Primer: CMAPCollections from Bioconductor annotation packages

February 8, 2013

1 reactome

The reactome.db package offers access to pathway annotations from the reactome database http://www.reactome.org/. This primer demonstrates how to use this annotaiton to generate a species-specific CMAP-Collection with Entrez gene identifiers.

First, we access the genes identifiers associated with each pathway. In a second step, we retrieve the names of the pathways and perform some basic filtering to remove duplicated or un-named sets.

```
> library(reactome.db)
> library(gCMAP)
> library(Matrix)
> ## use multiple cores if available
> options(mc.cores=2)
> ## retrieve entrez ids of pathway members
> pathways <- as.list(reactomePATHID2EXTID)
> ## retrieve names
> pathway.names <- unlist(mget(names(pathways), reactomePATHID2NAME))
> pathway.names <- pathway.names[ match(names( pathways),
+                                        names( pathway.names )) ]
> ## remove categories with duplicated or missing names
> filtered.names <- duplicated( names( pathway.names)) | is.na(pathway.names)
> pathways <- pathways[ ! filtered.names ]
> pathway.names <- pathway.names[ ! filtered.names]
```

Each pathway name contains the name of the respective species. We can use this information to generate species-specific reactome collections:

```
> head( pathway.names )

                                                        70326
                            "Homo sapiens: Glucose metabolism"
                                                        70221
           "Homo sapiens: Glycogen breakdown (glycogenolysis)"
                                                      1430728
                                      "Homo sapiens: Metabolism"
                                                        71387
                   "Homo sapiens: Metabolism of carbohydrates"
                                                      2496304
                "Saccharomyces cerevisiae: Glucose metabolism"
                                                      2496314
"Saccharomyces cerevisiae: Glycogen breakdown (glycogenolysis)"
```

```
> human <- grepl( "^Homo sapiens", pathway.names)
```

Next, we create new CMAPCollection, providing pathway names in the phenoData slot.

```
> pheno.data <- as(
+     data.frame(name=pathway.names[ human ],
+                row.names=names(pathways[ human ])
+               ),
+               "AnnotatedDataFrame")
> i.matrix <- Matrix::t( incidence( pathways[ human ] ) )
> reactome.hs <- CMAPCollection( i.matrix,
+                               phenoData=pheno.data,
+                               annotation='org.Hs.eg',
+                               signed=rep( FALSE, ncol(i.matrix)) )
```

To simplify this process in the future, we can define a helper function.

```
> pathway2cmap <- function(pathways, pathway.names, selected, anno){
+   pheno.data <- as(
+                   data.frame(name=pathway.names[ selected ],
+                              row.names=names(pathways[ selected ])
+                              ),
+                   "AnnotatedDataFrame")
+   i.matrix <- Matrix::t( incidence(pathways[ selected ]) )
+   CMAPCollection(i.matrix,
+                 phenoData=pheno.data,
+                 annotation=anno,
+                 signed=rep(FALSE, ncol(i.matrix)))
+ }
```

Now, generating `CMAPCollections` for other species is straightforward:

```
> mouse <- grepl( "^Mus musculus", pathway.names)
> reactome.mm <- pathway2cmap( pathways, pathway.names,
+                              selected=mouse, "org.Mm.eg")
```

2 KEGG

Similarly, the `KEGG.db` package offers the last public release of the KEGG gene annotation database `http://www.genome.jp/kegg/`. Analogous to the `reactome.db` package, species are identified by specific prefixes in the pathway identfiers. For example, human gene sets start with 'hsa', mouse sets begin with 'mmu' instead.'

```
> library(KEGG.db)
> ## retrieve entrez ids of pathway members
> pathways <- as.list(KEGGPATHID2EXTID)
> ## retrieve names
> pathway.names <- unlist(mget(sub("^...", "",names(pathways)), KEGGPATHID2NAME))
> ## species-specific CMAPCollections
> human <- grepl( "^hsa", names( pathways ))
> KEGG.hs <- pathway2cmap( pathways, pathway.names, selected=human, "org.Hs.eg")
> mouse <-  grepl( "^mmu", names( pathways ))
> KEGG.mm <- pathway2cmap( pathways, pathway.names, selected=mouse, "org.Mm.eg")
>  sessionInfo()

R version 2.15.2 (2012-10-26)
Platform: x86_64-unknown-linux-gnu (64-bit)

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=C                 LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

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

other attached packages:
 [1] KEGG.db_2.8.0        Matrix_1.0-10        reactome.db_1.42.0
 [4] RSQLite_0.11.2       DBI_0.2-5            reshape_0.8.4
 [7] plyr_1.8             gCMAP_1.1.7          DESeq_1.10.1
[10] lattice_0.20-13      locfit_1.5-8         GSEABase_1.20.2
[13] graph_1.36.2         annotate_1.36.0      AnnotationDbi_1.20.3
[16] Biobase_2.18.0       BiocGenerics_0.4.0

loaded via a namespace (and not attached):
 [1] GSEAlm_1.18.0         IRanges_1.16.4        RColorBrewer_1.0-5
 [4] XML_3.95-0.1          bigmemory_4.3.0       bigmemory.sri_0.1.2
 [7] bigmemoryExtras_1.0.0 genefilter_1.40.0     geneplotter_1.36.0
[10] grid_2.15.2           latticeExtra_0.6-24   limma_3.14.4
[13] parallel_2.15.2       splines_2.15.2        stats4_2.15.2
[16] survival_2.37-2       tools_2.15.2          xtable_1.7-0
```