

# *keggorthology*: the KEGG orthology as graph

VJ Carey

April 26, 2022

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>KOgraph</b>	<b>1</b>
<b>3</b>	<b>Application to gene filtering</b>	<b>3</b>
<b>4</b>	<b>Infrastructure considerations</b>	<b>4</b>
<b>5</b>	<b>Session info</b>	<b>4</b>

## 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: 30 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 4.2.0 RC (2022-04-21 r82226)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 20.04.4 LTS
```

```
Matrix products: default
BLAS: /home/biocbuild/bbs-3.16-bioc/R/lib/libRblas.so
LAPACK: /home/biocbuild/bbs-3.16-bioc/R/lib/libRlapack.so
```

```
locale:
```

```
[1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C
[3] LC_TIME=en_GB             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
[9] LC_ADDRESS=C              LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
```

attached base packages:

```
[1] stats4      stats      graphics  grDevices  utils      datasets  methods
[8] base
```

other attached packages:

```
[1] ALL_1.37.0          RBGL_1.73.0          keggorthology_2.49.0
[4] hgu95av2.db_3.13.0  org.Hs.eg.db_3.15.0  AnnotationDbi_1.59.0
[7] IRanges_2.31.0     S4Vectors_0.35.0    Biobase_2.57.0
[10] graph_1.75.0        BiocGenerics_0.43.0
```

loaded via a namespace (and not attached):

```
[1] Rcpp_1.0.8.3        XVector_0.37.0       zlibbioc_1.43.0
[4] bit_4.0.4           R6_2.5.1             rlang_1.0.2
[7] fastmap_1.1.0       blob_1.2.3           httr_1.4.2
[10] GenomeInfoDb_1.33.0 tools_4.2.0          png_0.1-7
[13] cli_3.3.0           DBI_1.1.2            bit64_4.0.5
[16] crayon_1.5.1        GenomeInfoDbData_1.2.8 bitops_1.0-7
[19] vctrs_0.4.1         KEGGREST_1.37.0     RCurl_1.98-1.6
[22] memoise_2.0.1       cachem_1.0.6         RSQLite_2.2.12
[25] compiler_4.2.0     Biostrings_2.65.0   pkgconfig_2.0.3
```