食品・植物・万物メタボロームレポジトリ

過去データの利用方法

2022 年 4 月 18 日版

国立遺伝学研究所
櫻井　望

1. 過去データについて

このドキュメントでは、以下の食レポジリーズより配布されている過去のデータの内容や使い方について解説しています。

食品メタボロームレポジトリ（食レポ）http://metabolites.in/foods
植物メタボロームレポジトリ（植レポ）http://metabolites.in/plants
万物メタボロームレポジトリ（物レポ）http://metabolites.in/things

食レポジリーズでは、ある時点で公開されていたデータを用いてユーザーが解析を行った結果について、その再現性を担保するため、大きなデータの更新がなされた場合に、それまでのデータをアーカイブし、配布しています。データの更新は、新規の分析データの追加以外にも、すでに公開されていたデータについて、サンプル情報などの修正や、化合物データベースの拡充に伴う検索結果の更新、新たな解析情報の追加などにより、不定期に実施されます。

2. 配布データ

食レポジリーズの主要な情報は、下記の 4 つのファイルに集約されています。

ファイル名	説明	形式	MariaDB
samples.txt	サンプルの情報です。サンプル名、サンプルの分類(カテゴリー)などを保持しています。	タブ区切り	✓
files.txt	分析データファイルの情報です。分析データのファイル名や、詳細な分析条件を記述した Metabolonote*サイトの ID、検出されたピーク数などの情報を保持しています。	タブ区切り	✓
peaks.txt	検出されたピークの情報です。m/z 値や保持時間、データベース検索結果の概要などを保持しています。	タブ区切り	✓
spectra.txt	マススペクトルの情報です。各ピークに紐づくマススペクトルデータを保持しています。	独自	

* Metabolonote は実験手法の詳細情報を専門で管理するデータベースです（参考文献1）

4 つのファイルのうち、samples.txt、files.txt、peaks.txt の 3 つは、タブ区切りテキスト形式であり、食レポシリーズのウェブシステムにおいて、リレーショナルデータベースシステムである MariaDB に読み込まれ、使用されます。spectra.txt は、独自のテキスト形式であり、MariaDB ではなく食レポシステムに直接読み込まれて使用されています。

MariaDB におけるテーブル定義は、以下のファイルに記述されています。データ更新の際、テーブル定義が変更されることもありますので、ご注意ください。

ファイル名	説明
create_tables.sql	MariaDB のテーブル定義情報です。samples.txt、files.txt、peaks.txt ファイルを、MariaDB へ読み込む際に使用されます。

3. データ内容

各配布ファイルに記載されたデータの詳細を解説します。

samples.txt

サンプル情報を保持しているファイルです。MariaDB では最初にできた食レポにちなみ、foods テーブルに読み込まれます。タブ区切りテキスト形式であり、各列は以下を示しています。

列番号	MariaDB フィールド名	説明
1	fid	サンプル ID。最初に食レポを作成したため、Food ID の意で fid となっています。
2	name_ja	サンプル名(日本語)
3	name_en	サンプル名(英語)
4	cat_ja	サンプルのカテゴリー分類(日本語)
5	cat_en	サンプルのカテゴリー分類(英語)
6	desc_ja	未使用
7	desc_en	未使用
8	dir	システム内でのデータファイルの配置場所

9	mnsid	Metabolonote*におけるサンプル ID
10	status	公開状況。現在はすべて「s1-published」となっています。

* Metabolonote は実験手法の詳細情報を専門で管理するデータベースです（参考文献1）

files.txt

分析データファイルの情報を保持しているファイルです。MariaDB では files テーブルに読み込まれます。タブ区切りテキスト形式であり、各列は以下を示しています。

列番号	MariaDB フィールド名	説明
1	id	空白。MariaDB に読み込む際に自動採番され、MariaDB 内ではユニークな ID として使われます。
2	fid	サンプル ID。最初に食レポを作成したため、Food ID の意で fid となっています。
3	type	ESI イオン化の極性（positive モードか negative モードか）を、pos または neg で示しています。
4	filename	分析データのファイル名を示すテキスト。先頭には、fid と type がアンダーバー「_」を介して付加されています。
5	mnmid	Metabolonote*1 の分析 ID。用いたサンプルと分析手法の詳細を一意に表す ID として使われます。
6	peaknum	検出されたピーク数
7	peaknum_n	N（窒素）原子が含まれると推定されたピークの数。評価されていない場合は-1 となっています。植レポの一部で使われています。
8	peaknum_s	S（硫黄）原子が含まれると推定されたピークの数。評価されていない場合は-1 となっています。植レポの一部で使われています。
9	peaknum_ms2	MS2 または MS/MS スペクトルを持つピークの数
10	peaknum_ms3	MS3 スペクトルを持つピークの数。評価されていない場合は-1 となっています。
11	peaknum_fs2	MS2 または MS/MS スペクトルを用いて、FlavonoidSearch*2 によるフラボノイドアグリコンの予測で 0 より大きいスコアが得られたピークの数。評価されていない場

		合は-1 となっています。
12	peaknum_fs3	MS3 スペクトルを用いて、FlavonoidSearch*2 によるフラボノイドアグリコンの予測で 0 より大きいスコアが得られたピークの数。評価されていない場合は-1 となっています。
13	peaknum_dbhit	化合物データベース検索結果を持つピークの数

*1 Metabolonote は実験手法の詳細情報を専門で管理するデータベースです（参考文献1）

*2 FlavonoidSearch は、マススペクトルデータからフラボノイドのアグリコンを予測するプログラムです（参考文献2）

peaks.txt

検出されたピークの情報を保持しています。配布ファイルの中で最もサイズが大きいファイルとなります。MariaDB では peaks テーブルに読み込まれます。タブ区切りテキストであり、各列は以下を示しています。

列番号	MariaDB フィールド名	説明
1	id	空白。MariaDB に読み込む際に自動採番され、MariaDB 内ではユニークな ID として使われます。
2	mnmid	Metabolonote*1 の分析 ID。用いたサンプルと分析手法の詳細を一意に表す ID として使われます。
3	mndid	Metabolonote*1 のデータ解析 ID の末尾部分。「D**」で表されます。mnmid と組み合わせ、用いた分析データとそのデータ解析方法の詳細を一意に表すために使われます。
4	pid	ピーク ID。mnmid および mndid と組み合わせて、検出されたピークを一意に表すために使われます。
5	type	ESI イオン化の極性（positive モードか negative モードか）を、pos または neg で示しています。
6	rt	クロマトグラフィーの保持時間（分）
7	mz	検出 *m/z* 値

8	int	ピーク強度。使用したデータ解析ソフト PowerGetBatch*2 で評価されたピークの面積値です。
9	intlog	ピーク強度のログ値。前列 int のピーク強度を10を底とする log 値に変換後、その分析データ内の全ピークの中央値が 0 となるようにセンタリングした値です。
10	adduct	推定されたアダクトイオン。複数の候補がある場合は、カンマ区切りで列挙され、最も可能性の高いものが最初に記載されています。
11	mzdi	中性の分子にした場合の質量値。検出された *m/z* 値とアダクトイオンの情報から計算されます。化合物データベース検索で用いられます。
12	annot	アノテーション情報。化合物データベースの検索結果を集計したものです。
13	skeleton	未使用
14	ms2	MS2 または MS/MS スペクトルの数
15	ms3	MS3 スペクトルの数
16	fs2	MS2 または MS/MS スペクトルを用いた FlavonoidSearch*3 のスコア。
17	fs3	MS3 スペクトルを用いた FlavonoidSearch*3 のスコア。複数の MS3 スペクトルがある場合はスコアの最大値を示します。
18	db_all	化合物データベース検索における構造異性体の数
19	db_kg	化合物データベース検索により KEGG で見つかった構造異性体の数
20	db_kn	化合物データベース検索により KNApSAcK で見つかった構造異性体の数
21	db_hm	化合物データベース検索により HMDB で見つかった構造異性体の数
22	db_lm	化合物データベース検索により LIPID MAPS で見つかった構造異性体の数
23	db_fl	化合物データベース検索により matebolomics.jp のフラボノイ

		ドデータベースで見つかった構造異性体の数
24	num_atom_n	N(窒素)原子の推定数。評価されていない場合は空欄。
25	num_atom_s	S(硫黄)原子の推定数。評価されていない場合は空欄。
26	pgroup	関連するピーク情報。同データ内に同じ化合物の異なるアダクトイオンなど、関連するピークがある場合に、その pid が自身も含めてカンマ区切りで示されています。 ＊2022 年 4 月 18 日以降のデータで追加されました。
27	num_shared	同様のピークを含むサンプル数。検出 *m/z* 値および保持時間を用いて、食レポシリーズ内を検索したときに、該当ピークが見つかるサンプルの数を示します。各ピークのサンプル特異性を考察する参考となります。 ＊2022 年 4 月 18 日以降のデータで追加されました。
28	option1	未使用

*1 Metabolonote は実験手法の詳細情報を専門で管理するデータベースです(参考文献1)

*2 PowerGetBatch は、ピーク検出、ピークの特徴づけ、ピークのサンプル間でのアラインメント、化合物データベース検索などを行うためのソフトウェアで、食レポシリーズのデータ解析全般で使われています(参考文献 3)

*3 FlavonoidSearch は、マススペクトルデータからフラボノイドのアグリコンを予測するプログラムです(参考文献2)

補足説明

➢ 1. 化合物データベース検索について

化合物データベース検索では、mzdi の値を用いて、食レポ・植レポでは 5 ppm、物レポでは 20 ppm の質量許容誤差を与え、以下の化合物データベースを対象に検索しています。

KEGG	http://www.genome.jp/kegg/
KNApSAcK	http://kanaya.naist.jp/KNApSAcK_Family/
HMDB	http://www.hmdb.ca/
LIPID MAPS	http://www.lipidmaps.org/
フラボノイドデータベース	http://metabolomics.jp/wiki/Category:FL

実際の検索では、MFSearcher サービス(http://webs2.kazusa.or.jp/mfsearcher, 参考文献4)の UC2 データベースが用いられています。UC2 データベースは、次のような特徴により、ヒットした化合物の多様性を考察する上で役立ちます。

- 同じ化合物(構造異性体)が、化合物データベース内、あるいは複数の化合物データベース間で、異なる電荷状態や塩などの付加体として重複して登録されている場合でも、それらを集約して1件として表す。
- 光学異性体は、1 件に集約されて表す。例えば L-アラニンと D-アラニンは 1 ヒット内に含まれます。

➢ 2. annot 欄について

同じ化合物が異なるデータベースに登録されている場合は、一番文字数の短い名前が自動的に採用されています。ヒットした構造異性体の名前がセミコロンで転結されていますが、合計の文字数が 200 文字を超える場合は、200 文字以降を切り捨て、末尾に「...」を付しています。名前が複数ある場合には、その数を「(* names)」として先頭に記載しています。

※データベースによっては、もともとセミコロンを使って複数の名前が登録されている場合があります。そのため、名前の数がデータベースヒット総数(db_all)と一致しない場合があります。

➢ 3. shared 欄について

検索に用いられた m/z 値と保持時間の許容誤差は以下の通りです。

レポジトリ	質量許容誤差(ppm)	保持時間許容誤差(分)
食レポ	5	1
植レポ	5	1
物レポ	20	0.5

spectra.txt

マススペクトルのデータを保持しています。このファイルは MariaDB には読み込まれず、食レポシリーズのシステムで直接使用されます。下記のような独自のテキスト形式となっています。

一つのマススペクトルは、以下のブロックで表されます。

・「>」で始まるヘッダー行
・それに引き続く、数字で始まるデータ行。複数行の場合があります。

各行はタブ区切りとなっており、各列は以下を表しています。

「>」で始まるヘッダー行

列番号	説明
1	「>」の文字。ヘッダー行であることを表します。
2	Metabolonote の分析 ID。peaks.txt の mnmid に対応します。用いたサンプルと分析手法の詳細を一意に表す ID として使われます。
3	ピーク ID。peaks.txt の pid に対応します。
4	マススペクトルのレベル。MS2 または MS/MS スペクトルは 2、MS3 スペクトルは 3 で表されます。
5	プリカーサーイオンの *m/z* 値
6	splash ID。マススペクトルから一意に計算される ID です。

数字で始まるデータ行

列番号	説明
1	プロダクトイオンの *m/z* 値
2	イオン強度の検出値
3	イオン強度の相対値。ベースピークイオン(マススペクトル中で最も強度が強かったイオン)の強度を 1000 とした際の相対強度を整数で表しています。
4	ニュートラルロスマス値。プリカーサーイオンとプロダクトイオンの *m/z* 値の差分です。
5	強度順に並べた時の順位

データ行は、イオン強度の強い順にソートされています。

MariaDB への読み込み

samples.txt、files.txt、peaks.txt について、MariaDB に読み込む方法を解説します。MariaDB に読み込むことで、大量のピーク検索などを高速に行えるため、大規模な解析等にご利用いただけます。

ここでは、MariaDB のインストールや基本的な取り扱い方法を習得している方を対象としています。

以下の解説では、データベース名、ユーザー名、パスワードの例として、以下の設定を使った作成例を解説します。これらの設定はご自分の環境に合わせて適宜変更してください。

データベース名	mydb1
ユーザー名	user1
パスワード	pass1

1. データベースの作成

データを読み込むデータベースを新規に作成し、アクセス権を付与します。

コマンドプロンプトなどの端末ソフトを開き、MariaDB のコンソールにログインしたのち、以下のコマンドを実行します。

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

以上で mydb1 データベースが作成され、user1 に対するアクセス権が設定されました。データベースが作成されたかどうかは、show databases コマンドで確認できます。

```
MariaDB> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| mydb1              |
| test               |
```

```
+--------------------+
3 rows in set (0.00 sec)
```

※ information_schema や mysql という名前のデータベースは、MariaDB が使用しているデータベースです。

2. テーブルの作成

次に、作成した mydb1 データベースに、データを流し込む受け皿となるテーブルを作成します。データの更新に伴い、create_tables.sql の定義が変更になっている場合があります。異なるバージョンの過去データを利用する場合には、必ず配布データに同梱されている create_tables.sql を使ってテーブルの作成をし直してください。

コマンドプロンプトなどの端末ソフトで、create_tables.sql ファイルが存在するディレクトリ（フォルダ）に移動します。移動には cd コマンドなどを使用します。

以下のコマンドを実行します。

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

パスワードを聞いてくるので、pass1 と入力します。

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

※ 画面上ではパスワードの文字が＊で隠されて表示されます。

以上で、create_tables.sql に記載されたテーブル定義に従って、mydb1 データベース内に必要なテーブルが作成されました。

テーブルが作成されたかどうかを確認するには、MariaDB にログインして、mydb1 を選択したのち、show tables コマンドで確認できます。

```
> mysql -u user1 -p
```

```
Enter password: *****
… 省略 …

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

3. データの読み込み

コマンドプロンプトなどの端末ソフトで、samples.txt、files.txt、peaks.txt が存在するディレクトリ(フォルダ)に移動します。移動には cd コマンドなどを使用します。

MariaDB のコンソールにログインして、mydb1 を選択します。

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

MariaDB> use mydb1;
```

その後、以下の二つのコマンドを順次実行します。

```
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: *
```

※ 配布ファイルの改行文字は Windows(CR+LF)となっています。このため、環境によってはデータをロードする際の line terminated by **** を適切に設定する必要があります。

最初のコマンドで foods テーブルの内容を全部削除し、次のコマンドで foods テーブルに samples.txt からデータをすべて読み込んでいます。最初に食レポシステムが作成されたため、テーブルの名前が foods になっていますが、流し込むデータファイルは samples.txt ですので、ご注意ください。

上記のように、Query OK…と表示されれば処理が完了です。

データがロードされたかどうかは、select コマンドで確認できます。

```
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)
```

※ limit 2 は、最初の 2 行だけを選択するための設定です。数字は任意に設定できます。limit の設定をしないと全件が表示されてしまい、表示終了までに時間がかかるため、小さい数を設定することを推奨します。

同様に、files テーブルに対して filise.txt のデータを、peaks テーブルに対して peaks.txt のデータをロードします。

```
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';
```

※ peaks.txt はファイルサイズが大きいため、ロードが完了するまでに数分かかる場合があります。

参考文献

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