How to use the homology packages

Jianhua Zhang Ben Wittner

Introduction

The homology packages are a group of data packages built based on a source data file provided by HomoloGene (`http://www.ncbi.nlm.nih.gov/HomoloGene/`). There is a separate package for each individual organism. This vignette illustrates how to use the homology packages to explore the homologous relationships among organisms. Two or more organisms have homologous relationships if their genes share an arbitrary threshold level of similarity determined by alignment of matching bases.

This vignette demonstrates the usage of the homology packages by performing two tasks to introduce users to the packages, although the potential use of the packages is well beyond the scope of the tasks.

Contents of the packages

The homology packages are built on organism bases with a separate data package for each of the species contained in the source file. The packages are named following a convention of *XXXhomology* where XXX are a three letter string consisting of the first letter of the genus name and the first two letters of the species name of a given organism. For example, hsahomology is the package for human (Homo sapiens).

As other Bioconductor data packages, the homology packages contains environment objects in the `data` subdirectory that will be loaded when the package is loaded. A full list of the environment objects is shown below (using the package for human as an example):

```
> library("annotate")
> library("hsahomology")
> hsahomology()

Quality control information for  hsahomology
Date built: Created: Wed Aug 29 13:38:12 2007

Mappings found for non-probe based rda files:
          hsahomologyACC2HGID found 18352
         hsahomologyDATA found 17211
```

```
hsahomologyHGID2ACC found 17114
hsahomologyHGID2LL found 17211
hsahomologyHGID found 17211
hsahomologyLL2HGID found 17247
hsahomologyORGCODE found 105
```

`XXXhomologyDATA` has HomoloGeneIDs as keys and lists of sub-lists containing data for other organisms that have been identified to have homology relationships as the corresponding values. The list for a given HomoloGene ID may have one or more sub-lists depending on whether homology relationships have been identified in only one or more other organisms.

Each sub-list has an element for the name of the organism (homoOrg), HomologGeneID (homoHGID), type of similarity (homoType. B - reciprocal best match between three or more organisms, b - reciprocal best match between two organisms, and c - curated homology relationship between two organisms), percent of identity (homoPS) measured as the percentage of base pair alignment between the matching sequences, and a url (homoURL) to the source if the relationship is a curated orthology. A sub-list with homoType = B or b will not have any value for homoURL and a sublist with homoType = c will not have any value for homoPS.

The `XXXhomologyLL2HGID` environment contains mappings between LocusLink identifiers and HGIDs (HomoloGene IDs used by HomoloGene to represent sequences represented by the LocusLink ids). This environment allows users to map LocusLink ids to HGIDs and then use the obtained HGIDs to locate homologous genes in other organisms using the environment that contains data for homologous genes found in other organisms using (XXXhomologyDATA).

0.1 Use *XXXhomology* to explore homologous relationships among organisms

0.1.1 Task 1

Given LocusLink id 25 in human (Homo sapiens), how would one find all other species containing homologous genes of best matches with percent similarity values above 80.00?

First we call a function `LL2homology` in *annotate* to obtain data for all the homologous genes of LocusLink id 25 in human.

```
> homoGenes <- LL2homology("hsahomology", "25")
```

`homoGenes` is a list of sub-lists with elements shown below:

```
> names(homoGenes[[1]][[1]])
```

```
[1] "homoOrg"  "homoType" "homoHGID" "homoPS"   "homoURL"
```

We are only interested in genes that are best best matches (homoType = B) to LocusLink id 25 with percent identity values greater than 80.00. The

following code finds and prints the names of the organisms and LocusLink IDs and HGIDs for genes that satisfy these conditions:

```
> goodG <- sapply(homoGenes[[1]], function(x) {
+     (x[["homoType"]] == "B" && x[["homoPS"]] > 80)
+ })
> geneList <- homoGenes[[1]][goodG]
> sapply(geneList, function(x) x[["homoOrg"]])

10090 10116  8364
"mmu" "rno" "xtr"

> sapply(geneList, function(x) x[["homoHGID"]])

 10090  10116   8364
387537 637464 734077
```

0.1.2 Task 2

Find all the genes in `Xenopus laevis` that are homologous to genes of `Danio rerio` with percent identity values greater than 90.00.

The organism code for `Danio rerio` is:

```
> library("xlahomology")
> subset(xlahomologyORGCODE, species_name == "Danio rerio")

   species_name tax_id tla
18  Danio rerio   7955 dre
```

The object containing homology data for `Xenopus laevis` is `xlahomology-HGID` in *xlahomology*, which is a vector of HGIDs. Genes in `Xenopus laevis` that satisfy the conditions can be obtained using the following code chunk.

```
> temp <- mget(xlahomologyHGID, xlahomologyDATA)
> tempFun <- function(x) {
+     for (i in x) {
+         if (!is.na(i[["homoOrg"]]) && i[["homoOrg"]] ==
+             "dre" && i[["homoPS"]] > 90) {
+             return(i)
+         }
+     }
+     return(NA)
+ }
> goodGenes <- sapply(temp, tempFun)
> goodGenes <- goodGenes[!is.na(goodGenes)]
```

`goodGenes` obtained above is a list of sub-lists. The names of the list are the HGIDs for genes in `Xenopus laevis` and the corresponding (homoData)

objects contain information about the homologous genes in "Danio rerio". The following code gets the HGIDs and percent similarity of these homologous genes:

```
> hgids <- unlist(sapply(goodGenes, function(x) x[["homoHGID"]]))
> ps <- unlist(sapply(goodGenes, function(x) x[["homoPS"]]))
```

1 Session Information

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

- R version 2.7.0 (2008-04-22), x86_64-unknown-linux-gnu
- Locale: LC_CTYPE=en_US;LC_NUMERIC=C;LC_TIME=en_US;LC_COLLATE=en_US;LC_MONETARY=C;LC_MESS
- Base packages: base, datasets, graphics, grDevices, grid, methods, stats, tools, utils
- Other packages: annotate 1.18.0, AnnotationDbi 1.2.0, Biobase 2.0.0, DBI 0.2-4, GO.db 2.2.0, graph 1.18.0, hgu95av2.db 2.2.0, hsahomology 2.0.2, Rgraphviz 1.18.0, RSQLite 0.6-8, xlahomology 2.0.2, XML 1.93-2, xtable 1.5-2
- Loaded via a namespace (and not attached): cluster 1.11.10