メタボロームＭＳスペクトル統合データベースの開発

慶應義塾大学大学院政策・メディア研究科
西岡　孝明

Integrated Database of Metabolome and Their Mass Spectra

Takaaki Nishioka
Graduate School of Governance and Media, Keio University

MassBank is a public repository for sharing mass spectral (MS) data among biological science research communities. It is a distributed database. Contributors share the cost of managing MassBank by preparing their spectral data in a common record format and by publishing the formatted data from their own data servers. MassBank is contributed from 19 research groups with a total of 30,312 MS data analyzed on 11,614 chemical compounds. KNApSAcK is a biological species-metabolite database which collects 101,500 species-metabolite pairs consisting of 20,018 biological species and 50,048 secondary metabolites. Metabolomics.JP is a new type of Wiki, on which Fragmentation Library, a library of the relationships between mass spectral peaks and chemical substructures, was constructed. KNApSAcK and the library will be integrated in a tool of chemical structure elucidation of unknown metabolites from their MS data.

１．はじめに

質量分析（MS）はどのような化合物でも高感度に検出できるので、プロテオームやメタボローム解析などの生物科学研究における基盤技術である。しかし、メタボローム解析では検出された代謝物質のうち同定できたものは数%（二次代謝物質）から 20 数%（一次代謝物質）に留まっている。このことがメタボローム研究の発展を阻害している。

代謝物質の同定は、検出したマススペクトル（MS データ）をあらかじめ測定しておいた既知代謝物質の MS データ（参照データ）と照合することによって、おこなっている。従って、同定率を向上するためには、できるだけ多数の参照データを研究者コミュニティで共有することである。さらに参照データが無い代謝物質は化学構造式を推定しなければならない。

本開発研究の目的は、代謝物質について測定した MS データを研究者コミュニティで共有するための public repository として MassBank を構築し、これらのデータを利用して質量分析で検出した代謝物質の同定あるいは未同定代謝物質の化学構造の推定をおこなうためのツールを開発することによって、メタボローム研究の発展に寄与することである。研究者コミュニティによるデータベースの維持と管理のためのツール類の開発をはじめ、質量分析の多様性を考慮した参照データの開発、化学構造式推定の基盤となる MS データと化学構造式の関係の解析、二次代謝物質データベースの開発、Wiki 上での知識集約をおこなった。

2. 研究開発の成果

2.1 分散型データベースの開発と実用化［1］

研究者コミュニティが低コストで維持管理できる分散型データベースMassBankを開発した。MSデータを公開する研究者は、共通のレコード形式（MassBank record format）でデータを作成し、自ら提供したデータサーバに登録して公開する。一方、MassBankのデータを利用する研究者はどこにデータサーバがあるのか知らないが、MassBankのWebページから全てのデータにアクセスすることができる。すなわち、MassBankシステムはデータサーバおよびアクセスポイントとして機能する。

分散したデータサーバにアクセスし、検索を高速で並列処理するサーバ間通信技術を確立した。また、どれか不具合のサーバがある場合の検出・対処や、分散しているデータサーバにインストールされているMassBankシステムの一斉update、の技術も開発した。

2.2 MassBankでデータを公開する研究者のためのツール開発［1］

データを公開する研究者がデータ量に応じてMassBankの維持、管理作業の大部分を負担することにした。これらの作業を容易におこなうことができないと、誰もデータ公開に協力しないと危惧されたので、公開のためのツールを開発して、提供した。

PCにLinux OSやMySQL、Apacheを含むMassBankシステムをインストールしてデータサーバをセットアップするインストーラの開発をおこなった。次に、測定したMS binary data（装置に依存したbinary data形式）からMassBankレコードを作成するためのツール開発をおこなった。MassBankレコードはピークデータと、装置や測定条件設定のパラメータの記述を必須データ項目としているので、Mass++プロジェクト（JST-CREST 2005-2010、小田吉哉代表)と連携して、MS binary dataを読みこみ、これら必須項目を自動抽出してMassBankレコード形式で出力するまでを自動化した。この出力に、測定した化合物の構造式情報をmolfileとして与えると、SMILESやInChIコードの作成、分子イオンの質量計算などをおこなって完全なMassBankレコードを作成するRecord Editorを開発した。完成したMassBankレコードをデータサーバ上に登録し、データの管理をおこなうAdministration Toolも開発した。MassBankインストーラ、Mass++、Record Editor、Administration Toolは研究者をサーバの設置、データ作成とサーバ上でのデータ管理という面倒な作業から解放した。

Windows PC版のMassBankシステムも開発した。異なるメーカのMS装置で測定したMSデータを1つのPCで一括管理することを可能にしたことが企業や大学の研究室で好評になり、downloadはLinux版 294件に対してWindows版1,855件（最近２年間）に上った。

2.3 MassBankは研究者コミュニティデータベースとして認められた

当初予定していた４つのMSデータベースであるPrimeデータベース（理研 PSC)、LipidSearch脂質関連データベース（東大・医)、Keio metabolome-MS spectral database（慶應大・先端生命研)、EI-MSデータベース（著作権が日本質量分析学会から東大・院・工学研究科に譲渡）はいずれもMassBankから公開された。これらを含めて2010年12月現在、19研究グループが11,614化合物について測定した30,312件のデータを８つのサーバ（うち中国1、ドイツ1）から公開している（図１)。

MassBank は日本質量分析学会の公式データベースとして認められ（2008 年）、学会活動を通してデータベース化の国内外への普及活動をしている。「データ公開マニュアル」日本語版（103 ページ）と英語版（94 ページ）を作成するとともに、学会と共催で東京と大阪で年 1 回づつ講習会を開催（いずれも定員を超える盛況であった）して普及に努めた。今後もデータの増加が期待される。

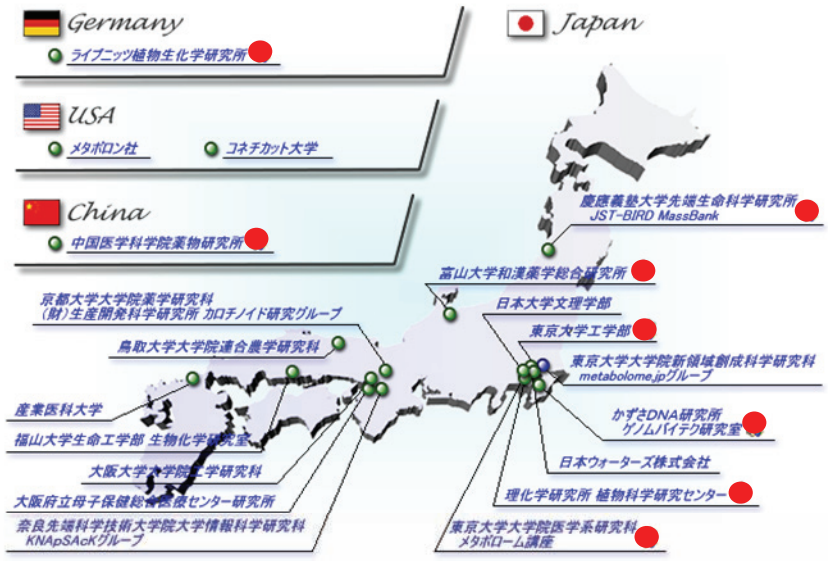

図 1．　MassBank データ公開の現状（●はサーバ設置を示す）。

2.4　装置や測定条件に依存しない参照 ESI-MS^2 データの開発と検索ツールの開発［1］

質量分析では（EI-MS を除いて）標準測定法が定められていないので、メタボローム研究者はそれぞれ任意に測定条件を決めて測定をしている。これらの MS データは、同じ装置、測定条件を採用している研究者にしか参照データとしては役立たない。最も利用者が多い LC-MS^2 で検出した代謝物質を同定するための、装置や測定条件に依存しない参照データを開発した。MassBank で公開されているさまざまな測定条件で分析した ESI-MS^2 データを化合物ごとに集約した「重ね合せ MS データ」を人工的に作成した。MS データの類似性（類似性スコア>0.6）に基づいて化合物同定をおこなうと、重ね合せ MS データを参照して正しく同定される場合（49%）は、従来の MS データを参照する場合（28%）と比べて、格段に改善された。化合物ごとに重ね合せた ESI-MS^2 データ 1,290 件を測定条件に依存しない参照データとして提供している。

この参照データとの類似性を利用して代謝物質の同定をおこなう、指定した化合物名や部分化学構造式を含む化合物の MS データを検索する、指定したピークが観測された MS データを検索する、などの利用ツールを開発した。平均アクセス数は 295 人/日（国外からのアクセスが多い）である。事実、国外の学会でも MassBank の名前を聞く機会が多い。

生物試料の一回の LC-MS 分析では、数千の代謝物質が検出される。このような大量のクエリ MS データを一挙に MassBank で同定するために、バッチ検索 SOAP-API を開発した。Mass++との連携によって、測定した MS binary data のデータセット名を Mass++上で指定するだけで同定結果を得ることができる。同定された代謝物質を KEGG SOAP-API を利用して KEGG PATHWAY 上に図示するメソッドも提供している。

2.5 ピークと部分化学構造式の経験的関係の獲得

MS で検出した代謝物質の同定率は公開されている MS データの化合物数によって制約されている。既知の二次代謝物質の数に比べると、現在 MassBank で公開されている MS データの化合物数はとても少ない。同定できなかった代謝物質（未同定代謝物質）をその MS データから推定するツールを開発することも本開発研究の重要な課題である。この推定のためには、「ピークと部分化学構造式の経験的関係」が不可欠である。この関係を次のように解析、収集した。

（1）MS データの化学的注釈づけ：ピークは *m/z* 値として表現される。しかしこの値は単なる数値であって、分析した化合物の化学構造情報を何も含んでいない。そこで、ピークは電荷を持った分子であるという原点に立ち返って、QqTOF-MS で測定した「重ね合せ ESI-MS^2 データ」（質量確度< 50ppm）について、各ピークの *m/z* 値から分子式を推定する化学的注釈付けをおこなった。*m/z* 300 より小さなピークについては唯一つの分子式を推定することができた（ユニークな分子式の数：陽イオン 2,305 分子、陰イオン 1,555 分子）。

次に解裂した化学結合を推定した。ピークの分子式から、分析した化合物の化学構造式から解裂した化学結合を推定した。この推定では、H+化あるいは脱 H+化した偶数電子を持つ分子イオンが単純解裂したと仮定した。分子式と化学結合を推定できた場合には、信頼性のある化学的注釈ができたものと判定した。*m/z* 300 より大きなピークについても、化学結合を推定することによって分子式の絞り込みをおこなうことができた場合があった。

（2）Fragmentation Library の構築：化学的注釈付け「重ね合せ ESI-MS^2 データ」1,290 件は、新たに開発した Wiki 上に構築した Fragmentation Library として収集した。通常の Wiki では各ページの内容は独立であるが、開発した Wiki では同一項目の内容は異なるページ間で関係づけられているので、例えば同じ分子式で表されるピークが観察されている化合物の検出とリストアップ、や統計情報を見ることができる。

（3）ピークと部分化学構造式の関係の獲得：化学的注釈で推定された、解裂した化学結合から直ちに「ピークと部分化学構造式の関係」が得られるとともに「中性脱離分子（= ピークの差）と部分化学構造式の関係」も得られた。後者は COOH 基などの官能基の存在を推定するために有用な関係である。

2.6 代謝物質データベースの開発

生物種-代謝物関係データベース KNApSAcK は科学文献をもとに情報を抽出し、2010 年 12 月現在、20,018 種の生物種における 50,048 種の代謝物について 101,500 対の生物種-代謝物の関係を整理し公開した[2]。またフラボノイドなど植物二次代謝物を階層的に分類したデータベース Metabolomics.JP も構築した[3]。

これらの代謝物質データベースはピークと部分化学構造式の関係とともに、将来開発される未同定代謝物質の化学構造式推定ツールの中核をなすものである。MS で検出された分子イオンに該当する二次代謝物質候補を KNApSAcK や Metabolomics.JP からリストアップし、ピークと部分化学構造式の関係に基づいて推定される部分化学構造式を有している候補を絞り込むことによって代謝物質を提示することができる。

3. まとめ

分散型データベースに基盤をおく public repository である MassBank は、データを公開する研究者がコストを分担しなければならないにもかかわらず、公開する研究グループが増加している。また、データベース型の新しい Wiki を開発した。この Wiki は MassBank における MS データの化学的注釈の作成、編集、表示を担う機能も有している。この２つの技術を融合すればデータベースを低コストに維持することが可能であることを実証することができた。また、装置や測定条件に依存しない参照データを開発したことによって、実用的な大規模 MS データベースの構築が可能であることを実証した。さらに MassBank で多様な化合物の MS データが公開されたことによって、ピークと部分化学構造式との関係の解析が可能になった。KNApSAcK や Metabolomics.JP の二次代謝物質データベースと組み合わせることによって、化合物推定ツールの開発にむけた具体的な道筋を示すことができた。

4. 研究開発実施体制

代表研究者　西岡 孝明（慶應義塾大学大学院政策・メディア研究科）

研究開発題目

(1) 生物種と代謝物質の関係データベースの構築
　グループリーダー　金谷 重彦（奈良先端科学技術大学院大学情報科学研究科）

(2) 代謝知識と代謝物質の関係データの収集と知識獲得
　グループリーダー　有田 正規（東京大学大学院理学系研究科）

(3) 統合システム開発およびマススペクトルデータベースの構築
　グループリーダー　西岡 孝明（慶應義塾大学大学院政策・メディア研究科）

5. 参考文献

[1] Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., Oda, Y., Kakazu, Y., Kusano, M., Tohge, T., Matsuda, F., Sawada, Y., Yokota-Hirai, M., Nakanishi, H., Ikeda, K., Akimoto, N., Maoka, T., Takahashi, H., Ara, T., Sakurai, N., Suzuki, H., Shibata, D., Neumann, S., Iida, T., Tanaka, K., Funatsu, K., Matsuura, F., Soga, T., Taguchi, R., Saito, K. and Nishioka, T. "MassBank: A public repository for sharing mass spectral data for life sciences", *J. Mass Spectrometry*, **45**(7), 703-714 (2010).

[2] Oishi, T., Tanaka, K., Hashimoto, T., Shinbo, Y., Jumtee, K., Bamba, T., Fukusaki, E., Suzuki, H., Shibata, D., Takahashi, H., Asahi, H., Kurokawa, K.,, Nakamura, Y., Hirai, A., Nakamura, K., Altaf-Ul-Amin, M., Kanaya, S., "An approach to peak detection in GC-MS chromatograms and application of KNApSAcK database in prediction of candidate metabolites," *Plant Biotechnol.*, **26**, 167-174, (2009).

[3] Arita M, Suwa K "Search extension transforms Wiki into a relational system: a case for flavonoid metabolite database" *BMC BioData Mining*, **1**:7 (2008).