

Neighbor_net analysis

Sahar Ansari and Sorin Draghici

Department of Computer Science, Wayne State University, Detroit MI 48201

October 30, 2018

Abstract

This package is intended to be the R implementation of the method introduced in [1]. Neighbor_net analysis aims to take advantage of the prior knowledge of gene-gene interactions and identifies the putative mechanisms at play in the given condition (e.g. a disease, a treatment, etc.). The captured network can be useful for the prediction of mechanisms of action of drugs or the responses of an organism to a specific impact.

1 Neighbor_net analysis

Neighbor_net (`neighborNet`) is a tool to identify the active mechanism involved in an investigated phenotype. This method uses two sources of data: one is the experiment data and the other is the gene-gene interactions knowledge.

1.1 Gene-gene interaction knowledge

Neighbor_net can accept any gene-gene interaction information obtained from different databases. The data has to be converted to a list format. Each element in the list represents the neighborhood of one gene.

We provided an example that includes the interactions exist in KEGG[2] and HPRD [3] databases.

```
> load(system.file("extdata/listofgenes.RData", package = "NeighborNet"))
```

The `listofgenes` is a list including the neighbors of the genes in the analysis:

```
> head(listofgenes)
```

```
$`216`
```

```
[1] "216"
```

```
$`3679`
```

```
[1] "3679" "1134" "1398" "1399" "5747" "6714" "60" "71" "9564"
```

```
$`55607`
```

```
[1] "55607" "71" "5575" "5504" "84687" "5499" "6198"
```

```
$`5552`
```

```
[1] "5552" "960" "5196" "56893" "6449" "3002" "213"
```

```
$`2886`
 [1] "2886" "2064" "3815" "2065" "2885" "7010" "5747" "9020" "1956" "3643"
[11] "5979"
```

```
$`5058`
 [1] "5058" "2064" "5335" "5604" "660" "9459" "4690" "4638"
 [9] "3984" "5879" "5829" "668" "4771" "10095" "9181" "1457"
[17] "1459" "985" "998" "4086" "4087" "4089" "57154" "7046"
[25] "58480" "6853" "1277" "5580" "58" "572" "10818" "9815"
[33] "7048" "2316" "2099" "2308" "834" "5894" "9020" "5605"
[41] "10746" "4215" "5609" "6416" "2317" "2318" "7074" "340156"
[49] "85366" "91807" "3985" "8874"
```

1.2 Experiment data

As an example, we provided five pre-processed data sets from GEO (GSE4183, GSE9348, GSE21510, GSE32323, GSE18671).

These data study the expression change between colorectal cancer and normal patients. The data was preprocessed using the *limma* package. Only probe sets with a gene associated to them have been kept.

```
> load(system.file("extdata/dataColorectal4183.RData", package = "NeighborNet"))
> load(system.file("extdata/dataColorectal9348.RData", package = "NeighborNet"))
> load(system.file("extdata/dataColorectal21510.RData", package = "NeighborNet"))
> load(system.file("extdata/dataColorectal32323.RData", package = "NeighborNet"))
> load(system.file("extdata/dataColorectal8671.RData", package = "NeighborNet"))
> head(dataColorectal4183)
```

	adj.P.Val	logFC	EntrezID
1	0.0005849192	2.165550	27253
2	0.0005849192	1.993385	7450
3	0.0005849192	1.402015	857
4	0.0015330474	1.887886	25937
5	0.0015330474	2.220579	29767
6	0.0015330474	3.536515	285

The next step is to select the genes that are differentially expressed, with p-value lower than 1% and absolute fold change more than 1.5.

```
> pvThreshold <- 0.01
> foldThreshold <- 1.5
> de1 <- dataColorectal4183$EntrezID [
+ dataColorectal4183$adj.P.Val < pvThreshold &
+ abs(dataColorectal4183$logFC) > foldThreshold
+ ]
> de2 <- dataColorectal9348$EntrezID [
+ dataColorectal9348$adj.P.Val < pvThreshold &
+ abs(dataColorectal9348$logFC) > foldThreshold
+ ]
```

```

> de3 <- dataColorectal21510$EntrezID [
+ dataColorectal21510$adj.P.Val < pvThreshold &
+ abs(dataColorectal21510$logFC) > foldThreshold
+ ]
> de4 <- dataColorectal32323$EntrezID [
+ dataColorectal32323$adj.P.Val < pvThreshold &
+ abs(dataColorectal32323$logFC) > foldThreshold
+ ]
> de5 <- dataColorectal8671$EntrezID [
+ dataColorectal8671$adj.P.Val < pvThreshold &
+ abs(dataColorectal8671$logFC) > foldThreshold
+ ]
>

```

Later, the differentially expressed genes from different datasets should be combined together:

```

> de <- unique( c(de1,de2,de3,de4,de5))

```

The reference contains all the genes measured in the analysis:

```

> ref <- unique( c(
+ dataColorectal4183$EntrezID,
+ dataColorectal9348$EntrezID,
+ dataColorectal21510$EntrezID,
+ dataColorectal32323$EntrezID,
+ dataColorectal8671$EntrezID
+ ))
> head(ref)

[1] "27253" "7450" "857" "25937" "29767" "285"

```

1.3 Neighbor_net analysis and resulted network

We have all the input for Neighbor_net analysis.

- the gene-gene knowledge in a list format -`listofgenes`
- the experiment data -`de` and -`ref`

```

> library("NeighborNet")
> library("graph")
> sig_genes <- neighborNet(de = de, ref = ref, listofgenes=listofgenes)
> sig_genes

```

```

A graphNEL graph with undirected edges
Number of Nodes = 144
Number of Edges = 251

```

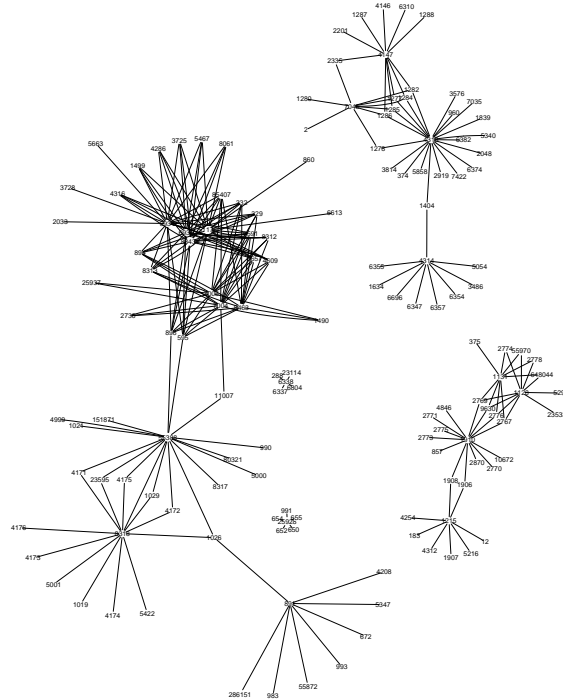


Figure 1: The active network that describes the putative mechanism involved in colorectal cancer.

1.4 Graphical representation of results

To visualize the identified network use the function `plot`(see Fig. 1):

```
> require("graph")
> attrs <- list(node= list(fontsize=40, fixedsize= FALSE), graph=list(overlap=FALSE), edge=list(
> nAttr <- list()
> nAttr$color <- c(rep("white", length(nodes(sig_genes))))
> names(nAttr$color) <- nodes(sig_genes)
> plot(sig_genes)
```

References

- [1] S. Ansari, M. Donato, N. Saberian, and S. Draghici. An approach to infer putative disease-specific mechanisms using neighboring gene networks. *Bioinformatics*, page btx097, 2017.
- [2] M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.
- [3] S. Peri, J. D. Navarro, R. Amanchy, T. Z. Kristiansen, C. K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T. Gandhi, M. Gronborg, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research*, 13(10):2363–2371, 2003.