Using Probe Information

Robert Gentleman

Overview

The Bioconductor project maintains a rich body of annotation data assembled into R libraries. For many different Affymetrix chips information is provided on both the sequence of the mRNA that was intended to be matched and the actual 25mers that were used for the bindings. In this vignette we show how to make use of the probe information.

A Simple Example

To demonstrate the use of probe level data we will use the `rae230a` chip (for rats). So we first need to load these libraries.

```
> library("annotate")
> library("rae230a.db")
> library("rae230aprobe")
```

Now, we do not have any data so all we are going to do is to examine the probe data and show how to use some of the different Bioconductor tools to access that information, and potentially check on the mapping information that has been given.

We will select a probe set,

```
> ps = names(as.list(rae230aACCNUM))
> myp = ps[1001]
> myA = get(myp, rae230aACCNUM)
> wp = rae230aprobe$Probe.Set.Name == myp
> myPr = rae230aprobe[wp,]
>
```

The probe data is stored as a *data.frame* with 6 columns. They are

sequence The sequence of the 25mer

x The x position of the probe on the array.

y The y position of the probe on the array.

Probe.Set.Name The Affymetrix ID for the probe set.

Probe.Interrogation.Position The location (in bases) of the 13th base in the 25mer, in the target sequence.

Target.Strandedness Whether the 25mer is a Sense or an Antisense match to the target sequence.

We note that it is not always the case that the sequence reported is found in the reference or if it is, it is not always at the location reported. One can check that using other tools available in the *annotate* package and in the *Biostrings* package.

```
> myseq = getSEQ(myA)
> nchar(myseq)

[1] 5775

> library("Biostrings")
> mybs = DNAString(myseq)
> match1 = matchPattern(as.character(myPr[1,1]), mybs)
> match1

Views on a 5775-letter DNAString subject
subject: GCCCGGGTCCCGCCTCTTCCTCAGCTTGG...TTAATAAAGGATTTACGGGATTTCCTTTC
views:
      start  end width
  [1]  5212 5236    25 [TGGGATTATGGCCTGTGTCACCACG]

> as.matrix(ranges(match1))

     [,1] [,2]
[1,] 5212   25

> myPr[1,5]

[1] 5224
```

And we can see that in this case the 13th nucleotide is indeed in exactly the place that has been predicted.

One additional thing to note is that Affymetrix does not accurately report the strandedness of the probes, so it is necessary to check the reverse complement of the sequence prior to assuming that the probe does not interrogate the correct gene.

```
> myp = ps[100]
> myA = get(myp, rae230aACCNUM)
> wp = rae230aprobe$Probe.Set.Name == myp
> myPr = rae230aprobe[wp,]
> myseq = getSEQ(myA)
> mybs = DNAString(myseq)
> Prstr = as.character(myPr[1,1])
> match2 = matchPattern(Prstr, mybs)
> ## expecting 0 (no match)
> length(match2)
```

```
[1] 0

> match2 = matchPattern(reverseComplement(DNAString(Prstr)), mybs)
> nchar(match2)

[1] 25

> nchar(myseq) - as.matrix(ranges(match2))

     [,1] [,2]
[1,]  273  652

> myPr[1,5]

[1] 262
```

Again, we see that the 13th nucleotide is exactly where predicted. It is relatively straightforward to check the other 25mers, and to develop different visualization tools that can be used to investigate the available data.

Other Sources of Information

There are other tools available that may also be of some interest. For instance, the Mental Health Research Institute at the University of Michigan have various custom cdf files for Affymetrix data analysis that have been updated using more current annotation information from GenBank and Ensembl.

http://brainarray.mhri.med.umich.edu/Brainarray/Database/CustomCDF/genomic_curated_CDF.asp

The Weizmann Institute of Science have a database that can be queried to get the sensitivity and specificity for the probes on the Affymetrix HG-U95av2 chip. Although the information here is limited to a particular chip, this general idea is something that an enterprising end-user might want to replicate for other chips.

http://genecards.weizmann.ac.il/geneannot/

1 Session Information

The version number of R and packages loaded for generating the vignette were:

```
R version 4.4.0 (2024-04-24 ucrt)
Platform: x86_64-w64-mingw32/x64
Running under: Windows Server 2022 x64 (build 20348)

Matrix products: default


locale:
```

```
[1] LC_COLLATE=C
[2] LC_CTYPE=English_United States.utf8
[3] LC_MONETARY=English_United States.utf8
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.utf8

time zone: America/New_York
tzcode source: internal

attached base packages:
[1] grid      stats4    stats     graphics  grDevices utils
[7] datasets  methods   base

other attached packages:
 [1] Biostrings_2.72.0    GenomeInfoDb_1.40.0  XVector_0.44.0
 [4] rae230aprobe_2.18.0  rae230a.db_3.13.0    org.Rn.eg.db_3.19.1
 [7] Rgraphviz_2.48.0     graph_1.82.0         xtable_1.8-4
[10] GO.db_3.19.1         hgu95av2.db_3.13.0   org.Hs.eg.db_3.19.1
[13] annotate_1.82.0      XML_3.99-0.16.1      AnnotationDbi_1.66.0
[16] IRanges_2.38.0       S4Vectors_0.42.0     Biobase_2.64.0
[19] BiocGenerics_0.50.0  BiocStyle_2.32.0

loaded via a namespace (and not attached):
 [1] sass_0.4.9              RSQLite_2.3.6
 [3] digest_0.6.35           evaluate_0.23
 [5] bookdown_0.39           fastmap_1.1.1
 [7] blob_1.2.4              jsonlite_1.8.8
 [9] DBI_1.2.2               BiocManager_1.30.23
[11] httr_1.4.7              UCSC.utils_1.0.0
[13] jquerylib_0.1.4         cli_3.6.2
[15] rlang_1.1.3             crayon_1.5.2
[17] bit64_4.0.5             cachem_1.0.8
[19] yaml_2.3.8              tools_4.4.0
[21] memoise_2.0.1           GenomeInfoDbData_1.2.12
[23] vctrs_0.6.5             R6_2.5.1
[25] png_0.1-8               lifecycle_1.0.4
[27] zlibbioc_1.50.0         KEGGREST_1.44.0
[29] bit_4.0.5               pkgconfig_2.0.3
[31] bslib_0.7.0             xfun_0.43
[33] knitr_1.46              htmltools_0.5.8.1
[35] rmarkdown_2.26          compiler_4.4.0
```