

keggorthology: the KEGG orthology as graph

VJ Carey

April 10, 2013

Contents

| | | |
|---|-------------------------------|---|
| 1 | Introduction | 1 |
| 2 | KOgraph | 1 |
| 3 | Application to gene filtering | 3 |
| 4 | Infrastructure considerations | 4 |
| 5 | Session info | 4 |

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

A graphNEL graph with directed edges

Number of Nodes = 358

Number of Edges = 357

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

```
[1] "KO.Feb10root"           "Metabolism"
[3] "Carbohydrate Metabolism" "Glycolysis / Gluconeogenesis"
[5] "Citrate cycle (TCA cycle)"
```

The upper component of the hierarchy is:

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

$KO.Feb10root
[1] "Metabolism"
[2] "Genetic Information Processing"
[3] "Environmental Information Processing"
[4] "Cellular Processes"
[5] "Organismal Systems"
[6] "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.Feb10root:PPAR signaling pathway`
$`KO.Feb10root:PPAR signaling pathway`$length
[1] 3

$`KO.Feb10root:PPAR signaling pathway`$path_detail
[1] "KO.Feb10root"           "Organismal Systems"      "Endocrine System"
[4] "PPAR signaling pathway"

$`KO.Feb10root:PPAR signaling pathway`$length_detail
$`KO.Feb10root:PPAR signaling pathway`$length_detail[[1]]
      KO.Feb10root->Organismal Systems
                        1
      Organismal Systems->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.Feb10root
```

```
[1] "NONE"
```

```
$Metabolism
```

```
[1] "01100"
```

```
$`Carbohydrate Metabolism`
```

```
[1] "01101"
```

```
$`Glycolysis / Gluconeogenesis`
```

```
[1] "00010"
```

```
$`Citrate cycle (TCA cycle)`
```

```
[1] "00020"
```

The depth of each term is also available.

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

```
$KO.Feb10root
```

```
[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 = getK0probes("Methionine")
> library(ALL)
> data(ALL)
> ALL[mp,]
```

```
ExpressionSet (storageMode: lockedEnvironment)
assayData: 32 features, 128 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: 01005 01010 ... LAL4 (128 total)
  varLabels: cod diagnosis ... date last seen (21 total)
  varMetadata: labelDescription
featureData: none
experimentData: use 'experimentData(object)'
  pubMedIds: 14684422 16243790
Annotation: hgu95av2
```

4 Infrastructure considerations

Based on keggorthology read of KEGG orthology, March 2 2010. Specifically, we run wget on ftp://ftp.genome.jp/pub/kegg/brite/ko/ko00001.keg and use parsing and modeling code given in inst/keggHTML to generate a data frame respecting the hierarchy, and then keggDF2graph function in keggorthology package to construct the graph.

5 Session info

```
> sessionInfo()
```

```
R version 3.0.0 (2013-04-03)
Platform: x86_64-apple-darwin10.8.0 (64-bit)
```

```
locale:
[1] C
```

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

other attached packages:

```
[1] ALL_1.4.14          RBGL_1.36.0          keggorthology_2.12.0
[4] hgu95av2.db_2.9.0   org.Hs.eg.db_2.9.0   RSQLite_0.11.2
[7] DBI_0.2-5           AnnotationDbi_1.22.1 Biobase_2.20.0
[10] BiocGenerics_0.6.0  graph_1.38.0
```

loaded via a namespace (and not attached):

```
[1] IRanges_1.18.0 stats4_3.0.0  tools_3.0.0
```