graphite

Gabriele Sales**, Enrica Calura† and Chiara Romualdi‡

Department of Biology, University of Padua

*gabriele.sales@unipd.it †enrica.calura@unipd.it ‡chiara.romualdi@unipd.it

May 2, 2019

Contents

1 Introduction

graphite (GRAPH Interaction from pathway Topological Environment) was designed to:

- provide networks derived from eight public pathway databases,
- automate the conversion of node identifiers (e.g. from Entrez IDs to gene symbols),

- facilitate the execution of topological pathway analyses on the provided networks.

The pathway databases available in this version are:

- KEGG [1]
- Biocarta
- Reactome [2]
- NCI/Nature Pathway Interaction Database [3]
- HumanCyc [9]
- Panther [10]
- smpdb [12]
- PharmGKB [11]

All provided pathways are annotated in human and, since version 1.14, other 13 species are available.

graphite pathways are collected and pre-processed at every BioConductor release (roughly every 6 months). This guarantees the synchronization of the provided data with all other BioConductor annotation packages.

The topological pathway analyses directly supported are:

- *SPIA* [5, 6, 4]
- *topologyGSA* [7]
- *clipper* [8]

Since version 1.24, every *graphite* pathway can be accessed through three separate views:

gene-only network where metabolite nodes have been removed and edges have been propagated through them;

metabolite-only network where protein nodes have been removed and edges have been propagated through them;

mixed network containing both proteins and metabolites, useful for metabolomic and transcriptomic data integration.

2 Pathway topology conversion to gene network

In order to gather curated information about pathways, we have collected data from different public databases that have emerged as reference points for the systems biology community. The KEGG database has been in development by Kanehisa Laboratories since 1995, and is now a prominent knowledge base for integration and interpretation of large-scale molecular data sets generated by genome sequencing and other high-throughput experimental technologies. KEGG is the only pathway database not in `BioPax` format, as it stores the information using the KGML format. KEGG pathways (KGML format) provides maps for both signaling and metabolic pathways [1]. Reactome, 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]. BioCarta (www.biocarta.com) is a developer, supplier and distributor company of reagents and assays for biopharmaceutical and academic research. Through an "open source"

approach, this community-fed forum constantly integrates emerging proteomic information from the scientific community. It also catalogs and summarizes important resources providing information for over 120,000 genes. BioCarta pathway data in `BioPax` format are available through the NCI website. NCI (NCI/Nature Pathway Interaction Database [3]) is a highly-structured, curated collection of information about known biomolecular interactions and key cellular processes assembled into signaling pathways. This was a collaborative project between the NCI and Nature Publishing Group (NPG). Finally, Panther [10] is data are a comprehensive, curated database of pathways, protein families, trees, subfamilies and functions available at http://pantherdb.org backed by the University of Southern California. HumanCyc is part of the BioCyc database collection of pathways [9]. Since version 1.24, *graphite* cointains also SMPDB [12] and PharmGKB [11], important resources for metabolomic data analyses.

2.1 Pathway definition

graphite pathways are derived by conversion of `KGML` and `BioPax` data formats. KEGG database provides a separate `xml` file, one for each pathway. Thus, a pathway is defined by all the reactions defined within each file. For the other databases, we define a pathway for each element of type `pathway` in the `BioPax` document.

2.2 Nodes with multiple elements

Within a pathway, a node can correspond to multiple gene products. These nodes with multiple elements can be divided into protein complexes (proteins linked by protein-protein interactions) and the groups containing alternative members (genes generally with similar biochemical functions). When considering signal propagation these groups should be considered differently; the first (hereafter group AND) should be expanded into a clique (all proteins connected to the others), while the second (hereafter group OR) should be expanded without connection among the contained elements.

In the KGML format there are two ways of defining nodes with multiple elements: protein complexes (group AND defined by entry `type="group"`) and group with alternative members (group OR defined by entry `type="gene"`). In the BioPax format only one type of group is allowed: protein complexes (group AND) with type `complex`. However, it often happens that `protein` tag contains multiple `xref` pointing to alternative elements of the process (group OR).

2.3 Metabolite and protein-mediated interactions

Metabolites mediated interactions are interactions for which a metabolite acts as a bridge between two proteins. Since metabolites are not measured during gene expression analysis (using microarray or RNA-seq), their removal from the original network is useful. However, the trivial elimination of the metabolites, without signal propagation, will strongly bias the topology, interrupting the signals that pass through them. If element A is linked to metabolite c and metabolite c is linked to element B, thus elements A should be linked to elements B.

Within the KGML format, there are two different ways of describing a metabolite mediated interaction: i) direct interaction `type="PPrel"` (A interacts whit B through metabolite c) and ii) indirect one `type="PCrel"` (A interacts whit metabolite c and c interacts whit B).

Since proper signal propagation is crucial for topological gene set analysis we decided to include additional rules for the propagation reconstructing a connection between two genes connected through a series of metabolites. Not all metabolites are considered for the propagation because some of them, such as Hydrogen, H2O, ATP, ADP etc., are highly frequent in map descriptions and the signal propagation through them would lead to degenerate too long chains of metabolites. The metabolites not considered for propagation are not characteristic of a specific reaction but secondary substrates/products widely shared among different processes.

Since version 1.24, *graphite* can be used also for metabolomics pathway analyses. We applied the same procedure descibed above to provide network with metabolites only propagating the signal through proteins.

Finally, *graphite* includes pathways containing both proteins and metabolites (without edges propagation) allowing also integrated pathway analyses of gene expression and metabolomic data.

2.4 Edge attributes

graphite allows the user to see the single/multiple relation types that characterize an edge. The type of edges have been preserved to remain as close as possible to the original annotations. Some new types have been introduced due to the needs of the topological conversion. For instance a *Process(indirect)* is introduced when the edge is generated propagating the signal from a gene to another gene passing through metabolites, or from a metabolite to another metabolite passing through proteins.

2.5 Loading pathways

Human pathways are natively distribuited with the package. Since the version 1.14.0, non-human pathway data are also available and can be dowloaded automatically using the `pathways` function.

`pathways` requires the name of the specie of interest and the name of the pathway database name as follows:

```
> humanReactome <- pathways("hsapiens", "reactome")
> names(humanReactome)[1:10]

 [1] "Interleukin-6 signaling"
 [2] "Apoptosis"
 [3] "Hemostasis"
 [4] "Intrinsic Pathway for Apoptosis"
 [5] "PKB-mediated events"
 [6] "PI3K Cascade"
 [7] "MAPK3 (ERK1) activation"
 [8] "Translesion synthesis by REV1"
 [9] "Translesion synthesis by Y family DNA polymerases bypasses lesions on DNA template"
[10] "Recognition of DNA damage by PCNA-containing replication complex"

> p <- humanReactome["ABC-family proteins mediated transport"]]
> p
```

```
"ABC-family proteins mediated transport" pathway
Native ID       = R-HSA-382556
Database        = Reactome
Species         = hsapiens
Number of nodes = 123
Number of edges = 1862
Retrieved on    = 17-04-2019
URL             = http://reactome.org/PathwayBrowser/#/R-HSA-382556
```

A pathway database is a list of pathways. We can access a pathway through its name, as above, or through its position in the list, as follow:

```
> p <- humanReactome[[21]]
> pathwayTitle(p)

[1] "Activation of BAD and translocation to mitochondria"
```

In the pathway, nodes represent genes/proteins:

```
> head(nodes(p))

[1] "UNIPROT:P10415" "UNIPROT:P31749" "UNIPROT:P31751" "UNIPROT:P48454"
[5] "UNIPROT:P63098" "UNIPROT:Q92934"
```

Edges can be characterized by multiple functional relationships:

```
> head(edges(p))

  src_type    src dest_type   dest  direction
1  UNIPROT P10415   UNIPROT P55957 undirected
2  UNIPROT P31749   UNIPROT Q92934   directed
3  UNIPROT P31751   UNIPROT Q92934   directed
4  UNIPROT P48454   UNIPROT P27348   directed
5  UNIPROT P48454   UNIPROT P31946   directed
6  UNIPROT P48454   UNIPROT P31947   directed
                                          type
1                                      Binding
2 Control(Out: ACTIVATION of BiochemicalReaction)
3 Control(Out: ACTIVATION of BiochemicalReaction)
4 Control(Out: ACTIVATION of BiochemicalReaction)
5 Control(Out: ACTIVATION of BiochemicalReaction)
6 Control(Out: ACTIVATION of BiochemicalReaction)
```

By default, the function `edges` and `nodes` provide the edges or the nodes of the pathway containing only proteins with signals propagated through metabolites. Since version 1.24, using the option *which* we can access the pathway with proteins and metabolites (*which* = *"mixed"*) or the pathway with only metabolites with edges propagated through proteins (*which* = "metabolites").

```
> head(nodes(p), which = "mixed")

[1] "UNIPROT:P10415" "UNIPROT:P31749" "UNIPROT:P31751" "UNIPROT:P48454"
[5] "UNIPROT:P63098" "UNIPROT:Q92934"

> head(edges(p), which = "mixed")
```

```
  src_type    src dest_type   dest  direction
1  UNIPROT P10415   UNIPROT P55957 undirected
2  UNIPROT P31749   UNIPROT Q92934   directed
3  UNIPROT P31751   UNIPROT Q92934   directed
4  UNIPROT P48454   UNIPROT P27348   directed
5  UNIPROT P48454   UNIPROT P31946   directed
6  UNIPROT P48454   UNIPROT P31947   directed
                                          type
1                                      Binding
2 Control(Out: ACTIVATION of BiochemicalReaction)
3 Control(Out: ACTIVATION of BiochemicalReaction)
4 Control(Out: ACTIVATION of BiochemicalReaction)
5 Control(Out: ACTIVATION of BiochemicalReaction)
6 Control(Out: ACTIVATION of BiochemicalReaction)
```

These same steps can be used to access the KEGG, Biocarta, Panther, HumanCyc, NCI, SMPDB and PharmGKB databases for Homo sapiens pathways (through the `kegg`, `biocarta`, `panther`, `humancyc` and `nci` `smpdb`, `pharmgkb` lists, respectively), but not all the databases are present for all the species. To know the pathway data available the user can use the `pathwayDatabases`.

```
> pathwayDatabases()
```

```
         species database
1      athaliana     kegg
2      athaliana pathbank
3        btaurus     kegg
4        btaurus reactome
5       celegans     kegg
6       celegans reactome
7    cfamiliaris     kegg
8    cfamiliaris reactome
9  dmelanogaster     kegg
10 dmelanogaster reactome
11        drerio     kegg
12        drerio reactome
13         ecoli     kegg
14         ecoli pathbank
15       ggallus     kegg
16       ggallus reactome
17      hsapiens biocarta
18      hsapiens humancyc
19      hsapiens     kegg
20      hsapiens      nci
21      hsapiens  panther
22      hsapiens pathbank
23      hsapiens pharmgkb
24      hsapiens reactome
25      hsapiens    smpdb
26     mmusculus     kegg
27     mmusculus pathbank
28     mmusculus reactome
```

```
29   rnorvegicus     kegg
30   rnorvegicus pathbank
31   rnorvegicus reactome
32   scerevisiae     kegg
33   scerevisiae pathbank
34   scerevisiae reactome
35       sscrofa     kegg
36       sscrofa reactome
37       xlaevis     kegg
```

3 Pathway 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 = 15
Number of Edges = 39
```

```
> edgeData(g)[1]

$`UNIPROT:P10415|UNIPROT:P55957`
$`UNIPROT:P10415|UNIPROT:P55957`$weight
[1] 1

$`UNIPROT:P10415|UNIPROT:P55957`$edgeType
[1] "Binding"
```

Similarly to the `edges` and `nodes`, by default, the function `pathwayGraph` provide the pathway with only proteins and edges propagated through metabolites. Since version 1.24, using the option *which* we can access the pathway with proteins and metabolites (*which* = *"mixed"*) or the pathway with only metabolites with edges propagated through proteins (*which* = "metabolites").

```
> g <- pathwayGraph(p, which = "mixed")
> g

A graphNEL graph with directed edges
Number of Nodes = 19
Number of Edges = 69
```

```
> edgeData(g)[1]

$`CHEBI:15377|CHEBI:18367`
$`CHEBI:15377|CHEBI:18367`$weight
[1] 1

$`CHEBI:15377|CHEBI:18367`$edgeType
[1] "Process(BiochemicalReaction)"
```

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 database uniquely represent biological entities, thus biological entities can be identified by homogeneous 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 among them or many-to-many relations. For detailed information about IDs, their structures and differences please consult those resources.

The function `converterIdentifiers` allows the user to convert the pathway IDs into different types, to fit the user needs. 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). We based the function of ID conversion through the species-specific Bioconductor *AnnotationDbi*, such as *org.Hs.eg.db*. Since the version 1.14, all the conversion supported in the annotation packages can be used. `converterIdentifiers` needs a list of pathways or a single pathway and a string describing the type of the identifier as provided by an Annotation package (for example,'"ENTREZID"'), while the values '"entrez"', '"symbol"' remains for backward compatibility.

```
> pSymbol <- convertIdentifiers(p, "SYMBOL")
> pSymbol

"Activation of BAD and translocation to mitochondria" pathway
Native ID       = R-HSA-111447
Database        = Reactome
Species         = hsapiens
Number of nodes = 19
Number of edges = 59
Retrieved on    = 17-04-2019
URL             = http://reactome.org/PathwayBrowser/#/R-HSA-111447

> head(nodes(pSymbol))

[1] "SYMBOL:BCL2"   "SYMBOL:AKT1"   "SYMBOL:AKT2"   "SYMBOL:PPP3CC"
[5] "SYMBOL:PPP3R1" "SYMBOL:BAD"
```

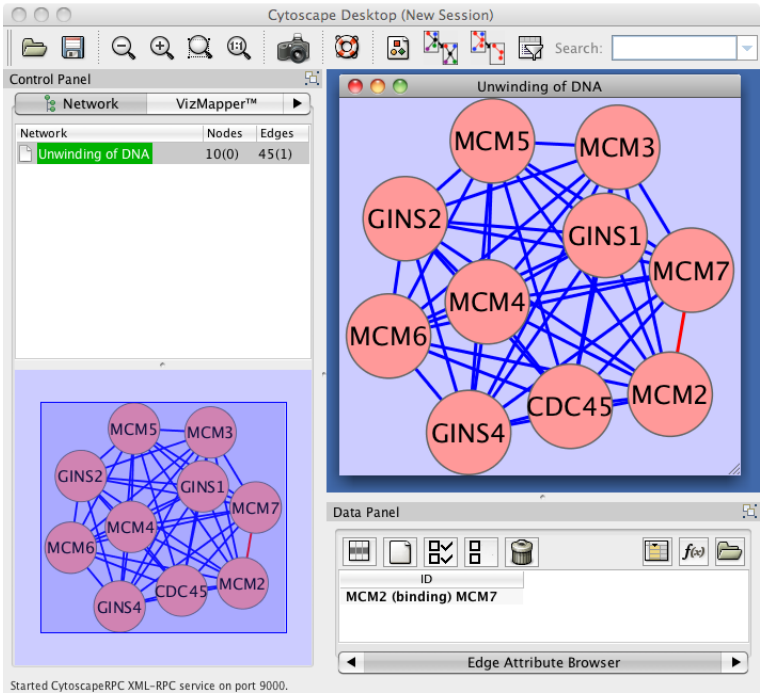
```
> reactomeSymbol <- convertIdentifiers(humanReactome[1:5], "SYMBOL")
```

Since the introduction of metabolites in *graphite* pathways (version 1.24), also the metabolite ID conversion is possible. The metabolite conversion is performed for the following identifiers:"PUBCHEM", "KEGGLYCAN", "KEGGDRUG", "KEGGCOMP", "CHEBI", "CAS".

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 *Rcy3* package to export the network to Cytoscape 3, through the function `cytoscapePlot`. 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. Since version 1.24, using the option *which* we can access the pathway with proteins and metabolites (*which* = *"mixed"*) or the pathway with only metabolites and edges propagated through proteins (*which* = "metabolites").

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



6 Topological pathway analysis

graphite gives access to three types of topological pathway analyses. For more details on the results obtained by these analyses see the corresponding R packages. Following developper instructions, all the methods with the exception of *clipper* are available only for pathways with proteins and with edges propagated through metabolites. *clipper* is the only analysis that can be used to study metabolomics data.

Note that, since version 1.24, given the introduction of nodes with mixed types, all the ID in the pathway have been prefixed with the type of identifier (e.g. ENTREZID:7157 or SYMBOL:TP53). Thus, also the gene expression or metabolite data, in order to match with the pathway nodes, should be in the same format (e.g. ENTREZID:7157 or SYMBOL:TP53).

6.1 SPIA

The analysis with *SPIA* requires the conversion of gene-gene 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`. Edges not included in SPIA have been coerced into the admitted SPIA types. Metabolite mediated interactions (edges of propagation through metabolites) annotated in graphite with the "indirect" type are mapped into the SPIA edge type "indirect effect". For more details see the *SPIA* package [5, 6, 4].

```
> library(SPIA)
> data(colorectalcancer)
> library(hgu133plus2.db)
> top$ENTREZ <- mapIds(hgu133plus2.db,
+                      top$ID, "ENTREZID", "PROBEID", multiVals = "first")
> top <- top[!is.na(top$ENTREZ) & !duplicated(top$ENTREZ), ]
> top$ENTREZ <- paste("ENTREZID", top$ENTREZ, sep = ":")
> tg1 <- top[top$adj.P.Val < 0.05, ]
> DE_Colorectal = tg1$logFC
> names(DE_Colorectal) <- tg1$ENTREZ
> ALL_Colorectal <- top$ENTREZ
> biocarta <- pathways("hsapiens", "biocarta")[1:10]
> biocarta <- convertIdentifiers(biocarta, "ENTREZID")
> prepareSPIA(biocarta, "biocartaEx")
> res <- runSPIA(de = DE_Colorectal, all = ALL_Colorectal, "biocartaEx")

Done pathway 1 : p38 mapk signaling pathway..
Done pathway 2 : regulation of eif-4e and p70s6..
Done pathway 3 : tnfr2 signaling pathway..
Done pathway 4 : melanocyte development and pig..
Done pathway 5 : il-7 signal transduction..
Done pathway 6 : ifn gamma signaling pathway..

> res[1:5,]

                                              Name pSize NDE      pNDE
1                       p38 mapk signaling pathway    21   7 0.1214148
2  melanocyte development and pigmentation pathway    12   4 0.2201876
3                          tnfr2 signaling pathway     9   1 0.8739299
4            regulation of eif-4e and p70s6 kinase    18   2 0.9101261
NA                                            <NA>    NA  NA        NA
           tA pPERT        pG     pGFdr pGFWER    Status
1  -4.2255669 0.697 0.2936114 0.4525128      1 Inhibited
2   5.9696008 0.429 0.3173470 0.4525128      1 Activated
3   0.6290229 0.119 0.3393846 0.4525128      1 Activated
4   0.0000000 1.000 0.9958346 0.9958346      1 Inhibited
NA         NA    NA        NA        NA     NA      <NA>
```

6.2 topologyGSA

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

In *graphite*, *topologyGSA* has a dedicated function, `runTopologyGSA`, which performs the analysis on a single pathway or on a pathway list.

```
> library(topologyGSA)
> data(examples)
> colnames(y1) <- paste("SYMBOL", colnames(y1), sep = ":")
> colnames(y2) <- paste("SYMBOL", colnames(y2), sep = ":")
> kegg <- pathways("hsapiens", "kegg")
> p <- convertIdentifiers(kegg[["Fc epsilon RI signaling pathway"]], "SYMBOL")
> runTopologyGSA(p, "var", y1, y2, 0.05)

        Pathway Variance Test

data: exp1, exp2 and g

lambda = 26.02199, df = 10, p-value = 0.003710726, equal variances: TRUE
```

The function `runTopologyGSA`, which easily performs the analysis on the entire pathway database, provides as result a list of two elements: a list with the results of the pathway analyses and the list of generated errors.

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

6.3 clipper

clipper is a package for topological analysis. It implements a two-step empirical approach based on the exploitation of graph decomposition into a junction tree to reconstruct the most relevant signal path. In the first step clipper selects significant pathways according to statistical tests on the means and the concentration matrices of the graphs derived from pathway topologies. Then, it "clips" the whole pathway identifying the signal paths having the greatest association with a specific phenotype.

In *graphite*, *clipper* has a dedicated function, `runClipper`, which performs the analysis on a single pathway or on a pathway list.

```
> library(ALL)
> library(a4Preproc)
> library(clipper)
> data(ALL)
> pheno <- as(phenoData(ALL), "data.frame")
> samples <- unlist(lapply(c("NEG", "BCR/ABL"), function(t) {
+   which(grepl("^B\\d*", pheno$BT) & (pheno$mol.biol == t))[1:10]
+ }))
> classes <- c(rep(1,10), rep(2,10))
> expr <- exprs(ALL)[,samples]
> rownames(expr) <- paste("ENTREZID",
+                         featureData(addGeneInfo(ALL))$ENTREZID,
+                         sep = ":")
> k <- as.list(pathways("hsapiens", "kegg"))
> selected <- k[c("Chronic myeloid leukemia",
+                 "Bladder cancer",
+                 "Cytosolic DNA-sensing pathway")]
```

```
> clipped <- runClipper(selected, expr, classes, "mean", pathThr = 0.1)
> resClip <- do.call(rbind, clipped$results)
> resClip[, c("startIdx", "endIdx", "maxIdx", "lenght",
+             "maxScore", "aScore", "involvedGenes")]

NULL
```

The function `runClipper`, which easily performs the analysis on the entire pathway database, provides as result a list of two elements: the list with the results of the pathway analyses and the list of eventually generated errors.

Since version 1.24, using the option *which* we can access the pathway with proteins and metabolites (*which* = *"mixed"*) or the pathway with only metabolites and edges propagated through proteins (*which* = "metabolites").

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

7 Build pathway

In *graphite*, it is also possible build a pathway object using `buildPathway`. This function creates a new object of type *Pathway* given a data frame describing its edges. The edges should be divided in three dataframes containing edges between proteins, edges between proteins and metabolites, and edges between metabolites, and the should be included in the new pathway using *proteinEdges*, *mixedEdges* and *metaboliteEdges* respectively. This practice will guarantee the compatibility with `convertIdentifiers`, this only works with the types of identifiers that are provided in the Annotation package (for example using the string,'"ENTREZID"').

```
> edges <- data.frame(src_type = "ENTREZID", src="672",
+                     dest_type = "ENTREZID", dest="7157",
+                     direction="undirected", type="binding")
> pathway <- buildPathway("#1", "example", "hsapiens", "database",
+                         proteinEdges = edges)
```

```
> edges <- data.frame(src_type = "ENTREZID", src="672",
+                     dest_type = "ENTREZID", dest="7157",
+                     direction="undirected", type="binding")
> edgemix <- data.frame(src_type = "CHEBI", src="77750",
+                     dest_type = "ENTREZID", dest="7157",
+                     direction="undirected", type="biochemicalReaction")
> edgemet <- data.frame(src_type = "CHEBI", src="15351",
+                     dest_type = "CHEBI", dest="77750",
+                     direction="undirected", type="biochemicalReaction")
> pathway <- buildPathway("#1", "example", "hsapiens", "database",
+                         proteinEdges = edges,
+                         mixedEdges = edgemix,
+                         metaboliteEdges = edgemet)
```

8 Parallelism

Some of *graphite* operations can be made significantly faster by exploiting the parallelism offered by recent hardware. Here is a list of the functions providing this option:

- `convertIdentifiers`
- `runClipper`
- `runTopologyGSA`

To exploit parallel processing you are going to need two ingredients. First, you have to specify the maximum number of cores that the package can use. For instance, if you have 6 cores on your computer you can set:

```
> options(Ncpus = 6)
```

Your code, then, should try to call *graphite* functions passing entire lists of pathways; you should not loop manually.

```
> original <- pathways("hsapiens", "reactome")
> # Do (will exploit parallelism)
> converted <- convertIdentifiers(original, "SYMBOL")
> # Don't (no parallelism here)
> converted <- lapply(original, convertIdentifiers, "SYMBOL")
```

Warning: Parallelism is not guaranteed to reduce run times. Splitting a given operation over multiple cores requires some coordination, and that has its own costs. In certain situations (small number of pathways / small number of cores) the overhead may actually make run times worse. Our suggestion is thus to:

- always set `Ncpus` to a value smaller or equal to the actual number of hardware cores (see `parallel::detectCores` for an indication);
- in any case, start from a small number like 2 or 4, and then work your way up measuring the actual speedups on your own hardware.

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. Bioinformatics, 2009, 25(1):75-82.

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

[8] Martini P, Sales G, Massa MS, Chiogna M, Romualdi C. Along signal paths: an empirical gene set approach exploiting pathway topology. Nucleic Acids Res. 2013 Jan 7;41(1):e19. doi: 10.1093/nar/gks866. Epub 2012 Sep 21.

[9] Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, Kaipa P, Karthikeyan AS, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Paley S, Popescu L, Pujar A, Shearer AG, Zhang P, Karp PD. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Research 38:D473-9 2010.

[10] PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Huaiyu Mi, Anushya Muruganujan and Paul D. Thomas Nucl. Acids Res. (2012) doi: 10.1093/nar/gks1118

[11] Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, Altman RB, Klein TE. Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther. 2012 Oct;92(4):414-7. doi: 10.1038/clpt.2012.96. Review. PubMed

[12] Jewison T, Su Y, Disfany FM, Liang Y, Knox C, Maciejewski A, Poelzer J, Huynh J, Zhou Y, Arndt D, Djoumbou Y, Liu Y, Deng L, Guo AC, Han B, Pon A, Wilson M, Rafatnia S, Liu P, Wishart DS. SMPDB 2.0: big improvements to the Small Molecule Pathway Database. Nucleic Acids Res. 2014 Jan;42(Database issue):D478-84. doi:10.1093/nar/gkt1067. Epub 2013 Nov 6.