How To Use PROFANCY

Qianlan Yao, Desi Shang, Chunquan Li

July 12, 2013

Contents

1 Overview

This vignette demonstrates how to easily use the PROFANCY package. This package can prioritize candidate disease metabolites by random walk analysis in the metabolites network with virtual pathway nodes. This system provides the rank of candidate diseases metabolites which coud be provided by users or this system.

2 Get all disease names provided by PROFANCY package

The section introduces how to get all disease names provided by PROFANCY package. PROFANCY package provides 71 diseases which with known disease metabolites. The function `getProvidedDisease-Name` could help users get the disease list. If the interest disease in this disease list, users could use the default seeds(disease known metabolites) in /Rfunction(getTopDiseaseMetabolites) without input their own. The following commands can get the disease list in R.

```
> #get path of the KGML files
> ProvidedDiseaseName<-getProvidedDiseaseName()
> print(ProvidedDiseaseName[1:5])

[1] Adenylosuccinate lyase deficiency
[2] Alzheimer's disease
[3] Subarachnoid hemorrhage
[4] Aromatic L-amino acid decarboxylase deficiency
[5] Primary biliary cirrhosis
71 Levels: 2-ketoadipic acidemia ... Transaldolase deficiency
```

```
> length(ProvidedDiseaseName)

[1] 71
```

3 The methods of prioritizing the disease metabolites

The function `getTopDiseaseMetabolites` could get the top ranked candidate metaoblites by prioritization of the disease candidate metabolites using known disease metabolites as seed node to perform random walk on metabolite network with virtual pathway nodes.

In this function, there are seven parameters. "diseaseName" is a character of the name of the disease users want to study. "network" is a charecter. Which type of metabolites network should be chose. Users could choose KEGG or EHMN network. The defult value is "KEGG" network."seed" is charecter vector. The seed metaboites are the known disease metabolites are used in random walk analysis on the network. If users have set the seedDefault parameter TURE and selceted diseaseName provided by system, users do not need to set this parameter. Otherwise, users should input seed metabolites. "candidates" is a charecter vector. If users set the candidateDefault parameter FALSE, users should input candidates. If users do not set this parameter, the candidate metabolites are all metabolites in the network except seed metabolites. "seedDefault" is a logical value(TRUE or FALSE). If users set TRUE and have selected the diseaseName provided by system, the seed metabolites are default known disease metabolites. Otherwise, users shoud input seed metabolites. "candidateDefault" is a logical value(TRUE or FALSE). If TRUE, the metabolites in network except seed metabolites will be prioritized. Otherwise, users shoud input candidate metabolites. "showTop" is an integer. The number of top ranked candidate metabolites users want to show.

The return value of this funcion is a dataframe which contained the top ranked candidate metaoblites and some information of them.

If users choose the disease name from provided in `getProvidedDiseaseName` and do not want to input seeds, then users should set the seedDefault TRUE. If users set candidateDefault is TRUE, users should not set the candidates parameter, the default candidate metabolites would be used in this condition.

The following commands could get the top ranked candidate metaoblites.

```
> ##Example 1: Users have chose disease name provided by system and network
> ## is set "EHMN". The seeds and candidates are default.
> ProvidedDiseaseName<-getProvidedDiseaseName() ##get all disease name provided.
> diseaseName<-ProvidedDiseaseName[17]  ##choose the disease you want to study.
> example1<-getTopDiseaseMetabolites(diseaseName=diseaseName,network="EHMN"
+           ,seedDefault=TRUE,showTop=30,candidateDefault=TRUE)

     [,1]
[1,] "Disease Name is: Prostate cancer"
[2,] "The network you used is : EHMN"
[3,] "The number of the seeds you input are: 0"
[4,] "The number of the seeds used in prioritizing the candidate metabolites are:  11"
[5,] "The seeds used in prioritizing the candidate metabolites are:  C00198;C01103;C00624;C00438;C0123
[6,] "The number of the candidate metabolites you input are:  0"
[7,] "The number of the candidate metabolites are prioritizing in this method are:  1555"

> print(example1[1:5,])

       Rank KEGGID
C00049    1 C00049
C01507    2 C01507
C04540    3 C04540
```

```
C00221    4 C00221
C04257    5 C04257
                                                                                MetaboliteName
C00049                                          L-asparticacid;L-aspartate;2-aminosuccinicacid
C01507                                                                              L-Iditol
C04540 N4-(acetyl-beta-D-glucosaminyl)asparagine;1-beta-aspartyl-N-acetyl-D-glucosaminylamine
C00221                                                                          beta-D-glucose
C04257                       N-Acetylmannosamine6-phosphate;N-Acetyl-D-mannosamine6-phosphate
            Score
C00049 0.015604047
C01507 0.010025750
C04540 0.009927151
C00221 0.009917939
C04257 0.009847467
```

```
> #write.table(example1,"example1.txt",quote=FALSE,row.names=FALSE,sep="\t")
>
>
> ##Example 2: The disease name is provided by users. The seeds and candidates
> ## are provided by users. The network is set "KEGG".
> diseaseName<-"prostate cancer"  ##the disease name provided by users.
> path1<-paste(system.file(package="PROFANCY"),"/localdata/ProstateCandidates.txt",sep="")
> candidateExample<-read.table(path1)
> candidateExample<-candidateExample[[1]]
> path2<-paste(system.file(package="PROFANCY"),"/localdata/ProstateSeeds.txt",sep="")
> seedExample<-read.table(path2)
> seedExample<-seedExample[[1]]
> example2<-getTopDiseaseMetabolites(diseaseName,network="KEGG",seed=seedExample,
+       candidates=candidateExample,seedDefault=FALSE,showTop=30,candidateDefault=FALSE)
```

```
     [,1]
[1,] "Disease Name is: prostate cancer"
[2,] "The network you used is : KEGG"
[3,] "The number of the seeds you input are:  10"
[4,] "The number of the seeds used in prioritizing the candidate metabolites are:  10"
[5,] "The seeds used in prioritizing the candidate metabolites are:  C00198;C01103;C00624;C00438;C0046
[6,] "The number of the candidate metabolites you input are:  105"
[7,] "The number of the candidate metabolites are prioritizing in this method are:  104"
```

```
> print(example2[1:5,])
```

```
       Rank KEGGID                                    MetaboliteName
C00794    1 C00794                  sorbitol;D-Sorbitol;D-Glucitol
C00137    2 C00137              myo-inositol;inositol;meso-Inositol
C00031    3 C00031                     dextrose;D-glucose;glucose
C00049    4 C00049   L-asparticacid;L-aspartate;2-aminosuccinicacid
C00270    5 C00270 N-acetylneuraminate;Neu5Ac;N-acetylneuraminicacid
            Score
C00794 0.012771185
C00137 0.005670056
C00031 0.005502823
C00049 0.004452486
C00270 0.003657003
```

```
> #write.table(example2,"example2.txt",quote=FALSE,row.names=FALSE,sep="\t")
>
>
>
> ##Example 3: Disease name chose in our provided. The seeds are provided by default.
> ## The candidates are provided by users. Network is set "EHMN".
> path1<-paste(system.file(package="PROFANCY"),"/localdata/ProstateCandidates.txt",sep="")
> candidateExample<-read.table(path1)
> candidateExample<-candidateExample[[1]]
> example3<-getTopDiseaseMetabolites(diseaseName="Prostate cancer",network="EHMN",
+           candidates=candidateExample,seedDefault=TRUE,showTop=30,candidateDefault=FALSE)

     [,1]
[1,] "Disease Name is: Prostate cancer"
[2,] "The network you used is : EHMN"
[3,] "The number of the seeds you input are: 0"
[4,] "The number of the seeds used in prioritizing the candidate metabolites are:  11"
[5,] "The seeds used in prioritizing the candidate metabolites are:  C00198;C01103;C00624;C00438;C0123
[6,] "The number of the candidate metabolites you input are:  105"
[7,] "The number of the candidate metabolites are prioritizing in this method are:  104"

> print(example3[1:5,])

       Rank KEGGID
C00049    1 C00049
C00137    2 C00137
C00794    3 C00794
C00074    4 C00074
C00093    5 C00093
                                                        MetaboliteName
C00049                    L-asparticacid;L-aspartate;2-aminosuccinicacid
C00137                              myo-inositol;inositol;meso-Inositol
C00794                                   sorbitol;D-Sorbitol;D-Glucitol
C00074                          phosphoenolpyruvate;Phosphoenolpyruvicacid
C00093 Glycerophosphoricacid;glycerol-3-phosphate;sn-Glycerol3-phosphate
             Score
C00049 0.0156040466
C00137 0.0073236474
C00794 0.0036650176
C00074 0.0012051852
C00093 0.0008217097

> #write.table(example3,"example3.txt",quote=FALSE,row.names=FALSE,sep="\t")
>
>
>
> ##Example 4: Disease name chose in our provided. The seeds are provided by users.
> ##The candidates are provided by default. Network is set "KEGG".
> path2<-paste(system.file(package="PROFANCY"),"/localdata/ProstateSeeds.txt",sep="")
> seedExample<-read.table(path2)
> seedExample<-seedExample[[1]]
> example4<-getTopDiseaseMetabolites(diseaseName="Prostate cancer",network="KEGG",
+                  seed=seedExample,seedDefault=FALSE,showTop=30,candidateDefault=TRUE)
```

```
     [,1]
[1,] "Disease Name is: Prostate cancer"
[2,] "The network you used is : KEGG"
[3,] "The number of the seeds you input are:  10"
[4,] "The number of the seeds used in prioritizing the candidate metabolites are:  10"
[5,] "The seeds used in prioritizing the candidate metabolites are:  C00198;C01103;C00624;C00438;C0046
[6,] "The number of the candidate metabolites you input are:  0"
[7,] "The number of the candidate metabolites are prioritizing in this method are:  3555"
```

```
> print(example4[1:5,])
```

```
       Rank KEGGID
C00794    1 C00794
C02888    2 C02888
C00025    3 C00025
C04133    4 C04133
C00365    5 C00365
                                                    MetaboliteName
C00794                                  sorbitol;D-Sorbitol;D-Glucitol
C02888            L-Sorbose1P;Sorbose1-phosphate;L-sorbose1-phosphate
C00025                    L-glutamicacid;L-glutamate;L-Glutaminicacid
C04133 N-Acetyl-L-glutamyl5-phosphate;N-Acetyl-L-glutamate5-phosphate
C00365               dUMP;Deoxyuridylicacid;Deoxyuridinemonophosphate
           Score
C00794 0.012771185
C02888 0.012671797
C00025 0.007321989
C04133 0.007188003
C00365 0.006037351
```

```
> #write.table(example4,"example4.txt",quote=FALSE,row.names=FALSE,sep="\t")
```

4 getSeed

This funtion can get seed metabolites (known disease metabolites). If seedDefault is TRUE, this system provided default seed metabolites. Otherwise, users should input the seed metabolites and the seed metabolites are inputting metabolite which are in the metabolites network.

```
> ##############    Get disease seed metabolites   ################
> ##Example 1: Users have chose disease name provided by system (Prostate cancer),
> ##network is set "EHMN". The seeds are provided by default.
> seed1<-getSeed(diseaseName="Prostate cancer",network="EHMN",seedDefault=TRUE)
> print(seed1[1:5])
```

```
[1] "C00198" "C01103" "C00624" "C00438" "C01239"
```

```
> ##Example 2: The disease name is provided by users. The seeds are provided by users.
> ##The network is set "KEGG".
> diseaseName<-"prostate cancer"  ##the disease name provided by users.
> path2<-paste(system.file(package="PROFANCY"),"/localdata/ProstateSeeds.txt",sep="")
> seedExample<-read.table(path2)
> seedExample<-seedExample[[1]]
```

```
> seed2<-getSeed(diseaseName="Prostate cancer",network="EHMN",seed=seedExample,seedDefault=FALSE)
> print(seed2[1:5])
[1] "C00198" "C01103" "C00624" "C00438" "C00460"
```

5 getCandidates

This funtion can get candidate metabolites to be prioritized. If candidateDefault is TRUE, this system provided default candidate metabolites which is the metabolites in metabolites network except seed metabolites. Otherwise, users could input candidate metabolites. Then the candidate metabolites are the inputting metabolites which are in the metabolites network.

```
> ##############    Get disease candidates   ################
> ##Example 1: Users have chose disease name provided by system (Prostate cancer),
> ## network is set to "EHMN".
> ##The candidates and seed metabolites are default.
> Candidates1<-getCandidates(diseaseName="Prostate cancer",network="EHMN",
+                  seedDefault=TRUE,candidateDefault=TRUE)
> print(Candidates1[1:5])
[1] "C01083" "C00031" "C05125" "C00068" "C01674"
> ##Example 2: The disease name is input by users. The seeds are provided by users.
> ##The network is set to "KEGG".The candidates are provided by users.
> path1<-paste(system.file(package="PROFANCY"),"/localdata/ProstateCandidates.txt",sep="")
> candidateExample<-read.table(path1)
> candidateExample<-candidateExample[[1]]
> path2<-paste(system.file(package="PROFANCY"),"/localdata/ProstateSeeds.txt",sep="")
> seedExample<-read.table(path2)
> seedExample<-seedExample[[1]]
> Candidates2<-getCandidates(diseaseName="Prostate cancer",network="KEGG",seed=seedExample,
+             seedDefault=FALSE,candidates=candidateExample,candidateDefault=FALSE)
> print(Candidates2[1:5])
[1] "C00794" "C00137" "C00049" "C00031" "C00270"
```

6 Data management

The environment variable `envData`, which is used as the database of the system, The environment variable envData save many information. We can use the function `ls` to see the variable and use `ls(envData)` to see information in it, which include `DiseaseInfList`, `EHMNAddPathInfNetwork`, `KEGGAddPathInfNetwork`, `MetaboliteInf`, `ProstateCandidates`, `ProstateSeeds` etc. We can use the function `get` to obtain one of them.

```
> ##data in environment variable envData
> ls(envData)
[1] "DiseaseInfList"        "EHMNAddPathInfNetwork" "KEGGAddPathInfNetwork"
[4] "MetaboliteInf"         "ProstateCandidates"    "ProstateSeeds"
```

We can obtain these data in the environment variable `envData` using the function `get`. The following command gets information of all metabolites in the variable `MetaboliteInf` in R.

```
> #get information of all metabolites in the network
> MetaboliteInf<-get("MetaboliteInf",envir=envData)
```