Motif import, export, and manipulation

Benjamin Jean-Marie Tremblay*

17 October 2021

Abstract
The universalmotif package offers a number of functions to manipulate motifs. These are introduced and explored here, including those relating to: import, export, motif modification, creation, visualization, and other miscellaneous utilities.

Contents
1 Introduction 2
2 The universalmotif class and conversion utilities 2
  2.1 The universalmotif class ........................................ 2
  2.2 Converting to and from another package’s class .......... 4
3 Importing and exporting motifs 6
  3.1 Importing .................................................. 6
  3.2 Exporting ................................................ 6
4 Modifying motifs and related functions 6
  4.1 Converting motif type ........................................ 6
  4.2 Merging motifs ............................................. 9
  4.3 Motif reverse complement .................................. 10
  4.4 Switching between DNA and RNA alphabets .............. 11
  4.5 Motif trimming ........................................... 12
  4.6 Rounding motifs ........................................... 13
5 Motif creation 14
  5.1 From a PCM/PPM/PWM/ICM matrix ....................... 14
  5.2 From sequences or character strings .................... 15
  5.3 Generating random motifs ................................ 15
6 Motif visualization 16
  6.1 Motif logos .............................................. 16
  6.2 Stacked motif logos ...................................... 19
  6.3 Plot arbitrary text logos .................................. 19
7 Higher-order motifs 20
8 Tidy motif manipulation with the universalmotif_df data structure 22
9 Miscellaneous motif utilities 25
  9.1 DNA/RNA/AA consensus functions ....................... 25
  9.2 Filter through lists of motifs ............................. 25

*benjamin.tremblay@uwaterloo.ca
1 Introduction

This vignette will introduce the universalmotif class and its structure, the import and export of motifs in R, basic motif manipulation, creation, and visualization. For an introduction to sequence motifs, see the introductory vignette. For sequence-related utilities, see the sequences vignette. For motif comparisons and P-values, see the motif comparisons and P-values vignette.

2 The universalmotif class and conversion utilities

2.1 The universalmotif class

The universalmotif package stores motifs using the universalmotif class. The most basic universalmotif object exposes the name, alphabet, type, type, strand, icscore, consensus, and motif slots; furthermore, the pseudocount and bkg slots are also stored but not shown. universalmotif class motifs can be PCM, PPM, PWM, or ICM type.

```r
library(universalmotif)
data(examplemotif)
examplemotif
```

```
#> Motif name: motif
#> Alphabet: DNA
#> Type: PPM
#> Strands: +-  
#> Total IC: 11.54
#> Pseudocount: 1  
#> Consensus: TATAWAW

#> T A T A W A W
#> A 0 1 0 1 0.5 1 0.5
#> C 0 0 0 0 0.0 0 0.0
#> G 0 0 0 0 0.0 0 0.0
#> T 1 0 1 0 0.5 0 0.5
```

A brief description of all the available slots:

- name: motif name
- altname: (optional) alternative motif name
- family: (optional) a word representing the transcription factor or matrix family
- organism: (optional) organism of origin
- motif: the actual motif matrix
- alphabet: motif alphabet
- type: motif ‘type’, one of PCM, PPM, PWM, ICM; see the introductory vignette
- icscore: (generated automatically) Sum of information content for the motif
- nsites: (optional) number of sites the motif was created from
- **pseudocount**: this value to added to the motif matrix during certain type conversions; this is necessary to avoid $-\text{Inf}$ values from appearing in PWM type motifs.
- **bkg**: a named vector of probabilities which represent the background letter frequencies.
- **bkgsites**: (optional) total number of background sequences from motif creation.
- **consensus**: (generated automatically) for DNA/RNA/AA motifs, the motif consensus.
- **strand**: strand motif can be found on.
- **pval**: (optional) P-value from *de novo* motif search.
- **qval**: (optional) Q-value from *de novo* motif search.
- **eval**: (optional) E-value from *de novo* motif search.
- **multifreq**: (optional) higher-order motif representations.
- **extrainfo**: (optional) any extra motif information that cannot fit in the existing slots.

The other slots will be shown as they are filled.

```r
library(universalmotif)
data(examplemotif)

## The various slots can be accessed individually using `[]`
examplemotif["consensus"]
#> [1] "TATAWAW"

## To change a slot, use `<-`
examplemotif["family"] <- "My motif family"
examplemotif
#> Motif name: motif
#> Family: My motif family
#> Alphabet: DNA
#> Type: PPM
#> Strands: +- 
#> Total IC: 11.54
#> Pseudocount: 1
#> Consensus: TATAWAW
#> T  A  T  A  W  A  W
#> A 0 1 0 1 0.5 1 0.5
#> C 0 0 0 0 0.0 0 0.0
#> G 0 0 0 0 0.0 0 0.0
#> T 1 0 1 0 0.5 0 0.5
```

Though the slots can easily be changed manually with `<-`, a number of safeguards have been put in place for some of the slots which will prevent incorrect values from being introduced.

```r
library(universalmotif)
data(examplemotif)

## The consensus slot is dependent on the motif matrix
examplemotif["consensus"]
#> [1] "TATAWAW"

## Changing this would mean it no longer matches the motif
examplemotif["consensus"] <- "GGGAGAG"
```

3
Below the exposed metadata slots, the actual ‘motif’ matrix is shown. Each position is its own column: row names showing the alphabet letters, and the column names showing the consensus letter at each position.

### 2.2 Converting to and from another package’s class

The universalmotif package aims to unify most of the motif-related Bioconductor packages by providing the `convert_motifs()` function. This allows for easy transition between supported packages (see `?convert_motifs` for a complete list of supported packages). Should you ever come across a motif class from another Bioconductor package which is not supported by the universalmotif package, but believe it should be, then feel free to bring it up with me.

The `convert_motifs` function is embedded in most of the universalmotif functions, meaning that compatible motif classes from other packages can be used without needed to manually convert them first. However keep in mind some conversions are final. Furthermore, internally, all motifs regardless of class are handled as universalmotif objects, even if the returning class is not. This will result in at times slightly different objects (though usually no information should be lost).

```r
library(universalmotif)
library(MotifDb)
data(examplemotif)
data(MA0003.2)
## convert from a `universalmotif` motif to another class
convert_motifs(examplemotif, "TFBSTools-PWMatrix")
#> Note: motif [motif] has an empty nsites slot, using 100.
#> An object of class PWMatrix
#> ID:
#> Name: motif
#> Matrix Class: Unknown
#> strand: *
#> Pseudocounts: 1
#> Tags:
#> list()
#> Background:
#> A C G T
#> 0.25 0.25 0.25 0.25
#> Matrix:
#> T A T A W A W
#> A -6.658211 1.989247 -6.658211 1.989247 0.9928402 1.989247 0.9928402
#> C -6.658211 -6.658211 -6.658211 -6.658211 -6.658211 -6.658211 -6.658211
#> T 1.989247 -6.658211 1.989247 -6.658211 0.9928402 -6.658211 0.9928402
## convert to universalmotif
convert_motifs(MA0003.2)
>
> Error in .local(x, i, ..., value): this slot is unmodifiable with [<-
#> Error in eval(expr, envir, enclos): * strand must be one of +, -, +-
Motif name: TFAP2A
Alternate name: MA0003.2
Family: Helix-Loop-Helix
Organism: 9606
Alphabet: DNA
Type: PCM
Strands: +
Total IC: 12.9
Pseudocount: 1
Consensus: NNNGCCSAGGSCA
Target sites: 5098
Extra info: [centrality_logp] -4343
[family] Helix-Loop-Helix
[medline] 10497269
...

```
N N N N G C C Y S A G G S C A
A 1387 2141 727 1517 56 0 0 62 346 3738 460 0 116 451 3146
C 1630 1060 1506 519 1199 5098 4762 1736 2729 236 0 0 1443 3672 690
G 851 792 884 985 3712 0 0 85 1715 920 4638 5098 3455 465 168
T 1230 1105 1981 2077 131 0 336 3215 308 204 0 0 84 510 1094
```

```
# convert between two packages
convert_motifs(MotifDb[1], "TFBSTools-ICMatrix")

Note: motif [ABF1] has an empty nsites slot, using 100.
[[1]]
An object of class ICMatrix
ID: M0001_1.02
Name: ABF1
Matrix Class: Unknown
strand: *
Pseudocounts: 1
Schneider correction: FALSE
Tags:
$dataSource
[1] "cisbp_1.02"

Background:
A  C  G  T
0.25 0.25 0.25 0.25
Matrix:
T
T
A  T  C  A  C
A 0.02119039 0.02119039 1.64861232 0.02119039 0.02119039 1.64861232 0.02119039
C 0.02119039 0.02119039 0.02119039 0.02119039 1.64861232 0.02119039 1.64861232
G 0.02119039 0.02119039 0.02119039 0.02119039 0.02119039 0.02119039 0.02119039
T 1.64861232 1.64861232 0.02119039 1.64861232 0.02119039 0.02119039 0.02119039
T
A 0.02119039
C 0.02119039
G 0.02119039
T 1.64861232
```
3 Importing and exporting motifs

3.1 Importing

The universalmotif package offers a number of `read_*()` functions to allow for easy import of various motif formats. These include:

- `read_cisbp()`: CIS-BP (Weirauch et al. 2014)
- `read_homer()`: HOMER (Heinz et al. 2010)
- `read_jaspar()`: JASPAR (Khan et al. 2018)
- `read_matrix()`: generic reader for simply formatted motifs
- `read_meme()`: MEME (Bailey et al. 2009)
- `read_motifs()`: native universalmotif format (not recommended; use `saveRDS()` instead)
- `read_transfac()`: TRANSFAC (Wingender et al. 1996)
- `read_uniprobe()`: UniPROBE (Hume et al. 2015)

These functions should work natively with these formats, but if you are generating your own motifs in one of these formats than it must adhere quite strictly to the format. An example of each of these is included in this package (see `system.file("extdata", package="universalmotif")`). If you know of additional motif formats which are not supported in the universalmotif package that you believe should be, or of any mistakes in the way the universalmotif package parses supported formats, then please let me know.

3.2 Exporting

Compatible motif classes can be written to disk using:

- `write_homer()`
- `write_jaspar()`
- `write_matrix()`
- `write_meme()`
- `write_motifs()`
- `write_transfac()`

The `write_matrix()` function, similar to its `read_matrix()` counterpart, can write motifs as simple matrices with an optional header. Additionally, please keep in mind format limitations. For example, multiple MEME motifs written to a single file will all share the same alphabet, with identical background letter frequencies.

4 Modifying motifs and related functions

4.1 Converting motif type

Any universalmotif object can transition between PCM, PPM, PWM, and ICM types seamlessly using the `convert_type()` function. The only exception to this is if the ICM calculation is performed with sample correction, or as relative entropy. If this occurs, then back conversion to another type will be inaccurate (and `convert_type()` would not warn you, since it won’t know this has taken place).

```
library(universalmotif)
data(examplemotif)

## This motif is currently a PPM:

eexamplemotif["type"]
#> [1] "PPM"
```

When converting to PCM, the `nsites` slot is needed to tell it how many sequences it originated from. If empty, 100 is used.
convert_type(examplemotif, "PCM")

#> Note: motif [motif] has an empty nsites slot, using 100.
#>
#> Motif name: motif
#> Alphabet: DNA
#> Type: PCM
#> Strands: +- 
#> Total IC: 11.54 
#> Pseudocount: 1 
#> Consensus: TATAWAW
#>
#> T A T A W A W 
#> A 0 100 0 100 50 100 50 
#> C 0 0 0 0 0 0 0 
#> G 0 0 0 0 0 0 0 
#> T 100 0 100 0 50 0 50

For converting to PWM, the pseudocount slot is used to determine if any correction should be applied:

examplemotif["pseudocount"]

#> [1] 1

convert_type(examplemotif, "PWM")

#> Note: motif [motif] has an empty nsites slot, using 100.
#>
#> Motif name: motif
#> Alphabet: DNA
#> Type: PWM
#> Strands: +- 
#> Total IC: 11.54 
#> Pseudocount: 1 
#> Consensus: TATAWAW
#>
#> T A T A W A W 
#> A -6.66 1.99 -6.66 1.99 0.99 1.99 0.99 
#> C -6.66 -6.66 -6.66 -6.66 -6.66 -6.66 -6.66 
#> G -6.66 -6.66 -6.66 -6.66 -6.66 -6.66 -6.66 
#> T 1.99 -6.66 1.99 -6.66 0.99 -6.66 0.99

You can either change the pseudocount slot manually beforehand, or pass one to convert_type().

convert_type(examplemotif, "PWM", pseudocount = 1)

#> Note: motif [motif] has an empty nsites slot, using 100.
#>
#> Motif name: motif
#> Alphabet: DNA
#> Type: PWM
#> Strands: +- 
#> Total IC: 11.54 
#> Pseudocount: 1 
#> Consensus: TATAWAW
#>
#> T A T A W A W 
#> A -6.66 1.99 -6.66 1.99 0.99 1.99 0.99 
#> C -6.66 -6.66 -6.66 -6.66 -6.66 -6.66 -6.66 
#> G -6.66 -6.66 -6.66 -6.66 -6.66 -6.66 -6.66 
#> T 1.99 -6.66 1.99 -6.66 0.99 -6.66 0.99
There are a couple of additional options for ICM conversion: `nsize_correction` and `relative_entropy`. The former uses the `TFBSTools:::schneider_correction()` function (and thus requires that the `TFBSTools` package be installed) for sample size correction. The latter uses the `bkg` slot to calculate information content. See the `IntroductionToSequenceMotifs` vignette for an overview on the various types of ICM calculations.

```r
examplemotif["nsites"] <- 10
convert_type(examplemotif, "ICM", nsize_correction = FALSE)

Motif name: motif
Alphabet: DNA
Type: ICM
Strands: +- 
Total IC: 11.54
Pseudocount: 1
Consensus: TATAWAW
Target sites: 10

T A T A W A W
A 0 2 0 2 0.5 2 0.5
C 0 0 0 0 0.0 0 0.0
G 0 0 0 0 0.0 0 0.0
T 2 0 2 0 0.5 0 0.5
```

```r
convert_type(examplemotif, "ICM", nsize_correction = TRUE)

Motif name: motif
Alphabet: DNA
Type: ICM
Strands: +- 
Total IC: 11.54
Pseudocount: 1
Consensus: TATAWAW
Target sites: 10

T A T A W A W
A 0.00 1.75 0.00 1.75 0.38 1.75 0.38
C 0.00 0.00 0.00 0.00 0.00 0.00 0.00
G 0.00 0.00 0.00 0.00 0.00 0.00 0.00
T 1.75 0.00 1.75 0.00 0.38 0.00 0.38
```

```r
examplemotif["bkg"] <- c(A = 0.4, C = 0.1, G = 0.1, T = 0.4)
convert_type(examplemotif, "ICM", relative_entropy = TRUE)

Motif name: motif
Alphabet: DNA
Type: ICM
Strands: +- 
Total IC: 11.54
Pseudocount: 1
Consensus: TATAWAW
Target sites: 10
```

```r
```
4.2 Merging motifs

The universalmotif package includes the `merge_motifs()` function to combine motifs. Motifs are first aligned, and the best match found before the motif matrices are averaged. The implementation for this is identical to that used by `compare_motifs()` (see the motif comparisons vignette for more information).

```r
library(universalmotif)

m1 <- create_motif("TTAAACCCC", name = "1")
m2 <- create_motif("AACC", name = "2")
m3 <- create_motif("AACCCCGG", name = "3")

view_motifs(c(m1, m2, m3),
            show.positions.once = FALSE, show.names = FALSE)

view_motifs(merge_motifs(c(m1, m2, m3), method = "PCC"))
```
This functionality can also be automated to reduce the number of overly similar motifs in larger datasets via the `merge_similar()` function.

```r
library(universalmotif)
library(MotifDb)

motifs <- filter_motifs(MotifDb, family = "bHLH")[1:100]
#> motifs converted to class 'universalmotif'
length(motifs)
#> [1] 100

motifs <- merge_similar(motifs)
length(motifs)
#> [1] 63
```

Comparison and merging parameters can be fine-tuned as users wish. See the `compare_motifs()` and `merge_motifs()` documentation for more details, as well as the “Motif comparison and P-values” vignette.

### 4.3 Motif reverse complement

Get the reverse complement of a motif.

```r
library(universalmotif)
data(examplemotif)

## Quickly switch to the reverse complement of a motif
## Original:

examplemotif
```

```r
> Motif name: motif
> Alphabet: DNA
> Type: PPM
> Strands: +- 
> Total IC: 11.54
> Pseudocount: 1
> Consensus: TATAWAW
> T A T A W A W
> A 0 1 0 1 0.5 1 0.5
> C 0 0 0 0 0.0 0 0.0
> G 0 0 0 0 0.0 0 0.0
> T 1 0 1 0 0.5 0 0.5
```
## Reverse complement:

```r
motif_rc(examplemotif)
```

```r
greater than

Motif name: motif
Alphabet: DNA
Type: PPM
Strands: +-  
Total IC: 11.54
Pseudocount: 1
Consensus: WTWATA

W T W T A T A
A 0.5 0 0.5 0 1 0 1
C 0 0 0 0 0 0 0 0
G 0 0 0 0 0 0 0 0
T 0.5 1 0.5 1 0 1 0
```

### 4.4 Switching between DNA and RNA alphabets

Since not all motif formats or programs support RNA alphabets by default, the `switch_alph()` function can quickly go between DNA and RNA motifs.

```r
library(universalmotif)
data(examplemotif)
```

```r
## DNA --> RNA

```

```r
switch_alph(examplemotif)
```

```r
greater than

Motif name: motif
Alphabet: RNA
Type: PPM
Strands: +-  
Total IC: 11.54
Pseudocount: 1
Consensus: UAUAWAW

U A U A  W A  W
A 0 1 0 1 0.5 1 0.5
C 0 0 0 0 0 0 0 0
G 0 0 0 0 0 0 0 0
U 1 0 1 0 0.5 0 0.5
```

```r
## RNA --> DNA

```

```r
motif <- create_motif(alphabet = "RNA")
motif
```

```r
greater than

Motif name: motif
Alphabet: RNA
Type: PPM
Strands: +-
4.5 Motif trimming

Get rid of low information content edges on motifs, such as \(\text{NNCGGGCNN}\) to \(\text{CGGGC}\). The ‘amount’ of trimming can also be controlled by setting a minimum required information content, as well as the direction of trimming (by default both edges are trimmed).

```r
library(universalmotif)

motif <- create_motif("NNGCSGCGGNN")
motif
#>  Motif name:  motif
#>  Alphabet:  DNA
#>  Type:  PPM
#>  Strands:  +-  
#>  Total IC:  13
#>  Pseudocount:  0
#>  Consensus:  NNGCSGCGGNN
#>  
#>  N  N  G  C  S  G  C  G  G  N  N
#>  A  0.25  0.25  0  0  0  0  0  0  0  0.25  0.25
#>  C  0.25  0.25  0  1  0.5  0  1  0  0  0  0.25  0.25
#>  G  0.25  0.25  1  0  0.5  1  0  1  1  0.25  0.25
#>  T  0.25  0.25  0  0  0  0  0  0  0  0  0.25  0.25

trim_motifs(motif)
#>
#>  Motif name:  motif
```
4.6 Rounding motifs

Round off near-zero probabilities.

```r
motif1 <- create_motif("ATCGATGC", pseudocount = 10, type = "PPM", nsites = 100)
motif2 <- round_motif(motif1)
view_motifs(c(motif1, motif2))
```
5 Motif creation

Though `universalmotif` class motifs can be created using the `new` constructor, the `universalmotif` package provides the `create_motif()` function which aims to provide a simpler interface to motif creation. The `universalmotif` class was initially designed to work natively with DNA, RNA, and amino acid motifs. Currently though, it can handle any custom alphabet just as easily. The only downsides to custom alphabets is the lack of support for certain slots such as the `consensus` and `strand` slots.

The `create_motif()` function will be introduced here only briefly; see `?create_motif` for details.

5.1 From a PCM/PPM/PWM/ICM matrix

Should you wish to make use of the `universalmotif` functions starting from a motif class unsupported by `convert_motifs()`, you can instead manually create `universalmotif` class motifs using the `create_motif()` function and the motif matrix.

```r
motif.matrix <- matrix(c(0.7, 0.1, 0.1, 0.1, 
                          0.7, 0.1, 0.1, 0.1, 
                          0.1, 0.7, 0.1, 0.1, 
                          0.1, 0.7, 0.1, 0.1, 
                          0.1, 0.1, 0.7, 0.1, 
                          0.1, 0.1, 0.7, 0.1, 
                          0.1, 0.1, 0.1, 0.7, 
                          0.1, 0.1, 0.1, 0.7), nrow = 4)

motif <- create_motif(motif.matrix, alphabet = "RNA", name = "My motif", 
pseudocount = 1, nsites = 20, strand = "+")

## The 'type', 'icscore' and 'consensus' slots will be filled for you

motif

#>  Motif name:  My motif
#>  Alphabet:   RNA
As a brief aside: if you have a motif formatted simply as a matrix, you can still use it with the universalmotif package functions natively without creating a motif with `create_motif()`, as `convert_motifs()` also has the ability to handle motifs formatted simply as matrices. However it is much safer to first format the motif beforehand with `create_motif()`.

### 5.2 From sequences or character strings

If all you have is a particular consensus sequence in mind, you can easily create a full motif using `create_motif()`. This can be convenient if you’d like to create a quick motif to use with an external program such as from the MEME suite or HOMER. Note that ambiguity letters can be used with single strings.

```r
motif <- create_motif("CCNSNGG", nsites = 50, pseudocount = 1)
```

```r
## Now to disk:
## write_meme(motif, "meme_motif.txt")
```

### 5.3 Generating random motifs

If you wish to, it’s easy to create random motifs. The values within the motif are generated using `rgamma()` to avoid creating low information content motifs. If background probabilities are not provided, then they are generated with `rpois()`.

```r
create_motif()
```

```r
## Motif name:  motif
```
You can change the probabilities used to generate the values within the motif matrix:

```r
create_motif(bkg = c(A = 0.2, C = 0.4, G = 0.2, T = 0.2))
```

With a custom alphabet:

```r
create_motif(alphabet = "QWERTY")
```

### 6 Motif visualization

#### 6.1 Motif logos

There are several packages which offer motif visualization capabilities, such as `seqLogo`, `motifStack`, and `ggseqlogo`. The `universalmotif` package has its own implementation via the function `view_motifs()`,
which renders motifs using the \texttt{ggplot2} package (similar to \texttt{ggseqlogo}). Here I will briefly show how to use these to visualize \texttt{universalmotif} class motifs.

\begin{verbatim}
library(universalmotif)
data(examplemotif)

## With the native `view_motifs` function:
view_motifs(examplemotif)
\end{verbatim}

The \texttt{view_motifs()} function generates \texttt{ggplot} objects; feel free to manipulate them as such. For example, flipping the position numbers for larger motifs (where the text spacing can become tight):

\begin{verbatim}
view_motifs(create_motif(15)) +
ggplot2::theme(
  axis.text.x = ggrepel::element_text(angle = 90, hjust = 1)
)
\end{verbatim}

A large number of options are available for tuning the way motifs are plotted in \texttt{view_motifs()}. Visit the documentation for more information.

Using the other Bioconductor packages to view \texttt{universalmotif} motifs is fairly easy as well:

\begin{verbatim}
## For all the following examples, simply passing the functions a PPM is
## sufficient
motif <- convert_type(examplemotif, "PPM")
## Only need the matrix itself
motif <- motif["motif"]

## seqLogo:
seqLogo::seqLogo(motif)
\end{verbatim}
```r
## motifStack:
motifStack::plotMotifLogo(motif)
```

```r
## ggseqlogo:
ggseqlogo::ggseqlogo(motif)
```

Warning: The `scale` argument of `guides()` cannot be `FALSE`. Use "none" instead as of ggplot2 3.3.4.

i The deprecated feature was likely used in the ggseqlogo package.

Please report the issue at <https://github.com/omarwagih/ggseqlogo/issues>.

This warning is displayed once every 8 hours.

Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
6.2 Stacked motif logos

The motifStack package allows for a number of different motif stacking visualizations. The universalmotif package, while not capable of emulating most of these, still offers basic stacking via `view_motifs()`. The motifs are aligned using `compare_motifs()`.

```r
library(universalmotif)
library(MotifDb)

motifs <- convert_motifs(MotifDb[50:54])
view_motifs(motifs, show.positions.once = FALSE, names.pos = "right")
```

6.3 Plot arbitrary text logos

The logo plotting capabilities of `view_motifs()` can be used for any kind of arbitrary text logo. All you need is a numeric matrix (the heights of the characters), with the desired characters as row names. The following example is taken from the `view_logo()` documentation.

```r
library(universalmotif)
data(examplemotif)

## Start from a numeric matrix:
toplot <- examplemotif["motif"]

# Adjust the character heights as you wish (negative values are possible):
```
7 Higher-order motifs

Though PCM, PPM, PWM, and ICM type motifs are still widely used today, a few ‘next generation’ motif formats have been proposed. These wish to add another layer of information to motifs: positional interdependence. To illustrate this, consider the following sequences:

Table 1: Example sequences.

<table>
<thead>
<tr>
<th>#</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CAAAACC</td>
</tr>
<tr>
<td>2</td>
<td>CAAAACC</td>
</tr>
<tr>
<td>3</td>
<td>CAAAACC</td>
</tr>
<tr>
<td>4</td>
<td>CTTTTCC</td>
</tr>
<tr>
<td>5</td>
<td>CTTTTCC</td>
</tr>
<tr>
<td>6</td>
<td>CTTTTCC</td>
</tr>
</tbody>
</table>

This becomes the following PPM:

Table 2: Position Probability Matrix.

<table>
<thead>
<tr>
<th>Position</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Based on the PPM representation, all three of CAAAACC, CTTTTCC, and CTATAACC are equally likely. Though looking at the starting sequences, should CTATAACC really be considered so? For transcription factor binding sites, this sometimes is not the case. By incorporating this type of information into the motif, it can allow for increased accuracy in motif searching. A few example implementations of this include: TFFM by Mathelier and Wasserman (2013), BaMM by Siebert and Soding (2016), and KSM by Guo et al. (2018).

The universalmotif package implements its own, rather simplified, version of this concept. Plainly, the standard PPM has been extended to include \( k \)-letter frequencies, with \( k \) being any number higher than 1. For example, the 2-letter version of the table 2 motif would be:

<table>
<thead>
<tr>
<th>Position</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>G</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>T</td>
<td>0.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

This format shows the probability of each letter combined with the probability of the letter in the next position. The seventh column has been dropped, since it is not needed: the information in the sixth column is sufficient, and there is no eighth position to draw 2-letter probabilities from. Now, the probability of getting CTATAACC is no longer equal to CTTTTCC and CAAAACC. This information is kept in the \texttt{multifreq} slot of \texttt{universalmotif} class motifs. To add this information, use the \texttt{add_multifreq()} function.

```r
library(universalmotif)

motif <- create_motif("CWWWCC", nsites = 6)
sequences <- DNAStringSet(rep(c("CAAAACC", "CTTTTCC"), 3))
motif.k2 <- add_multifreq(motif, sequences, add.k = 2)

## Alternatively:
# motif.k2 <- create_motif(sequences, add.multifreq = 2)

motif.k2
```

This format shows the probability of each letter combined with the probability of the letter in the next position. The seventh column has been dropped, since it is not needed: the information in the sixth column is sufficient, and there is no eighth position to draw 2-letter probabilities from. Now, the probability of getting CTATAACC is no longer equal to CTTTTCC and CAAAACC. This information is kept in the \texttt{multifreq} slot of \texttt{universalmotif} class motifs. To add this information, use the \texttt{add_multifreq()} function.
Motif name: motif
Alphabet: DNA
Type: PPM
Strands: +-  
Total IC: 10
Pseudocount: 0
Consensus: CWWWWCC
Target sites: 6
k-letter freqs: 2

C W W W W C C
A 0 0.5 0.5 0.5 0.5 0 0
C 1 0.0 0.0 0.0 0.0 1 1
G 0 0.0 0.0 0.0 0.0 0 0
T 0 0.5 0.5 0.5 0.5 0 0

To plot these motifs, use `view_motifs()`:

```r
view_motifs(motif.k2, use.freq = 2)
```

This information is most useful with functions such as `scan_sequences()` and `enrich_motifs()`. Though other tools in the `universalmotif` can work with multifreq motifs (such as `motif_pvalue()`, `compare_motifs()`), keep in mind they are not as well supported as regular motifs (getting P-values from multifreq motifs is exponentially slower, and P-values from using `compare_motifs()` for multifreq motifs are not available by default). See the sequences vignette for using `scan_sequences()` with the multifreq slot.

8 Tidy motif manipulation with the `universalmotif_df` data structure

For those who enjoy using the tidyverse functions for data handling, motifs can additionally represented as the modified `data.frame` format: `universalmotif_df`. This format allows one to modify motif slots for multiples motifs simultaneously using the `universalmotif_df` columns, and then return to a list of motifs afterwards to resume use with `universalmotif` package functions. A few key functions have been provided in relation to this format:

- `to_df()`: Generate a `universalmotif_df` object from a list of motifs.
- `update_motifs()`: After modifying the `universalmotif_df` object, apply these modifications to the actual `universalmotif` objects (contained within the `motif` column).
• **to_list()**: Return to a list of `universalmotif` objects for use with `universalmotif` package functions. Note that it is not required to use `update_motifs()` before using `to_list()`, as modifications will be checked for and applied if found.

• **requires_update()**: Boolean check as to whether the `universalmotif` objects and the `universalmotif_df` columns differ and require either a `update_motifs()` or `to_list()` call to re-sync them.

```r
library(universalmotif)
library(MotifDb)

## Obtain a `universalmotif_df` object
motifs <- to_df(MotifDb)

head(motifs)

#> motif name altname family organism consensus alphabet
#> 1 <mot:ABF1> ABF1 M0001_1.02 ABF1 Scerevisiae TTATCACT DNA
#> 2 <mot:AFT1> AFT1 M0002_1.02 AFT Scerevisiae NTTGCACCCNN DNA
#> 3 <mot:AFT2> AFT2 M0003_1.02 AFT Scerevisiae NHNNACCCYN DNA
#> 4 <mot:PK19..> PK19363.1 M0004_1.02 AP2 Csativa CCGCGCCCR DNA
#> 5 <mot:ORA59> ORA59 M0005_1.02 AP2 Athaliana MGGCCGGCN DNA
#> 6 <mot:WIN1> WIN1 M0006_1.02 AP2 Athaliana NCRCCGCNNNN DNA

#> strand icscore type pseudocount bkg dataSource
#> 1 +- 13.697468 PPM 1 0.25, 0..... cisbp_1.02
#> 2 +- 11.323910 PPM 1 0.25, 0..... cisbp_1.02
#> 3 +- 5.809302 PPM 1 0.25, 0..... cisbp_1.02
#> 4 +- 8.926117 PPM 1 0.25, 0..... cisbp_1.02
#> 5 +- 11.351632 PPM 1 0.25, 0..... cisbp_1.02
#> 6 +- 6.509679 PPM 1 0.25, 0..... cisbp_1.02

#> [Hidden empty columns: nsites, bkgsites, pval, qval, eval.]

Some tidy manipulation:

```r
library(dplyr)

motifs <- motifs %>%
  mutate(bkg = case_when(
    organism == "Athaliana" ~ list(c(A = 0.32, C = 0.18, G = 0.18, T = 0.32)),
    TRUE ~ list(c(A = 0.25, C = 0.25, G = 0.25, T = 0.25))
  ))

head(filter(motifs, organism == "Athaliana"))

#> motif name altname family organism consensus alphabet
#> 1 * <mot:ORA59> ORA59 M0005_1.02 AP2 Athaliana MGCCGCCN DNA
#> 2 * <mot:WIN1> WIN1 M0006_1.02 AP2 Athaliana NCRCCGCNNNN DNA
#> 3 * <mot:AT1G..> AT1G22985 M0007_1.02 AP2 Athaliana NNGCCGGNNN DNA
#> 4 * <mot:TEM1> TEM1 M0008_1.02 AP2 Athaliana WATGTTGCG DNA
#> 5 * <mot:ERF11> ERF11 M0009_1.02 AP2 Athaliana NNGCCGGNNN DNA
#> 6 * <mot:RAP2.6> RAP2.6 M0010_1.02 AP2 Athaliana NNGCCGGNN DNA

#> strand icscore type pseudocount bkg dataSource
#> 1 +- 11.351632 PPM 1 0.32, 0..... cisbp_1.02
#> 2 +- 6.509679 PPM 1 0.32, 0..... cisbp_1.02
#> 3 +- 5.148803 PPM 1 0.32, 0..... cisbp_1.02
#> 4 +- 5.148803 PPM 1 0.32, 0..... cisbp_1.02
#> 5 +- 4.227144 PPM 1 0.32, 0..... cisbp_1.02
#> 6 +- 4.227144 PPM 1 0.32, 0..... cisbp_1.02
```
Feel free to add columns as well. You can add 1d vectors which will be added to the `extrainfo` slots of motifs. (Note that they will be coerced to character vectors!)

```r
motifs <- motifs %>%
  mutate(MotifIndex = 1:n())

head(motifs)
```

```r
#> motif name altname family organism consensus alphabet
#> 1 * <mot:ABF1> ABF1 M0001_1.02 ABF1 Scerevisiae TTATCACT DNA
#> 2 * <mot:AFT1> AFT1 M0002_1.02 AFT Scerevisiae NTTGCACCCNN DNA
#> 3 * <mot:AFT2> AFT2 M0003_1.02 AFT Scerevisiae NHHNCACCYN DNA
#> 4 * <mot:PK19...> PK19363.1 M0004_1.02 AP2 Csativa CGCCGCCR DNA
#> 5 * <mot:ORA59> ORA59 M0005_1.02 AP2 Athaliana MGCCGCCN DNA
#> 6 * <mot:WIN1> WIN1 M0006_1.02 AP2 Athaliana NCRCCCGNN DNA

#> strand icscore type pseudocount bkg dataSource MotifIndex
#> 1 +- 13.697468 PPM 1 0.25, 0..... cisbp_1.02 1
#> 2 +- 11.323910 PPM 1 0.25, 0..... cisbp_1.02 2
#> 3 +- 5.809302 PPM 1 0.25, 0..... cisbp_1.02 3
#> 4 +- 8.926117 PPM 1 0.25, 0..... cisbp_1.02 4
#> 5 +- 11.351632 PPM 1 0.32, 0..... cisbp_1.02 5
#> 6 +- 6.509679 PPM 1 0.32, 0..... cisbp_1.02 6
```

If during the course of your manipulation you’ve generated temporary columns which you wish to drop, you can set `extrainfo = FALSE` to discard all extra columns. Be careful though, this will discard any previously existing `extrainfo` data as well.
9 Miscellaneous motif utilities

A number of convenience functions are included for manipulating motifs.

9.1 DNA/RNA/AA consensus functions

For DNA, RNA and AA motifs, the universalmotif will automatically generate a consensus string slot. Furthermore, create_motif() can generate motifs from consensus strings. The internal functions for these have been made available:

- consensus_to_ppm()
- consensus_to_ppmAA()
- get_consensus()
- get_consensusAA()

```r
library(universalmotif)

get_consensus(c(A = 0.7, C = 0.1, G = 0.1, T = 0.1))
#> [1] "A"

consensus_to_ppm("G")
#> [1] 0.001 0.001 0.997 0.001
```

9.2 Filter through lists of motifs

Filter a list of motifs, using the universalmotif slots with filter_motifs().

```r
library(universalmotif)
library(MotifDb)

## Let us extract all of the Arabidopsis and C. elegans motifs

motifs <- filter_motifs(MotifDb, organism = c("Athaliana", "Celegans"))
## Only keeping motifs with sufficient information content and length:

```r
motifs <- filter_motifs(motifs, icscore = 10, width = 10)
```

```
head(summarise_motifs(motifs))
```

```r
#> name altname family organism consensus alphabet strand icscore nsites
#> 1 ERF1 M0025_1.02 AP2 Athaliana NMGCCGCCRN DNA +- 12.40700 NA
#> 2 ATERF6 M0027_1.02 AP2 Athaliana NTGCCGGCGB DNA +- 11.77649 NA
#> 3 ATCBF3 M0032_1.02 AP2 Athaliana ATGTCGGYNN DNA +- 10.66970 NA
#> 4 AT2G18300 M0155_1.02 bHLH Athaliana NNNGCACGTGN DNA +- 11.50133 NA
#> 5 bHLH104 M0159_1.02 bHLH Athaliana GCCACGTGCC DNA +- 16.05350 NA
#> 6 hih-16 M0173_1.02 bHLH Celegans NNNAATATKGN DNA +- 10.32432 NA
```

### 9.3 Generate random motif matches

Get a random set of sequences which are created using the probabilities of the motif matrix, in effect generating motif sites, with `sample_sites()`.

```r
library(universalmotif)
data(examplemotif)
sample_sites(examplemotif)
```

```r
#> DNAStringSet object of length 100:
#> width seq
#> [1] 7 TATATAT
#> [2] 7 TATAAAA
#> [3] 7 TATAAAT
#> [4] 7 TATAAA
#> [5] 7 TATAAAT
#> ... ...
#> [96] 7 TATAAA
#> [97] 7 TATAAAA
#> [98] 7 TATAAA
#> [99] 7 TATAAA
#> [100] 7 TATAAA
```

### 9.4 Motif shuffling

Shuffle a set of motifs with `shuffle_motifs()`. The original shuffling implementation is taken from the linear shuffling method of `shuffle_sequences()`, described in the sequences vignette.

```r
library(universalmotif)
library(MotifDb)
motifs <- convert_motifs(MotifDb[1:50])
```
```
head(summarise_motifs(motifs))
```
```
#> name altname family organism consensus alphabet strand icscore
#> 1 ABF1 M0001_1.02 ABF1 Scerevisiae TTATCACT DNA +- 13.697468
#> 2 AFT1 M0002_1.02 AFT Scerevisiae NTTGCACCCNN DNA +- 11.323910
#> 3 AFT2 M0003_1.02 AFT Scerevisiae NNNNCACCYN DNA +- 5.809302
#> 4 PK19363.1 M0004_1.02 AP2 Csativa CGCCGCCR DNA +- 8.926117
#> 5 ORA59 M0005_1.02 AP2 Athaliana MGCCGCNN DNA +- 11.351632
#> 6 WIN1 M0006_1.02 AP2 Athaliana NCRCGCNNN DNA +- 6.509679
```
```
motifs.shuffled <- shuffle_motifs(motifs, k = 3)
head(summarise_motifs(motifs.shuffled))
```
```
#> name consensus alphabet strand icscore
#> 1 ABF1 [shuffled] TNNCNCMN DNA +- 4.684126
#> 2 AFT1 [shuffled] NNMGCCNNGY DNA +- 11.655862
#> 3 AFT2 [shuffled] NCNATCTATW DNA +- 12.331351
#> 4 PK19363.1 [shuffled] ATGNCNAT DNA +- 4.418793
#> 5 ORA59 [shuffled] NNNGNNCA DNA +- 4.980335
#> 6 WIN1 [shuffled] CNTGCCYGGY DNA +- 11.082030
```

### 9.5 Scoring and match functions

Motif matches in a set of sequences are typically obtained using logodds scores. Several functions are exposed to reveal some of the internal work that goes on.

- `get_matches()`: show all possible sequence matches above a certain score
- `get_scores()`: obtain all possible scores from all possible sequence matches
- `motif_score()`: translate score thresholds to logodds scores
- `prob_match()`: return probabilities for sequence matches
- `score_match()`: return logodds scores for sequence matches

```r
library(universalmotif)
data(examplemotif)
exexamplemotif
```
```
#> Motif name: motif
#> Alphabet: DNA
#> Type: PPM
#> Strands: +-  
#> Total IC: 11.54
#> Pseudocount: 1
#> Consensus: TATAWAW
#> T A T A W A W
#> A 0 1 0 1 0.5 1 0.5
#> C 0 0 0 0 0.0 0 0.0
#> G 0 0 0 0 0.0 0 0.0
#> T 1 0 1 0 0.5 0 0.5
```
```
## Get the min and max possible scores:
motif_score(examplemotif)
```
```
#> 0% 100%
#> -46.606 11.929
```
```
## Show matches above a score of 10:
```
get_matches(examplemotif, 10)

## Get the probability of a match:
prob_match(examplemotif, "TTTTTTT", allow.zero = FALSE)

## Score a specific sequence:
score_match(examplemotif, "TTTTTTT")

## Take a look at the distribution of scores:
plot(density(get_scores(examplemotif), bw = 5))

9.6 Type conversion functions

While `convert_type()` will take care of switching the current type for `universalmotif` objects, the individual type conversion functions are also available for personal use. These are:

- `icm_to_ppm()`
- `pcm_to_ppm()`
- `ppm_to_icm()`
- `ppm_to_pcm()`
- `ppm_to_pwm()`
- `pwm_to_ppm()`

These functions take a one dimensional vector. To use these for matrices:

```r
library(universalmotif)

m <- create_motif(type = "PCM")['motif']
```
```r
m
#> T K K W T W C M S C
#> A 12 1 2 62 8 44 8 62 0 0
#> C 0 3 3 1 1 7 89 38 30 74
#> G 15 47 50 1 4 3 0 0 70 22
#> T 73 49 45 36 87 46 3 0 0 4
apply(m, 2, pcm_to_ppm)
#> T K K W T W C M S C
#> [1,] 0.12 0.01 0.02 0.62 0.08 0.44 0.08 0.62 0.0 0.00
#> [2,] 0.00 0.03 0.03 0.01 0.01 0.07 0.89 0.38 0.3 0.74
#> [3,] 0.15 0.47 0.50 0.01 0.04 0.03 0.00 0.00 0.7 0.22
#> [4,] 0.73 0.49 0.45 0.36 0.87 0.46 0.03 0.00 0.0 0.04
```

Additionally, the `position_icscore()` can be used to get the total information content per position:

```r
library(universalmotif)

position_icscore(c(0.7, 0.1, 0.1, 0.1))
#> [1] 0.6307803
```

### Session info

```r
#> R version 4.4.0 beta (2024-04-15 r86425)
#> Platform: x86_64-pc-linux-gnu
#> Running under: Ubuntu 22.04.4 LTS
#>
#> Matrix products: default
#> BLAS: /home/biocbuild/bbs-3.19-bioc/R/lib/libRblas.so
#> LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.10.0
#>
#> locale:
#> [1] LC_CTYPE=en_US.UTF-8 LC_NUMERIC=C
#> [3] LC_TIME=en_US.UTF-8 LC_COLLATE=en_US.UTF-8
#> [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
#>
#> time zone: America/New_York
#> tzcode source: system (glibc)
#>
#> attached base packages:
#> [1] stats4 stats graphics grDevices utils datasets methods
#> [8] base
#>
#> other attached packages:
#> [1] TFBSTools_1.42.0 cowplot_1.1.3 dplyr_1.1.4
#> [4] ggtree_3.12.0 ggplot2_3.5.1 MotifDb_1.46.0
#> [7] GenomicRanges_1.56.0 Biostrings_2.72.0 GenomeInfoDb_1.40.0
#> [10] XVector_0.44.0 IRanges_2.38.0 S4Vectors_0.42.0
#> [13] BiocGenerics_0.50.0 universalmotif_1.22.0
#>
#> loaded via a namespace (and not attached):
```

29
#> [1] jsonlite_1.8.8    magrittr_2.0.3
#> [3] farver_2.1.1      rmarkdown_2.26
#> [5] fs_1.6.4          BiocIO_1.14.0
#> [7] zlibbioc_1.50.0   vctrs_0.6.5
#> [9] Cairo_1.6-2      memoise_2.0.1
#> [11] Rsamtools_2.20.0  R Curl_1.98-1.14
#> [13] base64enc_0.1-3   tinytex_0.50
#> [15] htmltools_0.5.8.1 S4Arrays_1.4.0
#> [17] curl_5.2.1        CNEr_1.40.0
#> [19] SparseArray_1.4.0 gridGraphics_0.5-1
#> [21] pracma_2.4.4      htmlwidgets_1.6.4
#> [23] plyr_1.8.9        cachem_1.0.8
#> [25] GenomicAlignments_1.40.0 lifecycle_1.0.4
#> [27] pkgconfig_2.0.3   Matrix_1.7-0
#> [29] R6_2.5.1          fastmap_1.1.1
#> [31] GenomInfoDbData_1.2.12 MatrixGenerics_1.16.0
#> [33] digest_0.6.35     aplot_0.2.2
#> [35] colorspace_2.1-0  TFMFvalue_0.0.9
#> [37] patchwork_1.2.0   AnnotationDbi_1.66.0
#> [39] RSQLite_2.3.6     seqLogo_1.70.0
#> [41] labeling_0.4.3    fansi_1.0.6
#> [43] htrr_1.4.7        abind_1.4-5
#> [45] compiler_4.4.0    bit64_4.0.5
#> [47] withr_3.0.0       BiocParallel_1.38.0
#> [49] DBI_1.2.2         highr_0.10
#> [51] R. utils_2.12.3    MASS_7.3-60.2
#> [53] poweRlaw_0.80.0    DelayedArray_0.30.0
#> [55] rjson_0.2.21      gtools_3.9.5
#> [57] caTools_1.18.2    tools_4.4.0
#> [59] splitstackshape_1.4.8 ape_5.8
#> [61] ggseqlogo_0.2      R. oo_1.26.0
#> [63] glue_1.7.0         restfulr_0.0.15
#> [65] mime_3.1-164      grid_4.4.0
#> [67] ade4_1.7-22       reshape2_1.4.4
#> [69] generics_0.1.3    BSgenome_1.72.0
#> [71] tibble_0.3.5       tzdb_0.4.0
#> [73] R.methodsS3_1.8.2  tidyrr_1.3.1
#> [75] data.table_1.15.4  hms_1.1.3
#> [77] utf8_1.2.4         pillar_1.9.0
#> [79] stringr_1.5.1     yulab.utils_0.1.4
#> [81] treeio_1.28.0     lattice_0.22-6
#> [83] rtracklayer_1.64.0 bit_4.0.5
#> [85] annotate_1.82.0    tidyselect_1.2.1
#> [87] DirichletMultinomial_1.46.0 GO.db_3.19.1
#> [89] knitr_1.46         grImport2_0.3-1
#> [91] bookdown_0.39      SummarizedExperiment_1.34.0
#> [93] xfun_0.43          Biobase_2.64.0
#> [95] matrixStats_1.3.0  stringi_1.8.3
#> [97] UCSC.utils_1.0.0   lazyeval_0.2.2
#> [99] ggfun_0.1.4        yaml_2.3.8
#> [101] evaluate_0.23     codeTools_0.2-20
#> [103] tibble_3.2.1       ggplotify_0.1.2
#> [105] cli_3.6.2          xtable_1.8-4
#> [107] munsell_0.5.1      Rcpp_1.0.12

30
References


