MotifDb

Paul Shannon

April 11, 2014

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[1].

Contents

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] 4204
```

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

[1] http://meme.sdsc.edu/meme/doc/meme.html

```
      jolma2013         stamlab FlyFactorSurvey      JASPAR_2014      JASPAR_CORE 
            843             683             614             592             459 
           hPDI        UniPROBE          ScerTF 
            437             380             196 
```

And 22 organisms (though the majority of the matrices come from just four):

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

```
     Hsapiens Dmelanogaster   Scerevisiae     Mmusculus     Athaliana 
         2015           870           641           528            53 
     Celegans   Vertebrata   Rnorvegicus         Zmays        Amajus 
           22            16            14            12             6 
     Psativum        Gallus     Hroretzi      Hvulgare   Nsylvestris 
            6             3             2             2             2 
   Ocuniculus   Pfalciparum     Phybrida     Taestivam       Xlaevis 
            2             2             2             2             2 
      Cparvum       Rrattus 
            1             1 
```

With these categories of metadata

```
> 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:

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

```
MotifDb object of length 2015
| Created from downloaded public sources: 2013-Aug-30
| 2015 position frequency matrices from 6 sources:
|                hPDI:  437
|         JASPAR_2014:  117
|         JASPAR_CORE:   66
|           jolma2013:  710
|             stamlab:  683
|            UniPROBE:    2
| 1 organism/s
|            Hsapiens: 2015
Hsapiens-jolma2013-BCL6B 
Hsapiens-jolma2013-CTCF 
Hsapiens-jolma2013-EGR1 
Hsapiens-jolma2013-EGR1-2 
Hsapiens-jolma2013-EGR2 
...
Hsapiens-stamlab-UW.Motif.0681 
Hsapiens-stamlab-UW.Motif.0682 
Hsapiens-stamlab-UW.Motif.0683 
Hsapiens-UniPROBE-Sox4.UP00401 
Hsapiens-UniPROBE-Oct_1.UP00399 
```

If you want all matrices associated with ***Sox*** transcription factors, regardless of dataSource or organism:

```
> 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
|               hPDI:    2
|        JASPAR_2014:    8
|        JASPAR_CORE:    5
|          jolma2013:   56
|           UniPROBE:   15
| 4 organism/s
|           Hsapiens:   48
|          Mmusculus:   37
|      Dmelanogaster:    2
|         Vertebrata:    1
Dmelanogaster-FlyFactorSurvey-Sox14_SANGER_10_FBgn0005612
Dmelanogaster-FlyFactorSurvey-Sox15_SANGER_5_FBgn0005613
Hsapiens-hPDI-SOX13
Hsapiens-hPDI-SOX14
Hsapiens-JASPAR_CORE-SOX9-MA0077.1
...
Mmusculus-UniPROBE-Sox30.UP00023
Mmusculus-UniPROBE-Sox4.UP00062
Mmusculus-UniPROBE-Sox5.UP00091
Mmusculus-UniPROBE-Sox7.UP00034
Mmusculus-UniPROBE-Sox8.UP00051
```

For all yeast transcription factors with a homeo domain

```
> 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-TOS8-MA0408.1
Scerevisiae-JASPAR_2014-YHP1-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 accompanined by close scrutiny of the data, such as these commands illustrate:

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

```
 [1] "Homeobox"                         "Hox9_act;Homeobox"
 [3] "LIM;Homeobox"                     "PAX;Homeobox"
 [5] "OAR;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))
```

```
[1] "Homeo"
[2] "Homeo::Nuclear Factor I-CCAAT-binding"
[3] "Homeodomain"
[4] "homeodomain"
```

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 knowedge of MotifDb metadata is required.

```
> 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:

```
> 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.

```
> 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':

```
> 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:

```
> 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
|          jolma2013:    3
|           UniPROBE:    1
| 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 0.20000000 0.13333333 0.0000000 0 0.0 0.2 0.06666667 0 0.1333333  0
C 0.26666667 0.06666667 0.8666667 0 0.0 0.0 0.00000000 0 0.6666667  0
G 0.06666667 0.80000000 0.0000000 1 0.2 0.8 0.93333333 1 0.0000000  1
T 0.46666667 0.00000000 0.1333333 0 0.8 0.0 0.00000000 0 0.2000000  0
          11
A 0.06666667
C 0.00000000
G 0.46666667
T 0.46666667

$`Hsapiens-JASPAR_2014-EGR1-MA0162.2`
           1         2          3          4          5         6 7          8
A 0.08958877 0.1228786 0.09464752 0.10892624 0.01901110 0.2375163 0 0.00000000
C 0.46736292 0.5586651 0.49355418 0.85109334 0.94435379 0.0000000 1 0.96703655
G 0.25155026 0.1108845 0.18358355 0.00000000 0.00000000 0.5580940 0 0.00000000
T 0.19149804 0.2075718 0.22821475 0.03998042 0.03663512 0.2043897 0 0.03296345
           9         10         11        12         13        14
A 0.00000000 0.29797650 0.00000000 0.1932115 0.00000000 0.2468995
C 0.82849217 0.68219648 0.97519582 0.0000000 0.80360640 0.4565111
G 0.04985313 0.00000000 0.00000000 0.5384302 0.11586162 0.1560868
T 0.12165470 0.01982702 0.02480418 0.2683584 0.08053198 0.1405026

$`Hsapiens-jolma2013-EGR1`
          1          2           3           4           5           6
A 0.2494781 0.51390568 0.003223727 0.105202754 0.000000000 0.002604167
C 0.2411273 0.39540508 0.969696970 0.005355777 0.980025773 0.992838542
G 0.1539666 0.03627570 0.007736944 0.854246366 0.007731959 0.000000000
T 0.3554280 0.05441354 0.019342360 0.035195103 0.012242268 0.004557292
            7           8           9         10       11         12        13
A 0.000000000 0.652638191 0.003253090 0.01906158 0.010000 0.68089431 0.2790573
C 0.928214732 0.343592965 0.995445673 0.01136364 0.938125 0.06910569 0.2485270
G 0.009363296 0.000000000 0.000000000 0.93181818 0.011875 0.14227642 0.1253348
T 0.062421973 0.003768844 0.001301236 0.03775660 0.040000 0.10772358 0.3470809
         14
A 0.2673936
C 0.1905504
G 0.1396677
T 0.4023884

$`Hsapiens-jolma2013-EGR1-2`
          1           2           3          4 5           6         7
A 0.2722977 0.737507906 0.006723716 0.01834431 0 0.000000000 0.0000000
C 0.2309510 0.249209361 0.987775061 0.00000000 1 0.992159228 0.9797136
G 0.1139988 0.001897533 0.001833741 0.98165569 0 0.000000000 0.0000000
T 0.3827525 0.011385199 0.003667482 0.00000000 0 0.007840772 0.0202864
            8            9           10         11         12         13
A 0.795439739 0.0000000000 0.0000000000 0.00000000 0.86166008 0.29390244
C 0.200000000 0.9993943065 0.0000000000 0.99220156 0.01317523 0.27926829
G 0.004560261 0.0000000000 0.9990732159 0.00000000 0.10540184 0.06341463
T 0.000000000 0.0006056935 0.0009267841 0.00779844 0.01976285 0.36341463
         14
A 0.3035714
```

```
C 0.1255952
G 0.1077381
T 0.4630952

$`Mmusculus-jolma2013-Egr1`
          1          2           3            4           5            6
A 0.3231418 0.32278481 0.618181818 0.0000000000 0.075444498 0.0000000000
C 0.3241961 0.30907173 0.366753247 0.9968454259 0.004324844 0.9994728519
G 0.1133368 0.03691983 0.003636364 0.0005257624 0.911100432 0.0005271481
T 0.2393253 0.33122363 0.011428571 0.0026288118 0.009130226 0.0000000000
            7 8           9          10         11          12         13
A 0.001578117 0 0.517114271 0.003149606 0.00422833 0.001579779 0.89181562
C 0.997369805 1 0.481305951 0.995275591 0.16732105 0.998420221 0.05738476
G 0.001052078 0 0.001579779 0.000000000 0.25581395 0.000000000 0.03621825
T 0.000000000 0 0.000000000 0.001574803 0.57263667 0.000000000 0.01458137
          14         15         16
A 0.44251055 0.31170886 0.26213080
C 0.32278481 0.19778481 0.31012658
G 0.04957806 0.04272152 0.09651899
T 0.18512658 0.44778481 0.33122363

$`Mmusculus-UniPROBE-Egr1.UP00007`
          1          2          3          4           5           6
A 0.2115466 0.14198757 0.03260499 0.11512588 0.003516173 0.004715059
C 0.2827083 0.72243721 0.87717185 0.07060553 0.990021152 0.982482238
G 0.2034722 0.05485440 0.01243161 0.78128969 0.002264928 0.009896878
T 0.3022730 0.08072082 0.07779155 0.03297890 0.004197748 0.002905824
            7           8           9         10         11         12
A 0.001626612 0.262351637 0.005889514 0.02289301 0.02303758 0.56763334
C 0.975937323 0.731731673 0.985755764 0.09046006 0.85994854 0.05739392
G 0.001661635 0.002729558 0.002081402 0.64932246 0.03791264 0.16679165
T 0.020774430 0.003187133 0.006273319 0.23732447 0.07910124 0.20818108
         13        14
A 0.1765973 0.1830489
C 0.3312648 0.1837744
G 0.1253083 0.2267928
T 0.3668295 0.4063840
```

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

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

```
                [,1]                      [,2]
providerName    Egr1                      EGR1
providerId      MA0162.1                  MA0162.2
dataSource      JASPAR_CORE               JASPAR_2014
geneSymbol      Egr1                      EGR1
geneId          13653                     1958
geneIdType      ENTREZ                    ENTREZ
proteinId       P08046                    P18146
proteinIdType   UNIPROT                   UNIPROT
organism        Mmusculus                 Hsapiens
sequenceCount      15                     12256
bindingSequence <NA>                      <NA>
bindingDomain   Zinc-coordinating         Zinc-coordinating
tfFamily        BetaBetaAlpha-zinc finger BetaBetaAlpha-zinc finger
experimentType  bacterial 1-hybrid        ChIP-seq
pubmedID        16041365                  16041365
                [,3]                    [,4]
providerName    Hsapiens-jolma2013-EGR1 Hsapiens-jolma2013-EGR1-2
providerId      EGR1                    EGR1
dataSource      jolma2013               jolma2013
geneSymbol      EGR1                    EGR1
geneId          1958                    1958
geneIdType      ENTREZ                  ENTREZ
proteinId       <NA>                    <NA>
proteinIdType   <NA>                    <NA>
organism        Hsapiens                Hsapiens
sequenceCount    1831                    1703
bindingSequence NMCGCCCMCGCANN          NACGCCCACGCANN
bindingDomain   <NA>                    <NA>
tfFamily        C2H2                    C2H2
experimentType  SELEX                   SELEX
pubmedID        23332764                23332764
                [,5]                     [,6]
providerName    Mmusculus-jolma2013-Egr1 SCI09/Egr1_pwm_primary.txt
providerId      Egr1                     UP00007
dataSource      jolma2013                UniPROBE
geneSymbol      Egr1                     Egr1
geneId          1958                     13653
geneIdType      ENTREZ                   ENTREZ
proteinId       <NA>                     P08046
proteinIdType   <NA>                     UNIPROT
organism        Mmusculus                Mmusculus
```

```
sequenceCount    2013                <NA>
bindingSequence NNMCGCCCMCTCANNN    <NA>
bindingDomain   <NA>                ZnF_C2H2
tfFamily        C2H2                <NA>
experimentType  SELEX               protein binding microarray
pubmedID        23332764            19443739
```

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:

```
> 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] 1206
```

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

Far more concise, and fully reliable as an interactive command (though *not* if used in a script[2]):

```
> 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.

```
> 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
```

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

```
$`Mmusculus-JASPAR_CORE-Egr1-MA0162.1`
           1          2         3 4   5   6          7 8         9 10
A 0.20000000 0.13333333 0.0000000 0 0.0 0.2 0.06666667 0 0.1333333  0
C 0.26666667 0.06666667 0.8666667 0 0.0 0.0 0.00000000 0 0.6666667  0
G 0.06666667 0.80000000 0.0000000 1 0.2 0.8 0.93333333 1 0.0000000  1
T 0.46666667 0.00000000 0.1333333 0 0.8 0.0 0.00000000 0 0.2000000  0
          11
A 0.06666667
C 0.00000000
G 0.46666667
T 0.46666667
```

```
> noquote (t (as.data.frame (values (egr1.motif))))
```

[2] See the help page of the base R command subset for detail), is the *subset* command

```
                [,1]
providerName    Egr1
providerId      MA0162.1
dataSource      JASPAR_CORE
geneSymbol      Egr1
geneId          13653
geneIdType      ENTREZ
proteinId       P08046
proteinIdType   UNIPROT
organism        Mmusculus
sequenceCount   15
bindingSequence <NA>
bindingDomain   Zinc-coordinating
tfFamily        BetaBetaAlpha-zinc finger
experimentType  bacterial 1-hybrid
pubmedID        16041365
```

Next we use the bioconductor *seqLogo* package to display this motif.

```
> seqLogo (as.list (egr1.motif)[[1]])
```

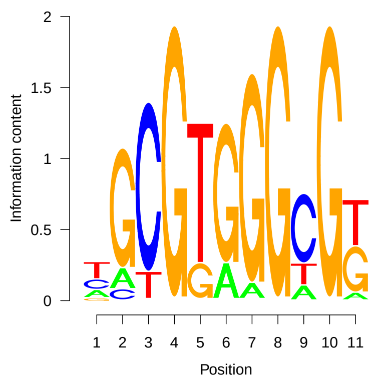


Figure 1: Mmusculus-JASPAR_CORE-Egr1-MA0162.1

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.

```
> egr1.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                      Hsapiens-jolma2013-EGR2 3.330669e-16
3                      Hsapiens-jolma2013-EGR1 6.261658e-14
4                    Hsapiens-jolma2013-EGR2-2 9.148238e-14
5                      Hsapiens-jolma2013-EGR3 1.879608e-13
6                    Hsapiens-jolma2013-EGR1-2 2.541301e-13
7                     Mmusculus-jolma2013-Egr3 3.194112e-13
8                      Hsapiens-jolma2013-EGR4 5.002665e-13
9                    Hsapiens-jolma2013-EGR4-2 1.278866e-12
```

```
10 Dmelanogaster-FlyFactorSurvey-sr_SANGER_5_FBgn0003499 2.638001e-12
11                  Mmusculus-JASPAR_2014-EGR2-MA0472.1 4.143796e-12
           sequence            match strand
1       NGCGTGGGCGK      NGCGTGGGCGK      +
2       NGCGTGGGCGK      TGCGTGGGCGK      -
3    --NGCGTGGGCGK-   NNTGCGTGGGCGKN      -
4   --NGCGTGGGCGK--  NNNGCGTGGGCGKNN      -
5   --NGCGTGGGCGK--  NNTGCGTGGGCGKNN      -
6    --NGCGTGGGCGK-   NNTGCGTGGGCGTN      -
7   ---NGCGTGGGCGK-  NNNTGCGTGGGCGTN      -
8  ---NGCGTGGGCGK-- NNNTGCGTGGGCGKNN      -
9  ---NGCGTGGGCGK-- NWNTGCGTGGGCGTNN      -
10   -NGCGTGGGCGK--   NKRNGKGGGCGKNN      +
11  -NGCGTGGGCGK---  NNGYGTGGGYGKNNN      -
```

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'))))
```

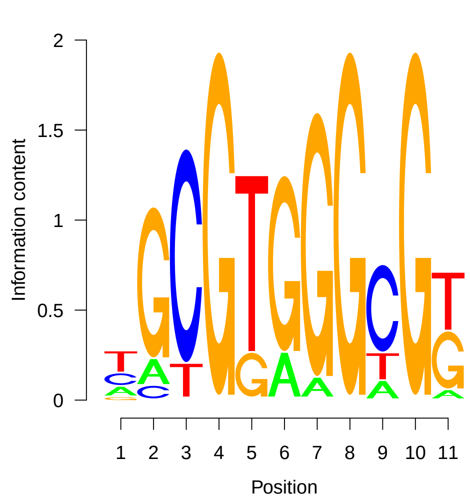
Zinc finger protein domains are classified into many *fold groups*; their respective cognate DNA sequence may classify similarly. That two fly matrices significantly match three reports of the mouse Egr1 motif suggests impressive conservation of this binding pattern, or convergent evolution.

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

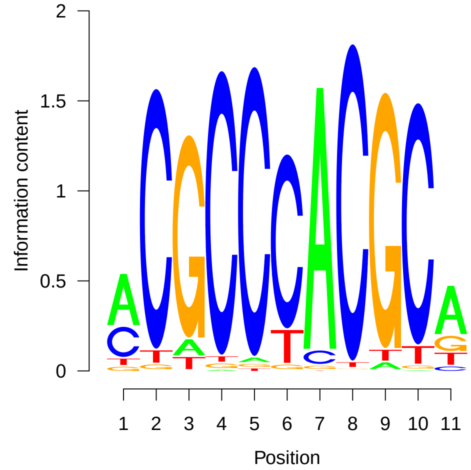
```
> noquote (t (as.data.frame (values (MotifDb)[grep ('FBgn0003499', values (MotifDb)$geneId),])))
```

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

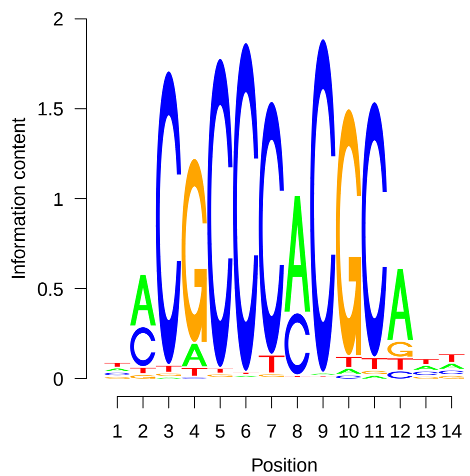
that the SOLEXA motif, based upon 2316 sequences, did not (in work not shown, it appears 22nd in the an expanded motifMatch hit list, with a eval of 10e-5). It is possible that the SOLEXA motif is more accurate, and that a close examination of this case, including sequence logos, position frequency matrices, 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.



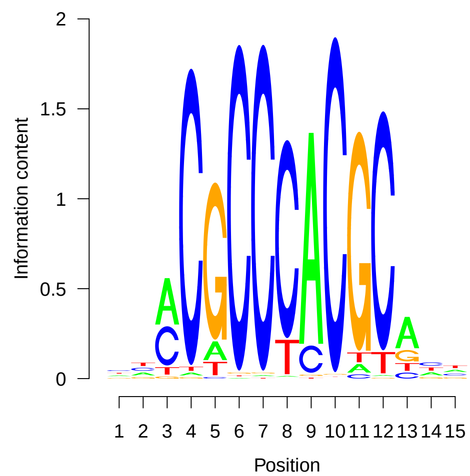
(a) Mmusculus-JASPAR_CORE-Egr1-MA0162.1



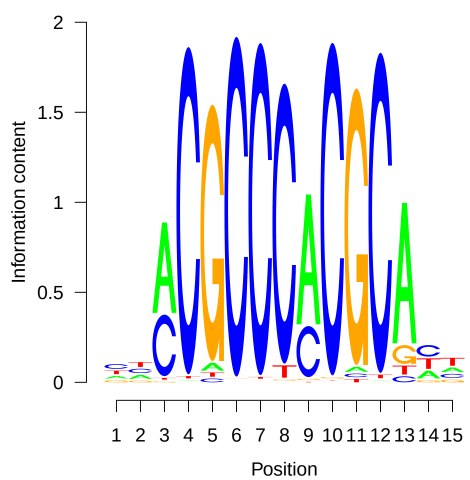
(b) Dme-FFS-sr_SANGER_5_FBgn0003499 (abbreviated)



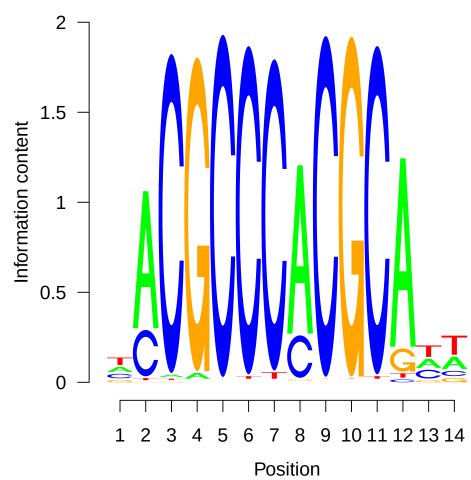
(c) Mmusculus-UniPROBE-Zif268.UP00400



(d) Dme-FFS-klu_SANGER_10_FBgn0013469



(e) Mmusculus-UniPROBE-Egr1.UP00007



(f) Dme-FFS-klu_SOLEXA_5_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.

```
> 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

- 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.