KEGGgraph: Application Examples

Jitao David Zhang

July 25, 2017

Abstract

In this vignette, we demonstrate the application of *KEGGgraph* as flexible module in analysis pipelines targeting heterogenous biological questions. For basic use of the *KEGGgraph* package, please refer to the vignette *KEGGgraph: a graph approach to KEGG PATHWAY in R and Bioconductor*.

1 Introduction

In many cases, *KEGGgraph* is used as stand-alone software package to parse KEGG pathway [Kanehisa *et al.*, 2008] into graph models and consequently to analyze them. However, *KEGGgraph* can also be combined with other tools (within or out of the scope of R and Bioconductor [Gentleman *et al.*, 2004, Carey *et al.*, 2005]), to build flexible analysis pipelines. In this vignette we demonstrate the cooperation between *KEGGgraph* and other tools with several examples.

2 Parse created or edited pathway

KGML-ED [Klukas and Schreiber, 2007] is a tool designed for the dynamic visualization, interactive navigation and editing of KEGG pathway diagrams. The program is able to modify or even create *de novo* new pathways and export them into KGML (KEGG XML) files. Since *KEGGgraph* captures the graph information of KEGG PATHWAY by parsing KGML files, it is also capable to parse the KGML files exported from KGML-ED program. The joint use of the two programs allows user edit KGML files from KEGG (for example, by adding/removing new nodes/edges), or create new pathways from scratch, and then analyse them.

Here we demontrate the parse of a toy pathway (network motif) created by KGML-ED, which illustrates a feed-forward loop (FFL) [Mangan and Alon, 2003, Alon, 2007].

```
> library(KEGGgraph)
> toyKGML <- system.file("extdata/kgml-ed-toy.xml", package="KEGGgraph")
```

```
> toyGraph <- parseKGML2Graph(toyKGML, genesOnly=FALSE)
> toyGraph

A graphNEL graph with directed edges
Number of Nodes = 5
Number of Edges = 6

> nodes(toyGraph)

[1] "CRP"       "GalS"      "galETK"    "cAMP"      "galactose"
```

We visualize the graph with *Rgraphviz* in Figure 1.

Similarly, users could create new or modify existing pathways by modifying KGML files with tools including KGML-ED, and these pathways can be consequently parsed, visualized and analyzed consequently by *KEGGgraph* and other tools for graphs implemented in Bioconductor. In this sense *KEGGgraph* is not only able to study existing biological pathways constructed by KEGG, but also a tool for modelling pathways and biological networks in R.

3 Microarray analysis

KEGG PATHWAY contains pathway maps for nearly 1000 organisms and has been intensively expanded with pathways including signaling transduction, cellular processes and human diseases in recent years. As the graph-theory based approach to this valuable knowledge-base, *KEGGgraph* can be built into analysis pipelines where the pathway information could be useful, especially with the graph attributes. To demonstrate this we show one example of microarray analysis, cooperating with another package SPIA [Tarca *et al.*, 2009], a package implementing the Signaling Pathway Impact Analysis (SPIA) algorithm.

The microarray data is decribed in the vignette of *SPIA*, here it is only briefly summarized: Affemetrix GeneChip (GEO accession: GSE4107), containing 10 normal samples and 12 colorectal cancer samples, is normalized with *affy* and *limma* package. The result of `topTable` in *limma* is used as the input data.

```
> if(require(SPIA)) {
+   data(colorectalcancer,package="SPIA")
+ } else {
+   data(colorectalcancerSPIA, package="KEGGgraph")
+ }
> library(SPIA)
> data(colorectalcancer)
> head(top)
```

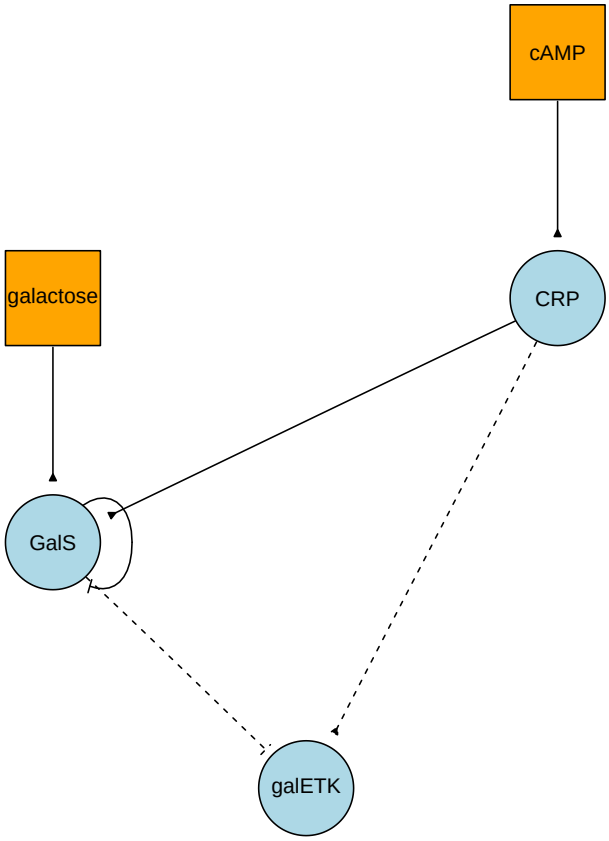

Figure 1: Visualization of the toy network (incoherent type-1 feed-forward loop, IL1-FFL) created by *KGML-ED* and parsed by *KEGGgraph*. The rectangle nodes represent compounds while ellipse ones represent gene/gene products. Solid and dashed red arrows represent activation and expression, whereas blue represent inhibition and repression. Note that self-inhibition of GalS is presented by a node and an edge pointing to its self.

```
              ID    logFC  AveExpr        t      P.Value    adj.P.Val        B
10738   201289_at 5.960206 6.226928 23.94388 1.789221e-17 9.782565e-13 25.40124
18604   209189_at 5.143502 7.487305 17.42995 1.560212e-14 2.843486e-10 21.02120
11143 201694_s_at 4.148081 7.038281 16.46040 5.153117e-14 7.043667e-10 20.14963
10490 201041_s_at 2.429889 9.594413 14.06891 1.293706e-12 1.414667e-08 17.66883
10913 201464_x_at 1.531126 8.221044 10.98634 1.685519e-10 1.151947e-06 13.61189
11463   202014_at 1.429269 5.327647 10.45906 4.274251e-10 2.418771e-06 12.80131
      ENTREZ
10738   3491
18604   2353
11143   1958
10490   1843
10913   3725
11463  23645
```

We only consider the probes annotated with EntrezGeneID and differentially expressed with FDR p-value less than 0.05.

```
> library(hgu133plus2.db)
> x <- hgu133plus2ENTREZID
> top$ENTREZ <- unlist(as.list(x[top$ID]))
> top <- top[!is.na(top$ENTREZ),]
> top <- top[!duplicated(top$ENTREZ),]
> tg1 <- top[top$adj.P.Val < 0.05,]
> DE_Colorectal <- tg1$logFC
> names(DE_Colorectal) <- as.vector(tg1$ENTREZ)
> ALL_Colorectal <- top$ENTREZ
```

The *SPIA* algorithm takes as input two vectors: log2 fold changes of differentially expressed genes (`DE_Colorectal`) and those of all EntrezGeneID annotated genes (`ALL_Colorectal`), and returns a table of pathways ranked from the most to the least significant. The results show that Colorectal cancer pathway (ID:05210) is significantly activated (ranked the 5th of the result table). To visualize the results, we visualize the pathway with differentially expressed genes marked with color.

For the first time we could use the `retrieveKGML` function to retrieve the KGML remotely from the KEGG FTP site.

```
> tmp <- "hsa05210.xml"
> retrieveKGML(pathwayid="05210", organism="hsa", destfile=tmp)
```

We have attached this file in the *KEGGgraph*, we parse it into graph and observe that around 30% percent of the genes in the pathway are differentially expressed. The path-

way map (figure) can be found at `http://www.genome.jp/dbget-bin/show_pathway?hsa05210`.

```
> colFile <- system.file("extdata/hsa05210.xml",
+                        package="KEGGgraph")
> g <- parseKGML2Graph(colFile)
> deKID <- translateGeneID2KEGGID(names(DE_Colorectal))
> allKID <- translateGeneID2KEGGID(ALL_Colorectal)
> isDiffExp <- nodes(g) %in% deKID
> sprintf("%2.2f%% genes differentially-expressed", mean(isDiffExp)*100)
```

```
[1] "29.76% genes differentially-expressed"
```

We visualize the pathway with pseudo-colors representing the log2 fold change of the differentially expressed genes in figure 2.

In the figure 2, we observe that:

- Not all the genes are connected with each other, 40.4761904761905% (34/84) nodes in the pathway have a degree of 0 – they are not conencted to any other node. There are several reasons for this. One of them is the genes with genetic alterations are indicated in the pathway map, but not interweaved into the pathway (cf. `http://www.genome.jp/dbget-bin/show_pathway?hsa05210`). Another important factor is that in some pathways, especially human disease pathways, KEGG records genes involved in the disease but does not provide any edge in that pathway, although one can find their interaction partners in other pathways (e.g., Grb2 has no edges in colorectal cancer pathway, however in other maps including MAPK signaling pathway, both up- and down-stream interactors are found). One solution to the later problem is to *merge* (*union*) the related pathways together, to this end *KEGGgraph* provides the function `mergeKEGGgraphs`.

- From the visualization in 2, it is difficult to recognize any patterns. To examine the pathway in more details, it is necessary to use *Divide and Conquer* strategy, namely to *subset* the graph into subgraphs. To this end *KEGGgraph* provides the function `subKEGGgraph`, which maintains the KEGG information while subsetting the graph.

Out of 34 nonconnected genes, 7 is up-regulated and 4 downregulated. Notably, out of 10 *Frizzled* homologues (Wnt receptors), four are up-regulated (*FZD2, FZD4, FZD7 and FZD10*), and only one is down-regulated(*FZD5*), in accordance with the former report [Vincan *et al.*, 2007].

Next we concentrate on upregulated genes, the result given by *SPIA* package indicates that colorectal pathway is activated.

We can use degree centrality (the sum of in- and out-degree of each node) as a measure of the relative importance of the node compared to others. In this subgraph, 8 nodes have

an degree equal or larger than three (in decreasing order: AKT3, TCF7L1, CCND1, FOS, MAPK3, MAPK1, JUN, MAPK10), and 5 of them are upreguated to different extent. This reinforces the conclusion of *SPIA* that the pathway is activated - it seems that the relative important nodes are up-regulated. Especially we note the upregulation of Jun and Fos, two transcription factors with many targeting genes, although their up-stream

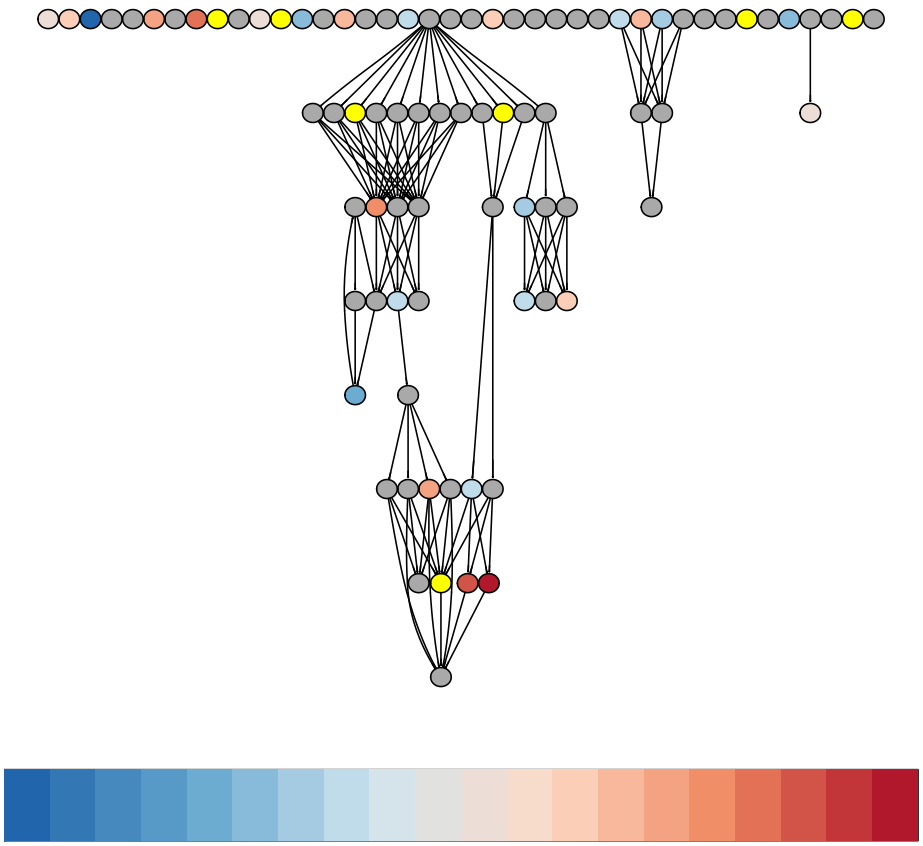

Figure 2: Overview of colorectal cancer pathway with pseudo-colored nodes representing expression change. Red nodes represented up-regulated mRNAs in cancer samples and blue ones are down-regulated. Grey nodes are detected not to be differentially expressed, yellow ones are not reported (possibly due to under detection threshold). The gradual colors indicate the log2 fold change of the expression - the darker the colors, the larger fold change. One distinct feature is the existence of many non-connected nodes, which could has been omitted when the graph information is discarded and only the membership of the gene is considered.

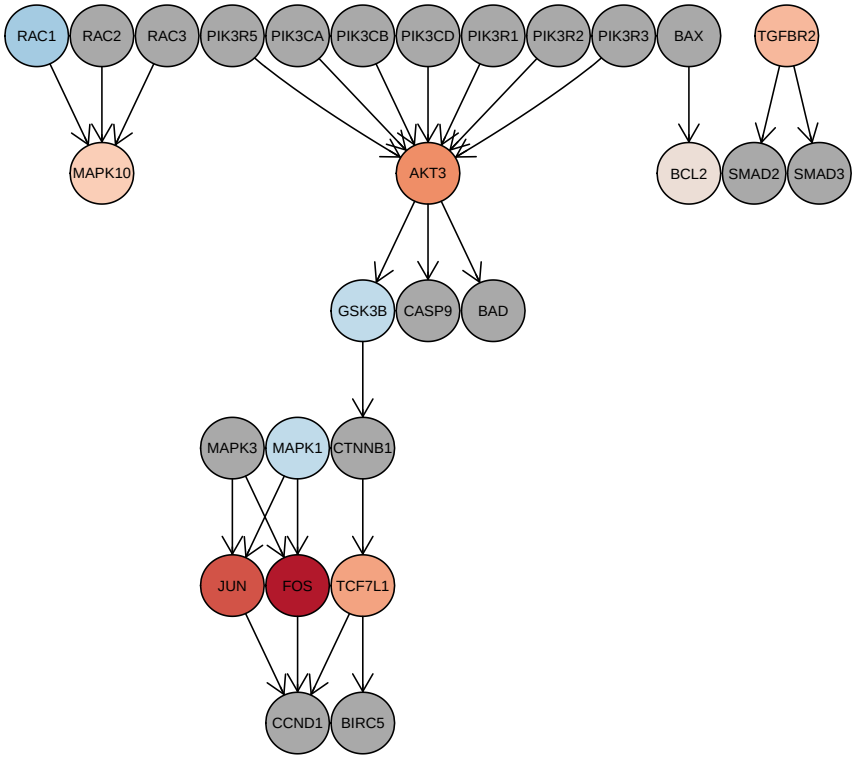

Figure 3: Up-regulated genes (seven, in red) and their neighborhood genes in colorectal cancer pathway, the color follows the legend in figure 2. Un-reported genes are hidden from the figure.

interactors (MAPK3 and MAPK1) are either not significantly differentially expressed or slightly down-regulated (MAPK1, logFC=-0.381).

Further analysis could done, for example, by merging the colorectal pathway with linked pathways (Wnt signaling, apoptosis, etc) and investigate the graph characteristics of differnetially expressed genes and their links.

References

[Gentleman *et al.*, 2004] Gentleman *et al.* (2004). Bioconductor: open software development for computational biology and bioinformatics, *Genome Biology*, **5**, R80.

[Carey *et al.*, 2005] Carey *et al.* (2005). Network structures and algorithms in Bioconductor, *Bioinformatics*, **21**, 135-136.

[Kanehisa *et al.*, 2008] Kanehisa *et al.* (2008). KEGG for linking genomes to life and the environment, *Nucleic Acids Research, Database issue*, **36**, 480-484.

[Klukas and Schreiber, 2007] Klukas and Schreiber. (2007). Dynamic exploration and editing of KEGG pathway diagrams, *Bioinformatics*, **23**, 344-350.

[Aittokallio and Schwikowski, 2006] Aittokallio and Schwikowski. (2006). Graph-based methods for analysing networks in cell biology, *Briefings in Bioinformatics*, **7**, 243-255.

[Mangan and Alon, 2003] Mangan and Alon. (2003) Structure and function of the feed-forward loop network motif, *Proc. Natl. Acad. Sci. U.S.A*, **100**, 11980-11985.

[Alon, 2007] Alon. (2007). An introduction to system biology: design principles of biological circuits, *Chapman and Hall*, 2007, 64.

[Tarca *et al.*, 2009] Tarca *et al.* (2009). A signaling pathway impact analysis for microarray experiments. *Bioinformatics*, **25**, 75-82.

[Vincan *et al.*, 2007] Vincan *et al.* (2007) Frizzled-7 dictates three-dimensional organization of colorectal cancer cell carcinoids, *Oncogene*, **26**, 2340-2352.