# Integration of networks and pathways with StarBioTrek package

*by Claudia Cava and Isabella Castiglioni*

**Abstract** High-throughput genomic technologies bring to light a comprehensive hallmark of molecular changes of a disease. It is increasingly evident that genes are not isolated from each other and the identification of a gene signature can only partially elucidate the de-regulated biological functions in a disease. The comprehension of how groups of genes (pathways) are related to each other (pathway-cross talk) could explain biological mechanisms causing diseases. Biological pathways are important tools to identify gene interactions and decrease the large number of genes to be studied by partitioning them into smaller groups. Furthermore, recent scientific studies have demonstrated that an integration of pathways and networks, instead of a single component of the pathway or a single network, could lead to a deeper understanding of the pathology.

**StarBioTrek** is an R package for the integration of biological pathways and networks which provides a series of functions to support the user in their analyses. In particular, it implements algorithms to identify pathways cross-talk networks and gene network drivers in pathways. It is available as open source and open development software in the Bioconductor platform.

## Introduction

In recent years new genomic technologies have made possible to define new marker gene signatures (Desmedt et al., 2009; Parker et al., 2009; Cava et al., 2014b). However, gene expression-based signatures present some constraints because they do not consider metabolic role of the genes and are affected by genetic heterogeneity across patient cohorts (Cava et al., 2015; Donato et al., 2013).

Pathway analysis can help the researchers in the identification of the biological roles of candidate genes exceeding these limitions (Folger et al., 2011). Indeed, considering the activity of entire biological pathways rather than the expression levels of individual genes can charactherize the whole tissue. In particular, there are several methods in computations and data used to perform the pathway analyses. They can be characterized in two different levels: gene-sets and pathway topology (García-Campos et al., 2015). Indeed, the existing tools integrating pathway data can be grouped into these groups based on the pathway definition.

In the first group we can include the tools that are based on gene sets definition as simple lists of biological molecules, in which the listed components share a common biological role. In this group, for example, we can include CoRegNet and Gene Set Enrichment Analysis (GSEA). CoRegNet reconstructs a co-regulatory network from gene expression profiles integrating, also, regulatory interactions, such as transcription factor binding site and ChIP data, presenting some analyses to identify master regulators of a given set of genes (Nicolle et al., 2015). One of the first and most popular methods is GSEA (Subramanian et al., 2005) that uses a list of ranked genes based on their differential gene expression between two labels. It then evaluates their distribution on a priori defined set of genes, thus generating an enrichment score (ES) for each set of genes.

In contrast, tools based on pathway topology do not only contain the components of a pathway but also describe the interactions between them. However, these methods still analyze the pathways as independent from each other and not considering the influence that a pathway can exert over another. In this second group we can include analysis methods that take into account the topological structure of a pathway, such as **NetPathMiner**, **ToPASeq**, and **XGR**. **NetPathMiner** (Mohamed et al., 2014) implements methods for the construction of various types of genome scale networks for network path mining. It generates a network representation from a pathway file supporting metabolic networks. Since the network is generated, the network edges can be weighted using gene expression data (e.g., Pearson correlation). Using machine learning methods and Markov mixture models, the pathways can be classified or clustered based on their association with a response label. The **ToPASeq** package implements seven different methods covering broad aspects for topology-based pathway analysis of RNA-seq data (Ihnatova and Budinska, 2015). With respect to other tools, **XGR** (Fang et al., 2016) is designed for enhanced interpretation of genomic data generating also SNP-modulated gene networks and pathways. However, compared to our tool, the others are not focused on the pathway cross-talk analyses.

In line with this scenario given the few methods focused on the pathway cross-talk network, the development of new methodologies to measure pathway activity and cross-talk among pathways integrating also the information of networks and gene expression data (e.g., TCGA data) could lead to a deeper knowledge of the pathology.

Furthermore, functional pathway representation attributes the same functional significance to each gene without considering the impact of gene interactions in performing that function. What kinds of interactions are there among genes in functional pathways? Specifically, biological system interactions are composed of multiple layers of dynamic interaction networks (Cantini et al., 2015). These dynamic networks can be decomposed, for example, into: co-expression, physical, co-localization, genetic, pathway, and shared protein domains.

We developed a series of algorithms (see (Cava et al., 2018; Colaprico et al., 2015; Cava et al., 2016)), implemented in **StarBioTrek** package able to work on all levels of the pathway analysis.

Starting from the gene expression data of two groups of samples (e.g., normal vs. disease), such algorithms aim at building a pathway cross-talk model by attributing a score for each pairwise pathway. Several scores are implemented in the tool using the gene expression levels inside the pathways. The interacting pathways are filtered considering pathways that are able to classify better the two groups of samples. In addition, the genes inside the pathways can be weighted defining key network drivers in the pathways as those gene drivers that are highly connected in biological networks.

In summary, **StarBioTrek** package proposes an approach that integrates knowledge on the functional pathways and multiple gene-gene (protein-protein) interactions into gene selection algorithms. The challenge is to identify more stable biomarker signatures, which are also more easily interpretable from a biological perspective. The integration of biological networks and pathways can also give further hypotheses of the mechanisms of driver genes.

## Package organization

**StarBioTrek** makes accessible data of biological pathways and networks in order to perform analyses without having to navigate and access different web-based databases, without the need to download data, and by integrating and locally processing the full data sets in a short time. Specifically, it allows the users to: i) query and download biological pathways and networks from several repositories such as KEGG, Reactome and GeneMANIA(Zuberi et al., 2013; Cava et al., 2017; Franz et al., 2018) importing several functions from graphite (Sales et al., 2012), and harmonize annotations for genes and proteins (query/ download/ annotation harmonization); (ii) integrate pathways and biological networks with a series of implemented algorithms.

## Get data

### Pathway and network data

The functions of **StarBioTrek** import a large amount of data (e.g., biological pathways and networks).

Specifically, the function `pathwayDatabases` can easily query some features of interest of the user such as species or specific pathway database from graphite (Sales et al., 2012). Then, the function `GetData` imports the selected data.

```
> library(graphite)
> pathwayDatabases()
  species database
1      athaliana     kegg
2      athaliana pathbank
3      athaliana reactome
4        btaurus     kegg
5        btaurus reactome
6       celegans     kegg
7       celegans reactome
8     cfamiliaris     kegg
9     cfamiliaris reactome
10 dmelanogaster     kegg
11 dmelanogaster reactome
12        drerio     kegg
13        drerio reactome
14         ecoli     kegg
15         ecoli pathbank
16       ggallus     kegg
17       ggallus reactome
18      hsapiens biocarta
```

```
19      hsapiens humancyc
20      hsapiens     kegg
21      hsapiens      nci
22      hsapiens  panther
23      hsapiens pathbank
24      hsapiens pharmgkb
25      hsapiens reactome
26      hsapiens    smpdb
27     mmusculus     kegg
28     mmusculus pathbank
29     mmusculus reactome
30   rnorvegicus     kegg
31   rnorvegicus pathbank
32   rnorvegicus reactome
33   scerevisiae     kegg
34   scerevisiae pathbank
35   scerevisiae reactome
36       sscrofa     kegg
37       sscrofa reactome
38       xlaevis     kegg
> path <- GetData(species="hsapiens", pathwaydb="kegg")
```

Since the user selected the data of interest, the function `GetPathData` allows us to obtain a list of genes grouped by functional role:

```
> pathwayALLGENE <- GetPathData(path_ALL=path[1:3])
[1] "Downloading............  Glycolysis / Gluconeogenesis   1 of 3 pathways"
[1] "Downloading............  Citrate cycle (TCA cycle)   2 of 3 pathways"
[1] "Downloading............  Pentose phosphate pathway   3 of 3 pathways"
```

The function `ConvertedIDgenes` will converter the gene nomenclature (e.g., ENTREZ ID) to Gene Symbol.

```
> pathway <- ConvertedIDgenes(path_ALL=path[1:10])
```

The function `getNETdata` of **StarBioTrek** imports biological networks from GeneMANIA. The biological networks can be selected among physical interactions, co-localization, genetic interactions, pathways, and shared protein domain networks. Furthermore, it supports 9 species ( Arabidopsis thaliana, Caenorhabditis elegans, Danio rerio, Drosophila melanogaster, Escherichia coli, Homo sapiens, Mus musculus, Rattus norvegicus, and Saccharomyces cerevisiae); for default it considers Homo sapiens. Specifically, the function call

```
> netw <- getNETdata(network="SHpd")
[1]"genemania.org/data/current/Homo_sapiens/Shared_protein_domains.INTERPRO.txt n.1 of 2"
[1]"genemania.org/data/current/Homo_sapiens/Shared_protein_domains.PFAM.txt n.2 of 2"
[1]"Preprocessing of the network n. 1 of 2"
[1]"Preprocessing of the network n. 2 of 2"
```

imports biological networks (i.e., shared protein domains interactions from INTERPRO and PFAM databases) for *Homo sapiens*. Otherwise, the user can select one of the 9 species or using the following parameters the user can select different network types: `PHint` for Physical interactions, `COloc` for Co-localization, `GENint` for Genetic interactions, `PATH` for Pathway, and `SHpd` for Shared protein domains. Finally, **StarBioTrek** provides the functions for the harmonization of gene nomenclature in the pathways and biological networks. Biological data are processed for downstream analyses mapping Ensembl Gene ID to gene symbols. Figure 1 shows an overview of network types supported by **StarBioTrek** with the function `getNETdata`.

## Analysing pathways

Starting from a gene expression matrix (DataMatrix), **StarBioTrek** groups the gene expression levels according to their biological roles in pathways for each sample.

```
> listpathgene <- GE_matrix(DataMatrix=tumo[,1:2], genes.by.pathway=pathway[1:10])
> str(listpathgene)
List of 2
```

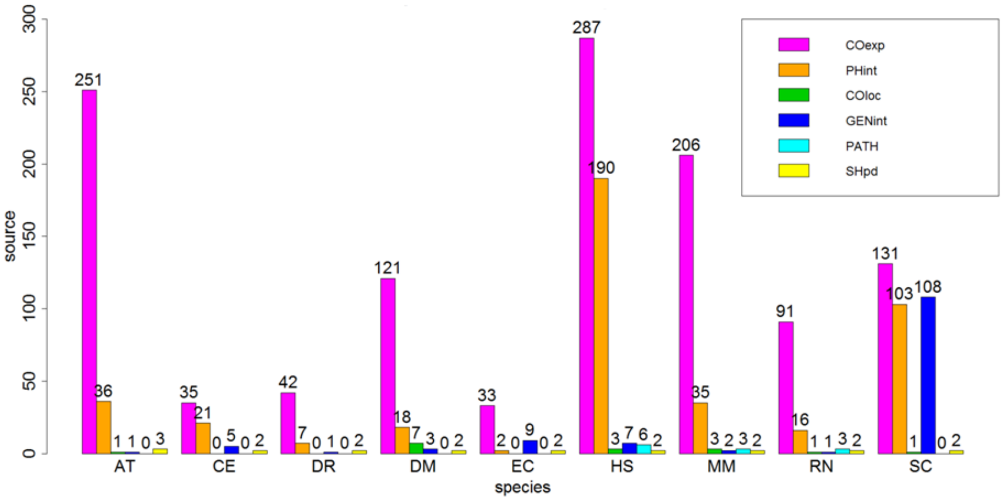

**Figure 1:** Network overview. Number of sources for network data. Barplot is divided by species: AT: *Arabidopsis thaliana*, CE: *Caenorhabditis elegans*, DR: *Danio rerio*, DM: *Drosophila melanogaster*, EC: *Escherichia coli*, HS: *Homo sapiens*, MM: *Mus musculus*, RN: *Rattus norvegicus*, SC: *Saccharomyces cerevisiae*, and by network type: COexp: Co-expression, PHint: Physical interactions, COloc: Co-localization, GENint: Genetic interactions, PATH: Pathway, SHpd: Shared protein domains

```
 $ Cell_cycle           :'data.frame':        114 obs. of  2 variables:
  ..$ TCGA-E9-A1RC-01A: num [1:114] 4218 695 4231 7029 1211 ...
  ..$ TCGA-BH-A0B1-01A: num [1:114] 3273 692 6733 6468 1290 ...
 $ p53_signaling pathway:'data.frame':        64 obs. of  2 variables:
  ..$ TCGA-E9-A1RC-01A: num [1:64] 989 1614 1592 3456 900 ...
  ..$ TCGA-BH-A0B1-01A: num [1:64] 816 1274 1770 3190 405 ...
```

This function allows the user to have in a short time the gene expression levels grouped by pathways.

### Pathway summary indexes

As described in Cava et al. (2014a) there are different measures to summarize the expression levels of each pathway, such as the mean:

```
> score_mean <- average(pathwayexpsubset=listpathgene)
```

or standard deviation:

```
> score_stdev <- stdv(gslist=listpathgene)
```

### Dissimilarity distances: Pathway cross-talk indexes

Dissimilarity distances have been proved useful in many application fields. Recent studies (Cava et al., 2013, 2014c) used with success dissimilarity representation among patients, considering the expression levels of individual genes. To our knowledge, dissimilarity representation is not used in pathway-based expression profiles. Our goal is to give a dissimilarity representation, which can express, through a function D(x,y), the dissimilarity between two pathways x and y, such as Euclidean distance between pairs of pathways:

```
> scoreeucdistat <- eucdistcrtlk(dataFilt=Data_CANCER_normUQ_fil,
                                 pathway_exp=pathway[1:10])
```

or discriminating score (Colaprico et al., 2015):

```
> crosstalkstdv <- dsscorecrtlk(dataFilt=Data_CANCER_normUQ_fil,
                                pathway_exp=pathway[1:10])
```

## Integration data

### Integration between pathways and networks from GeneMANIA

Biological pathways can involve a large number of genes that are not equivocally relevant for a functional role in the cell. Therefore, the integration of network and pathway-based methods can boost the power to identify the key genes in the pathways.

The function takes as arguments: a list of pathways as obtained by the function `ConvertedIDgenes` and the networks as obtained by the function `getNETdata`.

```
> listanetwork <- pathnet(genes.by.pathway=pathway[1:10], data=netw)
```

It creates a network of interacting genes for each pathway. The output of the function is a selection of interacting genes according to the network $N$ in a pathway $P$, namely a list with two columns where on the same row there are the two interacting genes.

The function `listpathnet` takes as inputs the output obtained by the function `pathnet` and the pathways as obtained by the function `ConvertedIDgenes`:

```
> listpath <- listpathnet(lista_net=listanetwork, pathway=pathway[1:10])
```

creating a list of vectors for each pathway containing only genes that have at least one interaction with other genes belonging to the pathway.

### Integration between pathways and networks from protein-protein interaction

The function `GetPathNet` allows us to obtain a list of interacting genes (protein-protein interactions from **graphite** package) for each pathway:

```
> pathwaynet <- GetPathNet(path_ALL=path[1:3])
```

using as its argument the output obtained by `GetData`.

## Analyzing networks and pathways: implemented algorithms

### Pathway cross-talk network

The first algorithm implemented in **StarBioTrek** explores a pathway cross-talk network from gene expression data to better understand the underlying pathological mechanism. The algorithm generates a network of pathways that shows a different behavior between two groups of samples (i.e., normal vs. disease).

Specifically,

```
# get pathways from KEGG database
path <- GetData(species="hsapiens", pathwaydb="kegg")
pathway <- ConvertedIDgenes(path_ALL=path)

# create a measure of pathway cross-talk (i.e., Euclidean distance) between pairwise
# of pathways starting from gene expression data (i.e.TCGA) with in the columns the
# samples and in the rows the genes
scoreeucdistat <- eucdistcrtlk(dataFilt=Data_CANCER_normUQ_fil, pathway=pathway)


# split samples' TCGA ID into normal and tumor groups
tumo <- SelectedSample(Dataset=Data_CANCER_normUQ_fil, typesample="tumour")
norm <- SelectedSample(Dataset=Data_CANCER_normUQ_fil, typesample="normal")

# divide the dataset in 60/100 for training and 40/100 for testing
nf <- 60

# a support vector machine is applied
res_class <- svm_classification(TCGA_matrix=scoreeucdistat[1:10,], nfs=nf,
                                normal=colnames(norm), tumour=colnames(tumo))
```

Since the AUC values are obtained for each pair of pathways, they can be ranked in order to obtain the pathway cross-talk interactions able to classify the two classes (i.e. normal vs. tumor samples) with the best performance. Such selection can be done considering AUC values:

```
cutoff=0.80
er <- res_class[res_class[,1]>cutoff, ]
```

The outputs are the only pathway interactions that are obtained with AUC values > 0.80. The implemented algorithm was used in (Colaprico et al., 2015) and (Cava et al., 2016) to screen pathway cross-talk associated to breast cancer.

The pseudocode of the implemented algorithm is summarized below.

```
Data: 1) a matrix of gene expression data (TCGA data). The samples are in the columns and the
      genes in the rows; 2) a matrix where the pathways are in the columns and the genes inside
      the pathways are in the rows
Result: pathway interactions that are able to classify two groups of samples with the best
        performances
Being $a$ and $b$ two pathways in a set of pathways $P$;
for all nodes(a,b) in P do
    a score distance between the nodes $a$ and $b$;
    if AUC > cut-off then
        keep ($a$,$b$) as edge;
    else
        remove ($a$,$b$) as edge;
    end
end
```

**Algorithm 1:** Algorithm implemented in (Colaprico et al., 2015) and (Cava et al., 2016)

### Driver genes for each pathway

Here, we propose an algorithm for the integrative analysis of networks and pathways. Our method is inspired on a well-validated method (the GANPA/LEGO) (Fang et al., 2012; Dong et al., 2016), based on the hypothesis that if one gene is functionally connected in the pathway with more genes than those expected (according to the functional networks), has a key role in that pathway. The algorithm, an extension of the GANPA/LEGO method, defines driver genes in a pathway if they are highly connected in a biological network.

The function

```
IPPI(patha=pathway_matrix[,1:10], netwa=netw_IPPI)
```

is used to identify driver genes for each pathway. The inputs of the function are pathways and network data. It calculates the degree centrality values of genes inside the network and the degree centrality of genes inside pathways.
The pseudocode of the implemented algorithm is summarized below.

```
Data: 1) a matrix where the pathways are in the columns and the genes inside the pathways are
      in the rows; 2) a data frame where the nodes are presented in the columns and the rows
      represent the edges
Result: a list of genes with high degree centrality for each pathway
Being ($i \in N$) & ($i \in P$) where $P$ is a vector containing the genes inside a pathway of size $k$ and
  $N$ is an indirect graph of size $m$;
for all nodes i in N do
    calculate the degree centrality $d_{iN}$;
end
for all nodes i in P do
    calculate the degree centrality $d_{iP}$, being the neighbors of $i$, $i_{ng} \in P$;
end
Calculate degree centrality expected $d_{iE}$ in $P$
if $d_{iE} < d_{iP}/k_p$ then
    $i \longleftarrow$potential gene drivers of $P$;
else
    $i \longleftarrow i+1$;
end
```

**Algorithm 2:** Algorithm implemented in (Cava et al., 2018)

In the first step, given the gene $i$ within the network $N$ with $m$ genes, the function computes the degree centrality $d_{iN}$ as the number of neighbor genes belonging to $N$ to which the gene $i$ is directly connected.

In the second step, given gene $i$ within the pathway $P$ with $k$ genes, the function then computes the degree centrality $d_{iP}$ considering only the relations among gene $i$ and the other genes in the networks belonging to pathway $P$. Overall, by combining the information of the network $N$ within the pathway $P$, is obtained a selection of interacting genes according to the network $N$.

Then, the function computes the degree centrality expected $d_{iE}$ by supposing equal probability for the existence of edges between nodes ($d_{iN}/m = d_{iE}/k$). Thus, $d_{iE} = d_{iN}\ xk/m$.

The function characterizes a gene as a 'network driver' in the pathway $P$, when $d_{iP}$ of involving gene, normalized to the size of the pathway ($k$), is higher than $d_{iE}$, $d_{iP}/k > d_{iE}$.

The speculation is that if one gene is functionally linked (according to the functional network) with more genes in the pathway than expected, its function is central in that pathway.

The function `IPPI` was used in (Cava et al., 2018) to find driver genes in the pathways that are also de-regulated in a pan-cancer analysis.

## Visualization

**StarBioTrek** presents several functions for the preparation to the visualization of gene-gene interactions and pathway cross-talk using the **qgraph** package (Epskamp et al., 2012). The function `plotcrosstalk` prepares the data:

```
> formatplot <- plotcrosstalk(pathway_plot=pathway[1:6],gs_expre=tumo)
```

It computes a Pearson correlation between the genes (according to a gene expression matrix, such as tumor) in which each gene is grouped in a gene set given by the user (e.g., pathway). Each gene is presented in a gene set if it is involved univocally in that gene set.

The functions of **qgraph**

```
> library(qgraph)
> qgraph(formatplot[[1]], minimum = 0.25, cut = 0.6, vsize = 5, groups = formatplot[[2]],
         legend = TRUE, borders = FALSE, layoutScale=c(0.8,0.8))
```

and

```
> qgraph(formatplot[[1]], groups=formatplot[[2]], layout="spring", diag = FALSE,
  cut = 0.6, legend.cex = 0.5, vsize = 6, layoutScale=c(0.8,0.8))
```

show the network with different layouts. The graphical output of the functions are presented in the Figure 2 and Figure 3. The color of interactions indicates the type of correlation: green edges are positive correlations and red edges are negative correlations. The thickness of the edge is proportional to the strength of correlation.

The outputs of the functions that compute the pairwise distance metrics can be easily used with heatmap plotting libraries such as heatmap or pheatmap as reported in the Figure 4.

Furthermore, the function `circleplot` of **StarBioTrek** implemented using the functions of **GOplot** (Walter et al., 2015) provides a visualization of driver genes (with a score indicating the role of genes in that pathway), as reported in Figure 5.

```
> formatplot <- plotcrosstalk(pathway_plot=pathway[1:6], gs_expre=tumo)
> score <- runif(length(formatplot[[2]]), min=-10, max=+10)
> circleplot(preplot=formatplot, scoregene=score)
```

## Case studies

In this section we will present two case studies for the usage of the **StarBioTrek** package. In particular, the first case study uses the first implemented algorithm reported above to identify pathway cross-talk network. The second case study uses the second implemented algorithm to identify gene drivers for each pathway.

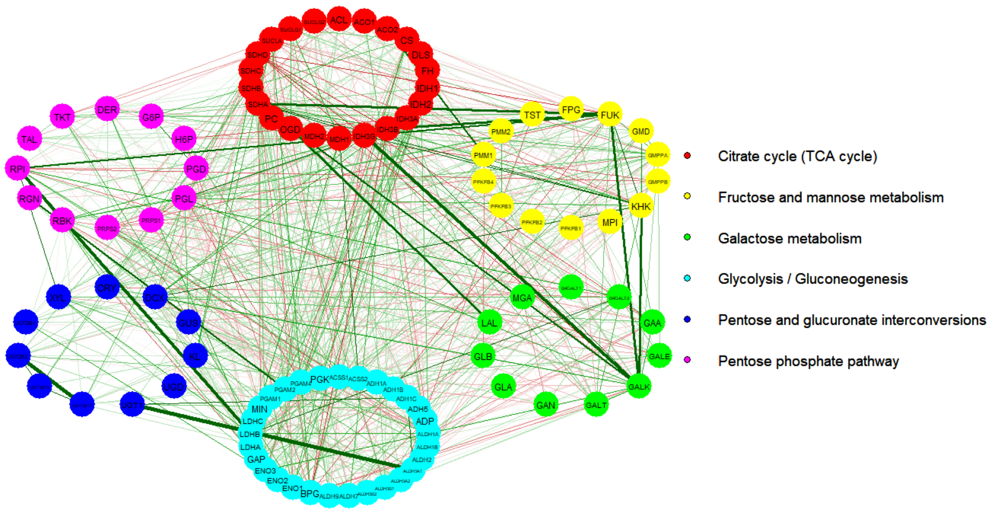

**Figure 2:** Graphical output of the function `plotcrosstalk` and `qgraph` with layout 1

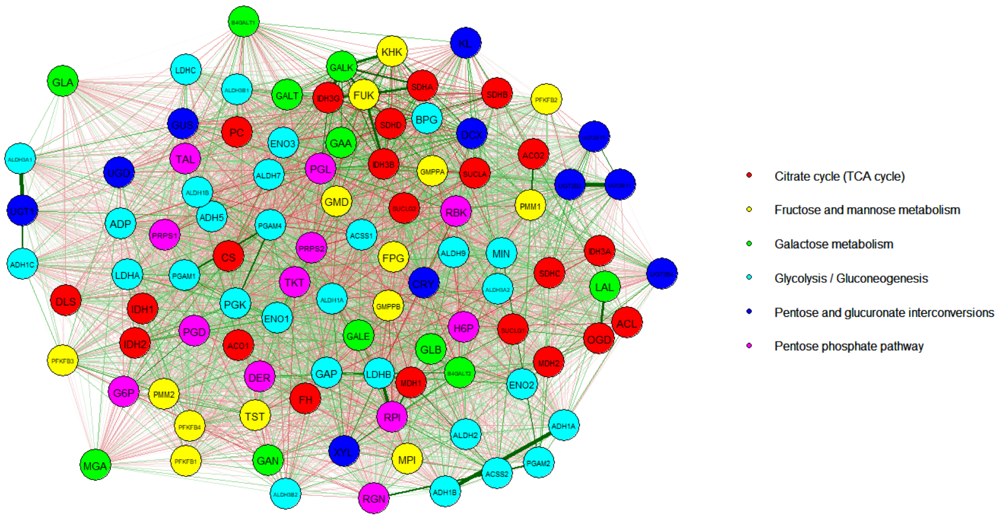

**Figure 3:** Graphical output of the function `plotcrosstalk` and `qgraph` with layout 2

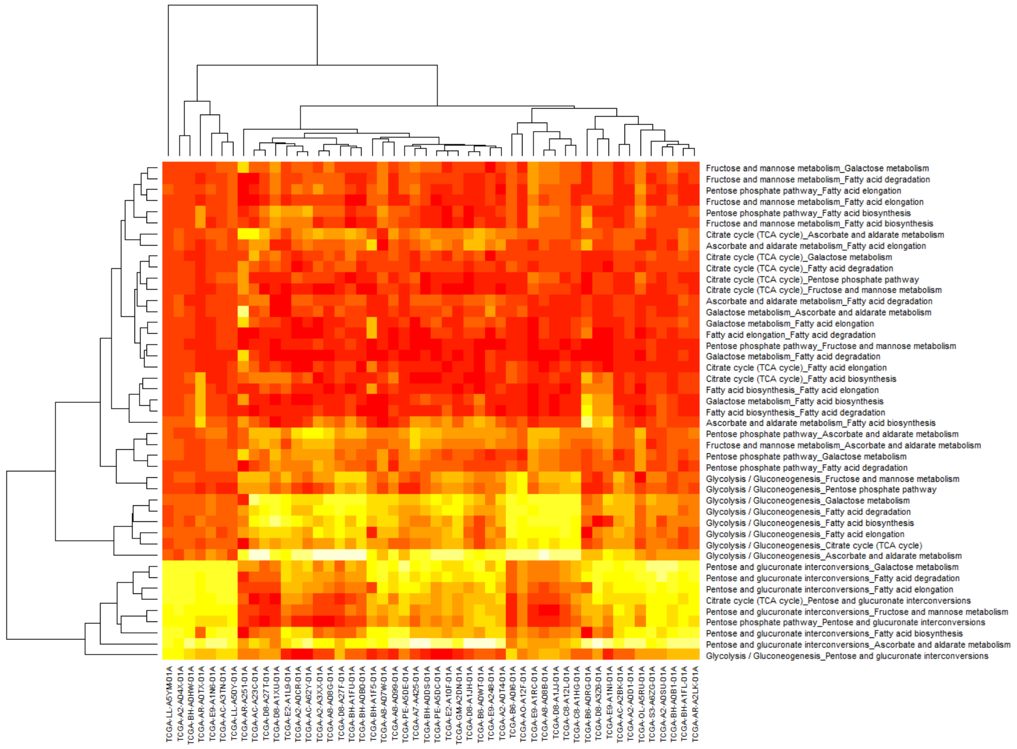

**Figure 4:** Heatmap of pathway cross-talk. Each row represent a distance metric between two pathways (the pathways are seperated by an underscore). The columns represent the samples.

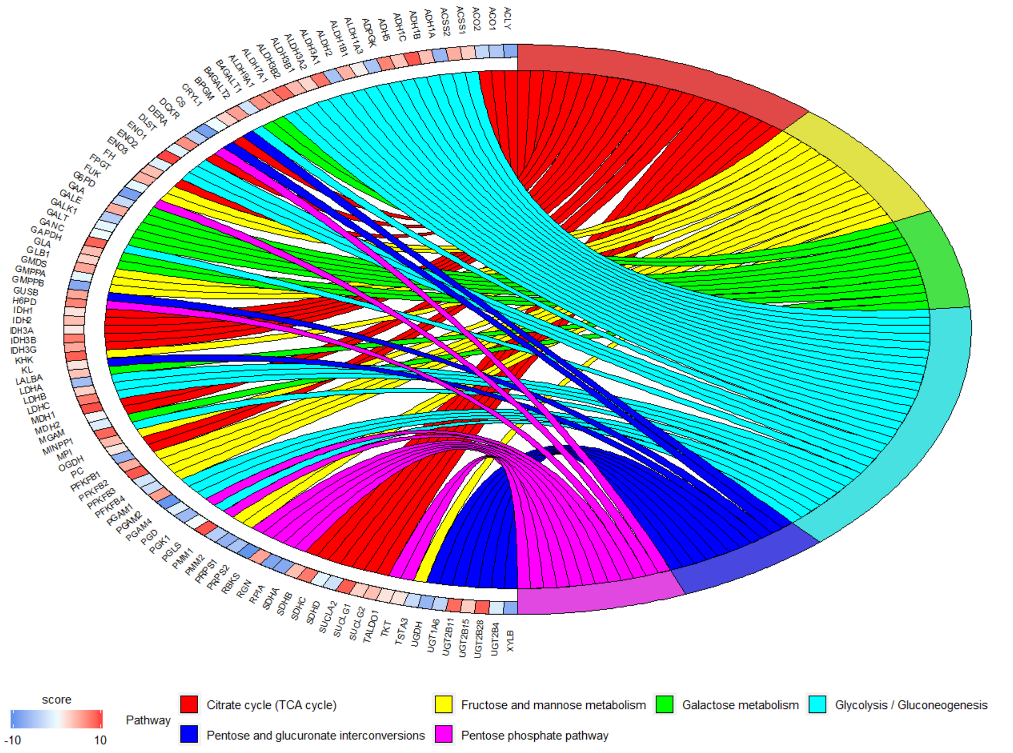

**Figure 5:** Circleplot of pathway cross-talk.The figure shows the relation between gene drivers and pathways. The pathways are represented with different colours. The intensity of colour of each block of genes is based on the score.

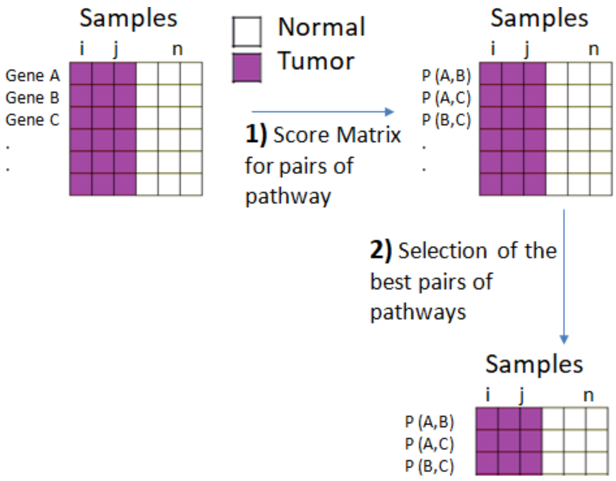

**Figure 6:** The computational approach. The matrix of gene expression data (samples in the columns and genes in the rows) is the input of our algorithm. The samples are grouped in two classes (e.g., normal vs. tumor). In the first step a matrix score is generated using all pairwise combinations of pathways. In the second step the score matrix is used as input for SVM classification. The pathway interactions with the best AUC performances are selected.

### Pathway cross-talk network in breast cancer

Starting from gene expression data of breast cancer samples and normal samples we grouped 15243 genes in pathways according to their functional role in the cell. Pathway data were derived from the function call:

```
path <- GetData(species="hsapiens", pathwaydb="kegg")
pathway <- ConvertedIDgenes(path_ALL=path)
```

For each pair of pathways we calculated a discriminating score as a measure of cross-talk. This measure can be used considering e.g. the pathways enriched with differentially expressed genes.

```
crosstalkscore <- dsscorecrtlk(dataFilt=Data_CANCER_normUQ_fil,
                               pathway_exp=pathway[1:10])
```

Discriminating score is given by |M1-M2|/S1+S2 where M1 and M2 are means and S1 and S2 standard deviations of expression levels of genes in a pathway 1 and in a pathway 2. In order to identify the best pathways for breast cancer classification (breast cancer vs. normal) we implemented a Support Vector Machine. We divided the original dataset in training data set (60/100) and the rest of original data in the testing set (40/100). In order to validate the classifier, we used a $k$-fold cross-validation ($k = 10$) obtaining Area Under the Curve (AUC).

```
tumo <- SelectedSample(Dataset=Data_CANCER_normUQ_fil, typesample="tumour")
norm <- SelectedSample(Dataset=Data_CANCER_normUQ_fil, typesample="normal")
nf <- 60
res_class <- svm_classification(TCGA_matrix=crosstalkscore, nfs=nf,
                                normal=colnames(norm), tumour=colnames(tumo))
```

Ranking AUC values obtained we selected the pathway cross-talk network with the best AUC. The approach of the algorithm is shown in Figure 6.

### Gene network drivers in pathways

In the second case study, we downloaded KEGG pathways

```
path <- GetData(species="hsapiens",pathwaydb="kegg")
pathway <- ConvertedIDgenes(path_ALL=path)
```

and network data for different network types from GeneMANIA

```
# for Physical interactions
netw <- getNETdata(network="PHint")
```

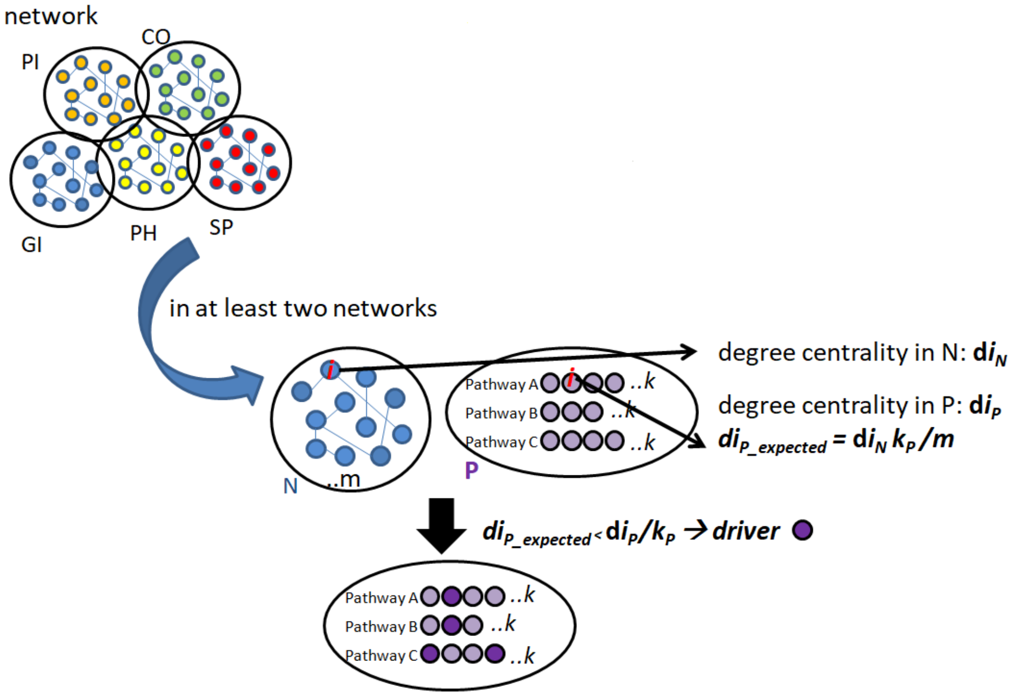
network
PI
CO
GI
PH
SP
in at least two networks
N
..m
i
Pathway A
Pathway B
Pathway C
..k
P
degree centrality in N: di_N
degree centrality in P: di_P
di_P_expected = di_N k_P /m
di_P_expected < di_P/k_P → driver
Pathway A
Pathway B
Pathway C
..k

**Figure 7:** The computational approach. The first step involves a network $N$ (e.g. physical interaction) of size $m$ and for each gene, $i$ in $N$ the algorithm calculates its degree centrality, DC ($d_{iN}$). The second step involves a set of functional pathways (e.g. pathway $P$) and for each gene $i$, the DC ($d_{iP}$) is calculated using the information on interacting genes from $N$. For the speculation of equal probability for existing edges between nodes, the algorithm computes the expected DC of gene $i$ in the pathway $P$. If the DC observed for the gene $i$ ($d_{iP}$) is higher than expected ($d_{iP}$ expected), $i$ could be a potential driver in the pathway $P$

```
# for Co-localization
netw <- getNETdata(network="COloc")

# for Genetic interactions
netw <- getNETdata(network="GENint")

# for Pathway interactions
netw <- getNETdata(network="PATH")

# for Shared_protein_domains
netw <- <getNETdata(network="SHpd")
```

We processed the data obtained by the function getNETdata in order to obtain a data format supported by the function IPPI. The function IPPI was applied for each of the 5 network types.

We obtained that genes with genetic interaction found the lowest number of potential gene network drivers. On the other hand, the network that includes proteins with shared protein domains found the highest number of potential driver genes. Finally, we defined a gene as a "network driver" in the pathway, when in at least two networks one gene is functionally connected in the pathway with more genes than those expected (according to the two networks).

The approach is shown in Figure 7.

## Conclusions

We have described **StarBioTrek**, an R package for the integrative analysis of biological networks and pathways. The package supports the user during the import and data analysis of data. **StarBioTrek** implements two algorithms: i) the identification of gene network drivers in the pathways; ii) the building of pathway cross talk network.

## Bibliography

L. Cantini, E. Medico, S. Fortunato, and M. Caselle. Detection of gene communities in multi-networks reveals cancer drivers. *Scientific reports*, 5:17386, 2015. URL https://doi.org/10.1038/srep17386. [p2]

C. Cava, I. Zoppis, M. Gariboldi, I. Castiglioni, G. Mauri, and M. Antoniotti. Copy–number alterations for tumor progression inference. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 104–109. Springer, 2013. URL https://doi.org/10.1007/978-3-642-38326-7_16. [p4]

C. Cava, G. Bertoli, and I. Castiglioni. Pathway-based expression profile for breast cancer diagnoses. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 1151–1154. IEEE, 2014a. URL https://doi.org/10.1109/embc.2014.6943799. [p4]

C. Cava, G. Bertoli, M. Ripamonti, G. Mauri, I. Zoppis, P. A. Della Rosa, M. C. Gilardi, and I. Castiglioni. Integration of mRNA expression profile, copy number alterations, and microRNA expression levels in breast cancer to improve grade definition. *PloS one*, 9(5):e97681, 2014b. URL https://doi.org/10.1371/journal.pone.0097681. [p1]

C. Cava, I. Zoppis, M. Gariboldi, I. Castiglioni, G. Mauri, and M. Antoniotti. Combined analysis of chromosomal instabilities and gene expression for colon cancer progression inference. *Journal of Clinical Bioinformatics*, 4(1):2, 2014c. URL https://doi.org/10.1186/2043-9113-4-2. [p4]

C. Cava, G. Bertoli, and I. Castiglioni. Integrating genetics and epigenetics in breast cancer: Biological insights, experimental, computational methods and therapeutic potential. *BMC Systems Biology*, 9 (1):62, 2015. URL https://doi.org/10.1186/s12918-015-0211-x. [p1]

C. Cava, A. Colaprico, G. Bertoli, G. Bontempi, G. Mauri, and I. Castiglioni. How Interacting Pathways Are Regulated by miRNAs in Breast Cancer Subtypes. *BMC Bioinformatics*, 17(12):348, 2016. URL https://doi.org/10.1186/s12859-016-1196-1. [p2, 6]

C. Cava, A. Colaprico, G. Bertoli, A. Graudenzi, T. C. Silva, C. Olsen, H. Noushmehr, G. Bontempi, G. Mauri, and I. Castiglioni. SpidermiR: An R/Bioconductor package for integrative analysis with miRNA data. *International journal of molecular sciences*, 18(2):274, 2017. URL https://doi.org/10.3390/ijms18020274. [p2]

C. Cava, G. Bertoli, A. Colaprico, C. Olsen, G. Bontempi, and I. Castiglioni. Integration of multiple networks and pathways identifies cancer driver genes in pan-cancer analysis. *BMC Genomics*, 19(1): 25, 2018. URL https://doi.org/10.1186/s12864-017-4423-x. [p2, 6, 7]

A. Colaprico, C. Cava, G. Bertoli, G. Bontempi, and I. Castiglioni. Integrative Analysis with Monte Carlo Cross-Validation Reveals miRNAs Regulating Pathways Cross-Talk in Aggressive Breast Cancer. *BioMed Research International*, 2015(831314):17, 2015. [p2, 4, 6]

C. Desmedt, A. Giobbie-Hurder, P. Neven, R. Paridaens, M.-R. Christiaens, A. Smeets, F. Lallemand, B. Haibe-Kains, G. Viale, R. D. Gelber, and others. The Gene expression Grade Index: a potential predictor of relapse for endocrine-treated breast cancer patients in the BIG 1–98 trial. *BMC Medical Genomics*, 2(1):40, 2009. URL https://doi.org/10.1186/1755-8794-2-40. [p1]

M. Donato, Z. Xu, A. Tomoiaga, J. G. Granneman, R. G. MacKenzie, R. Bao, N. G. Than, P. H. Westfall, R. Romero, and S. Draghici. Analysis and correction of crosstalk effects in pathway analysis. *Genome research*, 23(11):1885–1893, 2013. URL https://doi.org/10.1101/gr.153551.112. [p1]

X. Dong, Y. Hao, X. Wang, and W. Tian. LEGO: a novel method for gene set over-representation analysis by incorporating network-based gene weights. *Scientific Reports*, 6:18871, 2016. URL https://doi.org/10.1038/srep18871. [p6]

S. Epskamp, A. O. Cramer, L. J. Waldorp, V. D. Schmittmann, D. Borsboom, and others. qgraph: Network Visualizations of Relationships in Psychometric Data. *Journal of Statistical Software*, 48(4): 1–18, 2012. URL https://doi.org/10.18637/jss.v048.i04. [p7]

H. Fang, B. Knezevic, K. L. Burnham, and J. C. Knight. XGR Software for Enhanced Interpretation of Genomic Summary Data, Illustrated by Application to Immunological Traits. *Genome medicine*, 8(1): 129, 2016. URL https://doi.org/10.1186/s13073-016-0384-y. [p1]

Z. Fang, W. Tian, and H. Ji. A network-based gene-weighting approach for pathway analysis. *Cell research*, 22(3):565, 2012. URL https://doi.org/10.1038/cr.2011.149. [p6]

O. Folger, L. Jerby, C. Frezza, E. Gottlieb, E. Ruppin, and T. Shlomi. Predicting selective drug targets in cancer through metabolic networks. *Molecular systems biology*, 7(1):501, 2011. URL https://doi.org/10.1038/msb.2011.35. [p1]

M. Franz, H. Rodriguez, C. Lopes, K. Zuberi, J. Montojo, G. D. Bader, and Q. Morris. GeneMANIA Update 2018. *Nucleic acids research*, 46(W1):W60–W64, 2018. [p2]

M. A. García-Campos, J. Espinal-Enríquez, and E. Hernández-Lemus. Pathway analysis: State of the art. *Frontiers in Physiology*, 6:383, 2015. URL https://doi.org/10.3389/fphys.2015.00383. [p1]

I. Ihnatova and E. Budinska. ToPASeq: An R package for topology-based pathway analysis of microarray and RNA-Seq data. *BMC bioinformatics*, 16(1):350, 2015. URL https://doi.org/10.1186/s12859-015-0763-1. [p1]

A. Mohamed, T. Hancock, C. H. Nguyen, and H. Mamitsuka. NetPathMiner: R/Bioconductor Package for Network Path Mining through Gene Expression. *Bioinformatics*, 30(21):3139–3141, 2014. [p1]

R. Nicolle, F. Radvanyi, and M. Elati. CoRegNet: Reconstruction and integrated analysis of co-regulatory networks. *Bioinformatics*, 31(18):3066–3068, 2015. [p1]

J. S. Parker, M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, and others. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8):1160, 2009. URL https://doi.org/10.1200/jco.2008.18.1370. [p1]

G. Sales, E. Calura, D. Cavalieri, and C. Romualdi. Graphite-a Bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics*, 13(1):20, 2012. URL https://doi.org/10.1186/1471-2105-13-20. [p2]

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and others. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005. URL https://doi.org/10.1073/pnas.0506580102. [p1]

W. Walter, F. Sánchez-Cabo, and M. Ricote. GOplot: An R Package for Visually Combining Expression Data with Functional Analysis. *Bioinformatics*, 31(17):2912–2914, 2015. [p7]

K. Zuberi, M. Franz, H. Rodriguez, J. Montojo, C. T. Lopes, G. D. Bader, and Q. Morris. GeneMANIA prediction server 2013 update. *Nucleic Acids Research*, 41(W1):W115–W122, 2013. [p2]

*Claudia Cava*
*Institute of Molecular Bioimaging and Physiology, National Research Council (IBFM-CNR)*
*Via F.Cervi 93,20090*
*Segrate-Milan, Italy*
*ORCiD https://orcid.org/0000-0001-7191-5417*
claudia.cava@ibfm.cnr.it

*Isabella Castiglioni*
*Institute of Molecular Bioimaging and Physiology, National Research Council (IBFM-CNR)*
*Via F.Cervi 93,20090*
*Segrate-Milan, Italy*
*ORCiD https://orcid.org/0000-0002-5540-4104*
isabella.castiglioni@ibfm.cnr.it