Package 'MotifDb'

March 26, 2013

Type Package

Title An Annotated Collection of Protein-DNA Binding Sequence Motifs

Version 1.0.0

Date 2012-08-12

Author Paul Shannon

Maintainer Paul Shannon <pshannon@fhcrc.org>

Depends R (>= 2.15.0), methods, IRanges, Biostrings

Suggests RUnit, MotIV, seqLogo

Imports BiocGenerics, rtracklayer

Description More than 2000 annotated position frequency matrices from five public source, for multiple organisms

License Artistic-2.0

LazyLoad yes

biocViews GenomicSequence, MotifAnnotation

R topics documented:

export *export*

Description

Writes all matrices in the supplied list, in the specified format, to the specified connection.

Usage

```
## S4 method for signature 'MotifList,connection,character'
export(object, con, format, ...)
## S4 method for signature 'MotifList,character,character'
export(object, con, format, ...)
## S4 method for signature 'MotifList,missing,character'
export(object, con, format, ...)
```

Arguments

object	a MotifList object.
con	either a file connection or a filename or missing, implying stdout.
format	a character string, currently only 'meme' and 'transfac', which both produce the same result
...	ignore this

Value

The matrices list is written to the specified connection in the specified format.

Author(s)

Paul Shannon

See Also

MotifDb, query, subset, flyFactorSurvey, hPDI, jaspar, ScerTF, uniprobe

Examples

```
library (MotifDb)
  # identify all the SOX genes
sox.indices = grep ('^sox', values (MotifDb)$geneSymbol, ignore.case=TRUE)
matrices = MotifDb [sox.indices]
export (matrices, con='SoxGenes-meme.txt', format='meme')
```

MotifDb *MotifDb: An Annotated Collection of DNA-binding sequence motifs*

Description

Approximately 2000 position frequency matrices collected from public sources, with ample accompanying metadata, and search and export capabilities provided.

Details

MotifDb is an R object of class MotifList, whose entries are numeric matrices, accompanied by a 'parallel' metadata structure, a DataFrame, in which each row provides information about the corresponding matrix. This object is automatically created and fully populated by data from five public sources (see below) when the package is loaded into your R environment via the library call. The matrices are obtained from five public sources:

FlyFactorSurvey:	614
hPDI:	437
JASPAR_CORE:	459
ScerTF:	196
UniPROBE:	380

Representing primarily four organsisms:

Dmelanogaster:	739
Hsapiens:	505
Scerevisiae:	464
Mmusculus:	329
Rnorvegicus:	8
Celegans:	7
Zmays:	6
Athaliana:	5
Psativum:	3
Amajus:	3
Pfalciparum:	2
Gallus:	2
Xlaevis:	1
Vertebrata:	1
Taestivam:	1
Rrattus:	1
Phybrida:	1
Ocuniculus:	1
Nsylvestris:	1
Hvulgare:	1
Hroretzi:	1
Cparvum:	1

All the matrices are stored as position frequency matrices, in which each columm (each position)

sums to 1.0. When the number of sequences which contributed to the motif are known, that number will be found in the matrix's metadata. With this information, one can transform the matrices into either PCM (position count matrices), or PWM (position weight matrices), also known as PSSM (position-specific-scoring matrices). The latter transformation requires that a model of the background distribution be known, or assumed.

The names of the matrices are the same as rownames of the metadata DataFrame, and have been chosen to balance the needs of concision and full description, including the organism in which the motif was discovered, the data source, and the name of the motif in the data source from which it was obtained. For example: "Hsapiens-JASPAR_CORE-SP1-MA0079.2" and "Scerevisiae-ScerTF-GSM1-badis".

Subsets of the Matrices may be obtainted in several ways:

- By integer index, eg, MotifDb [[1]]
- By query, eg, as.list (query (MotifDb, 'FBgn0000014'))
- (Interactively only) by subset as.list (subset (MotifDb, geneSymbol=='Abda' & !is.na (pubmedID)))

The matrices are stored in a SimpleList which has semantics very similar to the familiar list of R base. To examine a matrix, however, you must sidestep the MotifDb show method. These three commands display quite different results:

```
> MotifDb [1]
MotifDb object of length 1
| Created from downloaded public sources: 2012-Jul6
| 1 position frequency matrices from 1 source:
|    FlyFactorSurvey:    1
| 1 organism/s
|     Dmelanogaster:    1
Dmelanogaster-FlyFactorSurvey-ab_SANGER_10_FBgn0259750
```

```
> MotifDb [[1]]
   1    2    3    4 5 6 7 8 9   10   11   12   13   14   15   16   17   18   19   20   21
A 0.0 0.50 0.20 0.35 0 0 1 0 0 0.55 0.35 0.05 0.20 0.45 0.20 0.10 0.40 0.40 0.25 0.50 0.30
C 0.3 0.15 0.25 0.00 1 1 0 0 0 0.10 0.65 0.70 0.45 0.25 0.10 0.25 0.25 0.10 0.10 0.25 0.25
G 0.4 0.05 0.50 0.65 0 0 0 1 1 0.00 0.00 0.05 0.05 0.15 0.05 0.20 0.05 0.15 0.55 0.15 0.45
T 0.3 0.30 0.05 0.00 0 0 0 0 0 0.35 0.00 0.20 0.30 0.15 0.65 0.45 0.30 0.35 0.10 0.10 0.00
```

```
> as.list (MotifDb [1])
$`Dmelanogaster-FlyFactorSurvey-ab_SANGER_10_FBgn0259750`
   1    2    3    4 5 6 7 8 9   10   11   12   13   14   15   16   17   18   19   20   21
A 0.0 0.50 0.20 0.35 0 0 1 0 0 0.55 0.35 0.05 0.20 0.45 0.20 0.10 0.40 0.40 0.25 0.50 0.30
C 0.3 0.15 0.25 0.00 1 1 0 0 0 0.10 0.65 0.70 0.45 0.25 0.10 0.25 0.25 0.10 0.10 0.25 0.25
G 0.4 0.05 0.50 0.65 0 0 0 1 1 0.00 0.00 0.05 0.05 0.15 0.05 0.20 0.05 0.15 0.55 0.15 0.45
T 0.3 0.30 0.05 0.00 0 0 0 0 0 0.35 0.00 0.20 0.30 0.15 0.65 0.45 0.30 0.35 0.10 0.10 0.00
```

There are fifteen kinds of metadata – though not all matrices have a full complement: not all of the public sources are complete in this regard. The information falls into these categories, using the *Dmelanogaster-FlyFactorSurvey-ab_SANGER_10_FBgn0259750* entry as an example (see below for the associated position frequency matrix):

1. providerName: "ab_SANGER_10_FBgn0259750"

2. providerId: "FBgn0259750"
3. dataSource: "FlyFactorSurvey"
4. geneSymbol: "Ab"
5. geneId: "FBgn0259750"
6. geneIdType: "FLYBASE"
7. proteinId: "E1JHF4"
8. proteinIdType: "UNIPROT"
9. organism: "Dmelanogaster"
10. sequenceCount: 20
11. bindingSequence: NA
12. bindingDomain: NA
13. tfFamily: NA
14. experimentType: "bacterial 1-hybrid, SANGER sequencing"
15. pubmedID: NA

References

- Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B, Wasserman WW, Sandelin A. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res. 2010 Jan;38(Database issue):D105-10. Epub 2009 Nov 11.
- Robasky K, Bulyk ML. UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. Nucleic Acids Res. 2011 Jan;39(Database issue):D124-8. Epub 2010 Oct 30.
- Spivak AT, Stormo GD. ScerTF: a comprehensive database of benchmarked position weight matrices for Saccharomyces species. Nucleic Acids Res. 2012 Jan;40(Database issue):D162-8. Epub 2011 Dec 2.
- Xie Z, Hu S, Blackshaw S, Zhu H, Qian J. hPDI: a database of experimental human protein-DNA interactions. Bioinformatics. 2010 Jan 15;26(2):287-9. Epub 2009 Nov 9.
- Zhu LJ, et al. 2011. FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system. Nucleic Acids Res. 2011 Jan;39(Database issue):D111-7. Epub 2010 Nov 19.

See Also

query, subset, export, flyFactorSurvey, hPDI, jaspar, ScerTF, uniprobe

Examples

```
  # are there any matrices for Sox4?  we find two
mdb.sox4 <- MotifDb [grep ('sox4', values (MotifDb)$geneSymbol, ignore.case=TRUE)]
  # the same two matrices can be obtained this way also
if (interactive ())
  mdb.sox4 <- subset (MotifDb, tolower(geneSymbol)=='sox4')
  # and like this
mdb.sox4 <- query (MotifDb, 'sox4')  # matches against all fields in the metadata
  # implicitly invoke the 'show' method
mdb.sox4
  # get their full names
```

```
names (mdb.sox4)
  # examine their metadata
values (mdb.sox4)
  # examine the matrices with names include
as.list (mdb.sox4)
  # export the matrices in meme format
destination.file = tempfile ()
export (mdb.sox4, destination.file, 'meme')
```

MotifList-class *MotifList*

Description

A direct subclass of SimpleList, having no extra slots, in which listData is a list of position frequency matrices (PFMs), and the elementMetadata slot is a DataFrame with fifteen columns describing each matrix. Upon loading the MotifDb class, one MotifList object is instantiated and filled with matrices and their metadata. There should be no need for users to explicitly create objects of this class. When you load the MotifDb package, a fully-populated instance of this class is created, with > 2000 matrices with metadata

Methods

subset(x): extract matrices by metadata.

export(x): write matrices

show(x): describe matrices compactly

query(x): find matrics

Author(s)

Paul Shannon

Examples

```
  # Examine the number of matrices contributed by each source.
print (table (values (MotifDb)$dataSource))
```

---

query *query*

Description

A very general search tool, returning all matrices whose metadata, in ANY column, is matched by the query string.

Usage

```
## S4 method for signature 'MotifList'
query(object, queryString, ignore.case=TRUE)
```

Arguments

object	a MotifList object.
queryString	a character string
ignore.case	a logical value, default TRUE

Value

A list of the matrices

Author(s)

Paul Shannon

See Also

MotifDb, subset, export, flyFactorSurvey, hPDI, jaspar, ScerTF, uniprobe

Examples

```
mdb <- MotifDb
matrices.human = query (mdb, 'hsapiens')
matrices.sox4 = query (mdb, 'sox4')
uniprobe.sox.matrices = query (query (mdb, 'uniprobe'), '^sox')
```

---

subset *subset*

Description

An analog of the base package subset method, this version will return all the matrices whose metadata match the (possibly intricate) logical expression in the "subset" argument.

Note: just as with the base subset method, this method is unreliable except when used interactively. Batch, script or other programmatic use of this function is to be avoided.

Usage

```
## S4 method for signature 'MotifList'
subset(x, subset, select, drop=FALSE, ...)
```

Arguments

x	a MotifList object.
subset	a logical expression whose terms are predicates on the column names of the metadata table
select, drop, ...	these are ignored, appearing here only in fidelity to the generic definition of the method.

Value

A list of the matrices whose metadata satisfies the supplied subset

Author(s)

Paul Shannon

See Also

MotifDb, query, export, flyFactorSurvey, hPDI, jaspar, ScerTF, uniprobe

Examples

```
mdb <- MotifDb
if (interactive ()) {
  matrices <- subset (mdb, dataSource=='UniPROBE')
  egr1.matrices <- subset (mdb, geneSymbol=='Egr1')
  jaspar.egr1.matrices <- subset (mdb, geneSymbol=='Egr1' &
                                  dataSource == 'JASPAR_CORE')
    # one of the mouse egr1 matrices has a geneSymbol 'Zif268', but
    # has the proper entrez geneId.
  all.egr1.matrices <- subset (mdb, geneId=='13653')
  }
```

Index