The **TFBSTools** package overview

Ge Tan

October 14, 2013

Contents

1 Introduction

Eukaryotic regulatory regions are characterized based a set of discovered transcription factor binding sites, which can be represented as sequence patterns with various degree of degeneracy.

This **TFBSTools** package is designed to be a compuational framework for transcription factor binding site analysis. It contains a set of integrated R S4 style classes, tools , JASPAR database interface functions. Most approaches can be described in three sequential phases. First, a pattern is generated for a set of target sequences known to be bound by a specific transcription factor. Second, a set of DNA sequences are analyzed to determine the locations of sequences consistent with the described binding pattern. Finally, in advanced cases, predictive statistical models of regulatory regions are constructed based on mutiple occurrences of the detected patterns.

TFBSTools aims to support all these functionalities in the environment **R**. However, only the JASPAR database interface functions are exported in this release to accompany *JASPAR2014*. More functions will be included in future release after well tested.

2 S4 classes in TFBSTools

The section will explain all the S4 classes defined in **TFBSTools**.

2.1 PFMatrix

PFMatrix is designed to store all the relevant information for one raw position frequency matrix (PFM). This object is compatible with one record from JASPAR database. For more details about this object, please consult the help page of this class.

```
> library(TFBSTools)
> pfm = PFMatrix(ID="MA0004.1", name="Arnt", matrixClass="Zipper-Type", strand="+",
+         bg=c(A=0.25, C=0.25, G=0.25, T=0.25),
+         tags=list(family="Helix-Loop-Helix", species="10090", tax_group="vertebrates",
+         medline="7592839", type="SELEX", ACC="P53762", pazar_tf_id="TF0000003",
+         TFBSshape_ID="11", TFencyclopedia_ID="580"),
+         matrix=matrix(c(4L,  19L, 0L,  0L,  0L,  0L,
+                 16L, 0L,  20L, 0L,  0L,  0L,
+                 0L,  1L,  0L,  20L, 0L,  20L,
+                 0L,  0L,  0L,  0L,  20L, 0L),
+                 byrow=TRUE, nrow=4, dimnames=list(c("A", "C", "G", "T")))
+         )
> ## coerced to matrix
> as.matrix(pfm)
> ## get the reverse complment matrix with all the same information except the strand.
> reverseComplement(pfm)
> ## access the slots of pfm
> ID(pfm)
> name(pfm)
> Matrix(pfm)
```

2.2 PFMatrixList

PFMatrixList is used to store a set of *PFMatrix* objects. Basically it is a SimpleList for easy manipulation the whole set of *PFMatrix*.

```
> pfm2 = pfm
> PFMatrixList(pfm1=pfm, pfm2=pfm2, use.names=TRUE)
```

3 Database interfaces for JASPAR2014 database

This section will demonstrate how to operate on the JASPAR 2014 database. JASPAR is a collection of transcription factor DNA-binding preferences, modeled as matrices. These can be converted into Position Weight Matrices (PWMs or PSSMs), used for scanning genomic sequences. JASPAR is the only database with this scope where the data can be used with no restrictions (open-source).

3.1 Search JASPAR2014 database

This search function fetches matrix data for all matrices in the database matching criteria defined by the named arguments and returns a PFMatrixList object. For more search criterias, please see the help page for (getMatrixSet).

```
> library(JASPAR2014)
> opts = list()
> opts[["species"]] = 9606
> opts[["name"]] = "RUNX1"
> #opts[["class"]] = "Ig-fold"
> opts[["type"]] = "SELEX"
> opts[["all_versions"]] = TRUE
> PFMatrixList = getMatrixSet(JASPAR2014, opts)
> opts2 = list()
> opts2[["type"]] = "SELEX"
> PFMatrixList2 = getMatrixSet(JASPAR2014, opts2)
```

3.2 Store, delete and initialize JASPAR2014 database

We also provide some functions to initialize an empty JASPAR2014 style database, store new *PFMatrix* or *PFMatrixList* into it, or delete some records based on ID.

```
> db = "jaspar.sqlite"
> initializeJASPARDB(db)
> storeMatrix(db, pfm)
> deleteMatrixHavingID(db, "MA0003")
```

4 Conclusion

The following is the session info that generated this vignette:

```
>   sessionInfo()

R version 3.0.2 (2013-09-25)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
```

```
[1] parallel  stats     graphics  grDevices utils     datasets
[7] methods   base

other attached packages:
[1] TFBSTools_1.0.0    IRanges_1.20.0     BiocGenerics_0.8.0

loaded via a namespace (and not attached):
[1] Biostrings_2.30.0 DBI_0.2-7         RSQLite_0.11.4
[4] XVector_0.2.0     grid_3.0.2        seqLogo_1.28.0
[7] stats4_3.0.2      tools_3.0.2
```