keggorthology: the KEGG orthology as graph

VJ Carey

October 17, 2016

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]
The upper component of the hierarchy is:

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

```r
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:

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

```r
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.

```r
> nodeData(KOgraph,"tag")[1:5]
```
$KO.Feb10root
[1] "NONE"

$Metabolism
[1] "01100"

$\text{"Carbohydrate Metabolism"}
[1] "01101"

$\text{"Glycolysis / Gluconeogenesis"}
[1] "00010"

$\text{"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

$\text{"Carbohydrate Metabolism"}
[1] 2

$\text{"Glycolysis / Gluconeogenesis"
[1] 3

$\text{"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.

```r
> library(hgu95av2.db)
> mp = getKOprobes("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](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

```r
> sessionInfo()
```

R version 3.3.1 (2016-06-21)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 16.04.1 LTS

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=en_US.UTF-8     LC_NAME=C
attached base packages:
[1] stats4  parallel  stats  graphics  grDevices  utils  datasets
[8] methods  base

other attached packages:
[1] ALL_1.15.0  RBGL_1.50.0  keggorthology_2.26.0
[4] hgu95av2.db_3.2.3  org.Hs.eg.db_3.4.0  AnnotationDbi_1.36.0
[7] IRanges_2.8.0  S4Vectors_0.12.0  Biobase_2.34.0
[10] graph_1.52.0  BiocGenerics_0.20.0

loaded via a namespace (and not attached):
[1] DBI_0.5-1  tools_3.3.1  RSQLite_1.0.0