Using Bioconductor's Annotation Libraries

Jianhua Zhang

Overview

The Bioconductor project maintains a rich body of annotation data assembled into R libraries. The purpose of this vignette is to discuss the structure, contents, and usage of these annotation data libraries. Executable code is provided as examples.

Contents

Bioconductor's annotation data libraries are constructed by assembling data collected from various public data repositories using Bioconductor's *AnnBuilder* package and distributed as regular R libraries that can be installed and loaded in the same way an R library is installed/loaded. Each annotation library is an independent unit that can be used alone or in conjunction with other annotation libraries. Figure 1 shows the relationship among libraries that are currently available. Platform specific libraries are a group annotation libraries assembled specifically for given platforms (e. g. Affymetrix HG_U95Av2). CHRLOC and LLMappings are groups of libraries containing data assembled at genome level for human, mouse, or rat. *KEGG* and *GO* are source specific libraries containing generic data for various genomes.

Each annotation library, when installed, contains a `data` and `man` subdirectory filled with assembled data and documentation about the data, respectively. Most assembled data are stored as binary R environment objects (hash table with key-value pairs) associating annotation values to a set of keys. For each environment object in the `data` directory, there is a corresponding help file in the `man` directory with detail descriptions of the data file and usage.

Each platform specific library contains R environment objects named following the convention of package name plus environment name. The package name is in lower case letters and the environment names are in capital letters. When a given environment maps platform specific keys to annotation data, only the name of the annotation data is used for the name of the environment. Otherwise, the environment name have a pattern of key name and value name joined by a "2" in between. For example, `hgu95av2ENTREZID` maps probe ids on an Affymetrix human genome U95Av2 chip to EntrezGene IDs while `hgu95av2GO2PROBE` maps Gene Ontology IDs to probe IDs. Names of the

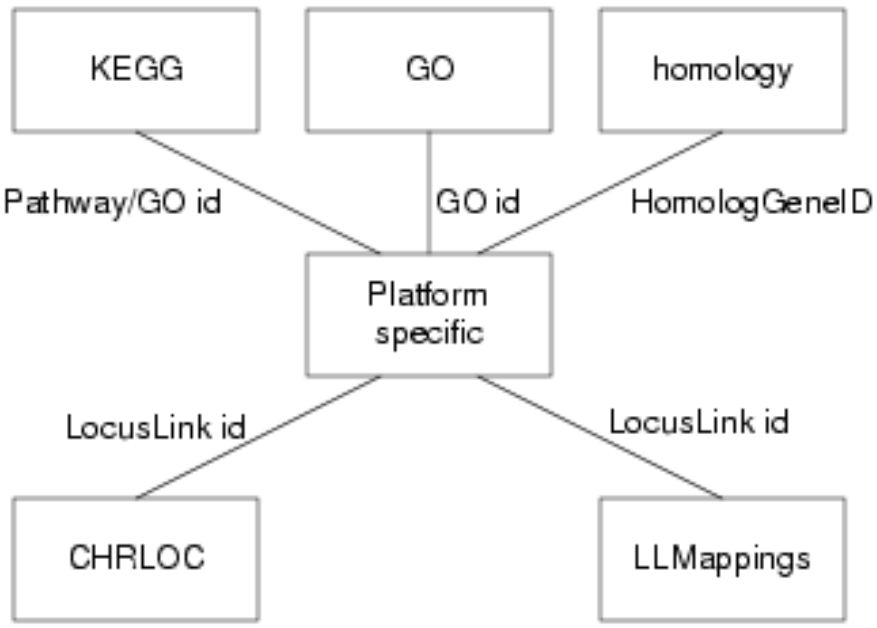

Figure 1: A diagram showing the overall structure and inter-relationship of Bioconductor annotation data packages. Boxes show a single or group of data packages. Lines between two boxes indicate connections through the key denoted by the text beside each line.

environment objects in a platform specific data package are not listed here to save space but are easily accessible as shown later in the section for usage.

Genome level annotation libraries are named in the form of XXXCHARLOC and XXXLLMappings, where XXX represents an organism name. Each data package in the CHRLOC group has LocusLink ids mapped to chromosomal location where transcription of genes corresponding to the LocusLink ids begins and ends. There is a START and END environment for LocusLink ids on each of the chromosomes (e. g. `humanCHRLOC1START` and `humanCHRLOC1END` for human chromosome number 1). Each library in the LLMappings group provides mappings/reverse mappings between LocusLink ids and some other commonly used public repository identifiers. Environment names are a cancatinations of the names of mapped ids with a 2 in between (e. g. `humanLLMappingsLL2GO`).

The KEGG library contains mappings between ids such as Locuslink and *GO* to *KEGG* pathway ids and thus pathway names. The *GO* library maintain the directed acyclic graph structure of the original data from Gene Ontology Consortium by providing mappings of GO ids to their direct parents or children for each of the three categories (molecular function, cellular component, and biological process). In addition, mappings between LocusLink and *GO* ids are provided with evidence code that specifies the type of evidence that supports the annotation of a gene to a particular *GO* term.

Usage

All the annotation libraries can be obtained from the Bioconductor web site (`http://www.bioconductor.org`). To illustrate their usages, we use the library for Affymetrix HG_U95Av2 chip (*hgu95av2*) as an example for platform specific data packages and the *GO* library for non-platform specific data packages. We assume that R (`www.r-project.org`) and Bioconductor's Biobase and annotation libraries have already been installed.

Package installation

After downloading libraries *hgu95av2* and *GO* from the link to MetaData of the Bioconductor web site, install the libraries by typing R INSTALL library name in the directory where the library is stored under Unix or click the menu bar *Packages* and then *Install package(s) from local zip files...* of an R session under Windows. Alternatively, if the *reposTools* library has already been installed/loaded, typing `install.packages2(library name)` installs the library for both Unix and Windows.

Typing library(library name) in an R session will load the library into R. For example,

```
> library(annotate)
> library(hgu95av2)
> library(GO)
```

Documentations

Each library contains documentations for the library in general and each of the individual environment objects contained by the library. Two documents at the library level can be accessed by typing a library name proceeded by a question mark (e. g. ?hgu95av2) and the library name followed by a pair of brackets (e. g. hgu95av2()), respectively. The former indicates how the data package was built from what version of public data sources and the latter lists all the environments contained by a library and provides information on the total number of keys within each of the environment objects contained by the library and how many of these keys are annotated.

The documentation for a given environment object can be accessed by typing the name of an environment object proceeded by a question mark (e. g. `hgu95av2GO`). The resulting documentation provides detail explanations to the environment object, data source used to build the object, and example code for accessing annotation data.

Accessing annotation data within a library

Annotation data of a given library are stored as environment objects in the form of key (items to be annotated) and value (annotation for an key item) pairs. Each environment object provides annotation for keys for a particular subject reflected by the name of the object. For example, `hgu95av2GO` annotates probes on the HGU95Av2 chip with ids of the Gene Ontology terms the probes correspond to.

The name of an environment object consists of package name (*hgu95av2*) and environment name (*GO*) to avoid confusion when multiple libraries are loaded to the system at the same time. Data contained by an environment can be accessed easily using Bioconductor's existing functions. For example, the following code stores all the keys contained by the `hgu95av2GO` environment object to variable *temp* and displays the first five keys on the screen:

```
> temp <- as.list(hgu95av2GO)
> temp[5]

$`738_at`
$`738_at`$`GO:0005737`
$`738_at`$`GO:0005737`$GOID
[1] "GO:0005737"

$`738_at`$`GO:0005737`$Evidence
[1] "IEA"

$`738_at`$`GO:0005737`$Ontology
[1] "CC"
```

```
$`738_at`$`GO:0005829`
$`738_at`$`GO:0005829`$GOID
[1] "GO:0005829"

$`738_at`$`GO:0005829`$Evidence
[1] "NR"

$`738_at`$`GO:0005829`$Ontology
[1] "CC"


$`738_at`$`GO:0008253`
$`738_at`$`GO:0008253`$GOID
[1] "GO:0008253"

$`738_at`$`GO:0008253`$Evidence
[1] "TAS"

$`738_at`$`GO:0008253`$Ontology
[1] "MF"


$`738_at`$`GO:0016787`
$`738_at`$`GO:0016787`$GOID
[1] "GO:0016787"

$`738_at`$`GO:0016787`$Evidence
[1] "IEA"

$`738_at`$`GO:0016787`$Ontology
[1] "MF"
```

To obtain annotation for a given set of keys, one may use the `mget` function. Suppose we have run an experiment using the HG_U95Av2 chip and found three genes represented by Affymetrix probe ids *738_at, 40840_at*, and *41668_r_at* interesting. To get the names of genes the three probe ids corresponding to, we do:

```
> mget(c("738_at", "40840_at", "41668_r_at"), hgu95av2GENENAME)

$`738_at`
[1] "5'-nucleotidase, cytosolic II"

$`40840_at`
[1] "peptidylprolyl isomerase F (cyclophilin F)"
```

```
$`41668_r_at`
[1] "TDP-glucose 4,6-dehydratase"
```

Similarly, identifiers of Gene Ontology terms corrsponding to the three probes can be obtained as shown below:

```
> temp <- mget(c("41561_s_at", "40840_at", "41668_r_at"),
+     hgu95av2GO)
```

In this case, the function `mget` returns a list of pre-defined S4 objects containing data for the ids, ontology, and evidence code of Gene Ontology terms corresponding to the three keys. The following code shows how to access the GO id, evidence code and ontology of the Gene Ontology term corresponding to probe id *40840_at*:

```
> temp <- get("738_at", hgu95av2GO)
> names(temp)

[1] "GO:0005737" "GO:0005829" "GO:0008253" "GO:0016787"

> temp[["GO:0008253"]][["Evidence"]]

[1] "TAS"

> temp[["GO:0008253"]][["Ontology"]]

[1] "MF"
```

As shown above, probe *40840_at* can be annotated by three Gene Ontology terms identified by *GO:0005829*, *GO:0008253*, and *GO:0016787*. The evidence code for *GO:0008253* is *TAS* (traceable author statement) and it belongs to ontology *MF* (molecular function).

Accessing annotation data across libraries

Often, data available in a given data package alone may not be sufficient and need to be sought across packages. Bioconductor's annotation data packages are linked by common public data identifiers to allow traverse between packages (Fig. 1). Using the example above, we know that probe id *738_at* are annotated by three Gene Ontology ids *GO:0005829*, *GO:0008253*, and *GO:0016787*. The Gene Ontology terms for various Gene Ontology ids, however, are stored in another package named *GO*. AS package *hgu95av2* and *GO* are linked by *GO* ids, one can annotate probe id *738_at* with Gene Ontology terms by linking data in the two packages using *GO* id as shown below:

```
> mget(names(get("738_at", hgu95av2GO)), GOTERM)
```

```
$`GO:0005737`
GOID: GO:0005737
Term: cytoplasm
Ontology: CC
Definition: All of the contents of a cell excluding the plasma
    membrane and nucleus, but including other subcellular
    structures.

$`GO:0005829`
GOID: GO:0005829
Term: cytosol
Ontology: CC
Definition: That part of the cytoplasm that does not contain
    membranous or particulate subcellular components.

$`GO:0008253`
GOID: GO:0008253
Term: 5'-nucleotidase activity
Ontology: MF
Definition: Catalysis of the reaction: a 5'-ribonucleotide +
    H2O = a ribonucleoside + phosphate.
Synonym: 5' nucleotidase activity

$`GO:0016787`
GOID: GO:0016787
Term: hydrolase activity
Ontology: MF
Definition: Catalysis of the hydrolysis of various bonds, e.g.
    C-O, C-N, C-C, phosphoric anhydride bonds, etc. Hydrolase
    is the systematic name for any enzyme of EC class 3.
```

It turns out that probe id *738_at* (corresponding to *GO:0008253*, and *GO:0016787*) has molecular function (MF) *5'-nucleotidase activity* and *hydrolase activity*.

1 Session Information

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

```
R version 2.6.0 (2007-10-03)
x86_64-unknown-linux-gnu

locale:
LC_CTYPE=en_US;LC_NUMERIC=C;LC_TIME=en_US;LC_COLLATE=en_US;LC_MONETARY=en_US;LC_MESSAGES=en_

attached base packages:
[1] tools     stats     graphics  grDevices utils     datasets
```

```
[7] methods   base

other attached packages:
 [1] XML_1.93-2          GO_2.0.1            Rgraphviz_1.16.0
 [4] graph_1.16.1        hgu95av2_2.0.1      annotate_1.16.1
 [7] xtable_1.5-2        AnnotationDbi_1.0.6 RSQLite_0.6-3
[10] DBI_0.2-4           Biobase_1.16.1

loaded via a namespace (and not attached):
[1] cluster_1.11.9
```