

# *keggorthology*: the KEGG orthology as **graph**

VJ Carey

May 19, 2021

## 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: 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 4.1.0 RC (2021-05-10 r80283)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows Server x64 (build 17763)
```

```
Matrix products: default
```

```
locale:
[1] LC_COLLATE=C
[2] LC_CTYPE=English_United States.1252
```

[3] LC\_MONETARY=English\_United States.1252

[4] LC\_NUMERIC=C

[5] LC\_TIME=English\_United States.1252

attached base packages:

[1] stats4 parallel stats graphics grDevices utils datasets

[8] methods base

other attached packages:

[1] ALL\_1.33.0 RBGL\_1.68.0 keggorthology\_2.44.0

[4] hgu95av2.db\_3.2.3 org.Hs.eg.db\_3.13.0 AnnotationDbi\_1.54.0

[7] IRanges\_2.26.0 S4Vectors\_0.30.0 Biobase\_2.52.0

[10] graph\_1.70.0 BiocGenerics\_0.38.0

loaded via a namespace (and not attached):

[1] Rcpp\_1.0.6 rstudioapi\_0.13 XVector\_0.32.0

[4] zlibbioc\_1.38.0 bit\_4.0.4 R6\_2.5.0

[7] rlang\_0.4.11 fastmap\_1.1.0 blob\_1.2.1

[10] httr\_1.4.2 GenomeInfoDb\_1.28.0 tools\_4.1.0

[13] png\_0.1-7 DBI\_1.1.1 bit64\_4.0.5

[16] crayon\_1.4.1 GenomeInfoDbData\_1.2.6 bitops\_1.0-7

[19] vctrs\_0.3.8 KEGGREST\_1.32.0 RCurl\_1.98-1.3

[22] memoise\_2.0.0 cachem\_1.0.5 RSQLite\_2.2.7

[25] compiler\_4.1.0 Biostrings\_2.60.0 pkgconfig\_2.0.3