NULL, dirac=NULL,
           conditioning=c("none"))
```

### Arguments

`nfam` number of families to simulate

`tre` a species tree in SIMMAP format.

`lambdamu` vector of size 1 or 2, for the duplication rate ($\lambda$) and loss rate ($\mu$). A vector of size 1 sets $\lambda$=$\mu$.

`retention` vector of length the number of WGD/WGT events in the tree, giving the retention rate at each event.

`geomMean` the mean of the prior geometric distribution for the number of genes at the root.

`dirac` value for the number of genes at the root, if fixed to the same value for all families.

`conditioning` type of filtering. No filtering implemented yet.

### Details

For the simmap format, see `MLEGeneCount`. For WGT events, the 2 extra copies are assumed to be retained independently with the same retention rate. With retention rate $q$, the probability to retain all 3 gene copies is then $q^2$, the probability to retain 2 gene copies is $2 * q * (1 - q)$, and the probability to retain the original gene only is $(1 - q)^2$.

The `geomMean` and `dirac` options are incompatible.

### Value

matrix with `nfam` rows, one per simulated family, and one column per node in the tree (tips and internal nodes).

### Author(s)

Cécile Ané

### Examples

```
# tree with 2 WGDs. The second is placed immediately after
#                   the split between C and AB:
tre.string <- "(D:{0,18.03},(C:{0,12.06},(B:{0,7.06},
  A:{0,7.06}):{0,2.49:wgd,0:0,2.50:wgd,0:0,1e-10}):{0, 5.97});"
tre.phylo4d = read.simmap(text=tre.string)
# do this to see how edges and nodes are numbered,
#                which WGD is the first, which is the second:
processInput(tre.phylo4d, startingQ=c(.6,.2))
rgenecount(nfam=10,tre.phylo4d,lambdamu=c(.03,.04),retention=c(.6,.2),dirac=1)
```

## `sampleData1` *Simulated gene count data with 1 WGD event*

### Description

Sample gene count data simulated with 1 WGD, 4 species (A, B, C, D) and 6000 families.

### Usage

```
data(sampleData1)
```

### Format

A data frame with 6000 observations on the following 4 species as 4 named variables: A, B, C, D.

### Details

These data were generated according to the following species tree (in simmap format version 1.1), with a single WGD event located on the internal edge leading to the MRCA of species A and B and retention rate 0.6: "(D:0,18.03, (C:0,12.06,(B:0,7.06,A:0,7.06):0,2.50 :wgd,0:0,2.50):0, 5.97);" The duplication and loss rates used for simulation were 0.02 and 0.03. Families with 0 or 1 copy were excluded. All families were started with only one ancestral gene at the root of the species tree.

### Examples

```
data(sampleData1)
dat <- sampleData1[1:100,] # reducing data to run examples faster

tree1WGD.str = "(D:{0,18.03}, (C:{0,12.06},(B:{0,7.06},A:{0,7.06})
              :{0,2.50 :wgd,0:0,2.50}):{0, 5.97});"
# tree with a single WGD event along the edge to MRCA of species A and B
tree1WGD = read.simmap(text=tree1WGD.str)

MLEGeneCount(tree1WGD, dat, dirac=1, conditioning="twoOrMore")
# to estimate retention, duplication and loss rates

MLEGeneCount(tree1WGD, dat, dirac=1, conditioning="twoOrMore",
             fixedRetentionRates=TRUE, startingQ=0.6)
# to estimate the duplication and loss rates only,
# based on a hypothesized retention rate 0.6 at the WGD.

filtered <- subset(dat, (A>0| B>0 | C>0) & D>0 )
# families with at least one copy in both clades at the root

MLEGeneCount(tree1WGD, filtered,dirac=1,conditioning="oneInBothClades")
# uses the appropriate filtering

## Analysis under a tree with no WGD
tree0WGD.str = "(D:{0,18.03}, (C:{0,12.06},(B:{0,7.06},A:{0,7.06})
             :{0,5.00}):{0, 5.97});"
tree0WGD = read.simmap(text=tree0WGD.str)
MLEGeneCount(tree0WGD, dat, dirac=1, conditioning="twoOrMore",
             fixedRetentionRates=TRUE)

## Analysis under a tree with 2 events: one WGD and one WGT
tree2events.str = "(D:{0,18.03}, (C:{0,12.06},(B:{0,7.06},A:{0,7.06}):
            {0,2.50 :wgt,0:0,2.50}):{0, 2.985: wgd,0:0,2.985});"
# oldest event: WGD on edge to MRCA of species A, B and C.
# recent event: WGT on edge to MRCA of species A, B
tree2events = read.simmap(text=tree2events.str)
MLEGeneCount(tree2events, dat, dirac=1, conditioning="twoOrMore")
```

## `sampleData2` *Simulated gene count data with two WGD events*

### Description

Sample gene count data simulated with 2 WGDs on the same branch, 4 species (A, B, C, D) and 6000 families.

### Usage

```
data(sampleData2)
```

### Format

A data frame with 6000 observations on the following 4 species as 4 named variables: A, B, C, D.

### Details

These data were generated according to the following species tree (in simmap format version 1.1), with both WGD events located along the internal edge leading species D, with retention rate 0.6 for the oldest event and 0.2 for the most recent event:

"(D:0,6.01:0.2,0:0,6.01:0.6,0:0,6.01, (C:0,12.06,(B:0,7.06,A:0,7.06):0,4.99):0,5.97);"

The duplication and loss rates used for simulation were 0.02 and 0.03. Families with 0 or 1 copy were excluded. All families were started with only one ancestral gene at the root of the species tree.

### Examples

```
data(sampleData2)
dat <- sampleData2[1:200,] # reducing data to run examples faster

tree2WGD.str="(D:{0,6.01:wgd,0:0,6.01:wgd,0:0,6.01}, (C:{0,12.06},
               (B:{0,7.06},A:{0,7.06}):{0,4.99}):{0,5.97});"
# both WGD events are located on the edge leading to species D
tree2WGD = read.simmap(text=tree2WGD.str)
MLEGeneCount(tree2WGD, dat, dirac=1, conditioning="twoOrMore")
```

# Index