浙江中醫藥大學
ZHEJIANG CHINESE MEDICAL UNIVERSITY

Seminar

Lichuang Huang

Supervisor: Cao Gang

Zhejiang Chinese Medical University

2022-10-26

1. Theme: free yourself and rescue time

2. New module

3. Progress

4. Next Schedule

Theme: free yourself and rescue time

Documentation

- R base
  1. character
  2. data.frame
  3. . . .
- Pubchem API
- RCurl
- dplyr
- pattern
- stringr
- . . .

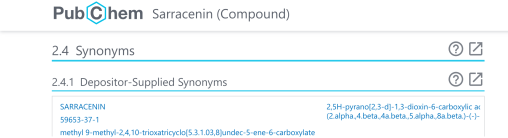
You get a mission

- A mission for 300 compounds.
  1. add CAS number.
  2. Format as pretty table.

Table 1: Compounds

seq	synonym
1	N-Lauroylglutamic acid
2	6-Keto-decanoylcarnitine
3	4-amino-5-dodecanoyloxy-5-oxopentanoic acid
4	2-(tridecanoylamino)butanedioic acid
5	7-[(5-carboxy-5-methylhexyl)amino]-2,2-dimethyl. . .
6	methyl (4S)-4-[(2-methylpropan-2-yl)oxycarbonyl. . .
7	methyl 12-nitro-15-oxohexadecanoate
8	(4S)-4-[butyl-[(2-methylpropan-2-yl)oxycarbonyl. . .
9	4-[(2-methylpropan-2-yl)oxycarbonylamino]butyl . . .
10	2-(2,2,6,6-tetramethylpiperidin-4-yl)oxyoctaned. . .
11	tert-butyl (4S,5S)-5-methyl-4-[(2-methylpropan-. . .
12	2-Morpholin-4-yl-succinic acid 1-nonyl ester

1. Approach 1
   1. copy and paste



PubChem Sarracenin (Compound)

2.4 Synonyms

2.4.1 Depositor-Supplied Synonyms

SARRACENIN

59653-37-1

methyl 9-methyl-2,4,10-trioxatricyclo[5.3.1.03,8]undec-5-ene-6-carboxylate

2,5H-pyrano[2,3-d]-1,3-dioxin-6-carboxylic ac
(2.alpha.,4.beta.,4a.beta.,5.alpha.,8a.beta.)-(-)-

2. Approach 2
   1. Install R. . .
   2. Read table (Excel)
   3. Use Pubchem API
   4. Match with pattern
   5. Format table
   6. Output

“Economic” benefits

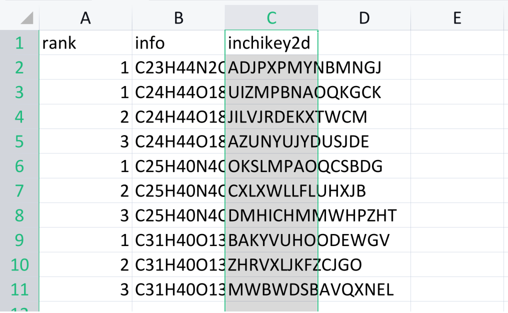
Time for shortcut:

1. For each compounds (30s):
   1. copy and paste (5s)
   2. search and find (20s)
   3. copy and paste (5s)
2. For all compounds
   - $(20 + 5 + 5) \times 300 \div 60 = 150\,(min)$
3. Reality
   - Your Emotional labor
     - $150 \times 2 = 300\,(min)$
   - Desertion
     - $300 \times 2 = 600\,(min)$
4. As pretty table
   - $600 + 20 = 620\,(min)$
5. Your gains
   1. Complex experience
   2. A thanks

Time for detour:

- Overall
  1. Install R $30\,(min)$
  2. Read table (Excel) $30\,(min)$
  3. Use Pubchem API $60\,(min)$
  4. Match with pattern $120\,(min)$
     - You find a new world
  5. Format and pretty table $120\,(min)$
  6. Output $30\,(min)$
- Your gains
  1. Entrance to programming
  2. Criticism
  3. The left no more in life
     - exact mass
     - compound name
     - chemistry irrelevant. . .
  4. . . .

Start with R: read table into R



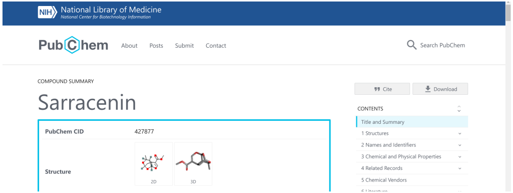
	A	B	C	D	E
1	rank	info	inchikey2d		
2	1	C23H44N2C	ADJPXPMYNBMNGJ		
3	1	C24H44O18	UIZMPBNAOQKGCK		
4	2	C24H44O18	JILVJRDEKXTWCM		
5	3	C24H44O18	AZUNYUJYDUSJDE		
6	1	C25H40N4C	OKSLMPAOQCSBDG		
7	2	C25H40N4C	CXLXWLLFLUHXJB		
8	3	C25H40N4C	DMHICHMMWHPZHT		
9	1	C31H40O13	BAKYVUHOODEWGV		
10	2	C31H40O13	ZHRVXLJKFZCJGO		
11	3	C31H40O13	MWBWDSBAVQXNEL		

Figure 1: Table in Excel

You spent more time and energy than double clicking
Give up or continue?

- "Read" package:
  1. base
     - read.table
  2. data.table
     - fread
  3. openxlsx
     - read.xlsx
- "Organize" package:
  1. dplyr
     - filter
     - select
     - relocate
     - mutate
     - arrange
     - rename
     - ......

Find data in Pubchem

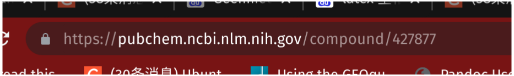


COMPOUND SUMMARY

Sarracenin

PubChem CID	427877
Structure	

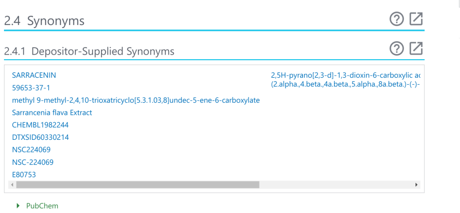
Figure 2: Pubchem Website



https://pubchem.ncbi.nlm.nih.gov/compound/427877

Figure 3: URL for Compounds

https://pubchem.ncbi.nlm.nih.gov/compound/ + "number"



2.4 Synonyms

2.4.1 Depositor-Supplied Synonyms

SARRACENIN
59653-37-1
methyl 9-methyl-2,4,10-trioxatricyclo[5.3.1.03,8]undec-5-ene-6-carboxylate
Sarracenia flava Extract
CHEMBL1982244
DTXSID60330214
NSC224069
NSC-224069
E80753

2,5H-pyrano[2,3-d]-1,3-dioxin-6-carboxylic ac
(2.alpha.,4.beta.,4a.beta.,5.alpha.,8a.beta.)-(-)-

Figure 4: Where is CAS number

https://pubchem.ncbi.nlm.nih.gov/rest/pug/compound/name/ + name + /synonyms/XML

- URL to get synonym
  1. Prefix
  2. your id
  3. suffix

Get data in Pubchem

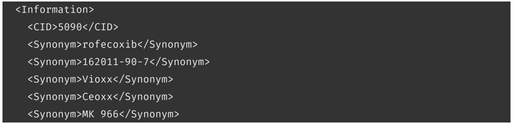
1. set URL

```
prefix <- "https..."
name <- "compound"
suffix <- ".../XML"
url <- paste0(prefix, name, suffix)
```

2. use “spider”

```
RCurl::getURL(url)
```

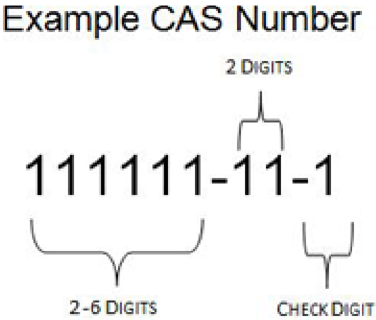
3. format...



```
<Information>
  <CID>5090</CID>
  <Synonym>rofecoxib</Synonym>
  <Synonym>162011-90-7</Synonym>
  <Synonym>Vioxx</Synonym>
  <Synonym>Ceoxx</Synonym>
  <Synonym>MK 966</Synonym>
```

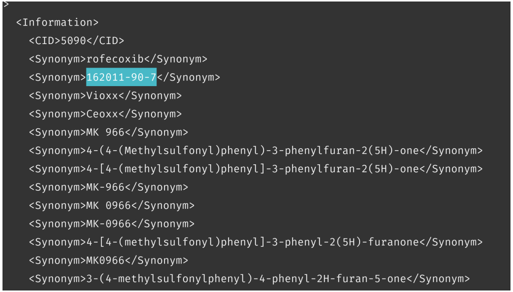
- for “spider” :
  1. base
     - paste0
     - ...
  2. RCurl
     - getURL
     - ......
  3. XML
     - xmlTreeParse
     - xmlParse
- Extra usage
  1. chemical informatic data
  2. Gene informatic data
  3. ...
  4. Beyond research

Match CAS number



CAS Registry Numbers contain up to 10 digits, divided by hyphens into three parts with right most digit a check digit

- CAS in your text



```
<Information>
  <CID>5090</CID>
  <Synonym>rofecoxib</Synonym>
  <Synonym>162011-90-7</Synonym>
  <Synonym>Vioxx</Synonym>
  <Synonym>Ceoxx</Synonym>
  <Synonym>MK 966</Synonym>
  <Synonym>4-(4-(Methylsulfonyl)phenyl)-3-phenylfuran-2(5H)-one</Synonym>
  <Synonym>4-[4-(methylsulfonyl)phenyl]-3-phenylfuran-2(5H)-one</Synonym>
  <Synonym>MK-966</Synonym>
  <Synonym>MK 0966</Synonym>
  <Synonym>MK-0966</Synonym>
  <Synonym>4-[4-(methylsulfonyl)phenyl]-3-phenyl-2(5H)-furanone</Synonym>
  <Synonym>MK0966</Synonym>
  <Synonym>3-(4-methylsulfonylphenyl)-4-phenyl-2H-furan-5-one</Synonym>
```

Figure 6: Your text

$< \ldots > + \text{number} + \text{number} + \text{number} < \ldots >$

Pattern

<...> + number + number + number <...>

1111-11-1

1. match number: [0-9]
2. match two number: [0-9]{2}
3. match more number [0-9]{1,}
4. match “-”: -
5. match CAS: [0-9]{1,}-[0-9]{2}-0-9
6. look before: (?<=...)
7. look after: (?=...)

```
stringr::str_extract(text,
"(?<=\>)[0-9]{1,}-[0-9]{2}-[0-9]{1,}(?=\<)")
```

- for string match:
  1. stringr
     - str_extract
- Extra usage
  1. Text search
  2. revise article
  3. ...

Pretty table

Table 2: Compounds

seq	synonym	CAS
1	N-Lauroylglutamic acid	...
2	6-Keto-decanoylcarnitine	...
3	4-amino-5-dodecanoyloxy-5-oxopentanoic acid	...
4	2-(tridecanoylamino)butanedioic acid	...
5	7-[(5-carboxy-5-methylhexyl)amino]-2,2-dimethyl...	...
6	methyl (4S)-4-[(2-methylpropan-2-yl)oxycarbonyl...	...
7	methyl 12-nitro-15-oxohexadecanoate	...
8	(4S)-4-[butyl-[(2-methylpropan-2-yl)oxycarbonyl...	...
9	4-[(2-methylpropan-2-yl)oxycarbonylamino]butyl ...	...
10	2-(2,2,6,6-tetramethylpiperidin-4-yl)oxyoctaned...	...
11	tert-butyl (4S,5S)-5-methyl-4-[(2-methylpropan-...	...
12	2-Morpholin-4-yl-succinic acid 1-nonyl ester	...

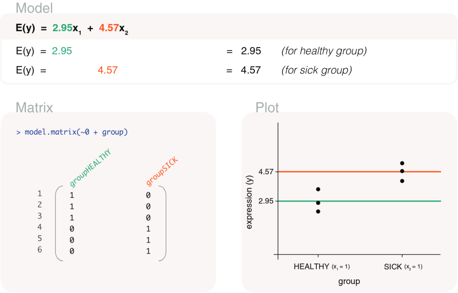
- for pretty table:
  1. gt
     - gt
     - ...
  2. knitr
     - kable
  3. write.table
- Extra usage
  1. ppt
  2. article
  3. ...

New module

Statistic system: binary_comparison

Table 3: Compounds

id	logFC	AveExpr	t	P.Value	adj.P.Val
2025	-39.001	2.604	-3.427	0.001	0.023
2071	36.610	-2.537	3.156	0.002	0.029
2097	-33.037	-3.986	-2.895	0.004	0.043
2095	24.035	-3.471	2.107	0.037	0.276
2096	19.966	1.764	1.747	0.083	0.449
209	-19.682	-2.278	-1.707	0.090	0.449
2076	18.118	2.034	1.591	0.114	0.486
2029	-16.703	-1.473	-1.416	0.159	0.587
2023	-16.095	5.772	-1.338	0.183	0.587
2024	14.873	3.637	1.300	0.196	0.587
2075	-13.884	0.386	-1.203	0.231	0.630
2091	12.416	-4.503	1.094	0.276	0.689



Model

$$E(y) = 2.95x_1 + 4.57x_2$$

$$E(y) = 2.95 \qquad = 2.95 \quad \textit{(for healthy group)}$$

$$E(y) = \qquad 4.57 \qquad = 4.57 \quad \textit{(for sick group)}$$

Matrix

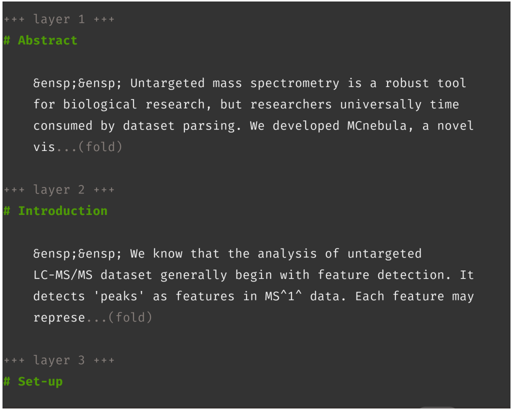
```
> model.matrix(~0 + group)
```

$$\begin{array}{c|cc} & \text{groupHEALTHY} & \text{groupSICK} \\ \hline 1 & 1 & 0 \\ 2 & 1 & 0 \\ 3 & 1 & 0 \\ 4 & 0 & 1 \\ 5 & 0 & 1 \\ 6 & 0 & 1 \end{array}$$

Plot

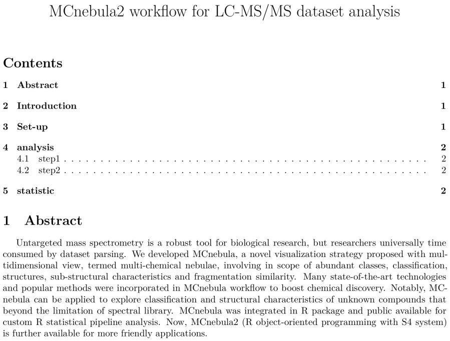
Figure 7: binary comparison

report system: pdf or html or other report



```
+++ layer 1 +++
# Abstract

    &ensp;&ensp; Untargeted mass spectrometry is a robust tool
    for biological research, but researchers universally time
    consumed by dataset parsing. We developed MCnebula, a novel
    vis...(fold)

+++ layer 2 +++
# Introduction

    &ensp;&ensp; We know that the analysis of untargeted
    LC-MS/MS dataset generally begin with feature detection. It
    detects 'peaks' as features in MS^1^ data. Each feature may
    represe...(fold)

+++ layer 3 +++
# Set-up
```

Figure 8: object in console



MCnebula2 workflow for LC-MS/MS dataset analysis

Contents

1 Abstract 1

2 Introduction 1

3 Set-up 1

4 analysis 2

4.1 step1 2

4.2 step2 2

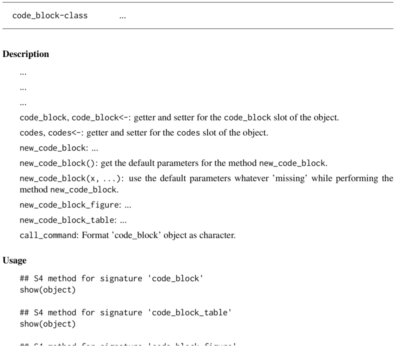
5 statistic 2

1 Abstract

Untargeted mass spectrometry is a robust tool for biological research, but researchers universally time consumed by dataset parsing. We developed MCnebula, a novel visualization strategy proposed with multidimensional view, termed multi-chemical nebulae, involving in scope of abundant classes, classification, structures, sub-structural characteristics and fragmentation similarity. Many state-of-the-art technologies and popular methods were incorporated in MCnebula workflow to boost chemical discovery. Notably, MCnebula can be applied to explore classification and structural characteristics of unknown compounds that beyond the limitation of spectral library. MCnebula was integrated in R package and public available for custom R statistical pipeline analysis. Now, MCnebula2 (R object-oriented programming with S4 system) is further available for more friendly applications.

Figure 9: pdf_report

MCnebula2 documentation: reference.pdf



code_block-class ...

Description

...

...

...

code_block, code_block<-: getter and setter for the code_block slot of the object.

codes, codes<-: getter and setter for the codes slot of the object.

new_code_block: ...

new_code_block(): get the default parameters for the method new_code_block.

new_code_block(x, ...): use the default parameters whatever 'missing' while performing the method new_code_block.

new_code_block_figure: ...

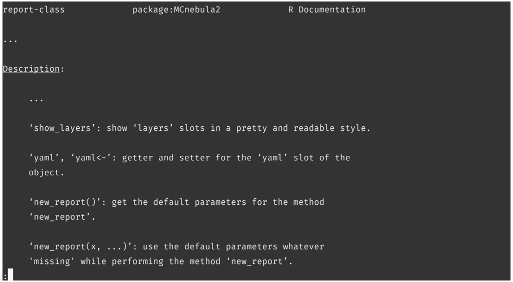
new_code_block_table: ...

call_command: Format 'code_block' object as character.

Usage

```
## S4 method for signature 'code_block'
show(object)

## S4 method for signature 'code_block_table'
show(object)

## S4 method for signature 'code_block_figure'
```

Figure 10: reference.pdf



```
report-class              package:MCnebula2              R Documentation

...

Description:

     ...

     ‘show_layers’: show ‘layers’ slots in a pretty and readable style.

     ‘yaml’, ‘yaml<-’: getter and setter for the ‘yaml’ slot of the
     object.

     ‘new_report()’: get the default parameters for the method
     ‘new_report’.

     ‘new_report(x, ...)’: use the default parameters whatever
     'missing' while performing the method ‘new_report’.
:
```

Figure 11: help(report)

Progress

Code refactoring: data processing

Formerly finished:

- S4 data structure ✓
- Extract data ✓
  1. project_conformation ✓
  2. project_metadata ✓
  3. project_api ✓
  4. project_dataset ✓
- Filter data
  0. collate_data ✓
  1. filter_formula ✓
  2. filter_structure ✓
  3. filter_ppcp ✓
  4. create_reference ✓

- oganize data
  1. create_features_annotation ✓
  2. create_stardust_classes ✓
  3. cross_filter_stardust ✓
     - quantity ✓
     - score ✓
     - identical ✓
  4. create_nebula_index ✓
  5. compute_spectral_similarity ✓
  6. create_parent_nebula ✓
  7. create_child_nebulae ✓

Code refactoring: visualization

Formerly finished:

- for oganizing visualization
  1. command ✓
     - new_command ✓
     - call_command ✓
  2. ggset ✓
     - new_ggset ✓
     - mutate_layer ✓
     - add_layer ✓
     - delete_layer ✓
     - modify_set_labs ✓
     - modify_unify_scale_limits ✓
     - modify_rm_legend ✓
     - modify_... ✓
     - call_command ✓
     - ...

- visualization
  1. create_child_layout ✓
  2. activate_nebulae ✓
  3. quantification ✓
     - features_quantification ✓
     - sample_metadata ✓
  4. annotate_nebula ✓
     - draw_structures ✓
     - draw_nodes ✓
  5. visualize ✓
     - visualize ✓
     - visualize_all ✓
     - show_node ✓
     - show_structure ✓

New module: statistic system

This time finished:

- binary_comparison ✓
    - features_quantification ✓
    - sample_metadata ✓
    - binary_comparison ✓
    - top_table ✓

New module: report system

This time finished:

- for each module
  - heading ✓
  - paragraph ✓
  - code_block ✓
  - picture ✓
  - table ✓
- section ✓
  - new_section ✓
  - new_heading ✓
  - new_code_block ✓
  - new_code_block_table ✓
  - new_code_block_figure ✓
  - call_command ✓
  - ...
- report
  - new_report ✓
  - show ✓
  - add_layer ✓
  - render ✓

Next Schedule

Overall

- Data filtering and arranging system. ✓
- Visualization system. ✓
- LC-MS runing in XCMS.
- Statistic system. ✓
- Report. ✓
- Documentation.
- Website.
- Figures for article.
- Rewrite article.