

# 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...TTAATAAAGGATTTACGGGATTTCTTTTC
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)
```



```

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

time zone: Etc/UTC
tzcode source: system (glibc)

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

other attached packages:
 [1] Biostrings_2.79.1      Seqinfo_1.1.0          XVector_0.51.0
 [4] rae230aprobe_2.18.0    rae230a.db_3.13.0      org.Rn.eg.db_3.22.0
 [7] Rgraphviz_2.55.0       graph_1.89.0           xtable_1.8-4
[10] GO.db_3.22.0           hgu95av2.db_3.13.0     org.Hs.eg.db_3.22.0
[13] annotate_1.89.0         XML_3.99-0.19          AnnotationDbi_1.71.2
[16] IRanges_2.45.0         S4Vectors_0.49.0       Biobase_2.71.0
[19] BiocGenerics_0.57.0    generics_0.1.4         BiocStyle_2.39.0

loaded via a namespace (and not attached):
 [1] sass_0.4.10            RSQLite_2.4.3          digest_0.6.37
 [4] evaluate_1.0.5         fastmap_1.2.0          blob_1.2.4
 [7] jsonlite_2.0.0         DBI_1.2.3              BiocManager_1.30.26
[10] httr_1.4.7             jquerylib_0.1.4        cli_3.6.5
[13] rlang_1.1.6            crayon_1.5.3           bit64_4.6.0-1
[16] cachem_1.1.0           yaml_2.3.10            tools_4.5.1
[19] memoise_2.0.1          buildtools_1.0.0       vctrs_0.6.5
[22] R6_2.6.1               png_0.1-8              lifecycle_1.0.4
[25] KEGGREST_1.49.2        bit_4.6.0              pkgconfig_2.0.3
[28] bslib_0.9.0            xfun_0.54              sys_3.4.3
[31] knitr_1.50             htmltools_0.5.8.1      rmarkdown_2.30
[34] maketools_1.3.2        compiler_4.5.1

```