植物・食品メタボローム解析メタデータのRDF化および測定生データの再解析に向けて

○長崎 英樹[1]、大澤 祥子[2]、荒 武[2、3]、福島 敦史[1]、高橋 みき子[1]、小林 紀郎[4]、櫻井 望[5]、平川 英樹[2]、有田 正規[1、4]

1.理化学研究所環境資源科学研究センターメタボローム情報研究チーム、2.かずさ DNA研究所ゲノム情報解析施設、
3.京都大学生存圏研究所、4.理化学研究所情報システム本部データ知識化開発ユニット、5.国立遺伝学研究所DDBJセンター

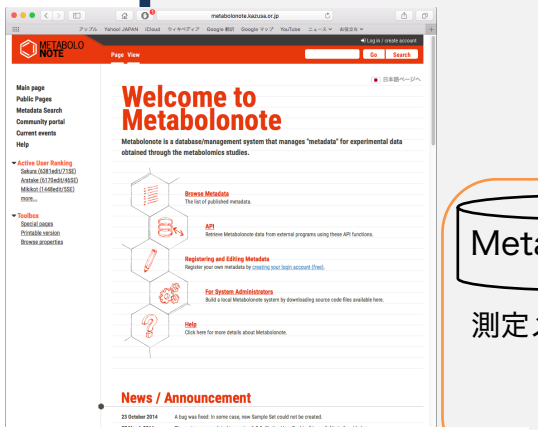
【要旨】
多様な代謝物の同定を行うメタボローム解析は、手法、解析機種、その設定も多種多様でそれらを収めたメタデータも複雑になる。統合化推進プロジェクトにおいて我々はメタボローム解析メタデータのResource Description Framework (RDF) というデータ形式でのアーカイブ化を計画している。RDFはデータに意味や属性を持たせるため、複雑なメタボロームのメタデータも検索等のデータ処理が容易になる。

かずさDNA研究所で所有するメタボローム解析メタデータは植物、食品の104プロジェクト、1,466サンプルでデータベース化されMetabolonote (https://metabolonote.kazusa.or.jp) として公開されてきたが、登録者の利便性を図って一部自由書式で登録されている。このため我々は改めてデータの意味付け、関連付けを人の目を介してRDF化を行った。

また、そのメタデータの測定生データは解析してから数年たつものが含まれているので、再解析を計画しているが、植物由来の生データだけでも数にして約2,700あるため、PowerGetBatch (Sakurai and Shibata, 2017) 等のソフトウェアをスーパーコンピュータ上で稼働させるなど効率化を図っている。

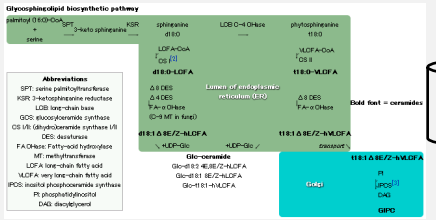
ライフサイエンスデータベース統合推進事業統合推進化プログラム
「物質循環を考慮したメタボロミクス情報基盤」の概要

MetaboBankの構成



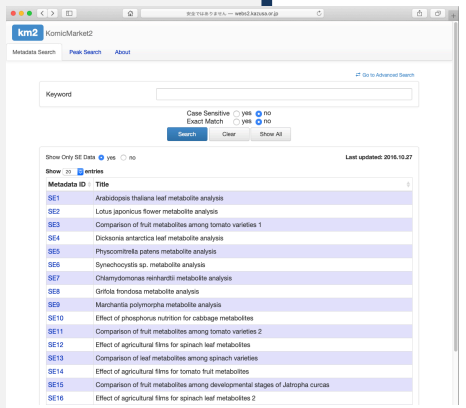
http://metabolonote.kazusa.or.jp

メタ代謝マップ



化合物文献

KNApSAcK



http://webs2.kazusa.or.jp/km2/

Metabolonote

測定メタデータ

旧MassBank

標準マススペクトル

KomicMarket2

データ・リポジトリ(代謝産物の解析結果を集積したdb)

MetaboBank

実験データ
メタデータ
提供

かずさDNA研究所

RIKEN
理化学研究所

Metabolonoteの特徴

メタボローム解析時のサンプル情報や分析情報(メタデータ)を専門に取り扱うデータベース

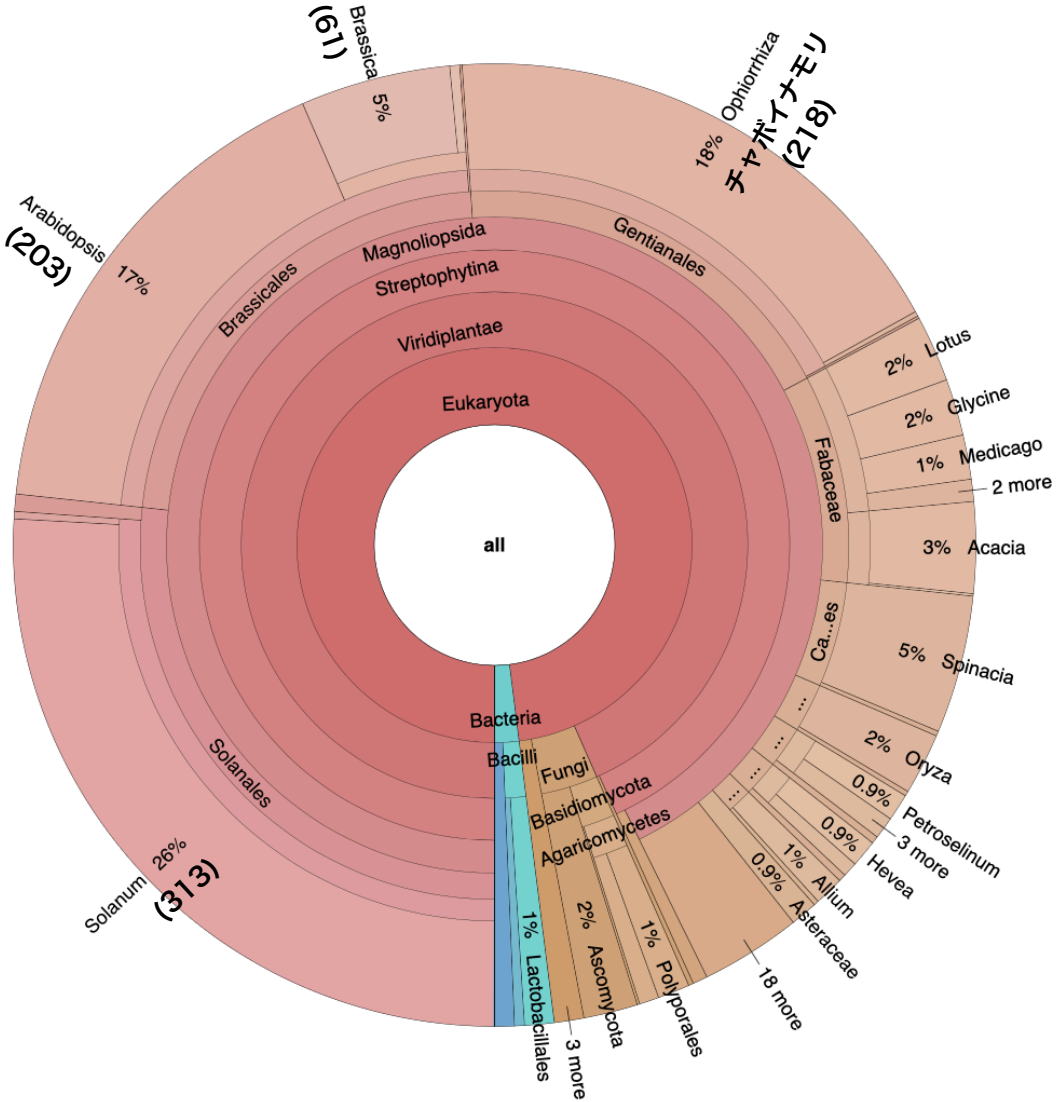
Wikiで作られていて作成や編集が容易

メタデータをXMLで一括出力可

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

Ara *et al*. Front Bioeng Biotechnol. 2015 Apr 7;3:38.

Metabolonoteのメタデータの生物種類分類



総数 1,210 うち植物 1,122

=> 数は多くないけど生物種の構成はユニーク

MetaboBank は Resource Description Framework (RDF)という形式で構築。

Metabolonoteの書式自由度が高く、データ間の紐付けが緩いリテラルのメタデータをMetaboBank用に、定義づけをやり直す。

Description	Harvested sample is frozen by liquid N2 and resulting powder (100mg) are solved in 300uL 80% methanol solution. 20uL sample is injected into HPLC after 0.2um membrane filter treatment. HPLC conditions: Agilent 1100 series (Agilent), Column: TSKgel-100V (4.6 x 250 mm, 5 micrometer; TOSOH), Solvent: A; 0.1% formic acid aq. B; ACN (addition 0.1% formic acid fc.), Gradient: (B);3 to 30% (0.0 to 25.0 min), 30 to 90% (25.0 to 40.0 min), 90% (40.0 to 45.0 min), 95% (45.1 to 50.0 min), 3% (50.1 to 57.0 min), Column temp.: 30 degree C, Flow rate=0.5mL/min, PDA: 200-650 nm (2 nm step). FT-ICR-MS conditions: Filter 1: FTMS + c norm !corona !pi res=50000 o(200.0-1500.0); 2: ITMS + c norm !corona !pi Dep MS/MS Most intense ion from (1).;3: ITMS + c norm o(200.0-1500.0)., Rejected mass=266.0000;294.0000;391.0000.

PerlスクリプトによるXMLファイルからデータ抽出、整理

```
<Page ID="" Title="/S01/M01">
<Template Name="M"><Field Name="M_ID">M01</Field><Field Name="M_Title">LC-FTICR-MS, ESI Positive analysis</Field><Field
Name="M_Method Set ID">MS1</Field><Field Name="M_Sample Amount">6.7 mg</Field><Field Name="M_Comment">[MassBase ID] MDLC1_05711</
Field></Template>
<Free_Text id="1">{{LinkTo_MassBase|MDLC1_05711}}</Free_Text>
</Page>
<Page ID="" Title="/S01/M01/D01">
<Template Name="D"><Field Name="D_ID">D01</Field><Field Name="D_Title">PowerGet data analysis for Bio-MassBank</Field><Field
Name="D_Data Analysis Set ID">DS1</Field><Field Name="D_Recommended decimal places of m/z">6|ITMS 2</Field></Template>
<Free_Text id="1">{{LinkTo_BioMassBank|keyword=SE1_S01_M01_D01}}
&lt;!-- {{LinkTo_KomicMarketTmp}} --&gt;
{{LinkTo_KM2|SE1}}</Free_Text>
</Page>
<Page ID="" Title="/S01/M01/D02">
<Template Name="D"><Field Name="D_ID">D02</Field><Field Name="D_Title">PowerGet data analysis for KomicMarket2</Field><Field
Name="D_Data Analysis Set ID">DS2</Field><Field Name="D_Recommended decimal places of m/z">6|ITMS 2</Field></Template>
<Free_Text id="1">&lt;!-- {{LinkTo_KomicMarketTmp}} --&gt;
{{LinkTo_KM2|SE1}}</Free_Text>
```

PGDBjやMassBaseなどかずさDNA研の他のdbへのリンクもまとめて出力

```
#S01/M01
M_ID:   M01
M_Title:    LC-FTICR-MS, ESI Positive analysis
M_Method Set ID:    MS1
M_Sample Amount:    6.7 mg
M_Comment:  [MassBase ID] MDLC1_05711


#S01/M01/D01
D_ID:   D01
D_Title:    PowerGet data analysis for Bio-MassBank
D_Data Analysis Set ID: DS1
D_Recommended decimal places of m/z:    6|ITMS 2


#S01/M01/D02
D_ID:   D02
D_Title:    PowerGet data analysis for KomicMarket2
D_Data Analysis Set ID: DS2
D_Recommended decimal places of m/z:    6|ITMS 2
```

…とスクリプト処理もしているが基本的にデータを手入力

Analytical Method Details Information

ID	MS1
Title	LC-FT-ICR-MS ESI positive method 1
Instrument	Agilent1100 HPLC (Agilent), LTQ-FT (Thermo Fisher Scientific)
Instrument Type	LC-FTICR-MS
Ionization	ESI
Ion Mode	Positive
Description	
Extruction	Harvested sample is frozen by liquid N2 and resulting powder (100mg) are solved in 300uL 80% methanol solution. 20uL sample is injected into HPLC after 0.2um membrane filter treatment.
HPLC	HPLC conditions: Agilent 1100 series (Agilent), Column: TSKgel-100V (4.6 x 250 mm, 5 micrometer; TOSOH), Solvent: A; 0.1% formic acid aq. B; ACN (addition 0.1% formic acid fc.), Gradient: (B);3 to 30% (0.0 to 25.0 min), 30 to 90% (25.0 to 40.0 min), 90% (40.0 to 45.0 min), 95% (45.1 to 50.0 min), 3% (50.1 to 57.0 min), Column temp.: 30 degree C, Flow rate=0.5mL/min, PDA: 200-650 nm (2 nm step).
Mass	FT-ICR-MS conditions: Filter 1: FTMS + c norm !corona !pi res=50000 o(200.0-1500.0); 2: ITMS + c norm !corona !pi Dep MS/MS Most intense ion from (1).;3: ITMS + c norm o(200.0-1500.0)., Rejected mass=266.0000;294.0000;391.0000.

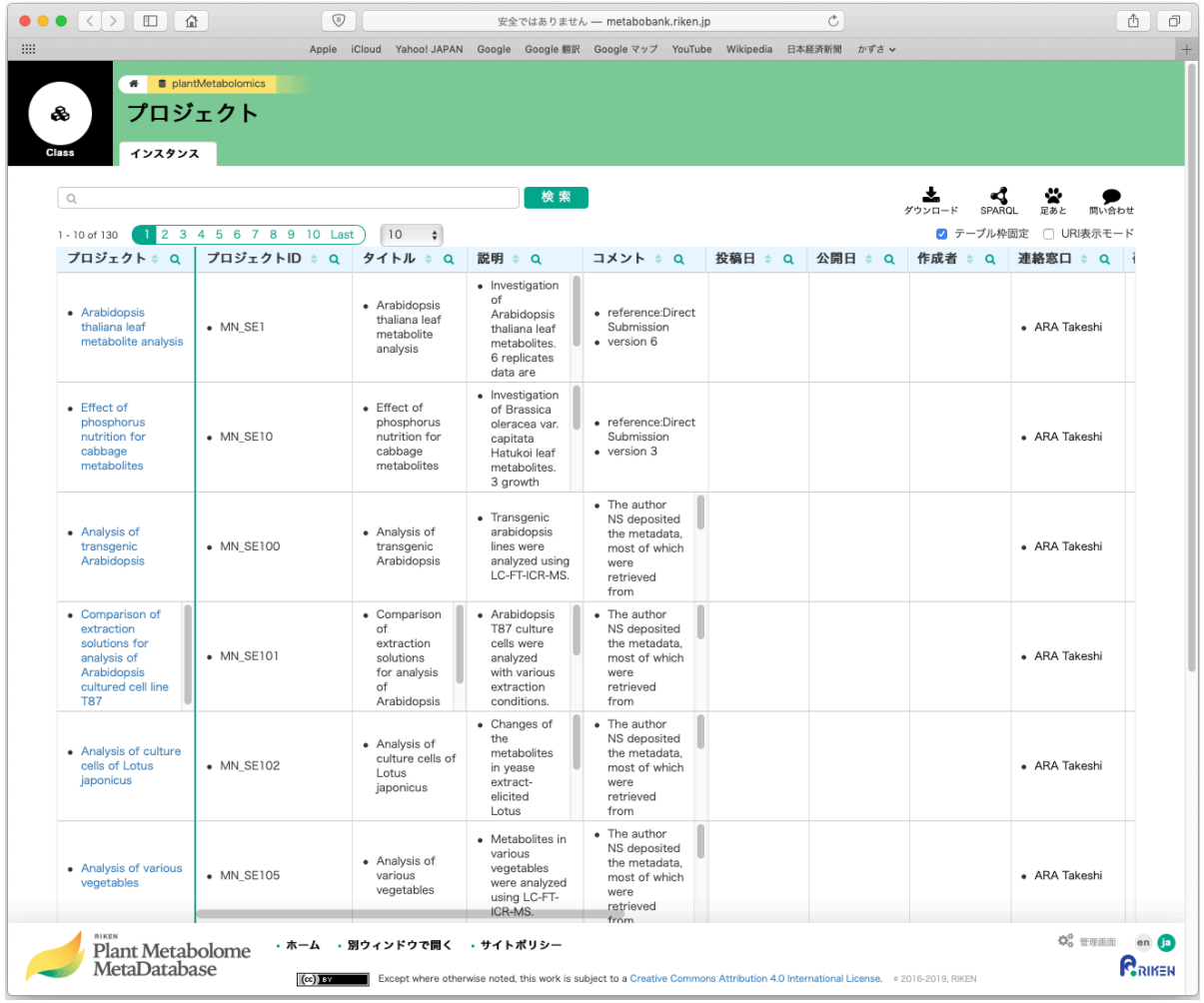
植物等のかずさ独	82プロジェクト
食品メタボローム	16プロジェクト
その他(東大など)	6プロジェクト
合計	104プロジェクト

共通の項目(クラス:述語)でリンクされた
50シートのRDF変換用エクセルファイル

Chromatography	comment	description	temperature gradient	column type	column temperature	column pressure	column name
クロマトグラフィ	コメント	説明	温度勾配	カラムの種類	カラム温度	カラム圧力	カラム名
	rdfs:comment	dcterms:description	temperatureGradient	columnType	columnTemperature	columnPressure	columnName
Chromatography	rdf:langString	rdf:langString	xsd:string	xsd:string	Temperature	Pressure	xsd:string
pm_chromato:MN_SE2_HPLC	"Elute monitoring by PDA equipped with Agilent1100 HPLC in 2 nm step."@en				[temp:30C]		"TSKgel-100V (4.6 x 250 mm, 5

=>エクセルファイルに収めたデータをRDF化 (turtle ファイル作成)

理研とのデータ連携協定に基づき理研のPlantMetabolomeMetaDatabaseから植物のみで構成された77プロジェクトが現在公開されている。



http://metabobank.riken.jp/pmm/db/plantMetabolomics

Metabolonoteの植物サンプル(1,122)のメタデータに紐づいている実験生データ(2,762)の再解析

http://www.kazusa.or.jp/komics/ja/tool-ja/235-powergetbatch.html

PowerGetBatch

液体クロマトグラフィー(LC)-高分解能質量分析(MS)のデータから、ピーク抽出、アダクト判定、サンプル間のアラインメントを行い、化合物データベース検索、FlavonoidSearchによる一次アノテーションを行う。

現在パラメータ調整など開発者と検討、再解析の優先順位等を検討している。後にスパコン等で大量再解析するとともに一部キュレーションをして精度を確認する。
構造物のdbも充実してきているため、以前より構造が決定できる代謝物が増えることが期待できる。

PowerGetBatch実行時間観測

PowerGetBatchをLinux上での並列処理が実行できるように調整 (各種パラメータは要検討)。
Metabolonoteにメタデータが登録されていたヤトロファ由来のmzXMLファイル12ファイルのうち、数が足りない分はコピーして名前を変更して使用した。

生データファイル数(mzXML)	ピーク検出 (min)	アライメント (min)
12	179.41	15.44
24	330.71	57.78
48	684.42	189.88
96	1337.14	*6904.02

Intel Xeon 2.60 GHz 20コア, 20GB RAM
*Intel Xeon 2.60 GHz 20コア, 60GB RAM

=>大規模データセットに対する計算速度の改善が必要

今後の課題

人手も足りないので以下のような点を効率化するシステムを構築して結果の評価、キュレーション、問題点の検討に時間と労力を割くようにする。

・データ移植、入力作業の効率化

・再解析パイプラインの改良

・再解析結果を比較、評価する仕組み (自動、半自動) の確立

・再解析結果の公開