

# *keggorthology*: the KEGG orthology as graph

VJ Carey

October 29, 2024

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>K0graph</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 K0graph

```
> library(keggorthology)
> library(graph)
> data(K0graph)
> K0graph
```

```
A graphNEL graph with directed edges
Number of Nodes = 358
Number of Edges = 357
```

```
> nodes(K0graph)[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 4.4.1 (2024-06-14)
Platform: x86_64-apple-darwin20
Running under: macOS Monterey 12.7.6
```

```
Matrix products: default
```

```
BLAS: /Library/Frameworks/R.framework/Versions/4.4-x86_64/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.4-x86_64/Resources/lib/libRlapack.dylib
```

```
locale:
```

[1] C/en\_US.UTF-8/en\_US.UTF-8/C/en\_US.UTF-8/en\_US.UTF-8

time zone: America/New\_York  
tzcode source: internal

attached base packages:

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

other attached packages:

[1] ALL\_1.47.0 RBGL\_1.82.0 keggorthology\_2.58.0  
[4] graph\_1.84.0 hgu95av2.db\_3.13.0 org.Hs.eg.db\_3.20.0  
[7] AnnotationDbi\_1.68.0 IRanges\_2.40.0 S4Vectors\_0.44.0  
[10] Biobase\_2.66.0 BiocGenerics\_0.52.0

loaded via a namespace (and not attached):

[1] crayon\_1.5.3 vctrs\_0.6.5 httr\_1.4.7  
[4] cli\_3.6.3 rlang\_1.1.4 DBI\_1.2.3  
[7] png\_0.1-8 UCSC.utils\_1.2.0 jsonlite\_1.8.9  
[10] bit\_4.5.0 Biostrings\_2.74.0 KEGGREST\_1.46.0  
[13] fastmap\_1.2.0 GenomeInfoDb\_1.42.0 memoise\_2.0.1  
[16] compiler\_4.4.1 RSQLite\_2.3.7 blob\_1.2.4  
[19] pkgconfig\_2.0.3 XVector\_0.46.0 R6\_2.5.1  
[22] GenomeInfoDbData\_1.2.13 tools\_4.4.1 bit64\_4.5.2  
[25] zlibbioc\_1.52.0 cachem\_1.1.0