科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１６　日現在

機関番号：　３２６１２
研究種目：　若手研究(B)
研究期間：　2012～2013
課題番号：　２４７１０２１５

研究課題名（和文）メタボロームの統合化データベースの開発

研究課題名（英文）Development of integrated database for metabolome data

研究代表者
杉本　昌弘（Masahiro, Sugimoto）

慶應義塾大学・政策・メディア研究科・准教授

研究者番号：３０４５８９６３

交付決定額（研究期間全体）：（直接経費）　　3,600,000円、（間接経費）　　1,080,000円

研究成果の概要（和文）：本研究ではデータの標準化が進んでいないメタボローム分野におけるデータの共有化と再利用を効率的に行うために、様々なユーザを想定した公開データベースと関連ツールの開発を行った。格納するデータは、定量値を利用するユーザ向けに代謝物プロファイル（代謝物の濃度パターン）、分子の同定を行う分析者向けに測定生データと、これらに関連した付帯情報（外部データへのリンクや、臨床サンプルなどを取り扱うための情報）を対象とする。また、代謝データを可視化するためのPathway可視化ツールの開発や、臨床情報の組み合わせによるデータ解析ツールも開発した。

研究成果の概要（英文）：This study aimed to develop a database of metabolome data because the standardization of metabolomics data is not well established compared to the other omics data. We have developed MMMDB that contains metabolomics profiles from single mice. We also developed data analysis methods for clinical information usually attached as metabolomic data sample information. The pathway visualization tools that enable users to edit the pathway easiliy.

研究分野：　ゲノム科学

科研費の分科・細目：　ゲノム科学・ゲノム生物学

キーワード：　メタボローム　データベース　パスウェイ　可視化　データ解析

１．研究開始当初の背景

ポストゲノムの時代、オミックス測定技術の進展により大規模なデータの蓄積が加速的に伸びてきた。同時に様々なデータベースが開発され、各研究成果は、オープンな情報として保存・公開されており、複数の研究成果を横断した検索やデータの再利用を行うことが可能となってきた。ゲノム、トランスクプトーム、プロテオームに関しては、様々な標準化が行われ、オントロジーやメタデータの整備が進んでおり、National Center for Biotechnology Information (NCBI)のサイトを始めとして、様々なデータリポジトリが既に存在する。

メタボロームに関しても、データベース開発は活発であり、様々なデータベースがあるが、大きく分けて 2 つのタイプがある。

① 文献情報を収集し、代謝 Pathway 上で代謝物や酵素の関係を統合したもの
② 分子単位で質量分析装置(MS)や核磁気共鳴(NMR)の測定データを格納したもの

まず①のデータベースの例としては、様々な生物種の情報（代謝分子だけでなく、Pathway や酵素情報も含む）を格納した KEGG (Kanehisa et al, Nucleic Acids Re, 38, D355-360, 2010)、Pathway 上で表示する分子の情報量の制御（例えば炭素の結合まで表示する場合や、それぞれの分子は１つの要素として表示する）が可能な MetaCyc (Caspi et al, Nucleic Acids Re, 38, D473-479, 2010)、ヒトのデータのみに特化した Reactome (Matthews, Nucleic Acids Res,D619-D622)、植物のデータのみに特化して特に 2 次代謝物の情報を充実させた MetaCrop (Grafahrend-Belau et al, Nucleic Acids Re, 38, D954-958, 2008)などがある。

次に②のデータベースの例としては、ガスクロマトグラフィー質量分析装置(GC/MS)から得られるマススペクトル情報を格納した GMD@CSB.DB (Kopka, J. et al. Bioinformatics, 21, 1635-1638, 2005)、高感度のタンデム質量分析装置(MS/MS)のスペクトルを格納した MELTLIN (Smith et al. Ther Drug Monit, 27, 747-751, 2005)や、複数の測定拠点のデータを効率的に格納して、検索負荷を低減させるために分散データベース形式を採用した MassBank (Horal et al. J Mass Spectrom, 45, 703-714, 2010)がある。

HMDB (Wishart et al, Nucleic Acids Res, 37, D603-610, 2009)と SMPD (Frolkis et al, Nucleic Acids Res, 38, D480-487, 2010)は(1)と(2)の両方を兼ねており、文献情報からヒトの体液中に含まれる代謝物と Pathway 情報を格納すると共に、様々なタイプの MS の測定データも格納している。また HMDB では、各文献上に記載のある分子濃度の絶対値も登録しているが、他のオミックスのデータ同様、単独分子の絶対量ではなく、むしろ測定データすべてのプロファイルの中で相対的にどのような値になっているかの情報がなければ生化学的な考察ができない。そこで、研究代表者も試験的にプロファイルデータとマススペクトルデータを格納するデータを開発してきた。

メタボローム分野におけるデータベースの問題点は、他のオミックスに比べて個々のデータベース間がお互いに独立しており、データベース間を横断的に検索して情報を抽出するには google といったキーワードレベルの検索に頼らざるを得ない。マススペクトルや分子を、何らかの属性から検索するためには、データの標準化や Application Programable Interface (API)の統一化されたルールなどを決める必要がある。しかし、メタボロームの分野において、標準化の試みはあるものの(Jenkins et al, Nature Biotechnology, 22, 1601-1606, 2004)(Goodacre et al, Metabolomics, 3, 231-241, 2007)まだ一般に浸透していない。これには様々な理由が考えられる。

(1)まだ測定技術が成熟しきっていないうちに、新しい測定機器や分析方法が頻繁に開発され、データ構造やフォーマットが更新される。

(2)他のオミックスデータに比べて表現型に近いために、様々な変動要因を同時に格納するメタデータ（付加的情報）の共通化・標準化することが極めて難しい。

(3)分子学的な制御因子（酵素など）だけではなく、メタ情報（測定対象の条件の詳細情報、例えば疾患の詳細や採取方法、測定条件なども含む）を十分に付加しなければデータとしての価値が薄く、これらの情報の構造化が極めて難しい。

このため、標準化とは別の視点で、ユーザが情報をアップロード・更新しやすい Wiki(ユーザが Web 上で簡易に文章の更新が行えるシステム)でのデータベース化も行われている(Arita, Briefings in Bioinformatics, 10, 295-296, 2009)(Arita, BioData Min, 17, 7, 2008)。しかし、純粋な Wiki では、自由度があまりに高く、データの冗長性の問題やキーワード検索以外の効率的なデータの取り出しが難しい。また、ゼロからデータを構築し、更新・維持してゆく労力も大きい。したがって、ユーザがデータの更新を行いやすく、現実的に継続的な運用と拡張を見据えた汎用性の広いデータベース（データだけではなく、システムやプログラムも含む）の開発が

必須である。

２．研究の目的

本研究ではデータの標準化が進んでいないメタボローム分野におけるデータの共有化と再利用を効率的に行うために、様々なユーザを想定した公開データベースと関連ツールの開発を行う。

格納するデータは、定量値を利用するユーザ向けに代謝物プロファイル（代謝物の濃度パターン）、分子の同定を行う分析者向けに測定生データと、これらに関連した付帯情報（外部データへのリンクや、臨床サンプルなどを取り扱うための情報）を対象とする。また、代謝データを可視化するための Pathway 可視化ツールの開発や、単にオミックスデータを解析するだけではなく、臨床情報との組み合わせによるデータ解析の機能も開発する。

３．研究の方法

データベースそのものは様々なデータを格納するように設計するが、実例として取り扱う測定データとしては、メタボロームの様々な測定系でも広く網羅的なデータが取得できる、キャピラリー電気泳動・質量飛行時間型質量分析装置（CE-TOFMS）のデータを対象とした。本装置は、イオン性代謝物を陽イオンと陰イオンの２度の測定で網羅的に測定できることができ、解糖系、TCA 回路、ペントースリン酸回路、アミノ酸の合成などにかかわる中間代謝物をプロファイリングできる。また、データは正常マウスの組織の抽出液と、血液データを用いた。

４．研究成果

（１）プロファイルデータの格納

商用で入手可能な標準物質のデータを用いて、補正した移動時間と *m/z* 値の比較を実施し、物質の同定を行った。428 の物質を検出でき、219 の物質に関して物質同定を行うことができた。肝臓、腎臓、膵臓、心臓など様々な臓器と（重量比にて補正）と血液のデータ（こちらは濃度をそのまま）のデータを格納した。これらは臓器ごとのプロファイルとしてデータを取り出したり、特定の条件に該当する物質だけにアクセスできる検索窓口の両方を設けた（図１）。また、同定した物質だけでなく未同定の物質もすべてデータベースに格納し、全てのデータをそのまま（大きなテキストファイルの塊として）ダウンロードすることもできるようにした。新規代謝物質が入手できれば、順次未知物質は、同定物質の方にプロパティを付与して移動させてゆくことができる。

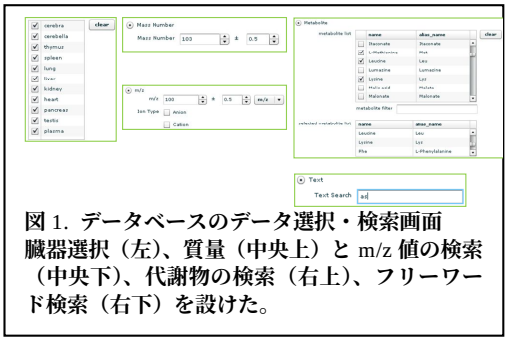

図 1. データベースのデータ選択・検索画面
臓器選択（左）、質量（中央上）と m/z 値の検索（中央下）、代謝物の検索（右上）、フリーワード検索（右下）を設けた。

（２）測定生データの格納・検索・閲覧・比較機能

測定データとしては物質定量ができたものは定量値を、同定できないものは、ピークサイズを内部標準のピークサイズで割って質量分析装置の感度補正を行ったデータを格納した。また、測定条件や測定装置が変わった場合の影響などを見られるように、測定対象に対してデータを帰属させず、測定対象-測定条件-測定データと帰属させた。これにより、バッチ（測定ごとに発生するデータのゆらぎ）の情報を調べることができる。データの閲覧画面では、Electropherogram や Mass Spectrum だけではなく、各臓器ごとの濃度の違いが分かる画面も用意した（図２と図３）。

図 2.DB 登録情報(代謝物情報、概要)
代謝物ごとのデータ表示画面（左上）、測定ごとのデータ表示画面（左中）、測定データの表示画面（左下）、分子構造表示画面（右上）、臓器ごとの濃度の違いの表示画面（右下）

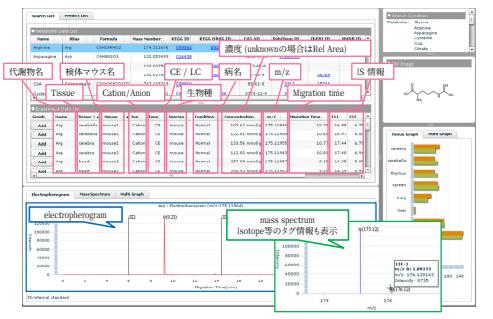

図 3.DB 登録情報(測定データ情報、詳細)
測定情報（左中央）では、代謝物名に関する測定条件の詳細（測定装置やそのパラメータ）とサンプル情報の詳細（マウスの飼育条件など）を示す。測定データ（左下）は、Electropherogram だけでなく、Mass Spectrum も表示し、対象物質だけでなく、同時に測定した内部標準物質も示す。

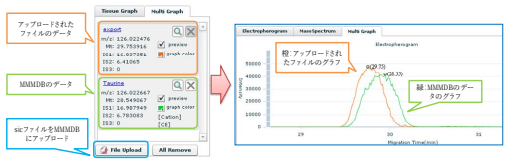

図 4.Electropherogram の重ね合わせ
測定したデータの一部（Electropherogram と Mass Spectrum）と、測定時の条件（内部標準物質の検出位置など）をアップロードすると、自動的に移動時間を補正して、登録されているデータと比較し、重ね合わせた結果を可視化する

また、格納されたデータと自分のデータを比較して測定データの同定ができるように、個々のデータをアップロードして比較を行う機能を実装した。この時、キャピラリー泳動の特徴で、移動時間が非線形にゆらぐ。この問題の解決のために、内部標準物質を２つ以上いれ、これらを(Sugimoto et al. Metabolomics, 2010)の方法用いて移動時間の補正を行い、補正後の Electropherogram を重ね合わせて表示を行う（図３、４）。更に、重ね合わせるだけでなく、時間補正後の時間や m/z の誤差から、ピークの類似度に関するスコアをだし、物質推定を定量的に行う指標も開発した（図５）。

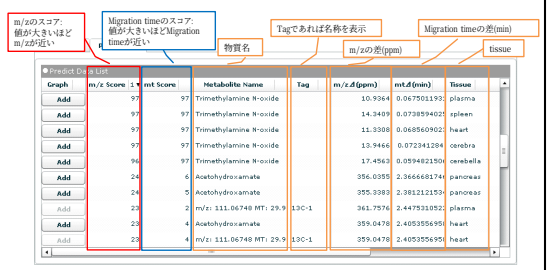

図 5.物質推定のスコアを表示する機能
アップロードしたデータと登録したデータを比較して、スコアリングする。左から m/z と時間の誤差をスコアに変換した値、候補となる物質名、他の物質と親子関係にあるピークかどうか（例えば、ある代謝物の同位体イオンとして検出されたデータかどうか）、m/z と時間の誤差の実測値、どの組織で検出されたことがあるデータかを表示する

（３）代謝 Pathway のデータ表示・編集機能

メタボロームのデータは、他のオミックスデータと異なり、単なる１変数の違いを比較するよりも、Pathway でその情報を可視化して、データのバランスがどのように変化していることを見ることが極めて重要である。また、Pathway の情報はそれぞれの生物種で異なる上に、日々新しい研究成果で修正が加わるため、自分で編集できることが望ましい。

しかし一方、このために特殊なソフトで編集機能を付加すると実質的に使えるツールとはならない。したがって、本研究では、パワーポイントなどの一般的な図の編集機能のあるソフトで図を作り、図の中の各オブジェクトに付けたプロパティと自己の定量データの間を関連つけるツールを開発した。これにより、一般的なソフトで Pathway を簡単に編集することができる。これにエクセルなどでデータ（代謝物を示す ID）を結びつけることができ、データを図の中にマッピングすることができる。本研究で作成した Pathway 可視化ツールでは、Pathway の全体像を表示するときは、色など抽象的な情報としてデータを表示し、クリッピングマップのように着目したい物質をクリックしたときに、詳細なデータや統計結果を表示させる機能を開発した。
（図６、図７）

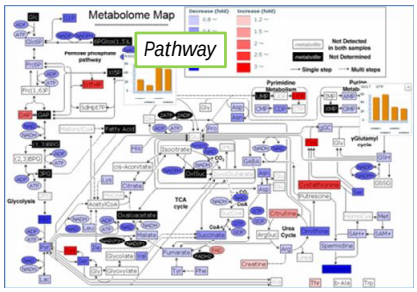

図 6.Pathway 表示ツールの全体図
プロファイルデータを Pathway 上で可視化する。ここでは２群のデータを対象として、２群間でのデータの変動の比率で各代謝物に色を付けて表示している。この図は、パワーポイントなど汎用的なソフトウェアで描け、各要素にプロパティを付与することでエクセルなど汎用的なデータ管理ソフトのデータと関連することができるようにした

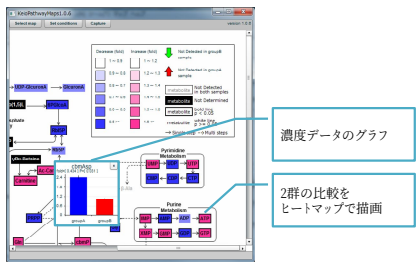

図 7.Pathway の詳細データ可視化機能
各代謝物をクリックして、物質ごとに詳細な定量値や統計値を示すグラフを表示する

（４）付随情報（臨床情報）の解析機能

本研究で対象とするデータは、メタボロームのデータに限らず、トランスクリプトームの定量値データと同様に、定量値の変数を多数持ったオミックスデータである。このようなデータベースでオミックスのデータそのものの解析機能は盛んではあるが、実際にそれだけでは不十分で、付随する情報と合わせた解析機能を開発する必要がある。そこで、本研究では、臨床検体（例えば、血液や尿）を測定した時に、付随する臨床情報の解析機能を開発した。

特に臨床サンプルに付随する臨床データで頻繁に発生する独自の問題点として、類似した変数（同じ内容の特徴であっても、精度の異なる別々の検査方法や測定方法で取得している）、データ中の症例数に偏りがある（大部分の方が健常者で、一部の方だけが疾患を持っている）、欠損値が多い（問診のような情報の収集方法でランダムに情報が欠損している場合と、ある変数の値がわかれば、より侵襲性の高い試験をしなくてもよい）という問題がある。これらの問題の影響を低減させて、疾患状態を精度よく予測する必要がある。

今回乳癌のリンパ節転移の術前化学療法後に腫瘍サイズが小さくなり pCR(臨床的な効果がある)を予測するデータ(Takada et al., Breast Cancer Res Treat, 2012)を用いた。方法としては、変数を一定の重みつけで線形結合する多重ロジスティック回帰(MLR)と、決定木形式で各変数に重みをつけて足し合わせてスコアを決定する(ADtree)を用いて、予測性の比較を行った。

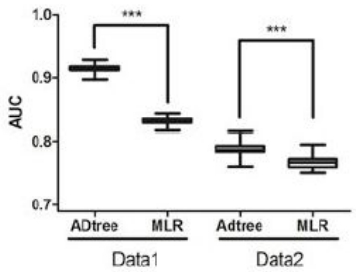

図 8 ADTree と MLR の比較
Y 軸は AUC 値で boxplot は、200 回試験を施行したときの AUC 値のばらつき（最大値、最小値、中央値と 4 分値）を示す。左側に Data1、右側に Data2 の結果を示す。

また、今回モデルを学習するデータ(Data1)と学習済のデータを評価する評価データ(Data2)を用いた試験を行った。これらのデータには欠損値が多く含まれている。AUC 値（疾患症例を正常症例から正確に見分けられるか否かの値、ROC 曲線以下の面積で 1 が最もよく、0 が最も悪いことを示す）図 8 に示す通り、Data1 でも Data2 でも MLR に比べて ADTree の値が高い結果が得られた。

MLR は、モデル作成の変数を選択する方法やパラメータがデータによって最適化させる必要がある。ここでは、最も一般的に使われている変数増減法で P=0.05 を変数の追加と除去の指標とした。

ADTree は一般的な指標はなく、パラメータとしてはツリーの中の boosting 数（＝変数の数）がある。また、今回 ADTree モデルを複数学習させてそれらの予測値を結合する Ensemble 学習も行い、これらの効果をみた。図 9 A)に示す通り、Ensemble 学習を行うと、ツリーの数が増えるほど AUC 値が高くなるが、ツリーが 10 あたりで AUC の向上は頭打ちとなる。一方図 9B)に示すとおり、boosting 数は 4 の段階で AUC 値が最も高くなり、その後は AUC 値が向上しない。したがって、ADTree と集団学習(Ensemble)を組み合わせた場合、１つのツリーの変数の数は少なく、10 程度のツリーを組み合わせると、欠損値が多く含まれるデータでも精度よく予測が行われることが分かった。

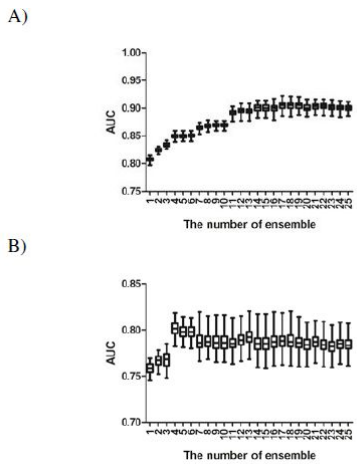

図 9 Ensemble と Boosting 数の影響
A)は Ensemble 値（=ADtree モデルの数）、B)は boosting 数（1 つの ADtree の中の要素数）が X 軸で、Y 軸は 200 回予測試験を施行したとき AUC 値のばらつきを示す。

本研究の目的は、他のオミックス分野よりも標準化が進んでいないメタボローム（代謝物）の測定データを対象として、様々な付帯情報を統合したデータベースを開発することであった。そこで、プロファイルデータを中心に格納するデータベースの開発を行った。

自動的に代謝物の位置を抽出し、別途測定した定量データを可視化する機能を実装した。また、多数のデータを比較・解析する場合は、単にサンプルの代謝濃度が含まれているだけではなく、サンプルに付随する情報も効率的に活用した解析が必要となる。そこで、特に臨床サンプルに付随する臨床データで頻繁に起きる欠損値があったとしても、疾患状態を予測するアルゴリズムも開発した。現段階では個々の要素は完成しており、データベースとして既に公開したものと、単独のツールとして開発しており、まだ Web 上で統一的に使えるプラットフォームにはなっていないため、今後は、これらの機能も順次 Web に統合して使えるよう加工し、公開してゆく予定である。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　1 件）

① Masahiro Sugimoto, Masahiro Takada, Masakazu Toi, Comparison of robustness against missing values of alternative decision tree and multiple logistic regression for predicting clinical data in primary breast cancer, 査読有, Conf Proc IEEE Eng Med Biol Soc, 2013, 3054-3057
DOI:10.1109/EMBC.2013.6610185

〔学会発表〕（計　3 件）

① Masahiro Sugimoto, A challenge filling in the gap between experimental and theoretical biology, Complex Biodynamics & Networks, 12/Nov/2013, Yamagata, Japan

② Masahiro Sugimoto, Bioinformatics provides novel insights into metabolomics, International Conference on Applied Informatics for Health and Life Science, 11/Sep/2013, Istanbul, Turkey

③ Masahiro Sugimoto, Masahiro Takada, Masakazu Toi, Comparison of robustness against missing values of alternative decision tree and multiple logistic regression for predicting clinical data in primary breast cancer, 35th Annual International IEEE EMBS Conference, 5/Jul/2013, Osaka, Japan

〔図書〕（計　1 件）

① 杉本昌弘、システムバイオロジー：メタボロームの展開・生命のビッグデータ利用の最前線 Frontier of Utilization of Big Data in Sciences、上田充美監修、シーエムシー出版社、2014 年、p32-40

〔産業財産権〕
○出願状況（計　０件）

○取得状況（計　０件）

〔その他〕
ホームページ等
http://mmmdb.iab.keio.ac.jp/

６．研究組織
(1)研究代表者
杉本　昌弘（SUGIMOTO, Masahiro）
慶應義塾大学・政策・メディア研究科・特任准教授
研究者番号：30458963

(2)研究分担者
なし

(3)連携研究者
なし