Abstract

Many kinds of biological activity are regulated by the binding of proteins to their cognate substrates. Of particular interest is the sequence-specific binding of transcription factors to DNA, often in regulatory regions just upstream of the transcription start site of a gene. These binding events play a pivotal role in regulating gene expression. Sequence specificity among closely related binding sites is nearly always incomplete: some variety in the DNA sequence is routinely observed. For this reason, these inexact binding sequence patterns are commonly described as motifs represented numerically as frequency matrices, and visualized as sequence logos. Despite their importance in current research, there has been until now no single, annotated, comprehensive collection of publicly available motifs. The current package provides such a collection, offering more than two thousand annotated matrices from multiple organisms, within the context of the Bioconductor project. The matrices can be filtered and selected on the basis of their metadata, used with other Bioconductor packages (MotIV for motif comparison, seqLogo for visualization) or easily exported for use with standard software and websites such as those provided by the MEME Suite.

Contents

1 Introduction and Basic Operations
2 Selection
   2.1 query
   2.2 grep
   2.3 subset
   2.4 The Egr1 Case Study
3 Motif Matching
4 Exporting to the MEME Suite
5 Future Work
6 References

1 Introduction and Basic Operations

The first step is to load the necessary packages:

> library (MotifDb)
> library (MotIV)
> library (seqLogo)

There are more than two thousand matrices, from five sources:

> length (MotifDb)

[1] 5078

> sort (table (values (MotifDb)$dataSource), decreasing=TRUE)

1http://meme.sdsc.edu/meme/doc/meme.html
And 22 organisms (though the majority of the matrices come from just four):

```r
> sort (table (values (MotifDb)$organism), decreasing=TRUE)

<table>
<thead>
<tr>
<th>Organism</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hsapiens</td>
<td>2328</td>
</tr>
<tr>
<td>Dmelanogaster</td>
<td>1008</td>
</tr>
<tr>
<td>Scerevisiae</td>
<td>701</td>
</tr>
<tr>
<td>Musculus</td>
<td>660</td>
</tr>
<tr>
<td>Athaliana</td>
<td>160</td>
</tr>
<tr>
<td>C elegans</td>
<td>44</td>
</tr>
<tr>
<td>Pfalciparum</td>
<td>28</td>
</tr>
<tr>
<td>Rnorvegicus</td>
<td>16</td>
</tr>
<tr>
<td>Vertebrata</td>
<td>16</td>
</tr>
<tr>
<td>Ncrassa</td>
<td>2</td>
</tr>
<tr>
<td>Zmays</td>
<td>5</td>
</tr>
<tr>
<td>Discoidium</td>
<td>3</td>
</tr>
<tr>
<td>Anidulans</td>
<td>3</td>
</tr>
<tr>
<td>Pfalciparum</td>
<td>2</td>
</tr>
<tr>
<td>Rnorvegicus</td>
<td>2</td>
</tr>
<tr>
<td>Vertebrata</td>
<td>2</td>
</tr>
<tr>
<td>Ncrassa</td>
<td>2</td>
</tr>
<tr>
<td>Ocunicus</td>
<td>2</td>
</tr>
<tr>
<td>Otauri</td>
<td>2</td>
</tr>
<tr>
<td>Phybrida</td>
<td>2</td>
</tr>
<tr>
<td>Taeastivam</td>
<td>2</td>
</tr>
<tr>
<td>Xlaevis</td>
<td>2</td>
</tr>
<tr>
<td>Acarolinensis</td>
<td>1</td>
</tr>
<tr>
<td>Apism</td>
<td>1</td>
</tr>
<tr>
<td>Aterreus</td>
<td>1</td>
</tr>
<tr>
<td>Gaculeatus</td>
<td>1</td>
</tr>
<tr>
<td>Hcapsulatum</td>
<td>1</td>
</tr>
<tr>
<td>Mdoresta</td>
<td>1</td>
</tr>
<tr>
<td>Mgallopavo</td>
<td>1</td>
</tr>
<tr>
<td>Mmorninus</td>
<td>1</td>
</tr>
<tr>
<td>Mtruncatula</td>
<td>1</td>
</tr>
<tr>
<td>Egruberi</td>
<td>1</td>
</tr>
<tr>
<td>Nhamaetococca</td>
<td>1</td>
</tr>
<tr>
<td>Swectensis</td>
<td>1</td>
</tr>
<tr>
<td>Pcaponis</td>
<td>1</td>
</tr>
<tr>
<td>Ppygeneus</td>
<td>1</td>
</tr>
<tr>
<td>Petraurelia</td>
<td>1</td>
</tr>
<tr>
<td>Rrattus</td>
<td>1</td>
</tr>
<tr>
<td>Tthermophila</td>
<td>1</td>
</tr>
<tr>
<td>Vvinifera</td>
<td>1</td>
</tr>
<tr>
<td>Xtropicalis</td>
<td>1</td>
</tr>
</tbody>
</table>

With these categories of metadata

```r
> colnames (values (MotifDb))
```

[1] "providerName"  "providerId"  "dataSource"  "geneSymbol"
[5] "geneId"         "geneIdType"  "proteinId"   "proteinIdType"
[9] "organism"       "sequenceCount" "bindingSequence" "bindingDomain"
[13] "tfFamily"       "experimentType" "pubmedID"

2 Selection

There are three ways to extract subsets of interest from the MotifDb collection. All three operate upon the MotifDb metadata, matching values in one or more of those fifteen attributes (listed just above), and returning the subset of MotifDb which meet the specified criteria. The three techniques: `query`, `subset` and `grep`

2.1 query

This is the simplest technique to use, and will suffice in many circumstances. For example, if you want all of the human matrices:

```r
> query (MotifDb, 'hsapiens')
```

MotifDb object of length 2328

<table>
<thead>
<tr>
<th>Created from downloaded public sources: 2013-Aug-30</th>
</tr>
</thead>
<tbody>
<tr>
<td>2328 position frequency matrices from 7 sources:</td>
</tr>
<tr>
<td>JASPAR_2014: 117</td>
</tr>
<tr>
<td>JASPAR_CORE: 66</td>
</tr>
<tr>
<td>UniPROBE: 2</td>
</tr>
<tr>
<td>cispb_1.02: 313</td>
</tr>
<tr>
<td>hPDI: 437</td>
</tr>
<tr>
<td>jolma2013: 710</td>
</tr>
<tr>
<td>stamlab: 683</td>
</tr>
<tr>
<td>1 organism/s</td>
</tr>
<tr>
<td>Hsapiens: 2328</td>
</tr>
<tr>
<td>Hsapiens-jolma2013-BCL68</td>
</tr>
<tr>
<td>Hsapiens-jolma2013-CTCF</td>
</tr>
<tr>
<td>Hsapiens-jolma2013-EG1</td>
</tr>
<tr>
<td>Hsapiens-jolma2013-EG1-2</td>
</tr>
<tr>
<td>Hsapiens-jolma2013-EG2</td>
</tr>
<tr>
<td>Hsapiens-stamlab-UW.Motif.0681</td>
</tr>
</tbody>
</table>
If you want all matrices associated with Sox transcription factors, regardless of dataSource or organism:

```r
> query (MotifDb, 'sox')
```

MotifDb object of length 88
| Created from downloaded public sources: 2013-Aug-30
| 88 position frequency matrices from 6 sources:
  | FlyFactorSurvey: 2
  | JASPAR_2014: 8
  | JASPAR_CORE: 5
  | UniPROBE: 15
  | hPDI: 2
  | jolma2013: 56
| 4 organism/s
  | Hsapiens: 48
  | Mmuscus: 37
  | Dmelanogaster: 2
  | Vertebrata: 1
Dmelanogaster-FlyFactorSurvey-Sox14_SANGER_10_FBgn0005612
Dmelanogaster-FlyFactorSurvey-Sox15_SANGER_6_FBgn0005613
Hsapiens-hPDI-Sox13
Hsapiens-hPDI-Sox14
Hsapiens-JASPAR_CORE-SOX9-MA0077.1
... Mmuscus-UniPROBE-Sox30.UP00023
Mmuscus-UniPROBE-Sox4.UP00062
Mmuscus-UniPROBE-Sox5.UP00091
Mmuscus-UniPROBE-Sox7.UP00034
Mmuscus-UniPROBE-Sox8.UP00061

For all yeast transcription factors with a homeo domain

```r
> query (query (MotifDb, 'cerevisiae'), 'homeo')
```

MotifDb object of length 24
| Created from downloaded public sources: 2013-Aug-30
| 24 position frequency matrices from 3 sources:
  | JASPAR_2014: 10
  | JASPAR_CORE: 10
  | UniPROBE: 4
| 1 organism/s
  | Scerevisiae: 24
Scerevisiae-UniPROBE-Cup9.UP00308
Scerevisiae-UniPROBE-Matalpha2.UP00307
Scerevisiae-UniPROBE-Pho2.UP00268
Scerevisiae-UniPROBE-Yox1.UP00274
Scerevisiae-JASPAR_CORE-CUP9-MA0288.1
... Scerevisiae-JASPAR_2014-STE12-MA0393.1
Scerevisiae-JASPAR_2014-TEC1-MA0406.1
Scerevisiae-JASPAR_2014-TOR1-MA0408.1
Scerevisiae-JASPAR_2014-YEP1-MA0426.1
Scerevisiae-JASPAR_2014-YOX1-MA0433.1

The last example may inspire more confidence in the precision of the result than is justified, and for a couple of reasons. First, the assignment of protein binding domains to specific categories is, as of 2012, an ad hoc and incomplete process. Second, the query commands matches the supplied character string to all metadata columns. In this case, ‘homeo’ appears both in the bindingDomain column and the tfFamily column, and the above query will return matches from both. Searching and filtering should always be accompanied by close scrutiny of the data, such as these commands illustrate:

```r
> unique (grep ('homeo', values(MotifDb)$bindingDomain, ignore.case=T, v=T))
```

```
[1] "Homeobox"    "Box9_act;Homeobox"
[3] "LIM;Homeobox" "PAX;Homeobox"
[5] "GAR;Homeobox" "Pou;Homeobox"
[7] "Distant similarity to homeodomain" "Homeo"
[9] "Homeo, PAX"  "Homeo, POU"

> unique (grep ('homeo', values(MotifDb)$tfFamily, ignore.case=T, v=T))
```

```
2.2 grep

This selection method (and the next, subset) require that you address metadata columns explicitly. This is a little more work, but the requisite direct engagement with the metadata is worthwhile. Repeating the 'query' examples from above, you can see how more knowledge of MotifDb metadata is required.

```r
> mdb.human <- MotifDb [grep ('Hsapiens', values (MotifDb)$organism)]
> mdb.sox <- MotifDb [grep ('sox', values (MotifDb)$geneSymbol, ignore.case=TRUE)]
> yeast.indices = grepl ('scere', values (MotifDb)$organism, ignore.case=TRUE)
> homeo.indices.domain = grepl ('homeo', values (MotifDb)$bindingDomain, ignore.case=TRUE)
> homeo.indices.family = grepl ('homeo', values (MotifDb)$tfFamily, ignore.case=TRUE)
> yeast.homeo.indices = yeast.indices & (homeo.indices.domain | homeo.indices.family)
> yeast.homeoDb = MotifDb [yeast.homeo.indices]
```

An alternate and somewhat more compact approach:

```r
> yeast.homeo.indices <- with(values(MotifDb),
+ grepl('scere', organism, ignore.case=TRUE) &
+ (grepl('homeo', bindingDomain, ignore.case=TRUE) |
+ grepl('homeo', tfFamily, ignore.case=TRUE)))
```

2.3 subset

MotifDb::subset emulates the R base data.frame subset command, which is not unlike an SQL select function. Unfortunately – and just like the R base subset function – this MotifDb method cannot be used reliably within a script: It is only reliable when called interactively. Here, with mixed success (as you will see), we use MotifDb::subset to reproduce the query and grep selections shown above.

```r
> if (interactive ())
+ subset (MotifDb, organism=='Hsapiens')
```

One can easily find all the 'sox' genes with the subset command, avoiding possible upper/lower case conflicts by passing the metadata's geneSymbol column through the function 'tolower':

```r
> if (interactive ())
+ subset (MotifDb, tolower (geneSymbol) == 'sox4')
```

Similarly, subset has limited application for a permissive 'homeo' search. But for the retrieval by explicitly specified search terms, subset works very well:

```r
> if (interactive ())
+ subset (MotifDb, organism=='Scerevisiae' & bindingDomain=='Homeo')
```

2.4 The Egr1 Case Study

We now do a simple geneSymbol search, followed by an examination of the sub-MotifDb the search returns. We are looking for all matrices associated with the well-known and highly conserved zinc-finger transcription factor, Egr1. There are two of these in MotifDb, both from mouse, and each from a different data source.
> # subset is convenient:
> if (interactive ())
> + as.list (subset (MotifDb, tolower (geneSymbol) == 'egr1'))
> # grep returns indices which allow for more flexibility
> indices = grep ("egr1", values (MotifDb)$geneSymbol, ignore.case=TRUE)
> length (indices)

[1] 6

There are a variety of ways to examine and extract data from this object, a MotifList of length 2.

>MotifDb [indices]

MotifDb object of length 6
| Created from downloaded public sources: 2013-Aug-30 |
| 6 position frequency matrices from 4 sources: |
| JASPAR_2014: 1 |
| JASPAR_CORE: 1 |
| UniPROBE: 1 |
| jolma2013: 3 |
| 2 organism/s |
| Hsapiens: 3 |
| Mmusculus: 3 |

Mmusculus-JASPAR_CORE-Egr1-MA0162.1
Hsapiens-JASPAR_2014-EGR1-MA0162.2
Hsapiens-jolma2013-EGR1
Hsapiens-jolma2013-EGR1-2
Mmusculus-jolma2013-Egr1
Mmusculus-UniPROBE-Egr1.UP00007

Now view the matrices as a named list:

> as.list (MotifDb [indices])

`$"Mmusculus-JASPAR_CORE-Egr1-MA0162.1"
1 2 3 4 5 6 7 8 9 10
A 0.2000000 0.1333333 0.0000000 0 0.0 0.2 0.0666667 0 0.1333333 0
C 0.2666667 0.0666667 0.8666667 0 0.0 0.0 0.0000000 0 0.6666667 0
G 0.0666667 0.8000000 0.0000000 1 0.2 0.8 0.9333333 1 0.0000000 1
T 0.4666667 0.0000000 0.1333333 0 0.8 0.0 0.0000000 0 0.2000000 0
11
A 0.0666667
C 0.0000000
G 0.4666667
T 0.4666667

`$"Hsapiens-JASPAR_2014-EGR1-MA0162.2"
1 2 3 4 5 6 7 8
A 0.0895888 0.1228786 0.0946475 0.1089262 0.0190111 0.2375163 0 0.0000000
C 0.4673629 0.4935418 0.8510933 0.9443538 0.0000000 1 0.9673656 0.0000000
G 0.2511502 0.1108845 0.1835836 0.0000000 0.0000000 0.5809490 0.0000000 0.0000000
T 0.1914980 0.2282147 0.3998042 0.0366352 0.2043890 0.0329634 0.0000000 0.0000000
9 10 11 12 13 14
A 0.0000000 0.2979765 0.0000000 0.1932115 0.0000000 0.2468995
C 0.8284921 0.6675868 0.9751958 0.0000000 0.8022923 0.2685357 0.4565111
G 0.0498531 0.0000000 0.0000000 0.5384302 0.1158616 0.1560868 0.1560868
T 0.1216547 0.0198270 0.0248021 0.2685357 0.0805631 0.1405039 0.1405039

`$"Hsapiens-jolma2013-EGR1"
1 2 3 4 5 6
A 0.2494781 0.5139057 0.0032237 0.1052028 0.0000000 0.0026042
C 0.2411273 0.3954051 0.9696969 0.0035577 0.9082577 0.9923836
G 0.1539666 0.0362757 0.0073944 0.8542536 0.0079319 0.0000000
T 0.3554280 0.0544135 0.0134226 0.3519550 0.0122452 0.0045729
7 8 9 10 11 12
A 0.0000000 0.6526382 0.0032531 0.0190616 0.0100000 0.0680943
C 0.9282147 0.3459296 0.9954467 0.0113636 0.938125 0.2485270
G 0.0093652 0.0000000 0.0000000 0.9318182 0.011875 0.1422564 0.1253348
T 0.0624197 0.0037698 0.0013025 0.0377660 0.040000 0.1077358 0.3476413
13
A 0.2673936
C 0.0105504
G 0.1396077
T 0.4023884

`$"Hsapiens-jolma2013-EGR1-2"
1 2 3 4 5 6 7
A 0.2722977 0.7375079 0.0067237 0.0183443 0 0.0000000 0.0000000

5
and finally, the metadata associated with these two matrices, transposed, for easy reading and comparison:

```r
t (as.data.frame (values (MotifDb [indices])))
```

```r
providerName Egr1 EGR1 Hsapiens-jolma2013-EGR1 Hsapiens-jolma2013-EGR1-2 
providerId MA0162.1 MA0162.2 EGR1 EGR1 
dataSource JASPAR_CORE JASPAR_2014 jolma2013 jolma2013 
geneSymbol Egr1 EGR1 EGR1 EGR1 
geneId 13653 1958 1958 1958 
geneIdType ENTREZ ENTREZ ENTREZ ENTREZ 
proteinId <NA> <NA> <NA> <NA> 
proteinIdType <NA> <NA> <NA> <NA> 
organism Mmusculus Hsapiens Hsapiens Hsapiens 
sequenceCount 15 12256 1831 1703 
bindingSequence <NA> <NA> <NA> <NA> 
`<NA>` bindingDomain Zinc-coordinating Zinc-coordinating Zinc-coordinating Zinc-coordinating 
tfFamily BetaBetaAlpha-zinc finger BetaBetaAlpha-zinc finger BetaBetaAlpha-zinc finger BetaBetaAlpha-zinc finger 
experimentType bacterial 1-hybrid ChiP-seq SELEX SELEX 
pubmedID 16041365 16041365 16041365 16041365 
```

```r
','.3] [,4] 
providerName Hsapiens-jolma2013-EGR1 Hsapiens-jolma2013-EGR1-2 
providerId EGR1 EGR1 <NA> <NA> 
dataSource jolma2013 jolma2013 <NA> <NA> 
geneSymbol EGR1 EGR1 <NA> <NA> 
geneId 1958 1958 <NA> <NA> 
geneIdType ENTREZ ENTREZ ENTREZ ENTREZ 
proteinId <NA> <NA> <NA> <NA> 
proteinIdType <NA> <NA> <NA> <NA> 
organism Hsapiens Hsapiens Hsapiens Hsapiens 
sequenceCount 1831 1703 1831 1703 
bindingSequence NMC00CMOCANN NMC00CMOCANN NMC00CMOCANN NMC00CMOCANN 
`<NA>` bindingDomain <NA> <NA> <NA> <NA> 
tfFamily C2H2 C2H2 C2H2 C2H2 
experimentType SELEX SELEX SELEX SELEX 
pubmedID 23332764 23332764 23332764 23332764 
```

```r
```

```r
```

```r
```
We used the `grep` function above to find rows in the metadata table whose `geneSymbol` column includes the string 'Egr1'. If you wish to identify matrices (and/or their attendant metadata) based upon a richer combination of criteria, for instance:

1. organism (`Mmusculus`)  
2. gene symbol (`Egr1`)  
3. data source (`JASPAR_CORE`)

the `grep` solution, while serviceable, becomes a little awkward:

```r
> geneSymbol.rows = grep ('Egr1', values (MotifDb)$geneSymbol, ignore.case=TRUE)  
> organism.rows = grep ('Mmusculus', values (MotifDb)$organism, ignore.case=TRUE)  
> source.rows = grep ('JASPAR', values (MotifDb)$dataSource, ignore.case=TRUE)  
> egr1.mouse.jaspar.rows = intersect (geneSymbol.rows, intersect (organism.rows, source.rows))  
> print (egr1.mouse.jaspar.rows)

[1] 2276
```

```r
> egr1.motif <- MotifDb [egr1.mouse.jaspar.rows]
```

Far more concise, and fully reliable as an interactive command (though not if used in a script):

```r
> if (interactive ()) {  
+   egr1.motif <- subset (MotifDb, organism=='Mmusculus' &  
+                        dataSource=='JASPAR_CORE' &  
+                        geneSymbol=='Egr1')  
+ }
```

Whichever method you use, this next chunk of code displays the matrix, and then the metadata for mouse JASPAR Egr1, the latter textually-transformed for easy reading within the size constraints of this page.

```r
> egr1.motif

MotifDb object of length 1  
| Created from downloaded public sources: 2013-Aug-30  
| 1 position frequency matrices from 1 source:  
| JASPAR_CORE: 1  
| 1 organism/s  
| Mmusculus: 1  
Mmusculus-JASPAR_CORE-Egr1-MA0162.1
```

```r
> as.list (egr1.motif)
```

\(^2\)See the help page of the base R command `subset` for detail, is the `subset` command
$\text{Mmusculus-JASPAR\_CORE-Egr1-MA0162.1}$

<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>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.20000000</td>
<td>0.13333333</td>
<td>0.0000000</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
<td>0.06666667</td>
<td>0</td>
<td>0.1333333</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0.26666667</td>
<td>0.06666667</td>
<td>0.8666667</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00000000</td>
<td>0</td>
<td>0.6666667</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0.06666667</td>
<td>0.80000000</td>
<td>0.0000000</td>
<td>1</td>
<td>0.2</td>
<td>0.8</td>
<td>0.93333333</td>
<td>1</td>
<td>0.0000000</td>
<td>1</td>
</tr>
<tr>
<td>T</td>
<td>0.46666667</td>
<td>0.00000000</td>
<td>0.1333333</td>
<td>0</td>
<td>0.8</td>
<td>0.0</td>
<td>0.00000000</td>
<td>0</td>
<td>0.2000000</td>
<td>0</td>
</tr>
</tbody>
</table>

11
| A        | 0.06666667 |
| C        | 0.00000000 |
| G        | 0.46666667 |
| T        | 0.46666667 |

\[\text{noquote (t (as.data.frame (values (egr1.motif))))}\]

\[
\begin{array}{ll}
\text{providerName} & \text{Egr1} \\
\text{providerId} & \text{MA0162.1} \\
\text{dataSource} & \text{JASPAR\_CORE} \\
\text{geneSymbol} & \text{Egr1} \\
\text{geneId} & \text{13653} \\
\text{geneIdType} & \text{ENTREZ} \\
\text{proteinId} & \text{PO8046} \\
\text{proteinIdType} & \text{UNIPROT} \\
\text{organism} & \text{Mmusculus} \\
\text{sequenceCount} & \text{15} \\
\text{bindingSequence} & \text{<NA>} \\
\text{bindingDomain} & \text{Zinc-coordinating} \\
\text{tfFamily} & \text{BetaBetaAlpha-zinc finger} \\
\text{experimentType} & \text{bacterial 1-hybrid} \\
\text{pubmedID} & \text{16041365}
\end{array}
\]

Next we use the bioconductor \textit{seqLogo} package to display this motif.

\[\text{seqLogo (as.list (egr1.motif)[[1]])}\]

![Motif Matching](image)

\text{Figure 1: Mmusculus-JASPAR\_CORE-Egr1-MA0162.1}

### 3 Motif Matching

We will look for the ten position frequency matrices which are the best match to JASPAR’s mouse EGR1, using the MotIV package. We actually request the top eleven hits from the entire MotifDb, since the first hit should be the target matrix itself, since that is of necessity found in the full MotifDb.

\[\text{egrl.hits <- motifMatch (as.list (egr1.motif) [1], as.list (MotifDb), top=11)}\]

Ungapped Alignment
Scores read
Database read
Motif matches : 11
> # 'MotIV.toTable' -- defined above (and hidden) -- will become part of MotIV in the upcoming release
> tbl.hits <- MotIV.toTable (egr1.hits)
> print (tbl.hits)

```
name   eVal
1  Mmusculus-JASPAR_CORE-Egr1-MA0162.1 1.110223e-16
2  Hsaapiens-jolma2013-EGR2  3.330669e-16
3  Hsaapiens-jolma2013-EGR1  6.261658e-14
4  Hsaapiens-jolma2013-EGR2-2 9.148238e-14
5  Hsaapiens-jolma2013-EGR3  1.879608e-13
6  Hsaapiens-jolma2013-EGR2-1 2.541301e-13
7  Athalana-cimpb_1.02-M2867_1.02 2.691181e-13
8  Mmusculus-jolma2013-Egr3  3.194112e-13
9  Hsaapiens-jolma2013-EGR4  5.002665e-13
10  Hsaapiens-jolma2013-EGR4-2 1.278866e-12
```

```
# Melanogaster-FlyFactorSurvey-or_SANGER_5_FBgn0003499  2.638001e-12
```

The `sequence` column in this table is the **consensus sequence** – with heterogeneity left out – for the matrix it describes.

**Puzzling:** the strand of the match reported above is opposite of what I expected, and opposite of what seqLogo displays. This is a question for the MotIV developers.

The six logos appear below, beginning with the logo of the query matrix, *Mmusculus-JASPAR_CORE-Egr1-MA0162.1*, including two other mouse matrices, and two zinc-finger fly matrices. Examining the three mouse matrices and their metadata reveals that all three (geneSymbol differences aside) describe the same protein:

```
> if (interactive ()){
+ noquote (t (as.data.frame (subset (values (MotifDb), geneId=='13653'))))
```

Zinc finger protein domains are classified into many **fold groups**; their respective cognate DNA sequence may classify similarly. That two fly matrices significantly match three reports of the mouse Egr1 motif suggests impressive conservation of this binding pattern, or convergent evolution.

Let us look at the metadata for the first fly match, whose geneId is **FBgn0003499**:

```
> noquote (t (as.data.frame (values (MotifDb)[grep ('FBgn0003499', values (MotifDb)$geneId),])))
```

````
providerName providerId dataSource geneSymbol geneId geneIdType proteinId proteinIdType organism sequenceCount bindingSequence bindingDomain tfFamily experimentType pubmedID
```

that the SOLEXA motif, based upon 2316 sequences, did not (in work not shown, it appears 22nd in the an expanded motifMatch hit list, with an eval of 10e-5). It is possible that the SOLEXA motif is more accurate, and that a close examination of this case, including sequence logos, position frequency matrices, and the search parameters of motifMatch, will be instructive. Repeating the search with **tomtom** might also be illuminating – either as confirmation of MotIV and the default
parameterization we used, or as a correction to it. Here we see the facilities for exploratory data analysis MotifDb provides, and the opportunities for data analysis which result.

(a) Mmusculus-JASPAR_CORE-Egr1-MA0162.1 (abbreviated)

(b) Dme-FFS-sr_SANGER_5_FBgn0003499

(a) Mmusculus-UniPROBE-Zif268.UP00400

(b) Dme-FFS-klu_SANGER_10_FBgn0013469
4 Exporting to the MEME Suite

Some users of this package may wish to export the data – both matrices and metadata – so that they may be used in other programs. The MEME suite, among others, is broadly useful, continuously improved and well-regarded throughout the bioinformatics community. The code below exports all of the MotifDb matrices as a text file in the MEME format, and all of the metadata as a tab-delimited text file.

```r
> matrix.output.file = tempfile () # substitute your preferred filename here
> meme.text = export (MotifDb, matrix.output.file, ‘meme’)
> metadata.output.file = tempfile () # substitute your preferred filename here
> write.table (as.data.frame (values (MotifDb)), file=metadata.output.file, sep=‘\t’,
+ row.names=TRUE, col.names=TRUE, quote=FALSE)
```

5 Future Work

This first version of MotifDb collects into one R package all of the best-known public domain protein-DNA binding matrices, with as much metadata as could be gleaned from the five providers. However, not all of these matrices are equally supported by data and by no means are all are accompanied by complete metadata.

With the passage of time our knowledge of protein-DNA binding sequence motifs will improve. They will be derived from more binding events, with more precision and specificity, and accompanied by more (and better understood) contextual detail. Cooperative binding, mentioned only in a few times in the current (July 2012) version of this package, will be well-represented. Metadata will improve. Better assignment of binding domains to consensus categories will be especially useful when it is available. Three-dimensional models of specific proteins binding to specific DNA may someday become commonplace.

6 References


