Annotations with NetAffx

Martin Morgan, Robert Gentleman

Created: 19 February 2008

Affymetrix provides annotations for all arrays they produce. The annotations are made available in Bioconductor with the *NetAffxResource* class in the AffyCompatible package; additional packages complement Affymetrix annotation information with data collected from additional public repositories. This document outlines a simple workflow to retrieve annotations available through NetAffx.

```
> library(AffyCompatible)
```

To use these facilities, one must be a registered Affymetrix user; see the Affymetrix user registration site for details.

The first step is to create an instance of the *NetAffxResource* class. Do this using the `NetAffxResource` function. Important arguments are *user* and *password* length 1 character vectors containing the registered user name and password. The password is printed, saved, and transmitted in clear text, and so is **not** secure. An additional argument is *directory*, which is the location where the NetAffx data base and downloaded files are stored. *directory* defaults to a session-specific temporary directory, meaning that if it is not supplied the data base and any downloaded annotations are removed when the R session ends. To create the *NetAffxResource* instance, evaluate a command like

```
> rsrc <- NetAffxResource(user="mtmorgan@fhcrc.org", password=password)
> rsrc
```

```
directory: /tmp/RtmpkeocVD
annotationsFile: NetAffxAnnotFileList.xml
affxUrl: https://www.affymetrix.com/analysis/downloads/netaffxapi/GetFileList.jsp
affxLicence: ********
user: mtmorgan@fhcrc.org
password: ********
```

This creates the resource, but does not validate the user name and password (the user name and password are verified when the NetAffx resource is first retrieved from Affymetrix, typically the first time the code in the following paragraph is evaluated).

A typical workflow involves querying `rsrc` for the names of available arrays, and the descriptions of annotations available for an array of interest:

```
> head(names(rsrc))

[1] "AG"                "ATH1-121501"       "AraGene-1_0-st-v1"
[4] "AraGene-1_1-st-v1" "AutoFocus_Array"   "Axiom_BioBank1"

> affxDescription(rsrc[["Bovine"]])

[1] "Annotations, CSV format"         "CDF Library File"
[3] "CIF Library File"                "Orthologs/Homologs, CSV Format"
[5] "PSI Library File"                "Probe Sequences, FASTA format"
[7] "Probe Sequences, tabular format"
```

Annotations usually include a comma-separated value (CSV) file that can be represented in R as a `data.frame`. The data frame usually includes a probe identifier column, and columns of additional information Affymetrix has collated from a variety of sources, as described on the NetAffx site. Additional annotation files usually include a (much larger physically, but containing comparable information) MAGE-ML representation of the CSV file, channel description files (CDF), other files describing probes preesent on chips, probe sequences in FASTA format, and possibly other files specific to the chip platform.

An R representation of the annotations of a particular array can be created with

```
> annos <- rsrc[["Porcine"]]
> annos

affxName: Porcine
affxAnnotation: AffxAnnotation(7)
```

A particular annotation can be selected from this using R commands to navigate the implied class structure:

```
> sapply(affxAnnotation(annos), force)[1:5]

[[1]]
affxType: Annot CSV
affxDescription: Annotations, CSV format
affxFile: AffxFile(1)

[[2]]
affxType: CDF
affxDescription: CDF Library File
affxFile: AffxFile(1)

[[3]]
affxType: CIF
affxDescription: CIF Library File
affxFile: AffxFile(1)
```

```
[[4]]
affxType: PSI
affxDescription: PSI Library File
affxFile: AffxFile(1)

[[5]]
affxType: Probe FASTA
affxDescription: Probe Sequences, FASTA format
affxFile: AffxFile(1)
```

```
> anno <- affxAnnotation(annos)[[3]]
> anno
```

```
affxType: CIF
affxDescription: CIF Library File
affxFile: AffxFile(1)
```

(The Porcine BLASTP Annotation file is chosen because it is small). The annotation file may also be obtained by subsetting the reseource with a second argument corresponding to the annotation description or index

```
> anno <- rsrc[["Porcine", "Annotations, CSV format"]]
> anno <- rsrc[["Porcine", 3]]
```

Annotation files can be retrieved with

```
> df <- readAnnotation(rsrc, annotation=anno)
```

This checks to see if the relevant annotation file is in the directory specified in the `rsrc` object. If the annotation file is not present, it is retrieved from the Affymetrix site. The argument *update=TRUE* forces retrieval. `readAnnotation` will read files with known type (e.g., CSV) into appropriate R objects (e.g., data frames), and return these from `readAnnotation`. Some file types (e.g., CDF) are not meant for representation as R objects, and for these `readAnnotation` returns the (local) path to the relevant file. For all annotations, the argument *content=FALSE* returns the local file path, without loading the content of the file into R.

Affymetrix does not specify the format of all files, so some files might reasonably be read into R but the `readAnnotation` code is not able to identify the appropriate format. The user is free to explore these annotation files using standard R commands, e.g.,

```
> anno <- rsrc[["Porcine", "PSI Library File"]]
> fl <- readAnnotation(rsrc, annotation=anno, content=FALSE)
> fl
```

```
[1] "/tmp/RtmpkeocVD/Porcine.psi.zip"
```

```
> ## a zip file, containing 'Porcine.psi'
> conn <- unz(fl, "Porcine.psi")
> readLines(conn, n=6)

[1] "#Probe Sets: 24123"    "1\tAFFX-BioB-5_at\t20" "2\tAFFX-BioB-M_at\t20"
[4] "3\tAFFX-BioB-3_at\t20" "4\tAFFX-BioC-5_at\t20" "5\tAFFX-BioC-3_at\t20"

> read.table(conn, header=FALSE, skip=1, sep="\t", nrows=5)

  V1             V2 V3
1  1 AFFX-BioB-5_at 20
2  2 AFFX-BioB-M_at 20
3  3 AFFX-BioB-3_at 20
4  4 AFFX-BioC-5_at 20
5  5 AFFX-BioC-3_at 20
```