keggorth: the KEGG orthology as `graph`

VJ Carey

October 28, 2009

Contents

1 Introduction

KEGG is the Kyoto Encyclopedia of Genes and Genomes. An important product of the KEGG group is a catalog of pathways. The KEGG Orthology (KO) organizes the pathways into a conceptual hierarchy. This package encodes the hierarchy as a graph, and provides some support for deriving sets of array feature identifiers from the hierarchy.

2 KOgraph

```
> library(keggorth)
> library(graph)
> data(KOgraph)
> KOgraph

A graphNEL graph with directed edges
Number of Nodes = 283
Number of Edges = 282

> nodes(KOgraph)[1:5]
```

```
[1] "KO.June07root"                "Metabolism"
[3] "Carbohydrate Metabolism"      "Glycolysis / Gluconeogenesis"
[5] "Citrate cycle (TCA cycle)"
```

The upper component of the hierarchy is:

```
> adj(KOgraph, nodes(KOgraph)[1])

$KO.June07root
[1] "Metabolism"
[2] "Genetic Information Processing"
[3] "Environmental Information Processing"
[4] "Cellular Processes"
[5] "Human Diseases"
```

Graph operations can be used to explore the orthology. For example, the context of the PPAR signaling pathway is found as follows:

```
> library(RBGL)
> sp.between(KOgraph, nodes(KOgraph)[1], "PPAR signaling pathway")

$`KO.June07root:PPAR signaling pathway`
$`KO.June07root:PPAR signaling pathway`$length
[1] 3

$`KO.June07root:PPAR signaling pathway`$path_detail
[1] "KO.June07root"          "Cellular Processes"     "Endocrine System"
[4] "PPAR signaling pathway"

$`KO.June07root:PPAR signaling pathway`$length_detail
$`KO.June07root:PPAR signaling pathway`$length_detail[[1]]
       KO.June07root->Cellular Processes
                                       1
    Cellular Processes->Endocrine System
                                       1
Endocrine System->PPAR signaling pathway
                                       1
```

Fixed-length identifiers are used to label pathways. These are available as the 'tag' nodeData attribute.

```
> nodeData(KOgraph, , "tag")[1:5]
```

```
$KO.June07root
[1] "NONE"

$Metabolism
[1] "01100"

$`Carbohydrate Metabolism`
[1] "01110"

$`Glycolysis / Gluconeogenesis`
[1] "00010"

$`Citrate cycle (TCA cycle)`
[1] "00020"
```

The depth of each term is also available.

```
> nodeData(KOgraph, , "depth")[1:5]

$KO.June07root
[1] 0

$Metabolism
[1] 1

$`Carbohydrate Metabolism`
[1] 2

$`Glycolysis / Gluconeogenesis`
[1] 3

$`Citrate cycle (TCA cycle)`
[1] 3
```

3 Application to gene filtering

Several functions are available for retrieving relevant information from the orthology. If you know a substring of the pathway name of interest, you can obtain the numerical tag(s).

```
> getKOtags("insulin")

Insulin signaling pathway
                  "04910"
```

We can get probe set identifiers corresponding to a term. The default chip annotation package used is hgu95av2.db.

```
> library(hgu95av2.db)
> mp = getKOprobes("Methionine")
> library(ALL)
> data(ALL)
> ALL[mp, ]

ExpressionSet (storageMode: lockedEnvironment)
assayData: 0 features, 128 samples
  element names: exprs
phenoData
  sampleNames: 01005, 01010, ..., LAL4  (128 total)
  varLabels and varMetadata description:
    cod:  Patient ID
    diagnosis:  Date of diagnosis
    ...: ...
    date last seen:  date patient was last seen
    (21 total)
featureData
  featureNames:
  fvarLabels and fvarMetadata description: none
experimentData: use 'experimentData(object)'
  pubMedIds: 14684422 16243790
Annotation: hgu95av2
```

4 Infrastructure considerations

The graph was built through manual massaging of an HTML page in which the orthology was expanded to level 3. This operation needs to be programmatic, and the version number needs to be made clear.

The HTML source from which this was built is in inst/keggHTML in the source package.

5 Session info

```
> sessionInfo()

R version 2.10.0 (2009-10-26)
x86_64-unknown-linux-gnu
```

```
locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=C              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

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

other attached packages:
 [1] ALL_1.4.7           RBGL_1.22.0         keggorth_1.8.0
 [4] hgu95av2.db_2.3.5   org.Hs.eg.db_2.3.6  RSQLite_0.7-3
 [7] DBI_0.2-4           AnnotationDbi_1.8.0 Biobase_2.6.0
[10] graph_1.24.0
```