平成 23 年度　研究開発実施報告書

ライフサイエンスデータベース統合推進事業「統合化推進プログラム」
平成 23 年度採択　研究代表者

金谷重彦
奈良先端科学技術大学院大学・情報科学研究科

メタボロームデータベースの開発

§1. 研究実施体制

(1)「奈良先端大」グループ
① 研究代表者:金谷重彦 （奈良先端科学技術大学院大学、情報科学科研究科、教授）
② 研究項目
・質量スペクトル DB（化合物 MSDB とメタボローム MSDB）の拡充
・代謝物質情報 DB の構築
(2)「かずさ DNA 研究所」グループ
① 主たる共同研究者: 櫻井望 （かずさ DNA 研究所 産業基盤開発研究部 研究員）
② 研究項目
・メタボローム MS データベースの開発
・代謝物質情報 DB の構築
(3)「理研」グループ
① 主たる共同研究者: 有田　正規 （理化学研究所植物科学研究センター、副グループディレクター）
② 研究項目
・メタボローム MS DB の開発
・ウィキを中心としたメタボローム統合データ統合技術開発
・メタボローム MS データベースの開発

§2. 研究実施内容

かずさ DNA 研究所、奈良先端大、理研との共同で、大量に蓄積してきた質量分析データをすみやかに Bio-MassBank にて公開し KNApSAcK や MassBank と連携したデータ利用を促進することを目的として、メタボローム共通フォーマットの作成、データ処理パイプラインの整備行い、データ公開を進めた。

メタボローム研究では生物試料を LC-MS, MS2 分析をおこなっている。生物試料1件あたり数百から数千の代謝物が検出されている。これら多数のマススペクトルを手作業で参照マススペクトルと照合して代謝物を同定するスペクトル検索はとても時間と労力を要する作業であった。とりわけ LC-MS2 データは各機器メーカに特有のデータ管理システム上にバイナリ形式で保存されていて、外部データベースを検索することによって同定することは想定されていなかった。そこでこの問題を解決するために、田中最先端研究所（FIRST プログラム）とエーザイ（株）が開発した Mass++プロジェクトと連携して、多量のバイナリ形式 MS2 データを入力として MassBank を一括してスペクトル検索を行うための SOAP-API インターフェイスを開発提供した。これによって 1 千スペクトルをクエリとした場合に約 30-40 分でスペクトル検索、同定をすることを実現した。このような検索機能は MassBank が世界に先駆けて実現したものであり、 NIST データベースにも未だ無い。

MassBank は日本質量分析学会の公式データベースとしても認定されている。平成 23 年度はこれまでに収集された ESI-MS2 データの監修を重点的に実施した（平成 23 年 7 月 30 日奈良先端大）。平成 24 年 3 月 10 日開催の本学会委員会で本ワーキンググループは「スペクトルデータ部会」に昇格することが全委員一致で決められた。スペクトルデータ部会は 6 番目の研究部会であり、本学会が MassBank を重要研究対象として認めたことになった。今後、MassBank の国際的普及に努めるだけでなく、現在 MassBank で公開されている高精度なマススペクトルデータを利用した新しい研究を学会と連携して実施する予定である。平成 24 年度からアジア地域で新らたに発行される質量分析英文誌や平成 24 年 9 月に京都で開催される国際質量分析会議（IMSC2012）の要旨集ではマススペクトルを MassBank で公開することが投稿規程に盛り込まれている。また、データの世界標準を決めることを目的に、2012年2月、成田に日米欧の主要 PI を集め、メタボローム国際ワークショップを開催した。計測データの公開と InChI と PubChem 番号等の ID を付与すること、アノテーションには信頼度をつけた上で既存のオントロジーを利用することに合意した。この結果をもとに日米欧間でメタボロミクスデータの共有を進めることになった。かずさ DNA 研究所が中心となり、奈良先端大チームと協議しながら検討を進め、2011 年 10 月 26 日に公開した（http://webs2.kazusa.or.jp/togodb/）フォーマットが基本的には採用され、2 生物種計 8 サンプルのデータを整備・提供し、計 1300 スペクトルのデータを MassBank より公開した。メタボロームオントロジーの整備は、化合物カテゴリー名 19 種、アノテーション状態 2 種、エビデンス名 3 種を選定し、データ処理に利用した。計 10545 件のデータについてサンプルや分析の概要情報をまとめ、新フォーマットへ変換する優先順位をつけてメンバー内に提供した。既存のデータ型式として MassBase に 5058 件、KomicMarket に 12 件のデータを追加し公開した。このように本プロジェクトで開発されている MassBankシステムは、世界の研究者との連携をも考慮した標準質量スペクトルデータベースとしての役割を担うに至った。

2011 年度は東北大震災によって東日本で停電や節電のためにサービスができなかったデータサーバが多かった。そこで関西（奈良先端大）にミラーサーバを設置する計画をたてて、PC を一台購入した。MassBank で公開しているデータはいずれも Creative Commons License 表示をしているので、最初に分散している各データサーバから定期的にデータのバックアップコピーを収集するツールを開発した。これらバックアップコピーはライプチッヒ植物生化学研究所（ドイツ）のデータサーバにも置いて、EU 地域でのサービス向上に利用している。平成 24 年度はミラーサーバのシステム開発をおこなう計画である。

質量スペクトルにより代謝物の化学構造を推定する目的で、ESI-MS2 データ（マススペクトル）で観察されたピーク（product ion）について、そのイオンが由来した部分化学構造との関係（「ピークと部分化学構造との関係」）を解析、収集してきた。一般に与えられた分子式に対応する部分化学構造は複数ありうるが、ピークと部分化学構造との関係のうち1：1の対応関係にあるものを精選したところ、269分子式と94部分化学構造との関係（630ペア）が得られた（1つの部分化学構造から複数のイオンに開裂するので、分子式の数が部分化学構造の数よりも多くなっている）。開発した化学構造式ツールは、先ず ESI-MS で観測された分子イオンピーク（精密質量）に該当する代謝物を KNApSAcK データベースから選び、次いで、ESI-MS2 で観察されたピークについて「ピークと部分化学構造との関係」から推定される部分化学構造を有する代謝物を絞り込むことによって、代謝物を推定しようとするものである。このツールは MassBank に「Metabolite Identification」として実装した。未だ「ピークと部分化学構造との関係」の数が少ないので、このツールで推定することができる代謝物はまだ限られている。平成 24 年度では多様な二次代謝物について、文献などから「ピークと部分化学構造との関係」を広く収集することによって推定対象とする代謝物を広げていく計画である。

メタボロームデータベースを統合化するためには、質量スペクトル情報以外の情報として、代謝物についての化学、生物学情報を整理する必要がある。そこで、代謝物の化学構造情報、代謝物-生物種の関係、ならびに代謝物-生物活性の情報についてのデータベース設計を行い、文献データをもとに実際にデータを収集し、本データベースを公開するとともに論文として公表した(1-3)。これらの情報の充実を 24 年度以降も粛々と進めて行く予定である。全ての代謝物情報、すなわち、メタボローム情報の統合化をめざし、理化学研究所では、CAS, InChI コード逆引き辞書を作成した。複数データベースの統合には CAS, InChI コードが最適という研究プロジェクト内の合意を踏まえ、基礎代謝物 1800 化合物について CAS, InChI コードから構造情報を逆引きする wiki ページを構築した。KNApSAcK および Flavonoid DB のデータも追加する予定である。

§3. 成果発表等

(3—1)データベースおよびウェブツールの構築と公開

① 公開中のデータベース・ウェブツール等

データベース名： **KNApSAcK Core DB**

概要： 本DBは、文献情報をもとに生物種代謝物の関係を収集し、公開している。現在までに、10万種の生物種-代謝物関係が蓄積されており、メタボロミクス研究の標準データベースとなっている。

公開日：H16 年 4 月 1 日

URL： http://kanaya.naist.jp/knapsack_jsp/top.html

アクセス数：アクセス数についてはカウントしていないが毎年数十の論文で引用されるにいたっている。

データベース名： **Metabolome Activity DB**

概要： 本 DB は、文献情報をもとに代謝物の活性情報を収集し、公開している。現在までに、5s千種の代謝物-活性の関係が蓄積されており、メタボロミクス研究の標準データベースとなっている。

公開日：H24 年 2 月 1 日

URL： http://kanaya.naist.jp/MetaboliteActivity/top.jsp

アクセス数：1030 アクセス/年。

データベース名： **MassBank**

概要： 本DBは、代謝物質あるいはそれらの関連物質をEI-, FAB-, ESI-, MALDI-MS, MS2などを用いて分析したマススペクトルを収集、公開している。データの公開状況（2012 年 3 月末）は次のとおりである。21 研究グループ（日本 15、米国 3、ドイツ 2、中国 1）が 13,534 化合物について分析した合計 29,644 マススペクトルデータを 9 つのデータサーバから公開している。それらのうち、ESI-MS、MS2 データは 2,304 化合物について測定した 16,440 件である。

公開日：H18 年 12 月 6 日

URL： http://www.massbank.jp/

アクセス数：2011 年 10 月（一ヶ月間）のアクセス数は 10,063（ユニークな IP アドレス数/月）であった。

データベース名： **Bio-MassBank**

概要： 本 DB は、植物の組織や微生物試料を LC-, GC-, CE-MS で分析して得られたマススペクトルを代謝物を同定できた、できないにかかわらず収集する。未同定代謝物のマススペクトルをその化学構造を表現する化学的 descriptor として利用することによって、異なる試料間で同じあるいは類似したマススペクトルがあれば同じ未同定代謝物が含まれている、と考えることができる。このように代謝物を同定することができなくても、その存在を知ることができる。現在、シロイヌナズナの葉、ミヤコグサの花を LC-MS, MS2 分析したデータそれぞれ 664 件、636 件を公開。

公開日：H23 年 12 月 16 日

URL： http://bio.massbank.jp/

アクセス数： 統計をとらず

データベース名： MassBase

概要： 本 DB は、主に質量分析の未加工データ（生データ）、生データをテキスト型式に変換したデータ、同定・推定（アノテーション）を行わないピークデータを、蓄積・公開するための DB である。

公開日：H20 年 1 月 7 日 （H23 年度中の最新アップデートは H24 年 2 月 29 日）

URL： http://webs2.kazusa.or.jp/massbase/

アクセス数：

H20 年 9 月から H24 年 4 月 24 日　　7262 回

データベース名： KomicMarket

概要： 本 DB は、化合物ピークのアノテーションを行った質量分析データ（主に LC-MS）を蓄積・公開するための DB である。

公開日：H19 年 8 月 7 日 （H23 年度中の最新アップデートは H24 年 2 月 27 日）

URL： http://webs2.kazusa.or.jp/komics/

アクセス数：

H20 年 11 月から H24 年 4 月 24 日　　18,870 回

データベース名： MFSearcher

概要： 本 DB は、精密質量分析で得られた精密質量値から、組成式を迅速に推定するためのウェブサービスである。また、KNApSAcK, KEGG, PubChem へ精密質量値からの検索も高速に行うことができる。ハイスループットなピークアノテーションに利用されている。

公開日：H21 年 6 月

URL： http://webs2.kazusa.or.jp/mfsearcher/

アクセス数：

公開日から H24 年 4 月 23 日現在　　29,133,347 回

データベース名： MS-MS FragmentViewer
概要： 本DBは、フラボノイド標品のMS/MSフラグメントを化学構造に帰属させ、分子開裂モデルを提示しているデータベースである。化合物ピークのアノテーションに利用されている。
公開日：H20 年 8 月 25 日
URL： http://webs2.kazusa.or.jp/msmsfragmentviewer/
アクセス数：
H20 年 12 月から H24 年 4 月 24 日　　18,796 回

データベース名： metabolomics.jp
概要： 本 DB は、メタボロミクスを中心とした有田研究室の活動全般を対象としたポータルサイトである。フラボノイド、基礎代謝物、生薬、植物系統分類データベースのほか、ファイトレメディエーションを含めた放射線情報、講義資料も掲載している。
公開日：平成19年
URL： metabolomics.jp
アクセス数：
ウェブページ http://metabolomics.jp/awstats/awstats.pl?config=metabolomics.jp から閲覧可能

（3－2） 原著論文発表

① 発行済論文数（国内（和文） 0 件、国際（欧文） 1 件）：
② 未発行論文数（“accepted”、“in press”等）（国内（和文） 0 件、国際 （欧文） 2 件）
③ 論文詳細情報

1. Afendi,FM,, Okada,T., Yamazaki, M, Morita, A., Nakamura, Y.,Nakamura, K., Ikeda, S, Takahashi, H., Amin, M. Latifah K. Darusman, LK., Saito, K. Kanaya,S., KNApSAcK Family Databases: Integrated Metabolite.Plant Species Databases for Multifaceted Plant Research, Plant Cell Physiol. 53(2): e1(1.12) doi:10.1093/pcp/pcr165
2. Afendi FM., Katsuragi, T, Kato, A., Nishihara, N., Nakamura, K., Nakamura,Y., Tanaka,K., Hirai Morita,A. , Amin, A., Takahashi, H., Kanaya, S., Systems Biology Approaches and Metabolomics for Understanding Japanese Traditional Kampo Medicine, Curr. Pharmacogenomics Personalized Med. 10 (in press) (2012), 10
3. Wada,M., Takahashia,H., Amin, M., Nakamura, K., Hirai, MY., Ohta, D, Kanaya,S., Prediction of operon-like gene clusters in the Arabidopsis thaliana genome based on co-expression analysis of neighboring genes, Gene, (2012), (accepted)

