Package 'geno2proteo'

January 24, 2018

Type Package

Title Finding the DNA and Protein Sequences of Any Genomic or Proteomic Loci

Version 0.0.3

Date 2018-01-24

Author Yaoyong Li

Maintainer Yaoyong Li<liyaoyong85@gmail.com>

biocViews Genetics, Proteomics, Sequencing, Annotation, GenomeAnnotation, GenomeAssembly

Description Using the DNA sequence and gene annotation files provided in 'ENSEMBL' <https://www.ensembl.org/index.html>, the functions implemented in the package try to find the DNA sequences and protein sequences of any given genomic loci, and to find the genomic coordinates and protein sequences of any given protein locations, which are the frequent tasks in the analysis of genomic and proteomic data.

License Artistic-2.0

Depends R (>= 3.0)

Imports S4Vectors, BiocGenerics, GenomicRanges, IRanges, R.utils, RUnit

SystemRequirements Perl (>= 2.0.0)

LazyLoad yes

NeedsCompilation no

Repository CRAN

Date/Publication 2018-01-24 12:25:39 UTC

R topics documented:

`geno2proteo-package` *Finding the DNA and Protein Sequences of Any Genomic or Proteomic Loci*

Description

Using the DNA sequence and gene annotation files provided in 'ENSEMBL' <https://www.ensembl.org/index.html>, the functions implemented in the package try to find the DNA sequences and protein sequences of any given genomic loci, and to find the genomic coordinates and protein sequences of any given protein locations, which are the frequent tasks in the analysis of genomic and proteomic data.

Details

The DESCRIPTION file:

Package:	geno2proteo
Type:	Package
Title:	Finding the DNA and Protein Sequences of Any Genomic or Proteomic Loci
Version:	0.0.3
Date:	2018-01-24
Author:	Yaoyong Li
Maintainer:	Yaoyong Li<liyaoyong85@gmail.com>
biocViews:	Genetics, Proteomics, Sequencing, Annotation, GenomeAnnotation, GenomeAssembly
Description:	Using the DNA sequence and gene annotation files provided in 'ENSEMBL' <https://www.ensembl.c
License:	Artistic-2.0
Depends:	R (>= 3.0)
Imports:	S4Vectors, BiocGenerics, GenomicRanges, IRanges, R.utils, RUnit
SystemRequirements:	Perl (>= 2.0.0)
LazyLoad:	yes

Index of help topics:

`generatingCDSaaFile`	`Generating a file containing the DNA and AA sequences`
`geno2proteo-package`	`Finding the DNA and Protein Sequences of Any Genomic o`
`genomicLocsToProteinSequence`	`Obtaining the protein sequences and DNA sequences of`
`genomicLocsToWholeDNASequence`	`Obtaining the DNA sequences of a list of genomic loc`
`proteinLocsToGenomic`	`Obtaining the genomic coordinates for a list of protei`
`proteinLocsToProteinSeq`	`Obtaining the amino acid sequences of a list of prote`

~~ An overview of how to use the package and the most important functions ~~

The package needs three data files. One contains the genetic table, another one contains the DNA sequences of a certain genome, and the third one contains the gene annotations in the same format

as the GTF file used in ENSEMBL. The standard genetic table was provided in this package. The other two data files can be downloaded from ENSEMBL web site. For details about those data files and how to obtain them, see the introduction document of this package. Of course you can create your own data files, as long as you observe the format of the files which are specified in the introduction document as well.

The package also needs Perl installed and being available for use in order to run some of the functions.

Once you have the three data files, you need to run the function `generatingCDSaaFile` to generate another data file, which will be used by some of the functions in this package.

Four main functions were implemented in this package so far. The function `genomicLocToProteinSequence` will find the protein sequences and DNA sequences of the coding regions within a list of genomic loci given as input. `genomicLocToWholeDNASequence` will obtain the whole DNA sequences of any genomic loci given in the input data. `proteinLocsToGenomic` will find the genomic coordinates for a list of sections in proteins as input. `proteinLocsToProteinSeq` will find the the protein sequences of a list of the protein sections.

Author(s)

Yaoyong Li

Maintainer: Yaoyong Li<liyaoyong85@gmail.com>

`generatingCDSaaFile` *Generating a file containing the DNA and AA sequences of all the protein coding regions (CDSs) in a genome*

Description

This function will find the DNA and protein sequences for each CDS region listed in the ENSEMBL gene annotation file (gtf file) provided, and store the CDS regions and the corresponding DNA and protein sequences in an output data file. The output data file will be needed by some of the functions in this package.

Usage

```
generatingCDSaaFile(geneticCodeFile_line, gtfFile, DNAfastaFile,
outputFolder = "./", perlExec='perl')
```

Arguments

`geneticCodeFile_line`

A text file containing a genetic coding table of the codons and the amino acids coded by those codons. A file containing the standard genetic coding table is provided in this package, which is the file "geneticCode_standardTable_lines.txt" in the folder "geno2proteo/extdata/" which will be located in the folder that you install the package. This will be available after you have installed the package.

Alternatively you can create your own genetic table if needed. The format of the genetic table used in this package is one line for one codon. The first column is the codon (namely 3 DNA letters) and the second column is the name (in a single letter) of the amino acid coded by the codon. You may have more columns if you want but only the first two columns are used by the package, and the columns are separated by a tab.

`gtfFile` A text file in GTF format containing the gene annotations of the species that you are interested in. You may obtain this file from the ENSEMBL web site. The text file can be compressed by GNU Zip (e.g. gzip). For the details about how to get the data file from ENSEMBL web site, see the documentation of this package.

`DNAfastaFile` A text file in fasta format containing the DNA sequence for the genome that you want to use for analysing the data. The text file can be compressed by GNU Zip (e.g. gzip). You may download the file directly from the ENSEMBL web site. For the details about how to get the data file from ENSEMBL web site, see the documentation of this package.

`outputFolder` A folder where the result file will be stored. The default value is the current folder "./".

`perlExec` Its value should be the full path of the executable file which can be used to run Perl scripts (e.g. "/usr/bin/perl" in a linux computer or "C:/Strawberry/perl/bin/perl" in a Windows computer). The default value is "perl".

Details

This function generates a data file containing the genomic locations, DNA sequences and protein sequences of all of the coding regions (CDSs), which will be used by some of other functions in this packages.

Value

This function does not return any specific value but generates a file in the output folder containing the DNA and AA sequences of the CDS regions. The first part of the name of the result file is same as the gene annotation file (namely the gtf file), and the last part is '_AAseq.txt.gz'.

Author(s)

Yaoyong Li

Examples

```
# the data folder in this package
dataFolder = system.file("extdata", package="geno2proteo")

geneticCodeFile_line = file.path(dataFolder,
                                 "geneticCode_standardTable_lines.txt")
gtfFile = file.path(dataFolder,
            "Homo_sapiens.GRCh37.74_chromosome16_35Mlong.gtf.gz")
DNAfastaFile =  file.path(dataFolder,
        "Homo_sapiens.GRCh37.74.dna.chromosome.16.fa_theFirst3p5M.txt.gz")
```

```
outputFolder = tempdir(); # using the current folder as output folder
# calling the function.
generatingCDSaaFile(geneticCodeFile_line=geneticCodeFile_line,
    gtfFile=gtfFile, DNAfastaFile=DNAfastaFile, outputFolder=outputFolder)

filename00 = sub(".*/", "", gtfFile)
# get the output file's name.
outputFile = paste(outputFolder, "/", filename00, "_AAseq.txt.gz", sep="")
# read the content of the output file into a data frame.
aaSeq = read.table(outputFile, sep="\t", stringsAsFactors=FALSE)
```

`genomicLocsToProteinSequence` *Obtaining the protein sequences and DNA sequences of the coding regions within a list of loci in genome*

Description

`genomicsLocToProteinSequence` takes a list of genomic loci given in the input and tries to find the protein sequences and DNA sequences of the coding regions of genome which are within those genomic loci.

Usage

```
genomicLocsToProteinSequence(inputLoci, CDSaaFile)
```

Arguments

`inputLoci`	A data frame containing the genomic loci as the input. Each row is for one genomic locus. The first column is for the chromosome, the 2nd and 3rd columns are for the start and end coordinates of the locus in the chromosome, and the 4th column is for the strand ("+" or "-" for forward and reverse strand, respectively). Other columns are optional and will not be used by the function. Note that the chromosome name can be either in the ENSEMBL style, e.g. 1, 2, 3, ..., and X, Y and MT, or in another popular style, namely chr1, chr2, chr3, ..., and chrX, chrY and chrM. But they cannot be mixed in the input of one function call.
`CDSaaFile`	The data file generated by the package's function `generatingCDSaaFile`, containing the genomic locations, DNA sequences and protein sequences of all coding regions in a specific genome which is used in your analysis.

Value

A data frame containing the original genomic loci specified in the input and the protein sequence and the DNA sequence of the coding regions within each of the loci. In detail, the returned data frame contains the original genomic loci specified in the input and after them, the five added columns:

- Column "transId" lists the ENSEMBL IDs of the transcripts whose coding regions overlap with locus specified and the overlapping coding regions are exactly the same among those transcripts.
- Column "dnaSeq" contains the DNA sequence in the overlapping coding regions.
- Column "dnaBefore" contains the DNA letters which are in the same codon as the first letter in the DNA sequence in the column "dnaSeq".
- Column "dnaAfter" contains the DNA letters which are in the same codon as the last letter in the DNA sequence in the previous column 'dnaSeq'.
- Column "pepSeq" contains the protein sequence translated from the DNA sequences in the three preceding columns, "dnaBefore", "dnaSeq" and "dnaAfter".

Author(s)

Yaoyong Li

Examples

```
dataFolder = system.file("extdata", package="geno2proteo")
inputFile_loci=file.path(dataFolder,
    "transId_pfamDomainStartEnd_chr16_Zdomains_22examples_genomicPos.txt")
CDSaaFile=file.path(dataFolder,
    "Homo_sapiens.GRCh37.74_chromosome16_35Mlong.gtf.gz_AAseq.txt.gz")

inputLoci = read.table(inputFile_loci, sep="\t", stringsAsFactors=FALSE)

proteinSeq = genomicLocsToProteinSequence(inputLoci=inputLoci,
                                          CDSaaFile=CDSaaFile)
```

---

genomicLocsToWholeDNASequence

Obtaining the DNA sequences of a list of genomic loci

Description

The function takes a list of genomic loci and tries to find the whole DNA sequences within each of the loci.

Usage

```
genomicLocsToWholeDNASequence(inputLoci, DNAfastaFile,
                              tempFolder = "./", perlExec='perl')
```

Arguments

inputLoci	A data frame containing the genomic loci as the input. Each row is for one genomic locus. The first column is the chromosome name, the 2nd and 3rd columns are the start and end coordinates of the locus in the chromosome, and the 4th column specifies the strand of chromosome ("+" and "-" for forward and reverse strand, respectively). Other columns are optional and will not be used by the function. Note that the chromosome name can be either in the ENSEMBL style, e.g. 1, 2, 3, ..., and X, Y and MT, or in another popular style, namely chr1, chr2, chr3, ..., and chrX, chrY and chrM. But they cannot be mixed in the input of one function call.
DNAfastaFile	The name of a fasta file containing the whole DNA sequence of the genome used. For details about this data file see the documentation of this package.
tempFolder	A temporary folder into which the program can write some temporary files which will be deleted when the function running is finished. The default value is the current folder.
perlExec	Its value should be the full path of the executable file which can be used to run Perl scripts (e.g. "/usr/bin/perl" in a linux computer or "C:/Strawberry/perl/bin/perl" in a Windows computer). The default value is "perl".

Details

This function obtains the whole DNA sequences of a list of genomic loci. Note that, in contrast, another function genomicLocToProteinSequence in this package can return the DNA sequences of the coding regions within the given genomic loci.

Value

The function returns a data frame containing the original genomic loci as in the input and after them, one additional column for the DNA sequence of the corresponding genomic locus.

Author(s)

Yaoyong Li

Examples

```
dataFolder = system.file("extdata", package="geno2proteo")
inputFile_loci=file.path(dataFolder,
    "transId_pfamDomainStartEnd_chr16_Zdomains_22examples_genomicPos.txt")
DNAfastaFile =  file.path(dataFolder,
    "Homo_sapiens.GRCh37.74.dna.chromosome.16.fa_theFirst3p5M.txt.gz")

inputLoci = read.table(inputFile_loci, sep="\t", stringsAsFactors=FALSE)

tmpFolder = tempdir()

DNASeqNow = genomicLocsToWholeDNASequence(inputLoci=inputLoci,
                        DNAfastaFile=DNAfastaFile, tempFolder=tmpFolder)
```

`proteinLocsToGenomic` *Obtaining the genomic coordinates for a list of protein sections*

Description

The function takes a list of protein sections and the corresponding ENSEMBL ID of these proteins, and tries to find the genomic coordinates of these protein sections.

Usage

```
proteinLocsToGenomic(inputLoci, CDSaaFile)
```

Arguments

`inputLoci`	A data frame containing the protein sections as the input. The 1st column must be the ENSEMBL ID of either the protein or the transcript encoding the protein (or the equivalent of ENSEMBL ID if you have created your own gene annotation GTF file). But you have to use only one of two formats (namely either protein ID or transcript ID), and cannot use both of them in the input of one function call. The 2nd and 3rd columns give the coordinate of the first and last amino acids of the section along the protein sequence. Other columns are optional and will not be used by the function.
`CDSaaFile`	The data file generated by the package's function `generatingCDSaaFile`, containing the genomic locations, DNA sequences and protein sequences of all coding regions in a specific genome which is used in your analysis.

Value

The function returns a data frame containing the original protein locations specified in the input and before them, the six added columns for the corresponding genomic coordinates of the protein sections:

- The 1st, 2nd, 3rd and 4th columns give the chromosome name, the coordinates of the start and end positions, and the strand in the chromosome, which specify the genomic locus corresponding to the protein section.
- The 5th and 6th columns give the first and last coding exons in the given transcript which correspond to the given protein section.

Author(s)

Yaoyong Li

Examples

```
dataFolder = system.file("extdata", package="geno2proteo")
inputFile_loci=file.path(dataFolder,
    "transId_pfamDomainStartEnd_chr16_Zdomains_22examples.txt")
CDSaaFile=file.path(dataFolder,
    "Homo_sapiens.GRCh37.74_chromosome16_35Mlong.gtf.gz_AAseq.txt.gz")

inputLoci = read.table(inputFile_loci, sep="\t", stringsAsFactors=FALSE)

genomicLoci = proteinLocsToGenomic(inputLoci=inputLoci, CDSaaFile=CDSaaFile)
```

`proteinLocsToProteinSeq`

Obtaining the amino acid sequences of a list of protein sections

---

Description

Given a list of sections in proteins defined by the ENSEMBL IDs of those proteins and thestart and end coordinates of those sections along the amino acid sequences of the proteins, the function returns the amino acid sequences of those sections.

Usage

```
proteinLocsToProteinSeq(inputLoci, CDSaaFile)
```

Arguments

`inputLoci`	A data frame containing the coordinates of the protein sections in the protein sequences. The 1st column must be the ENSEMBL ID of either the protein or the transcript that the protein corresponds to (or the equivalent of ENSEMBL ID if you have created your own gene annotation GTF file). But you have to use onnly one of two formats (namely protein ID or transcript ID), and cannot use both of them in the input of one function call. The 2nd and 3rd columns give the coordinate of the first and last amino acids of the section in the protein sequence. Other columns are optional and will not be used by the function.
`CDSaaFile`	The data file generated by the package's function `generatingCDSaaFile`, containing the genomic locations, DNA sequences and protein sequences of all coding regions in a specific genome which is used in your analysis.

Value

The function returns a data frame containing the original protein locations specified in the input and after them, one added columnfor the amino acid sequences of the protein sections.

Author(s)

Yaoyong Li

Examples

```
dataFolder = system.file("extdata", package="geno2proteo")
inputFile_loci=file.path(dataFolder,
    "transId_pfamDomainStartEnd_chr16_Zdomains_22examples.txt")
CDSaaFile=file.path(dataFolder,
    "Homo_sapiens.GRCh37.74_chromosome16_35Mlong.gtf.gz_AAseq.txt.gz")

inputLoci = read.table(inputFile_loci, sep="\t", stringsAsFactors=FALSE)

ProtSeqNow = proteinLocsToProteinSeq(inputLoci=inputLoci,
                                     CDSaaFile=CDSaaFile)
```

Index