NOA-package

Qiang Huang

August 7, 2013

1 Introduction

1.1 NOA

NOA

NOA (abbreviated to Network Ontology Analysis) is a freely available collection of Gene Ontology tools aiming to analyze functions of gene network instead of gene list. Network rewiring facilitates the function changes between conditions even with the same gene list. Therefore, it is necessary to annotate the specific function of networks by considering the fundamental roles of interactions from the viewpoint of systems biology. NOA is such a novel functional enrichment analysis method capable to handle both dynamic and static networks. The application of NOA in biological networks shows that NOA can not only capture changing functions in rewiring networks but also find more relevant and specific functions in traditional static networks.

NOA method is a collection of tools for whole-net NOA, sub-net NOA, whole-net GLM (Gene List Method), and sub-net GLM. For whole-net methods, users can input either a gene list or a gene network. Differently, for sub-net methods, in addition to the input of test gene list or test gene network, reference gene list or reference network is also necessary. It is worth to mention that reference set is required to contain test set due to the definition of reference set.

1.2 NOA package

To make the NOA method easily used, we develop its related R package, also named NOA. In this package, we extend the supported species to twenty and more supported gene IDs. NOA package implements Network Ontology Analysis and Gene List Method based on the hypergeometric distribution hypothesis testing.

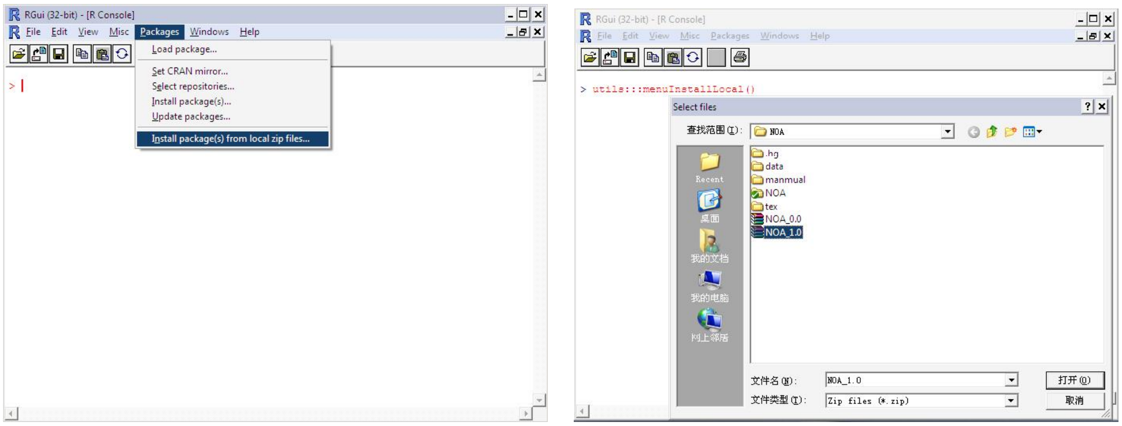
Note: This version package is only for testing, not published, if there is any problem, please contact Qiang Huang (hq04405114@gmail.com) directly.

Functions in this package:

- NOA: NOA method for network ontology analysis and gene list method.
- id.mapping: Different gene id mapping method.
- gene.term: Gene-to-term annotation method.
- species.information: The supported species information in NOA package.
- idmapping.information: The id mapping information for different species.
- result.information: The detail information about the NOA results.
- Readme: Simple introduction about NOA package.

The main reference is [2]:

Jiguang Wang, Qiang Huang, Zhi-Ping Liu, Yong Wang, Ling-Yun Wu, Luonan Chen, and Xiang-Sun Zhang. NOA: a novel Network Ontology Analysis method. Nucleic Acids Research, doi: 10.1093/nar/gkr251, 2011.



```
> utils:::menuInstallLocal()
```

文件名(N): NOA_1.0

文件类型(T): Zip files (*.zip)

(a)

(b)

Figure 1: The local installing process using Rgui.

2 Local install

Until now, NOA package is only avaiable in our NOA web site (http://app.aporc.org/NOA). The following is the specific install process for NOA package.

Download the Zip file from our web server, install NOA package by Rgui, and choose ”Packages” → ”install package(s)from local zip files...” terms and the corresponding .zip file. The installing process is simply shown in Figure 1.

3 Supporting information

To do the enrichment analysis, we should prepare the annotated relationships between genes and terms, and the mappings between the input gene id and the annotated gene id. Therefore, we need install the annotated packages firstly.

3.1 Annotation packages

The NOA R package supports twenty species. The annotation data are stored in the corresponding R packages which can be downloaded from Bioconductor [1]. If you have install NOA package, you can use the following commands to install all the annotation packages:

```
library(NOA)
install.annotation.packages()
```

The corresponding annotation R packages for all the species also can be installed by following commands:

```
source(”http://bioconductor.org/biocLite.R”)
biocLite(”org.Hs.eg.db”)
biocLite(”org.Sc.sgd.db”)
biocLite(”org.Dm.eg.db”)
biocLite(”org.Ce.eg.db”)
```

ABBREVIATION	**IDs**
ACCNUM	GenBank accession numbers
ALIAS	Common gene symbol or Alias Gene Names
COMMON	Yeast common names
ENSEMBL	Ensembl gene accession numbers
ENSEMBLPROT	Ensembl protein accession numbers
Entrez	Entrez Gene identifiers
GENENAME	Gene name
ORF	ORF identifiers
PUBMED	PubMed identifiers
REFSEQ	RefSeq identifiers
SGD	SGD accession numbers
SMART	SMART identifiers
SYMBOL	Gene abbreviations or Gene symbol
TAIR	TAIR identifiers
UNIPROT	Uniprot accession numbers

Table 1: Gene ID abbreviations.

```
biocLite(”org.Mm.eg.db”)
biocLite(”org.At.tair.db”)
biocLite(”org.Rn.eg.db”)
biocLite(”org.Dr.eg.db”)
biocLite(”org.Bt.eg.db”)
biocLite(”org.Cf.eg.db”)
biocLite(”org.Ag.eg.db”)
biocLite(”org.EcSakai.eg.db”)
biocLite(”org.Gg.eg.db”)
biocLite(”org.Pt.eg.db”)
biocLite(”org.Pf.plasmo.db”)
biocLite(”org.Mmu.eg.db”)
biocLite(”org.Sco.eg.db”)
biocLite(”org.Ss.eg.db”)
biocLite(”org.Xl.eg.db”)
biocLite(”org.EcK12.eg.db”)
```

3.2 Gene IDs

To support the different input gene IDs, the ID maps of different species which are in the annotation R packages are used here, too. For different gene IDs, we use different abbreviations. The corresponding relations are shown in Table 1.

3.3 Species information

The corresponding relationships of R packages, supported gene ids with the species are shown in Table 2.

R package	Species	Supported Gene IDs
org.Hs.eg.db	Human	ACCNUM,ALIAS,ENSEMBL,ENSEMBLPROT,SYMBOL,UNIPROT,REFSEQ,Entrez
org.Sc.sgd.db	Yeast	ALIAS,ENSEMBL,COMMON,GENENAME,SGD,SMART,UNIPROT,REFSEQ,PUBMED,Entrez,ORF
org.Dm.eg.db	Fly	ACCNUM,ALIAS,ENSEMBL,ENSEMBLPROT,SYMBOL,UNIPROT,REFSEQ,Entrez
org.Ce.eg.db	Worm	ACCNUM,ALIAS,ENSEMBL,ENSEMBLPROT,SYMBOL,UNIPROT,REFSEQ,Entrez
org.Mm.eg.db	Mouse	ACCNUM,ALIAS,ENSEMBL,ENSEMBLPROT,SYMBOL,UNIPROT,REFSEQ,Entrez
org.At.tair.db	Arabidopsis	Entrez,SYMBOL,REFSEQ,TAIR
org.Rn.eg.db	Rat	ACCNUM,ALIAS,ENSEMBL,ENSEMBLPROT,SYMBOL,UNIPROT,REFSEQ,Entrez
org.Dr.eg.db	Zebrafish	ACCNUM,ALIAS,ENSEMBL,ENSEMBLPROT,SYMBOL,UNIPROT,REFSEQ,Entrez
org.Bt.eg.db	Bovine	ACCNUM,ALIAS,ENSEMBL,ENSEMBLPROT,SYMBOL,UNIPROT,REFSEQ,Entrez
org.Cf.eg.db	Canine	ACCNUM,ALIAS,ENSEMBL,ENSEMBLPROT,SYMBOL,UNIPROT,REFSEQ,Entrez
org.Ag.eg.db	Anopheles	ACCNUM,ENSEMBL,ENSEMBLPROT,GENENAME,SYMBOL,UNIPROT,REFSEQ,Entrez
org.EcSakai.eg.db	E coli strain Sakai	ACCNUM,ALIAS,SYMBOL,REFSEQ,Entrez
org.Gg.eg.db	Chicken	ACCNUM,ALIAS,ENSEMBL,ENSEMBLPROT,SYMBOL,UNIPROT,REFSEQ,Entrez
org.Pt.eg.db	Chimp	ACCNUM,ENSEMBL,ENSEMBLPROT,SYMBOL,UNIPROT,REFSEQ,Entrez
org.Pf.plasmo.db	Malaria	ALIAS,SYMBOL,ORF
org.Mmu.eg.db	Rhesus	ACCNUM,ENSEMBL,ENSEMBLPROT,SYMBOL,UNIPROT,REFSEQ,Entrez
org.Sco.eg.db	Streptomyces coelicolor	ACCNUM,ALIAS,SYMBOL,REFSEQ,Entrez
org.Ss.eg.db	Pig	ACCNUM,ALIAS,SYMBOL,UNIPROT,REFSEQ,Entrez
org.Xl.eg.db	Xenopus	ACCNUM,SYMBOL,UNIPROT,REFSEQ,Entrez
org.EcK12.eg.db	E coli strain K12	ACCNUM,ALIAS,SYMBOL,REFSEQ,Entrez

Table 2: Species information.

3.4 File format

To do the gene list based enrichment analysis, a gene list file should be given. To do the network based enrichment analysis, a gene interaction network should be given. The reference gene list or network file can also be given which is optional.
Gene list file format: each line contains a gene.

gene1
gene2
gene3
...

Gene interaction network file format: each line contains a gene interaction which is separated by a space or tab.

gene1 gene2
gene3 gene4
...

Where gene1, gene2, gene3, gene4, ... are gene names, each gene interaction is represented by a pair of genes.

4 How to use?

4.1 Example

Arguments

NOA(upfile,species,idinput)
NOA(upfile,species,idinput,reffile)

NOA(upfile,species,idinput,reffile,evidences)

- upfile: The test gene network or gene list file.
- species: The species for the test gene network or gene list.
- idinput: The gene id in the test gene network or gene list.
- reffile: The reference gene network or gene list file.
- evidences: The GO ontology evidences for the test gene network or list.

Example

```
library(NOA)
data("subnet")
data("refnet")
write.table(subnet,file="upfile",quote=FALSE,row.names=FALSE,col.names=FALSE)
write.table(refnet,file="reffile",quote=FALSE,row.names=FALSE,col.names=FALSE)
species="Yeast"
idinput="SGD"
result = NOA("upfile",species,idinput,"reffile")
```

4.2 Result information

The NOA function returns a list as its result. The result list contains two sub lists named as "noaresult" and "glmresult" which are corresponding to noa method and glm (gene list method) results, respectively. When the input is a gene list file, the "noaresult" list is NULL.

- noaresult: all the information of NOA method, if the input is a gene list, it is NULL.
- glmresult: all the information of GLM method.

You can use the following commands to see the names of result lists and its sub lists.

```
names(result)
names(result$noaresult)
names(result$glmresult)
```

The "noaresult" and "glmresult" lists have the same variable names. The specific result information can be found by command:

```
result.information()
```

Where the specific explains for the variables in the sub lists are as follows:

- genepair: the input gene nework edges or gene list of upfile.
- method: 1 means that GLM(gene list method); 2 means that NOA.
- geneset: the unique gene set contains in the upfile.

Species	Input IDs	Mapped ID
Human	ACCNUM,ALIAS,ENSEMBL,ENSEMBLPROT,SYMBOL,UNIPROT,REFSEQ,Entrez	Entrez
Fly	ACCNUM,ALIAS,ENSEMBL,ENSEMBLPROT,SYMBOL,UNIPROT,REFSEQ,Entrez	Entrez
Worm	ACCNUM,ALIAS,ENSEMBL,ENSEMBLPROT,SYMBOL,UNIPROT,REFSEQ,Entrez	Entrez
Mouse	ACCNUM,ALIAS,ENSEMBL,ENSEMBLPROT,SYMBOL,UNIPROT,REFSEQ,Entrez	Entrez
Rat	ACCNUM,ALIAS,ENSEMBL,ENSEMBLPROT,SYMBOL,UNIPROT,REFSEQ,Entrez	Entrez
Zebrafish	ACCNUM,ALIAS,ENSEMBL,ENSEMBLPROT,SYMBOL,UNIPROT,REFSEQ,Entrez	Entrez
Bovine	ACCNUM,ALIAS,ENSEMBL,ENSEMBLPROT,SYMBOL,UNIPROT,REFSEQ,Entrez	Entrez
Canine	ACCNUM,ALIAS,ENSEMBL,ENSEMBLPROT,SYMBOL,UNIPROT,REFSEQ,Entrez	Entrez
Chicken	ACCNUM,ALIAS,ENSEMBL,ENSEMBLPROT,SYMBOL,UNIPROT,REFSEQ,Entrez	Entrez
Yeast	ALIAS,ENSEMBL,COMMON,GENENAME,SGD,SMART,UNIPROT,REFSEQ,PUBMED,Entrez,ORF	ORF
Malaria	ALIAS,SYMBOL,ORF	ORF
Arabidopsis	Entrez,SYMBOL,REFSEQ,TAIR	TAIR
Anopheles	ACCNUM,ENSEMBL,ENSEMBLPROT,GENENAME,SYMBOL,UNIPROT,REFSEQ,Entrez	Entrez
E coli strain Sakai	ACCNUM,ALIAS,SYMBOL,REFSEQ,Entrez	Entrez
Chimp	ACCNUM,ENSEMBL,ENSEMBLPROT,SYMBOL,UNIPROT,REFSEQ,Entrez	Entrez
Rhesus	ACCNUM,ENSEMBL,ENSEMBLPROT,SYMBOL,UNIPROT,REFSEQ,Entrez	Entrez
Streptomyces coelicolor	ACCNUM,ALIAS,SYMBOL,REFSEQ,Entrez	Entrez
Pig	ACCNUM,ALIAS,SYMBOL,UNIPROT,REFSEQ,Entrez	Entrez
Xenopus	ACCNUM,SYMBOL,UNIPROT,REFSEQ,Entrez	Entrez
E coli strain K12	ACCNUM,ALIAS,SYMBOL,REFSEQ,Entrez	Entrez

Table 3: Gene ID mapping information

- refgenepair: the reference gene network edges or gene list of reffile. If reffile is NULL, for GLM, the refgenepair is all the genes in the species, for NOA, the reference network is the clique network.
- refgeneset: the unique gene set contains in the refgenepair.
- mapgeneset: the mapped gene set for geneset. In enrichment analysis, we should map input gene id into the annotated gene id.
- maprefgeneset: the mapped gene set for refgeneset.
- unmappedgene: the gene set that can not map to the annotated id.
- term: the enrichment analysis result which is a matrix. Each row is corresponding to one term.

5 Other information

In NOA package, we also export the id.mapping and gene.term method to use solely.

5.1 Gene ID mapping

There are several gene IDs to do the id mapping in a species. Simply, we only map all the gene IDs (Input IDs) into a specific gene ID (Mapped ID) for each species. The specific information can be found in Table 3.
Arguments
id.mapping(genelist,id1,id2,species)

- genelist: The input gene list as a character vector.
- id1: The input gene id.
- id2: The mapped gene id.

- species: The species of input gene list.

Example

```
library(NOA)
data(”subnet”)
genelist = unique(as.matrix(subnet))
species=”Yeast”
id1=”SGD”
id2=”ORF”
maps = id.mapping(genelist,id1,id2,species)
maps
```

5.2 Gene annotation

The gene annotation is used to extract the annotated GO terms for a given gene list. Here, we need the input gene ID is the mapped ID described as above. If your gene list is not the corresponding gene ID, id mapping is need.
Arguments
gene.term(geneset,species)
gene.term(geneset,species,evidences)

- geneset: The gene set for annotation.
- species: The species of the gene set.
- evidences: The annotation evidences for gene set, which is a character vector. The default is all evidences. More information can be found in http://www.geneontology.org/GO.evidence.shtml.

Example

```
library(NOA)
data(”subnet”)
genelist = unique(as.matrix(subnet))
species = ”Yeast”
id1 = ”SGD”
id2 = ”ORF”
maps = id.mapping(genelist,id1,id2,species)
annotation = gene.term(maps[,id2],species)
annotation
```

References

[1] Robert C Gentleman, Vincent J. Carey, Douglas M. Bates, and others. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

[2] J. Wang, Q. Huang, Z.P. Liu, Y. Wang, L.Y. Wu, L. Chen, and X.S. Zhang. Noa: a novel network ontology analysis method. *Nucleic acids research*, 39(13):e87–e87, 2011.