Tutorial of KGMN result visualization and analysis

Zhiwei Zhou

2022-06-05

Introduction

Unknown metabolite annotation is one of long-standing challenges in untargeted metabolomics. We developed an approach, namely, knowledge-guided multi-layer network (KGMN), to enable global metabolite annotation from knowns to unknowns in untargeted metabolomics. The KGMN approach integrated three-layer networks, including knowledge-based metabolic reaction network (Network 1), knowledge-guided MS/MS similarity network (Network 2), and global peak correlation network (Network 3). This tutorial will help users to visualize, reproduce and investigate putatively annotated known and unknown metabolites from KGMN.

1. Installation

The analysis and visualization of KGMN results mainly relies on R package – MetDNA2Vis, and its depended R packages; The Cytoscape software is used for manually visualize networks, and interactively investigate results of KGMN; The ChemDraw software is involved for drawing chemical structures.

- Install R packages

```
# Install related packages
if(!require(devtools)){
install.packages("devtools")
}

if(!require(BiocManager)){
install.packages("BiocManager")
}

# Install CRAN/Bioconductor packages
required_pkgs <- c("dplyr","tidyr","readr","CHNOSZ","igraph",
  "magrittr","ggplot2","ggraph","tidygraph")
list_installed <- installed.packages()

new_pkgs <- required_pkgs[!(required_pkgs %in% list_installed[,'Package'])]
```

```
if (length(new_pkgs) > 0) {
  BiocManager::install(new_pkgs)
} else {
  cat('Required CRAN/Bioconductor packages installed\n')
}


# Install ZhuLab packages
devtools::install_github("ZhuMetLab/SpectraTools")
devtools::install_github("ZhuMetLab/MetDNA2Vis")
```

- Cytoscape software (Version 3.8 or higher required): https://cytoscape.org/
- ChemDraw software (Version 19.0 or higher required): https://perkinelmerinformatics.com/products/research/chemdraw

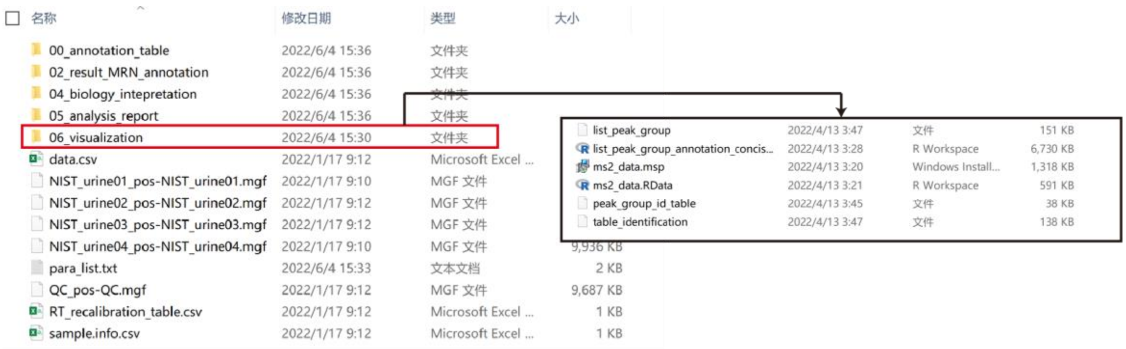
2. Step-by-step instruction for visualization

In this part, we introduce how to visualize multi-layer networks from KGMN. It will help users to reproduce figures in KGMN manuscripts. Here, the Human NIST urine (Positive data, used in KGMN manuscript) was used as a demo dataset. This data set have been processed and exported by **MetDNA2 web server** (version 1.0.4). The raw data files and results can be downloaded at **here**. The more details of sample extraction and data preprocessing can be found in our KGMN manuscript.

The step-by-step demonstration is provided as below.

2.1 Download demo data and unzip the archive.

- All required intermediate files for visualization is provided in '06_visualization' folder.



名称	修改日期	类型	大小
00_annotation_table	2022/6/4 15:36	文件夹	
02_result_MRN_annotation	2022/6/4 15:36	文件夹	
04_biology_intepretation	2022/6/4 15:36	文件夹	
05_analysis_report	2022/6/4 15:36	文件夹	
06_visualization	2022/6/4 15:30	文件夹	
data.csv	2022/1/17 9:12	Microsoft Excel ...	
NIST_urine01_pos-NIST_urine01.mgf	2022/1/17 9:10	MGF 文件	
NIST_urine02_pos-NIST_urine02.mgf	2022/1/17 9:12	MGF 文件	
NIST_urine03_pos-NIST_urine03.mgf	2022/1/17 9:12	MGF 文件	
NIST_urine04_pos-NIST_urine04.mgf	2022/1/17 9:10	MGF 文件	9,936 KB
para_list.txt	2022/6/4 15:33	文本文档	2 KB
QC_pos-QC.mgf	2022/1/17 9:12	MGF 文件	9,687 KB
RT_recalibration_table.csv	2022/1/17 9:12	Microsoft Excel ...	1 KB
sample.info.csv	2022/1/17 9:12	Microsoft Excel ...	1 KB

list_peak_group	2022/4/13 3:47	文件	151 KB
list_peak_group_annotation_concis...	2022/4/13 3:28	R Workspace	6,730 KB
ms2_data.msp	2022/4/13 3:20	Windows Install...	1,318 KB
ms2_data.RData	2022/4/13 3:21	R Workspace	591 KB
peak_group_id_table	2022/4/13 3:45	文件	38 KB
table_identification	2022/4/13 3:47	文件	138 KB

2.2 Preparing.

- Set the working directory (‘your_path/06_visualization’) and load required packages. Then, please check required files whether existed.

```
# load packages
library(MetDNA2Vis)
library(CHNOSZ)
library(dplyr)

# check required files
checkFiles4Vis()

## Check required files ...
## Check required files: done!
```

2.3 Reconstruct and export global multi-layer networks.

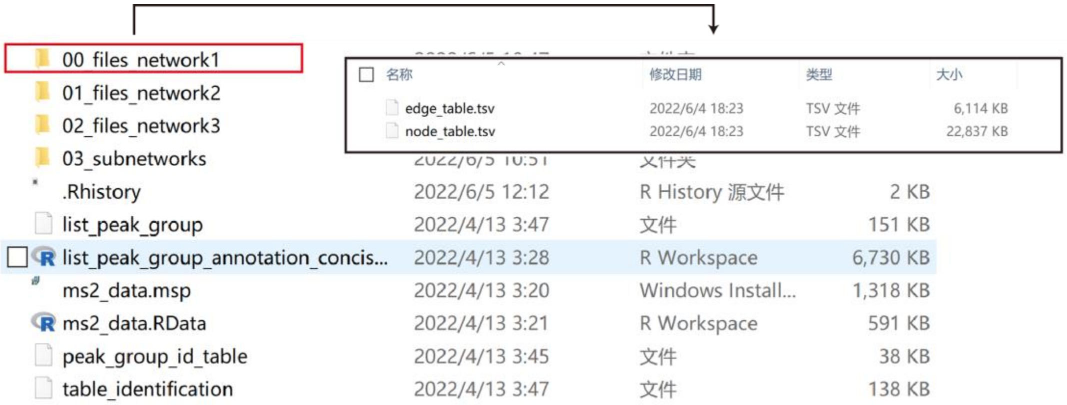
2.3.1 Network 1

The network 1 is the knowledge-guided metabolic reaction network. For knowns, the KEGG reaction pair network is directly used. For unknowns, an extended KEGG reaction pair network was used. The network expansion is performed with in-silico enzymic reactions (via Biotransformer), and further connected with KEGG reaction pair network. The details of network construction and expansion are described in our KGMN manuscript. It should be note that the KEGG reaction pair network and extended network were built in advance.

To export the network 1, it is easily to run reconstructNetwork1 function as below:

```
# export network 1 for visualization
reconstructNetwork1(is_unknown_annotation = TRUE)
```

The networks files will be exported in ‘00_network1’ folder. It contains two files, including “edge_table.tsv” and “node_table.tsv” (Figure 2.3.1). These tables can be import into Cytoscape software for visualization.



名称	修改日期	类型	大小
edge_table.tsv	2022/6/4 18:23	TSV 文件	6,114 KB
node_table.tsv	2022/6/4 18:23	TSV 文件	22,837 KB

00_files_network1			
01_files_network2			
02_files_network3			
03_subnetworks	2022/6/5 10:51	文件夹	
.Rhistory	2022/6/5 12:12	R History 源文件	2 KB
list_peak_group	2022/4/13 3:47	文件	151 KB
list_peak_group_annotation_concis...	2022/4/13 3:28	R Workspace	6,730 KB
ms2_data.msp	2022/4/13 3:20	Windows Install...	1,318 KB
ms2_data.RData	2022/4/13 3:21	R Workspace	591 KB
peak_group_id_table	2022/4/13 3:45	文件	38 KB
table_identification	2022/4/13 3:47	文件	138 KB

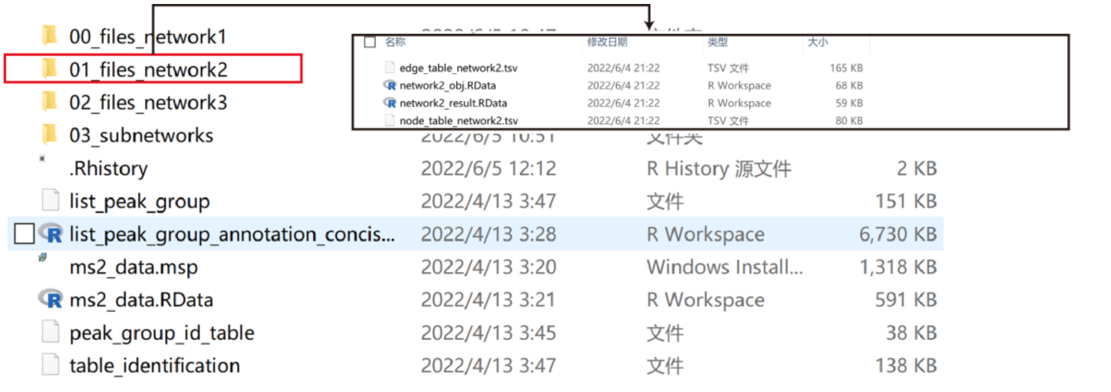
2.3.2 Network 2

The network 2 is a knowledge-guided MS/MS network. Although it calls MS/MS network, differing to MS/MS network (mainly based on MS2), the linkage (edge) of network2 has a prerequisite. It requires a reasonable reaction relationship and definitive structure candidate first. As a result, their retention time can also be predicted. In other words, two linked nodes indicate 4 messages. Their candidates of these nodes have (1) reasonable reaction relationships, (2) low m/z errors, (3) low RT error against with predicted RT values, and (4) MS/MS similarity. It should be note that optimized network2 required to be reconstructed from KGMN exported results, because the global peak correlation network remove and collapse some error nodes and edges in prior analysis. This process usually requires 10-20 min to complete.

To export the network 2, it is easily to run reconstructNetwork2 function as below:

```
# Modify format of KGMN result
annotation_table <- reformatTable1()

# Export global network2 files
reconstructNetwork2(annotation_table = annotation_table,
  is_unknown_annotation = TRUE)
```

The networks files will be exported in ‘01_network2’ folder. The “edge_table.tsv” and “node_table.tsv” in this folder can be imported to Cytoscape.



名称	修改日期	类型	大小
edge_table_network2.tsv	2022/6/4 21:22	TSV 文件	165 KB
network2_obj.RData	2022/6/4 21:22	R Workspace	68 KB
network2_result.RData	2022/6/4 21:22	R Workspace	59 KB
node_table_network2.tsv	2022/6/4 21:22	TSV 文件	80 KB

00_files_network1			
01_files_network2			
02_files_network3			
03_subnetworks			
.Rhistory	2022/6/5 12:12	R History 源文件	2 KB
list_peak_group	2022/4/13 3:47	文件	151 KB
list_peak_group_annotation_concis...	2022/4/13 3:28	R Workspace	6,730 KB
ms2_data.msp	2022/4/13 3:20	Windows Install...	1,318 KB
ms2_data.RData	2022/4/13 3:21	R Workspace	591 KB
peak_group_id_table	2022/4/13 3:45	文件	38 KB
table_identification	2022/4/13 3:47	文件	138 KB

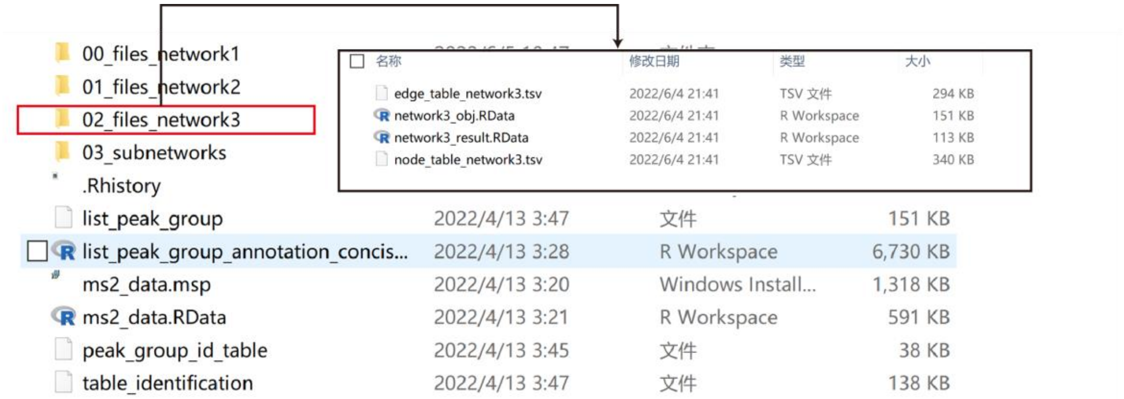
2.3.3 Network 3

The network 3 is the global peak correlation network. This network recognized abiotic peaks derived from peaks from network 2, including adducts, isotopes, neutral losses, and in-source fragments (ISF). The network 3 is used to optimize the annotation and linkage of network 2. The optimization has been completed in KGMN analysis. The details of network 3 construction and optimization can be found in our manuscript.

To export the network 3, it is easily to run reconstructNetwork3 function as below:

```
# export network3
reconstructNetwork3()
```

The networks files will be exported in ‘**02_files_network3**’ folder. The “edge_table.tsv” and “node_table.tsv” in this folder can be imported to Cytoscape for visualization.



名称	修改日期	类型	大小
edge_table_network3.tsv	2022/6/4 21:41	TSV 文件	294 KB
network3_obj.RData	2022/6/4 21:41	R Workspace	151 KB
network3_result.RData	2022/6/4 21:41	R Workspace	113 KB
node_table_network3.tsv	2022/6/4 21:41	TSV 文件	340 KB

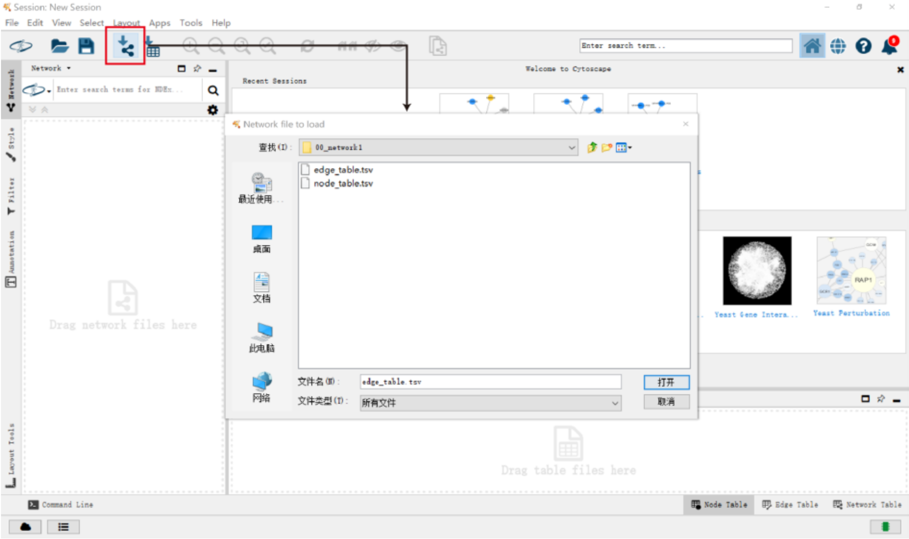
00_files_network1			
01_files_network2			
02_files_network3			
03_subnetworks			
.Rhistory			
list_peak_group	2022/4/13 3:47	文件	151 KB
list_peak_group_annotation_concis...	2022/4/13 3:28	R Workspace	6,730 KB
ms2_data.msp	2022/4/13 3:20	Windows Install...	1,318 KB
ms2_data.RData	2022/4/13 3:21	R Workspace	591 KB
peak_group_id_table	2022/4/13 3:45	文件	38 KB
table_identification	2022/4/13 3:47	文件	138 KB

2.4 Visualize global networks with Cytoscape

Above networks (Network 1-3) can be imported to Cytoscape software tool for visualization. The process of network visualization is generally similar. Here, we use the above network 1 as a demonstration. The version of Cytoscape used here is 3.8.2.

Below is the step-by-step instruction:

1. **Import edge file.** Select the “edge_table.tsv” file and open it in the box.



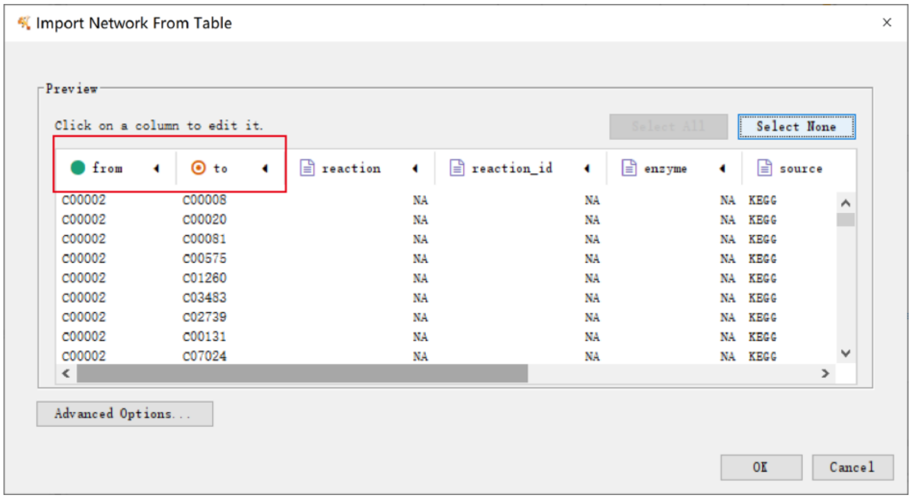
Network file to load

查找(I): 00_network1

edge_table.tsv
node_table.tsv

文件名(N): edge_table.tsv
文件类型(T): 所有文件

2. **Assign column attributes.** Click the ‘from’ column and select it as “source node”. Similarly, click the “to” column and select it as “target node”. After assigning attributes, click **OK** to construct a network.



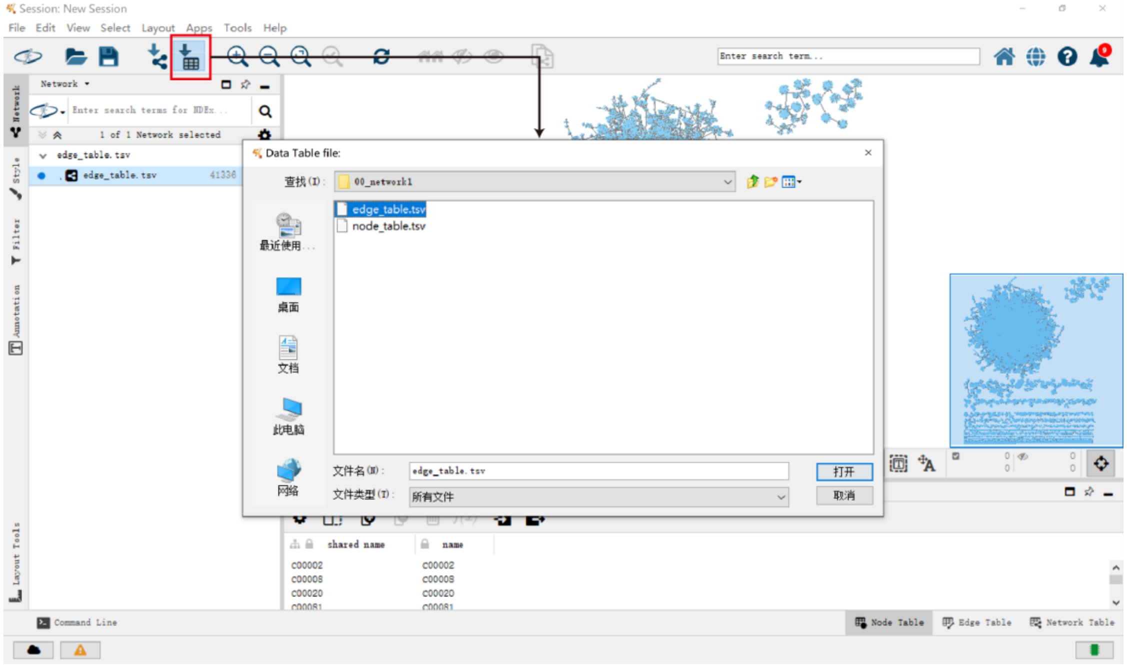
Import Network From Table

Preview

Click on a column to edit it.

from	to	reaction	reaction_id	enzyme	source
C00002	C00008	NA	NA	NA	KEGG
C00002	C00020	NA	NA	NA	KEGG
C00002	C00081	NA	NA	NA	KEGG
C00002	C00575	NA	NA	NA	KEGG
C00002	C01260	NA	NA	NA	KEGG
C00002	C03483	NA	NA	NA	KEGG
C00002	C02739	NA	NA	NA	KEGG
C00002	C00131	NA	NA	NA	KEGG
C00002	C07024	NA	NA	NA	KEGG

3. **Import node file.** Select the “node_table.tsv” file and open it in the box.



Data Table file:

查找(I): 00_network1

edge_table.tsv
node_table.tsv

最近使用...
桌面
文档
此电脑
网络

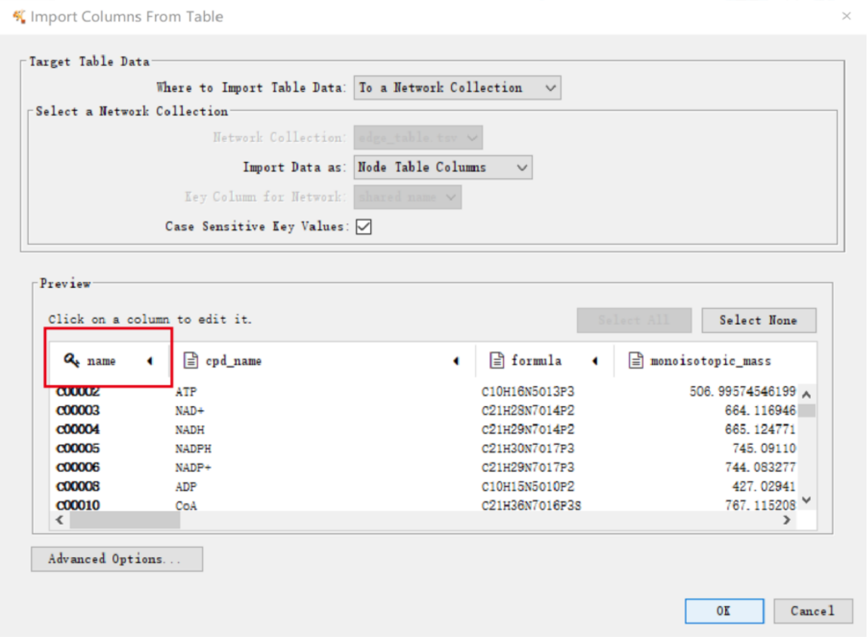
文件名(N): edge_table.tsv

文件类型(T): 所有文件

打开
取消

shared name	name
c00002	c00002
c00008	c00008
c00020	c00020
c00081	c00081

4. Select the “name” column as a key. Then, click the **OK** button.



Import Columns From Table

Target Table Data

Where to Import Table Data: To a Network Collection

Select a Network Collection

Network Collection: edge_table.tsv

Import Data as: Node Table Columns

Key Column for Network: shared name

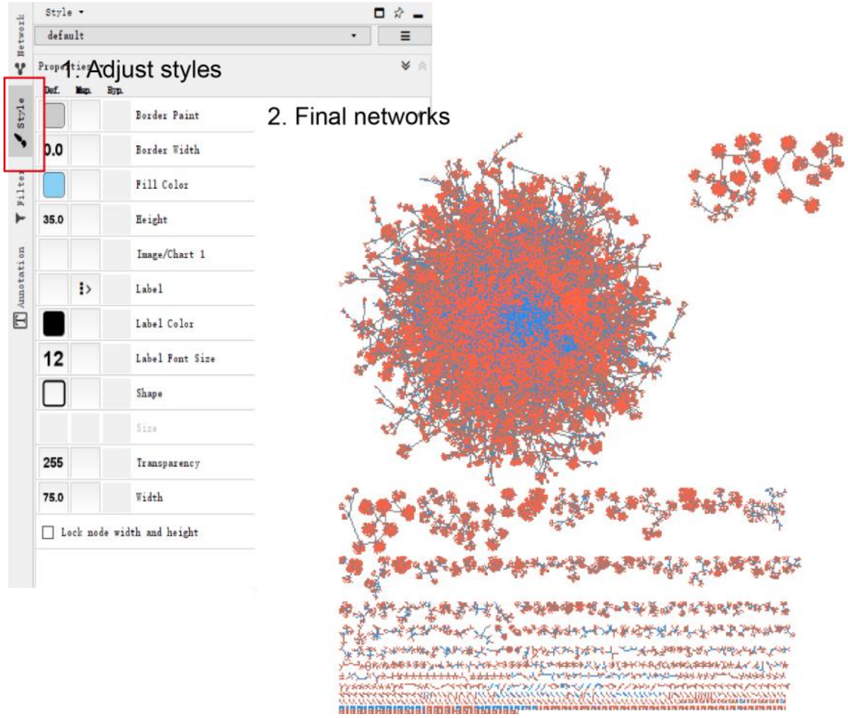
Case Sensitive Key Values: ☑

Preview

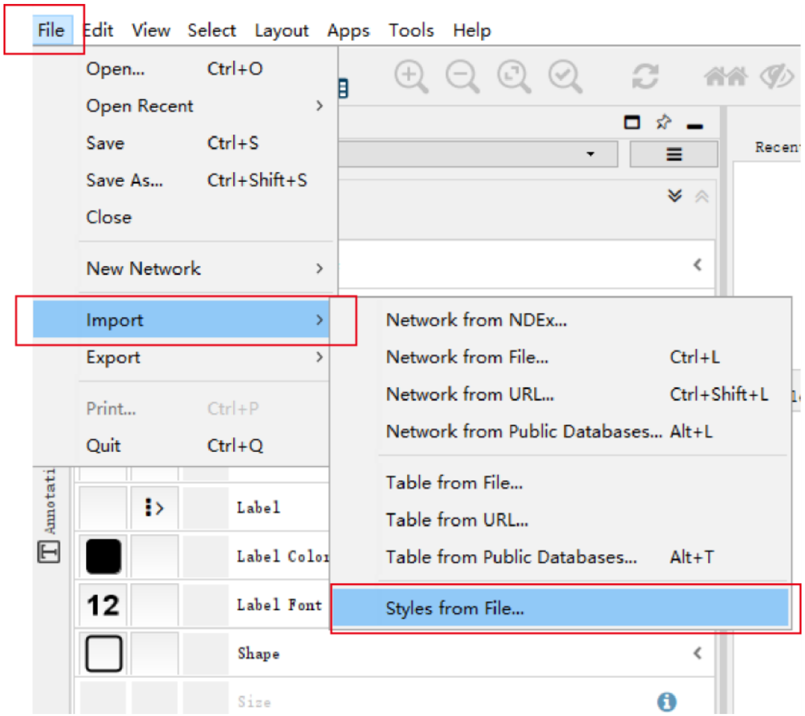
Click on a column to edit it.

name	cpd_name	formula	monoisotopic_mass
c00002	ATP	C10H16N5013P3	506.99574546199
c00003	NAD+	C21H28N7014P2	664.116946
c00004	NADH	C21H29N7014P2	665.124771
c00005	NADPH	C21H30N7017P3	745.09110
c00006	NADP+	C21H29N7017P3	744.083277
c00008	ADP	C10H15N5010P2	427.02941
c00010	CoA	C21H36N7016P3S	767.115208

5. **Modify the style for visualization.** Click the Style type, you can adjust node shapes and colors, edge types and colors.



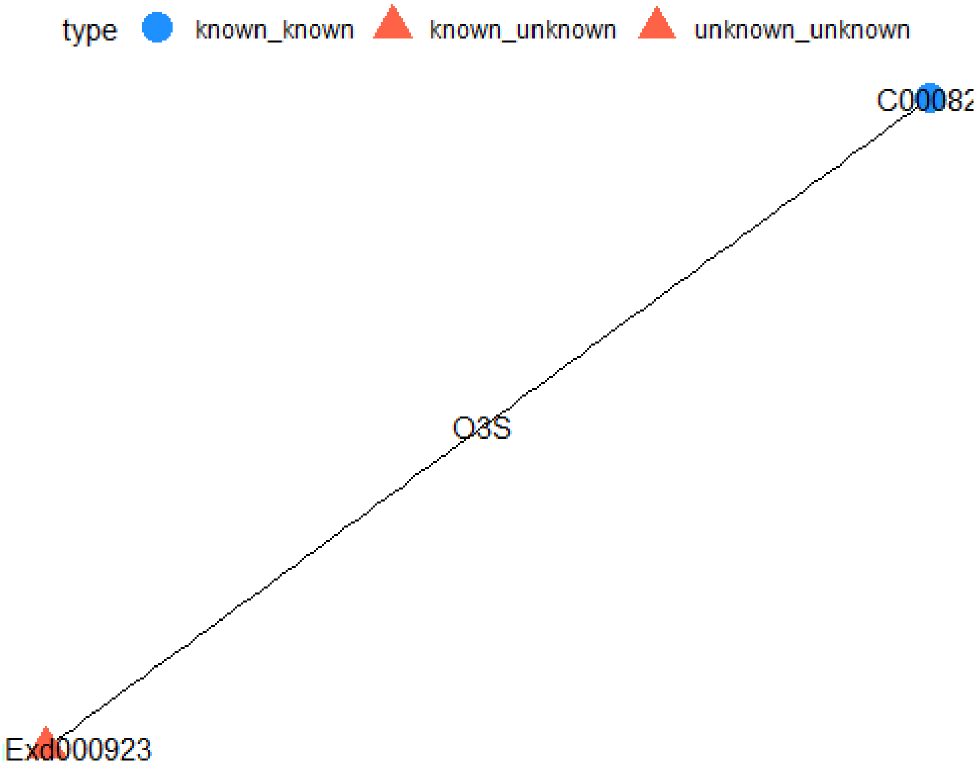
To help users reproduce our plot quickly, users can directly import our style file. The styles of different networks are provided **here**.



2.5 Select and export interesting subnetwork

Through above procedures, users can easily visualize global network 1-3. With such global networks, users can find interesting subnetworks in Cytoscape. The Cytoscape supports a interactively investigation. **It should be note that the targeted subnetwork selection is customized.** Users can directly find interesting nodes from KGMN annotation results, or considering more information, like in-silico MS/MS, chemical structure and/or statistics analysis. For example, in KGMN manuscript, we combined MASST to select an unknown subnetwork of M262T526 (**Figure 5e in manuscript**). This unknown peak is putatively annotated as O-sulfotyrosine, and this annotation is from M182T541-Tyrosine. This subnetwork consists of 2 peaks and 2 metabolites. Here, we mainly introduce how to export and visualize this subnetwork. First, export network 1 of this subnetwork. **Note:** the export and visualization require intermediate results from global networks. Therefore, please run global peaks export first. To export the subnetwork 1, please directly run retrieveSubNetwork1 function as below.

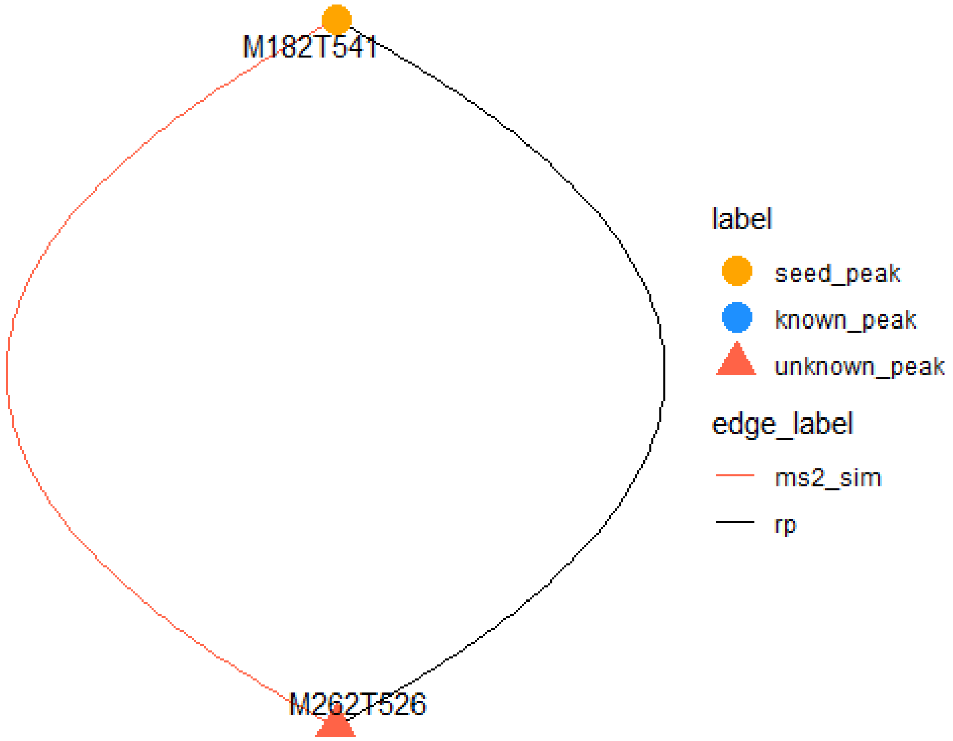
```
# network 1 of unknown peak subnetwork
# Note: the folder_output should keep same among different layer subnetworks
retrieveSubNetwork1(centric_met = c('C00082', 'KeggExd000923'),
  is_unknown_annotation = TRUE,
  folder_output = c('M182T541_M262T526'))
```



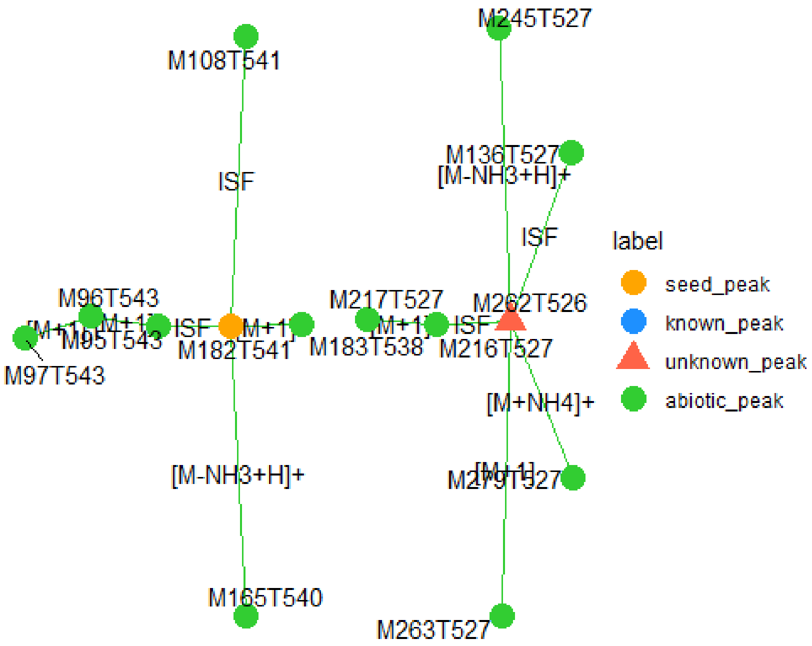
The networks files will be exported in ‘03_subnetworks/your_defined_folder/network 1’ folder. Here, the exported folder is “M182T541_M262T526”. The “edge_table.tsv” and “node_table.tsv” in this folder can be imported to Cytoscape for visualization. **Note:** if you run in RStudio, the preview plot of subnetwork 1 will be directly shown in the plot panel.

Similarly, export network 2 and network 3 of this subnetwork can be completed through running retrieveSubNetwork2 and retrieveSubNetwork3 functions, respectively. The preview plots of subnetwork 2 and subnetwork 3 will be shown in the plot panel if you run in RStudio.

```
# network 2 of unknown peak subnetwork
retrieveSubNetwork2(from_peak = 'M182T541',
  end_peak = 'M262T526',
  folder_output = c('M182T541_M262T526'))

## Using `sugiyama` as default layout
```

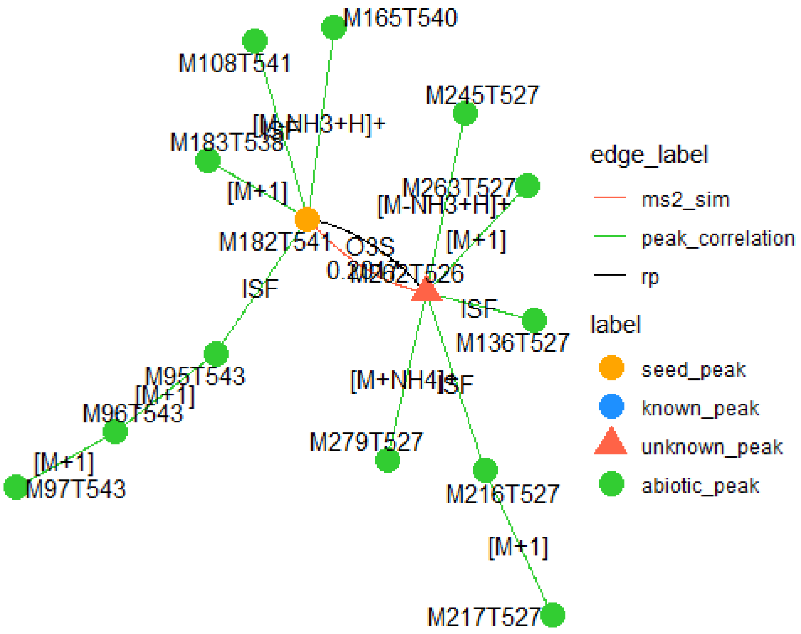


```
# network 3 of unknown peak subnetwork
retrieveSubNetwork3(base_peaks = c('M182T541', 'M262T526'),
  base_adducts = c('[M+H]+', '[M+H]+'),
  folder_output = c('M182T541_M262T526'))

## Using `stress` as default layout
```

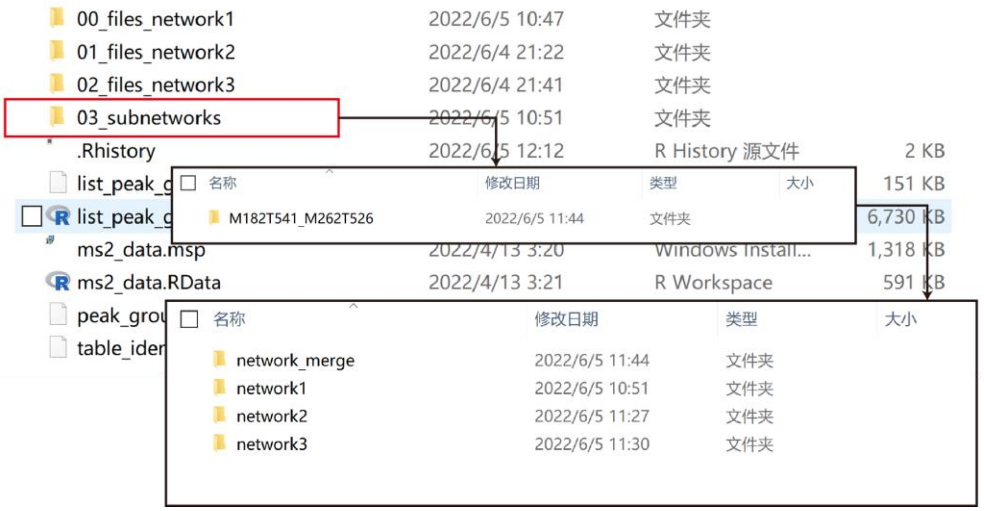


The network 2 and network 3 of the subnetwork can be further merged through running mergeSubnetwork function. The 'network_merge' folder contains node table and edge table for reproduce the merged network.

```
# merge subnetwork
mergeSubnetwork(from_peak = 'M182T541',
  end_peak = 'M262T526',
  folder_output = 'M182T541_M262T526')

## Using `stress` as default layout
```



Finally, the folder of subnetwork is organized like below. Each folder contains related files of each network for further visualization in other tools (e.g. Cytoscape).



00_files_network1	2022/6/5 10:47	文件夹	
01_files_network2	2022/6/4 21:22	文件夹	
02_files_network3	2022/6/4 21:41	文件夹	
03_subnetworks	2022/6/5 10:51	文件夹	
.Rhistory	2022/6/5 12:12	R History 源文件	2 KB
list_peak_g			151 KB
list_peak_g			6,730 KB
ms2_data.msp	2022/4/13 3:20	Windows Install...	1,318 KB
ms2_data.RData	2022/4/13 3:21	R Workspace	591 KB
peak_grou			
table_ider			

□ 名称	修改日期	类型	大小
M182T541_M262T526	2022/6/5 11:44	文件夹	

□ 名称	修改日期	类型	大小
network_merge	2022/6/5 11:44	文件夹	
network1	2022/6/5 10:51	文件夹	
network2	2022/6/5 11:27	文件夹	
network3	2022/6/5 11:30	文件夹	

3. The script for visualization

Here is a script contains above codes to help to reproduce above analysis quickly.

```
# load packages
library(CHNOSZ)
library(dplyr)
library(MetDNA2Vis)

# set working directory
setwd('D:/project/00_zhulab/01_metdna2/00_data/20220602_visualization_kgmn/Demo_MetDNA2_NIST_urine_pos/06_visualization/')

# Export global networks
# construct network 1
reconstructNetwork1(is_unknown_annotation = TRUE)

# construct network 2
annotation_table <- reformatTable1()
reconstructNetwork2(annotation_table = annotation_table)

# construct network 3
reconstructNetwork3()

# Export subnetworks --------------------------------------------------------
# network 1 of unknown peak subnetwork
# Note: the folder_output should keep same among different layer subnetworks
retrieveSubNetwork1(centric_met = c('C00082', 'KeggExd000923'),
  is_unknown_annotation = TRUE,
  folder_output = c('M182T541_M262T526'))



# network 2 of unknown peak subnetwork
retrieveSubNetwork2(from_peak = 'M182T541',
  end_peak = 'M262T526',
  folder_output = c('M182T541_M262T526'))

# network 3 of unknown peak subnetwork
retrieveSubNetwork3(base_peaks = c('M182T541', 'M262T526'),
  base_adducts = c('[M+H]+', '[M+H]+'),
  folder_output = c('M182T541_M262T526'))



# merge subnetwork
mergeSubnetwork(from_peak = 'M182T541',
  end_peak = 'M262T526',
  folder_output = 'M182T541_M262T526')
```