Package 'hoardeR'

January 27, 2015

Type Package

Title hoardeR package

Version 0.0-2

Date 2014-03-06

Author Daniel Fischer

Maintainer Daniel Fischer <daniel.fischer@mtt.fi>

Depends R (>= 3.0.2), httr (>= 0.2), XML (>= 3.98-1.1), stringr (>= 0.6.2)

Description Information retrieval from NCBI databases, with main focus on Blast.

License GPL (>= 2)

NeedsCompilation no

Repository CRAN

Date/Publication 2014-06-07 00:37:44

R topics documented:

hoardeR-package *Collect and Retrieve Annotation Data for Various Genetic Data.*

Description

The hoardeR package is designed for collecting and retrieving data from various sources. The current main focus is on setting up a connection to the NCBI Blast service. Also, the gene information for Ensemble Genes can be retrieved from NCBI. Methods for visualizing the results is currently under development. The latest 'night-build' can be retrieved from

https://github.com/fischuu/hoardeR

Details

Package:	hoardeR
Type:	Package
Version:	0.0-2
Date:	2014-03-06
License:	GPL
LazyLoad:	yes

Author(s)

Daniel Fischer

Maintainer: Daniel Fischer <daniel.fischer@mtt.fi>

blastSeq *Sending Genomic Sequences to NCBI Blast service*

Description

This function sends genomic sequences to the NCBI Blast service.

Usage

```
blastSeq(seq, n_blast=20, delay_req=3, delay_rid=60, email=NULL,
         xmlFolder=NULL, keepInMemory=TRUE, database="chromosome",
         verbose=TRUE)
```

Arguments

seq	The fasta sequence that should be blasted (String).
n_blast	Amount of parallel blast requests, in case seq is a vector.
delay_req	Seconds between the single Blast requests.
delay_rid	Seconds between the single result requests.
email	User email, required information from NCBI (String).
xmlFolder	Path to the result folder.
keepInMemory	Logical, shall the results be kept in the memory.
database	The NCBI database to use.
verbose	Shall the program give extensive feedback.

Details

This function sends fasta sequences to the NCBI blast service. The defaults for the delays are required by NCBI and must not be smaller than the default values. Also, NCBI asks the user to provide an email address.

The input seq can be a vector of strings. In that case the sequences are one after another processed. The option n_blast sets then the upper threshold of how many blast requests are send to the NCBI Blast service at a time and kept running there parallel. It is here in the users obligation not to misuse the service with too many parallel requests.

The xmlFolder parameter specifies the folder to where the XML results will be stored. In case the folder does not exist, R will create it.

In case the option keepInMemory is set to TRUE the Blast results will be kept in memory, otherwise they will be just written to the HDD. Especially if many sequences are used to Blast it is recommended to drop the result from the memory.

Value

An xml file that contains the the NCBI result.

Author(s)

Daniel Fischer

Examples

```
## Not run:
blastSeq("ACGTGCATCGACTAGCTACGACTACGACTATC")

## End(Not run)
```

getEnsgInfo *Retrieve Gene Information From The NCBI Database.*

Description

This function retrieves for a given Ensemble Number the corresponding information from the NCBI database.

Usage

```
getEnsgInfo(ensg)
```

Arguments

ensg	Ensemble ID (String).

Details

This function retrieves for a given Ensemble Number the corresponding information from the NCBI database. The object ensg can also be a vector of Ensemble IDs.

Value

A matrix with information.

Author(s)

Daniel Fischer

Examples

```
## Not run:
ensg <- c("ENSG00000174482", "ENSG00000113494")
getEnsgInfo(ensg)

## End(Not run)
```

getGeneLocation *Extracting Gene Locations.*

Description

This function extracts the gene locations from an imported gtf file.

Usage

```
getGeneLocation(gtf)
```

Arguments

gtf	An imported gtf object.

Details

This function extracts the information from an imported gtf object.

Value

A matrix.

Author(s)

Daniel Fischer

Examples

```
## Not run:
getGeneLocation(gtf)

## End(Not run)
```

getGeneSeq *Extracting a gene sequence from NCBI database.*

Description

This function retrieves a gene sequence from the NCBI database.

Usage

```
getGeneSeq(chr, start, end, organism)
```

Arguments

chr	Chromosome number, numeric/string
start	Start position, numeric
end	End position, numeric
organism	Name of the organism, string

Details

Extracting a gene sequence from NCBI database.

Value

A string that contains the genomic sequence.

Author(s)

Daniel Fischer

Examples

```
## Not run:
# Extracting for Sus Scrofa, build version 3:
getGeneSeq(1,2134,14532,"susScr3")

## End(Not run)
```

---

importFA *Importing a Fasta File.*

Description

This function imports a standard fasta file.

Usage

```
importFA(file)
```

Arguments

file	Specifies the filename/path.

Details

This function imports a standard fasta file. It assumes that label and sequence lines are alternating, meaning in the odd lines is the sequence name given, starting with > and in the even rows are the corresponding sequences.

Value

A vector containg the sequences. The vector names coorespond to the sequence names given in the fasta file.

Author(s)

Daniel Fischer

Examples

```
## Not run:
importFA("/home/data/myFasta.fa")

## End(Not run)
```

`importGFF3` *Import a GFF3 File*

Description

This function imports a gff3 file.

Usage

```
importGFF3(gff)
```

Arguments

`gff`	File name of the gff3 file

Details

This function imports a gff file and splits the last column which is usually tricky to handle as the order of the variables is not always the same.

Value

A data frame containing the columns of the gtf file, including the splitted last column.

Author(s)

Daniel Fischer

`importGTF` *Import a GTF File*

Description

This function imports a gtf file.

Usage

```
importGTF(gtf)
```

Arguments

`gtf`	File name of the gtf file

Details

This function imports a gtf file and splits the column 9 which is usually tricky to handle as the order of the variables is not always the same.

Value

A data frame containing the columns of the gtf file, including the splitted last column.

Author(s)

Daniel Fischer

`importXML` *Import a XML File*

Description

This function imports a xml file produced from `blastSeq`.

Usage

```
importXML(seqNames, folder, which=c(1,2), idTH = 0.8, verbose=TRUE)
```

Arguments

`seqNames`	Sequence names.
`folder`	Fodler , where the xml files are stored.
`which`	Which of the provided sequence names should be imported.
`idTH`	Identity threshold, see details.
`verbose`	Logical, function give status messages.

Details

This function imports a xml files produced from the `blastSeq` function. The `idTh` options sets the limit, what the minimum id threshold is until a hit will be taken into the result data frame.

Value

A data frame containing the results.

Author(s)

Daniel Fischer

Index