Food, Plant, Thing Metabolome Repository

# Use of Previous Data

April 18, 2022

National Institute of Genetics
Nozomu Sakurai

# 1. Introduction

This document is a guide for using the archives of previous data of the Food/Plant/Things Metabolome Repositories (XMRs) below.

Food Metabolome Repository (FoodMR) http://metabolites.in/foods
Plant Metabolome Repository (PlantMR) http://metabolites.in/plants
Thing Metabolome Repository (ThingMR) http://metabolites.in/things

To ensure the reproducibility of the data analysis performed by the users of XMRs, we provide a set of previous data when we perform a major update of the data. We irregularly update the data not only by adding new data but also by curating metadata, updating the compound database search results due to the enlargement of the source databases, the addition of novel analyzed results, and so on.

# 2. Archived files

The following four files contain the primary data in XMRs.

| Filename | Description | Format | MariaDB |
|---|---|---|---|
| samples.txt | The file contains the sample information, such as sample names and categories of the samples. | Tab-separated text | ✓ |
| files.txt | The file contains information related to the analyzed data, such as filenames, IDs of the Metabolonote website that provides metadata details, and numbers of peaks detected. | Tab-separated text | ✓ |
| peaks.txt | The file contains peak information, such as m/z value, | Tab-separated text | ✓ |

| | | | |
|---|---|---|---|
| | retention time, and results of compound database search. | | |
| spectra.txt | The file contains mass spectral data. | text | |

The former three files, namely samples.txt, files.txt and peaks.txt, provided in a tab-separated text format, are loaded in a relational database system "MariaDB" and used in the XMR web system. The last file, spectra.txt, is described in our original text format and loaded directly into the XMR web system.

The table definition of the MariaDB is provided in the file below. Please note that the table definition may change by the data update.

| Filename | Description |
|---|---|
| create_tables.sql | The file contains the definitions of MariaDB tables for loading the data described in samples.txt, files.txt, and peaks.txt. |

# 3. Details of the data

The details of the data in the files are as follows.

## samples.txt

The samples.txt file contains the sample information in tab-separated text format. The file is loaded in the "foods" table in MariaDB because we first developed Food Metabolome Repository. The details of the columns are below.

| Column No. | MariaDB field name | Description |
|---|---|---|
| 1 | fid | The sample ID. The field name stands for Food ID because we first released Food Metabolome |

| | | |
|---|---|---|
| | | Repository. |
| 2 | name_ja | The sample name in Japanese |
| 3 | name_en | The sample name in English |
| 4 | cat_ja | The category of the sample in Japanese |
| 5 | cat_en | The category of the sample in English |
| 6 | desc_ja | not used |
| 7 | desc_en | not used |
| 8 | dir | The path to the data file in the system |
| 9 | mnsid | The sample ID in Metabolonote* |
| 10 | status | 公開状況。現在はすべて「s1-published」となっています。 |

* Metabolonote is a database system specified for managing metadata of metabolomics experiments (Reference 1).

# files.txt

The files.txt file contains the information of analyzed data files. The file is loaded in the “files” table in MariaDB. The details of the columns are below.

| Column No. | MariaDB field name | Description |
|---|---|---|
| 1 | id | Blank. The unique id is automatically assigned when the data are loaded into MariaDB. |
| 2 | fid | The sample ID. The field name stands for Food ID. |
| 3 | type | The polarity of ESI ionization. pos: positive, neg: negative. |
| 4 | filename | The label for the analyzed data file. The fid and type are attached with underscores “_” at the head of the label. |
| 5 | mnmid | The analytical method ID of Metabolonote*1 that specifies the details of the analytical methods used. |

| | | |
|---|---|---|
| 6 | peaknum | The number of detected peaks. |
| 7 | peaknum_n | The number of peaks estimated as containing nitrogen (N) atom(s) in the chemical structure. The estimation was performed for a part of the samples in PlantMR. The value would be -1 if the estimation was not performed. |
| 8 | peaknum_s | The number of peaks estimated as containing sulfur (S) atom(s) in the chemical structure. The estimation was performed for a part of the samples in PlantMR. The value would be -1 if the estimation was not performed. |
| 9 | peaknum_ms2 | The number of peaks with MS2 or MS/MS spectrum data. |
| 10 | peaknum_ms3 | The number of peaks with MS3 spectrum data. |
| 11 | peaknum_fs2 | The number of peaks with a hit score of FlavonoidSearch*2 above zero is calculated with the MS2 or MS/MS spectrum. The value would be -1 if the calculation was not performed. |
| 12 | peaknum_fs3 | The number of peaks with a hit score of FlavonoidSearch*2 above zero is calculated with the MS3 spectrum. The value would be -1 if the calculation was not performed. |
| 13 | peaknum_dbhit | The number of peaks with compound database search results. |

*1 Metabolonote is a database system specified for managing metadata of metabolomics experiments (Reference 1).

*2 FlavonoidSearch is a computational tool for predicting flavonoid aglycones using mass spectral data (Reference 2).

## peaks.txt

The peaks.txt file contains the peak information and is the largest among the distributed data set. The file is loaded in the “peaks” table in MariaDB. The details of the columns are below.

| Column No. | MariaDB field name | Description |
|---|---|---|
| 1 | id | Blank. The unique id is automatically assigned when the data are loaded into MariaDB. |
| 2 | mnmid | The analytical method ID of Metabolonote*1 that specifies the details of the analytical methods used. |
| 3 | mndid | The tail of data analysis ID (D**) of Metabolonote*1. By using with the mnmid, the ID specifies the details of the data analysis procedures used. |
| 4 | pid | The peak ID. By using with the mnmid and mndid, the ID specifies the peak detected in the sample. |
| 5 | type | The polarity of ESI ionization. pos: positive, neg: negative). |
| 6 | rt | The retention time (min) of the chromatography. |
| 7 | mz | The *m/z* value detected. |
| 8 | int | The peak intensity as the peak area estimated by PowerGetBatch*2. |
| 9 | intlog | The log10-transformed peak intensity. The value of column 8 was transformed to log base ten and centralized by the median value of all peaks detected in the analysis. |
| 10 | adduct | The estimated adduct ion. If several candidates were predicted, they are described in comma-separated format, with the most probable one being the first. |

|  |  |  |
|---|---|---|
| 11 | mzdi | The mass value of the neutralized molecule based on the *m/z* value (column 7) and the estimated adduct ion (column 10). This value is used for the compound database search. |
| 12 | annot | The annotation of the peak based on the compound database search. |
| 13 | skeleton | not used |
| 14 | ms2 | The number of MS2 or MS/MS spectra obtained for the peak. |
| 15 | ms3 | The number of MS3 spectra obtained for the peak. |
| 16 | fs2 | The hit score of FlavonoidSearch *3 calculated with the MS2 or MS/MS spectrum. |
| 17 | fs3 | The hit score of FlavonoidSearch *3 calculated with the MS2 or MS/MS spectrum. The maximum value is provided if several MS3 spectra were obtained. |
| 18 | db_all | The number of constitutional isomers found by the compound database search. |
| 19 | db_kg | The number of constitutional isomers found in KEGG. |
| 20 | db_kn | The number of constitutional isomers found in KNApSAcK. |
| 21 | db_hm | The number of constitutional isomers found in HMDB. |
| 22 | db_lm | The number of constitutional isomers found in LIPID MAPS. |
| 23 | db_fl | The number of constitutional isomers found in the flavonoid database in metabolomics.jp. |
| 24 | num_atom_n | The number of nitrogen (N) atoms estimated. A blank means that no estimation is performed. |

|  |  |  |
|---|---|---|
| 25 | num_atom_s | The number of sulfur (S) atoms estimated. A blank means that no estimation is performed. |
| 26 | pgroup | The information on related peaks. When multiple variations of adduct ions were predicted for the same compound, the peak IDs, including itself, were presented in the comma-separated text.<br>* Since April 18, 2022. |
| 27 | num_shared | The sample numbers in the repository that contain similar peaks to the peak in the row. Similar peaks were searched in the repository using the *m/z* value and the retention time.<br>* Since April 18, 2022. |
| 28 | option1 | not used |

*1 Metabolonote is a database system specified for managing metadata of metabolomics experiments (Reference 1).

* PowerGetBatch is software for peak detection, peak characterization, peak alignment between the samples, compound database searching, and so on, used in the data processing of XMRs (Reference 3).

*3 FlavonoidSearch is a computational tool for predicting flavonoid aglycones using mass spectral data (Reference 2).

# Supplementary explanation

## ➢ Compound database search

The compound database search was performed using the value in “mzdi” (column 11), giving mass tolerances of 5 ppm for FoodMR and PlantMR and 20 ppm for ThingMR using the compound databases below.

|  |  |
|---|---|
| KEGG | http://www.genome.jp/kegg/ |
| KNApSAcK | http://kanaya.naist.jp/KNApSAcK_Family/ |
| HMDB | http://www.hmdb.ca/ |

| | |
|---|---|
| LIPID MAPS | http://www.lipidmaps.org/ |
| flavonoid database | http://metabolomics.jp/wiki/Category:FL |

The cross-database search was performed using the MFSearcher web service（http://webs2.kazusa.or.jp/mfsearcher, Reference 4）and the UC2 database in the service. The use of the UC2 database is advantageous for grasping the variation of candidates for the features below.

- When the same compound (constitutional isomer) is redundantly registered as ones with different charges, salts, and so on among the compound databases, they are compiled in a single result.
- The optical isomers are compiled in a single result. For example, L-alanine and D-alanine are included in a single result.

## ➢ annot (column 12)

When the same constitutional isomers are registered in multiple compound databases, the shortest name of the compound is selected as a representative. The shortest name for each candidate was concatenated with semi-colon “;”. If the concatenated name is more than 200 characters, any characters after the first 200 were truncated, and three periods “...” were attached at the tail. The number of names is provided at the head as “(* names)” when multiple candidates are found.

**＊There are cases where semi-colon is used in the name of compound databases. Therefore, the number of names might not be the same as the number of candidates (db_all, column 18).**

## ➢ shared (column 27)

The tolerances of mass value and retention time used for the search were as follows.

| Repository | Mass tolerance（ppm) | Retention time tolerance (min) |
|---|---|---|
| FoodMR | 5 | 1 |
| PlantMR | 5 | 1 |

|  |  |  |
|---|---|---|
| ThingMR | 20 | 0.5 |

# spectra.txt

The spectra.txt file contains the mass spectral data. This file is not loaded to MariaDB but is directly used in the XMR system. The mass spectral data are described in the text format below.

A single mass spectrum is presented as a block below:

- A header line starts with “>”
- Data lines after the header line start with a number

The lines are tab-separated, and each column provides information below.

## The header line starts with “>”

| Column No. | Description |
|---|---|
| 1 | A character “>” represents the header line. |
| 2 | The analytical method ID of Metabolonote, which corresponds to the mnmid in the peak.txt file. The ID specifies the sample and details of analytical methods. |
| 3 | The peak ID, which corresponds to the pid in the peak.txt file. |
| 4 | The level of mass spectra. 2: MS2 or MS/MS, 3: MS3 |
| 5 | The *m/z* value of the precursor ion. |
| 6 | The splash ID which is uniquely assigned to the mass spectral data. |

## The data line starts with a number

| Column No. | Description |
|---|---|
| 1 | The *m/z* value of the product ion. |
| 2 | The intensity of the product ion. |
| 3 | The relative intensity of the product ion. The intensity of the base peak ion (the most intense ion in the spectrum) is represented as 1000. |
| 4 | The neutral loss mass, which is the mass difference between the product ion and the precursor ion. |
| 5 | The order of the product ion when sorted by the intensity. |

The data lines were sorted by the intensity.

# Loading to MariaDB

This section describes how to load the data in samples.txt, files.txt, and peaks.txt into MariaDB. The use of MariaDB is advantageous at the search rate, especially for the search in peak data.

This section is described for the users who are familiar with the use of MariaDB.

The following names are used in this section as an example of a setting. Please modify the settings according to the user’s environment and purpose.

| | |
|---|---|
| Database name | mydb1 |
| User name | user1 |
| Pass word | pass1 |

## 1. Creation of database

First, prepare a database and an account to access the database in MariaDB.

Open the terminal software and log in to MariaDB. Execute the following command.

```
MariaDB> create database mydb1 default character set utf8;
MariaDB> grant all privileges on mydb1.* to 'user1'@'localhost' identified by
'pass1';
```

A database named “mydb1” is created, and the user account “user1” is created with an access password “pass1”. You can check the creation of the database using the “show databases” command.

```
MariaDB> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| mydb1              |
| test               |
+--------------------+
3 rows in set (0.00 sec)
```

※ The databases named “information_schema” and “mysql” are system-used databases automatically created in MariaDB.

## 2. Creation of tables

Second, create tables in the mydb1 database for storing the loaded data. Please note that the table definition in the “create_tables.sql” file may change by the data update. It is recommended to drop and re-create the tables before using the different versions of the previous data using the right version of the create_tables.sql file included in the dataset.

Open the terminal software and go to the directory where the “create_tables.sql” file exists. Use the “cd” command to change the current directory (folder).

Execute the command below.

```
> mysql -u pass1 -p mydb1 < create_tables.sql
```

Then, enter the pass word “pass1.”

```
> mysql -u pass1 -p mydb1 < create_tables.sql
Enter password: *****
```

* The input characters are hidden with asterisks on the display.

The tables are created in the mydb1 database according to the definition described in the create_tables.sql file.

You can check the creation of the tables by logging in to MariaDB, selecting the mydb1 database, and using the “show tables” command as below.

```
> mysql -u user1 -p
Enter password: *****
… omitted …

MariaDB> use mydb1;
Database changed
MariaDB> show tables;
+-----------------+
| Tables_in_mydb1 |
+-----------------+
| files           |
| foods            |
| peaks            |
+------------------+
3 rows in set (0.00 sec)
```

# 3. Loading the data

Lastly, load the data from the files.

Open the terminal software, and go to the directory where the samples.txt, files.txt, and peaks.txt files exist. Use the “cd” command to change the directory.

Login to MariaDB and select the mydb1 database.

```
>mysql -u user1 -p
Enter password: *****
… omitted …

MariaDB> use mydb1;
```

Then, execute the following two commands in this order.

```
MariaDB> delete from foods;
Query OK, 0 rows affected (0.00 sec)

MariaDB> load data local infile ”samples.txt” into table foods fields terminated by
’¥t’ lines terminated by ’¥r¥n’;
Query OK, *** rows affected, * warnings (0.** sec)
Records: *** Deleted: 0 Skipped: * Warnings: *
```

* The newline character used in the distributed files is CR + LF (Windows). Therefore, you may have to specify the “line terminate by” properly, depending on your

computational environment.

The first command deletes all the records on the foods table. The second command loads data from samples.txt into the foods table. Please note that the name of the table is “foods,” standing for the first release of the XMR series, but the file is named “sample.txt.”

The process is finished with the message “Query OK…” as above.

You can check the loading of the data using the “select” command.

```
MariaDB> select * from foods limit 2;
+--------+----------------------------------------------------------------------
--------------+-----------------------------------------------------------------
----------+-------------+-------------+---------+---------+-------------------+-
-----------+--------------+
| fid    | name_ja
| name_en                                                                      |
cat_ja      | cat_en      | desc_ja | desc_en | dir                | mnsid      |
status      |
+--------+----------------------------------------------------------------------
--------------+-----------------------------------------------------------------
----------+-------------+-------------+---------+---------+-------------------+-
-----------+--------------+
| 221001 | イロハモミジ / 葉 / Acer palmatum Thunb. / ムクロジ科 カエデ属
| Japanese maple / Leaf / Acer palmatum Thunb. / Sapindaceae Acer            | 植
物;葉      | Plant;Leaf  |         |         | data/SE221/221001 | SE221_S001 |
s1-published |
| 221002 | ホトケノザ / 地上部 / Lamium amplexicaule / シソ科 オドリコソウ亜科 オ
ドリコソウ属 | Common henbit / Shoot / Lamium amplexicaule / Lamiaceae Lamioideae
Lamium | 植物;地上部 | Plant;Shoot |         |         | data/SE221/221002 |
SE221_S002 | s1-published |
+--------+----------------------------------------------------------------------
--------------+-----------------------------------------------------------------
----------+-------------+-------------+---------+---------+-------------------+-
-----------+--------------+
2 rows in set (0.00 sec)
```

＊The option “limit 2” is for limiting the selection to the first two rows. You can set an arbitral number. When the limit option is ignored, all the records are displayed on the terminal software, and it will take a longer time. Therefore, we recommend setting a small number for the limit option to check the data loading.

Similarly, load the data in files.txt into the files table and the data in peaks.txt into the peaks table.

```
MariaDB> delete from files;
MariaDB> load data local infile ”files.txt” into table files fields terminated by
’¥t’ lines terminated by ’¥r¥n’;

MariaDB> delete from peaks;
MariaDB> load data local infile ”peaks.txt” into table peaks fields terminated by
’¥t’ lines terminated by ’¥r¥n’;
```

＊Because of the huge data size of peaks.txt, it may take several minutes to finish the data loading.

# References

1. Metabolonote http://metabolonote.kazusa.or.jp/

Ara T, Enomoto M, Arita M, Ikeda C, Kera K, Yamada M, Nishioka T, Ikeda T, Nihei Y, Shibata D, Kanaya S and Sakurai N (2015) Metabolonote: a wiki-based database for managing hierarchical metadata of metabolome analyses. *Front Bioeng Biotechnol* 3: 38

2. FlavonoidSearch
http://www.kazusa.or.jp/komics/software/FlavonoidSearch

Akimoto N, Ara T, Nakajima D, Suda K, Ikeda C, Takahashi S, Muneto R, Yamada M, Suzuki H, Shibata D and Sakurai N (2017) FlavonoidSearch: A system for comprehensive flavonoid annotation by mass spectrometry. *Sci Rep* 7: 1243

3. PowerGetBatch

http://www.kazusa.or.jp/komics/software/PowerGetBatch

Sakurai N and Shibata D (2017) Tools and databases for an integrated metabolite annotation environment for liquid chromatography-mass spectrometry-based untargeted metabolomics. *Carotenoid Science* 22: 16-22

Sakurai N, Ara T, Enomoto M, Motegi T, Morishita Y, Kurabayashi A, Iijima Y, Ogata Y, Nakajima D, Suzuki H and Shibata D (2014) Tools and Databases of the KOMICS Web Portal for Preprocessing, Mining, and Dissemination of Metabolomics Data. *BioMed Research International* 2014: 1-11

4. MFSearcher http://webs2.kazusa.or.jp/mfsearcher/

Sakurai N, Narise T, Sim J-S, Lee C-M, Ikeda C, Akimoto N, Kanaya S (2018) UC2 search: Using unique connectivity of uncharged compounds for metabolite annotation by database searching in mass spectrometry-based metabolomics. Bioinformatics 34: 698-700

Sakurai N, Ara T, Kanaya S, Nakamura Y, Iijima Y, Enomoto M, Motegi T, Aoki K, Suzuki H and Shibata D (2013) An application of a relational database system for high-throughput prediction of elemental compositions from accurate mass values. *Bioinformatics* 29: 290-291