GRAPH Interaction from pathway Topological Environment

Gabriele Sales (gabriele.sales@unipd.it)
Enrica Calura (enrica.calura@gmail.com)
Chiara Romualdi (chiara.romualdi@unipd.it)

July 4, 2011

1 Introduction

graphite (GRAPH Interaction from pathway Topological Environment) an R package built to i) provide networks from four databases (KEGG, [1]; Biocarta, www.biocarta.com; Reactome, [2]; NCI/Nature Pathway Interaction Database, [3]); ii) discriminate between different types of gene groups; iii) propagate pathway connections through chemical compounds; iv) allow the selection of edge attributes and the conversion of nodes identifiers to EntrezGene IDs and HUGO Symbols; finally, v) to run *SPIA*, *DEGraph*, and *topologyGSA* analyses directly on *graphite* networks.

2 Pathways

In order to gather curated information about human pathways, we have collected data from the four public databases that have emerged as reference points for the systems biology community. Reactome (using the BioPax format), backed by the EBI, is one of the most complete repository; it is frequently updated and provides a semantically rich description of each pathway [2]. KEGG Pathways (KGML format) provides maps for both signaling and metabolic pathways [1]. Finally, we selected BioCarta (www.biocarta.com) and NCI (NCI/Nature Pathway Interaction Database [3]), both publishing their data using the BioPax format.

A pathway network can be retrieved using the name of the pathway,

```
> names(biocarta)[1:10]
```

```
 [1] "acetylation and deacetylation of rela in nucleus"
 [2] "actions of nitric oxide in the heart"
 [3] "activation of camp-dependent protein kinase pka"
 [4] "activation of csk by camp-dependent protein kinase inhibits signaling through the t cell receptor"
```

```
 [5] "activation of pkc through g-protein coupled receptors"
 [6] "adp-ribosylation factor"
 [7] "agrin in postsynaptic differentiation"
 [8] "ahr signal transduction pathway"
 [9] "akap95 role in mitosis and chromosome dynamics"
[10] "akt signaling pathway"
```

```
> p <- biocarta[["acetylation and deacetylation of rela in nucleus"]]
> p
```

```
"acetylation and deacetylation of rela in nucleus" pathway from BioCarta
Number of nodes     = 6
Number of edges     = 9
Type of identifiers = native
Retrieved on        = 2011-05-12
```

or its position in the list of pathways:

```
> p <- biocarta[[1]]
> p@title
```

```
[1] "acetylation and deacetylation of rela in nucleus"
```

In the network, nodes represent genes and edges functional relationships among them. Nodes can have heterogeneous IDs (see section 4 for more details):

```
> nodes(p)
```

```
[1] "EntrezGene:4792"           "EntrezGene:5970"
[3] "EntrezGene:8841"           "EnzymeConsortium:2.3.1.48"
[5] "p50_0-0"                   "ubiquitin"
```

Edges can be characterized by multiple functional relationships:

```
> edges(p)
```

```
                       src                       dest  direction
1           EntrezGene:4792                   p50_0-0 undirected
2           EntrezGene:5970           EntrezGene:4792 undirected
3           EntrezGene:5970 EnzymeConsortium:2.3.1.48 undirected
4           EntrezGene:5970                   p50_0-0 undirected
5           EntrezGene:8841           EntrezGene:5970   directed
6           EntrezGene:8841                   p50_0-0   directed
7 EnzymeConsortium:2.3.1.48           EntrezGene:8841   directed
8 EnzymeConsortium:2.3.1.48 EnzymeConsortium:2.3.1.48 undirected
9 EnzymeConsortium:2.3.1.48                   p50_0-0 undirected
                     type
```

```
1                  binding
2                  binding
3                  binding
4                  binding
5 catalysisOut(ACTIVATION)
6 catalysisOut(ACTIVATION)
7  catalysisIn(ACTIVATION)
8                  binding
9                  binding
```

This same steps can be used to access the Reactome, KEGG and NCI databases (through the `reactome`, `kegg` and `nci` lists, respectively).

3 Graph

The function `pathwayGraph` builds a `graphNEL` object from a pathway `p`:

```
> g <- pathwayGraph(p)
> g
```

```
A graphNEL graph with directed edges
Number of Nodes = 6
Number of Edges = 14
```

```
> edgeData(g)[1]
```

```
$`EntrezGene:4792|p50_0-0`
$`EntrezGene:4792|p50_0-0`$weight
[1] 1

$`EntrezGene:4792|p50_0-0`$edgeType
[1] "binding"
```

4 Identifiers

Gene annotations databases are widely used as public repositories of biological information. Our current knowledge on biological elements is spread out over a number of databases (such as: Entrez Gene , RefSeq, backed by the NCBI http://www.ncbi.nlm.nih.gov/, UniProt, ENSEMBL backed by the EBI http://www.ebi.ac.uk/ to name just a few), specialised on a subset of specific biological entities (for instance, UniProt focuses on proteins while Entrez Gene focuses on genes). Key identifiers (IDs) in the internal structure of each such database uniquely represent biological entities, thus

biological entities can be identified by heterogeneous IDs according to the selected database they refer to. Due to their different origins and specificity, switching from an ID to another is possible but not trivial: there could be either no correspondence between them or many-to-many relations. For detailed information about IDs, their structures and differences please consult those resources. For our purposes, we have chosen EntrezGene IDs and Gene Symbols because of their widespread use and simplicity. The function `converterIdentifiers` allows the user to map such variety of IDs to a single type. This mapping process, however, may lead to the loss of some nodes (not all identifiers may be recognized) and has an impact on the topology of the network (one ID may correspond to multiple IDs in another annotation or vice versa).

```
> pEntrez <- convertIdentifiers(p, "entrez")
> pEntrez
```

```
"acetylation and deacetylation of rela in nucleus" pathway from BioCarta
Number of nodes     = 8
Number of edges     = 27
Type of identifiers = Entrez Gene
Retrieved on        = 2011-05-12
```

```
> nodes(pEntrez)
```

```
[1] "4792" "5970" "8841" "1387" "2033" "2648" "8850" "9575"
```

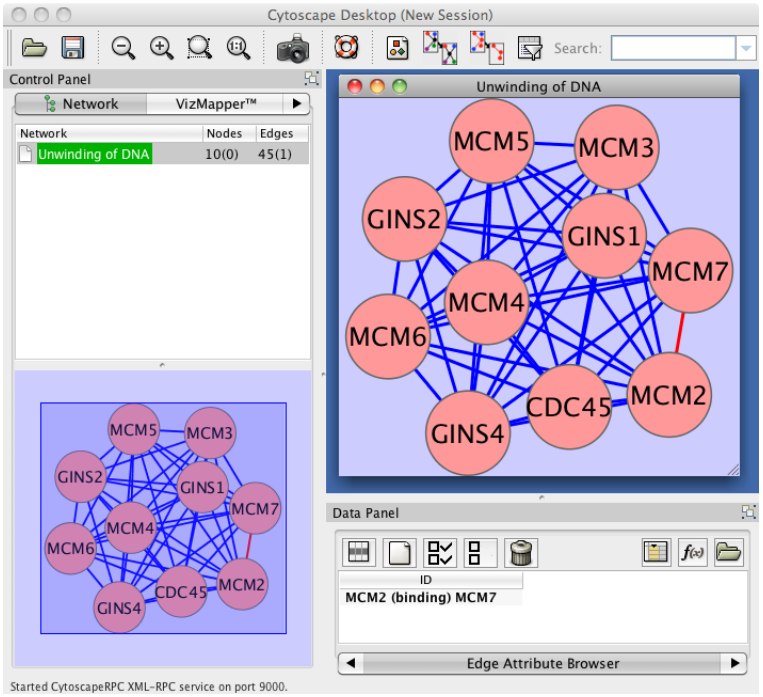
```
> pSymbol <- convertIdentifiers(p, "symbol")
> nodes(pSymbol)
```

```
[1] "NFKBIA" "RELA"   "HDAC3"  "CREBBP" "EP300"  "KAT2A"  "KAT2B"  "CLOCK"
```

5 Cytoscape Plot

Several pathways have a huge number of nodes and edges, thus there is the need of an efficient system of visualization. To this end *graphite* uses the *Rcytoscape* package to export the network to Cytoscape. Cytoscape is a Java based software specifically built to manage biological network complexity and for this reason it is widely used by the biological community.

```
> cytoscapePlot(convertIdentifiers(reactome$`Unwinding of DNA`, "symbol"))
```



6 Topological pathway analysis

graphite gives access to three types of topological pathway analyses recently proposed. For more details on the results obtained by these analyses see the corresponding R packages.

6.1 SPIA

The analysis with *SPIA* requires the conversion of the networks in a suitable format. This conversion is performed by the function `prepareSPIA` that must be executed before the analysis command `runSPIA`. The *SPIA* data will be saved in the current working directory; every time you change it, you should also re-run `prepareSPIA`.

```
> library(SPIA)
> data(colorectalcancer)
> 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
> prepareSPIA(biocarta[1:2], "biocartaEx")
> runSPIA(de=DE_Colorectal, all=ALL_Colorectal, "biocartaEx")

Done pathway 1 : acetylation and deacetylation ..
Done pathway 2 : actions of nitric oxide in the..                                                    Name pSize N
1             actions of nitric oxide in the heart    43  12 0.1424714
2 acetylation and deacetylation of rela in nucleus     7   3 0.1512768
          tA pPERT        pG     pGFdr    pGFWER    Status
1 -0.5368375 0.717 0.3351907 0.4174951 0.6703814 Inhibited
2 -0.2016161 0.933 0.4174951 0.4174951 0.8349901 Inhibited
```

For more details see the *SPIA* package [5, 6, 4].

6.2 DEGraph

DEGraph implements recent hypothesis testing methods which directly assess whether a particular gene network is differentially expressed between two conditions.

```
> library(DEGraph)

Scalable Robust Estimators with High Breakdown Point (version 1.3-00)

> data("Loi2008_DEGraphVignette")
> p <- convertIdentifiers(biocarta[["actions of nitric oxide in the heart"]], "entrez")
> res <- runDEGraph(p, exprLoi2008, classLoi2008)
> res$`1`

$p.value
                     T2 T2 (1 Fourier components)
               0.4801202                 0.4510231

$graph
A graphNEL graph with directed edges
Number of Nodes = 2
Number of Edges = 3

$k
[1] 1
```

For more details see the *DEGraph* package [7].

6.3 topologyGSA

topologyGSA uses graphical models to test the pathway components and to highlight those involved in its deregulation.

```
> library(topologyGSA)
> data(examples)
> p <- convertIdentifiers(kegg[["hsa04664"]], "symbol")
> runTopologyGSA(p, "var", exp1, exp2, 0.05)

$alpha.obs
[1] 0.007421451

$cli.moral
$cli.moral[[1]]
[1] "GRB2"

$cli.moral[[2]]
[1] "SYK"   "BTK"   "PLCG2"

$cli.moral[[3]]
[1] "SYK" "LYN" "BTK"


$check
[1] TRUE

$graph
A graphNEL graph with undirected edges
Number of Nodes = 5
Number of Edges = 5

$lambda.obs
[1] 26.02199

$lambda.theo
[1] 18.30704
```

For more details see the *topologyGSA* package [8].

References

[1] Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 1999 Jan 1;27(1):29-34.

[2] Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B, Kanapin A, Lewis S, Mahajan S, May B, Schmidt E, Vastrik I, Wu G, Birney E, Stein L, D'Eustachio P. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res. 2009 Jan;37(Database issue):D619-22. Epub 2008 Nov 3.

[3] Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH. PID: the Pathway Interaction Database. Nucleic Acids Res. 2009 Jan;37(Database issue):D674-9. Epub 2008 Oct 2.

[4] Draghici, S., Khatri, P., Tarca, A.L., Amin, K., Done, A., Voichita, C., Georgescu, C., Romero, R. A systems biology approach for pathway level analysis. Genome Research, 17, 2007.

[5] Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, Kim CJ, Kusanovic JP, Romero R. A novel signaling pathway impact analysis. Bioinformatics. 2009 Jan 1;25(1):75-82.

[6] Adi L. Tarca, Sorin Draghici, Purvesh Khatri, et. al, A Signaling Pathway Impact Analysis for Microarray Experiments, 2008, Bioinformatics, 2009, 25(1):75-82.

[7] L. Jacob, P. Neuvial, and S. Dudoit. Gains in power from structured two-sample tests of means on graphs. Technical Report arXiv:q-bio/1009.5173v1, arXiv, 2010.

[8] Massa MS, Chiogna M, Romualdi C. Gene set analysis exploiting the topology of a pathway. BMC System Biol. 2010 Sep 1;4:121.