既存データからも新たな知見を得られる、質量分析インフォマティクスの確立をめざす

教授　有田（ありた） 正規（まさのり）　情報研究系　生命ネットワーク研究室

東京大学理学部卒業、東大院理学系情報科学専攻満期退学。理学（博士）。電子技術総合研究所（現、産業技術総合研究所）、東大院新領域（情報生命科学専攻）、理学系（生物化学専攻）を経て、2013年国立遺伝学研究所（遺伝研）教授。2018年4月より遺伝研 DDBJ センター長を兼任。

「質量分析インフォマティクス」という概念を作り、質量分析の装置や標準物質、照合用のデータベースを限定しないメタボローム解析の技術開発を進める、有田正規教授。そのために利用する生物は、細菌、かび、植物、藻類からヒトまで、実に多岐に渡る。根底にあるのは、「物質が自然界でどのようにめぐり、代謝されるのかを知りたい」という知的好奇心。データ共有の必要性や科学研究のあり方に対する意見も、積極的に発信している。

計算機科学を学び、いち早くメタボローム解析領域へ

数学と化学が得意だったので、いずれかを専攻しようと思い、東京大学理学部に進みました。ところが、入ってみると授業のレベルも学生のレベルも驚くほど高く、「少し方向を変えよう」と考えて計算機科学の道を選びました。まず手がけたのは、ショウジョウバエの胚発生のシミュレーションです。当時（1995年頃）は、堀田凱樹先生が東大理学部の物理学教室でショウジョウバエの遺伝子ネットワーク解析に取り組んでおられ、私もいろいろ話をお伺いしました。

博士後期課程に進むときに、東大医科学研究所に新設されたヒトゲノム解析センターに所属を変え、さらに米国のワシントン州立大学に留学して「代謝の経路」を意識した研究を始めることになりました。ちょうど、ゲノム解読の波が来て、配列情報がドッと出てきたところでしたが、私の興味はゲノムにはなく、代謝の情報を誰もが利用できるデータに変換することにありました。留学先では代謝マップを電子化するための技術開発を進め、「代謝の原子レベル再構築」という論文タイトルで博士号を取得しました。取得後は、いったん電子技術総合研究所（現　産業技術総合研究所）に就職したのですが、2003 年に東大に戻り、2013年に遺伝研に赴任しました。

必要なのは、質量分析インフォマティクスという概念

私が実現したいのは、各国で生産されるメタボローム解析のデータが揃っており、誰でも自由にダウンロードして研究に活用できる環境です。DNA の塩基配列情報には「4 種の塩基で記載する」という原則があり、データ形式もほぼ標準化されています。

よって、ゲノム解析では、どのようなシーケンサーを使ってもほぼ同じ配列情報が得られます。もちろん、厳密には DNA のメチル化といった化学修飾もあるので、それほど単純とはいえませんが。一方のメタボローム解析は、まったく標準化されていないと言わざるを得ません。質量分析の手法や機種ごとにデータのフォーマットまで変わり、得られたデータを先行研究のデータや既存のデータベースの内容と比較することすら難しいのです。

私は、このような問題を情報科学の観点から解決し、質量分析装置から得られるデータ（スペクトル）から汎用性のある「知識」を導きたいと考えています。ここでいう知識とは、代謝物の意味づけ、代謝反応の流れ、それに対する応答などのことです。「アミノ酸は物質Aの制御を受けて変動します」、「植物が成長するときに重要な代謝物はBです」といった知見を、研究者がメタボロームのデータベースから得られるようにしたいのです。

そのために「質量分析インフォマティクス」という概念を広めようとしています。質量分析の手法、標準物質、データの種類（生データ、二次データ、論文等）にこだわるのではなく、多くの代謝情報を検索、比較、解釈して情報を洗練できるサイクルといえば理解いただけるでしょうか。生物や化学の研究者は新規物質をみつけるのが好きだと思いますが、質量分析インフォマティクスを使えば、データベースの中からでも「これまでに知られていなかった新規の代謝物」をみつけることが可能になります。

質量分析インフォマティクスを実現するソフト開発

質量分析インフォマティクスを実現するために、さまざまな技術を開発しています。たとえば、最近の理化学研究所や千葉大学との共同プロジェクトでは、シロイヌナズナやイネなど 12 の植物を測定したシグナル（MS/MS スペクトル）を比較しながら、代謝物の炭素数を決定、組成式を算出、代謝物の大まかな分類、化合物の骨格を決定、ネットワーク可視化、といった流れを自動化したパイプラインを開発しました。

私たちはまず、安定同位体 C-13 のみで育てた植物体 A と、通常条件下で育てた植物体 B から、それぞれ測定試料を調整し、質量分析計で計測しました。こうすると、同じ代謝物（たとえばグルコース）でも、A に含まれるグルコースの方が質量が大きくなります。このような質量の違いを捉えることで、代謝物の炭素数を決定できます。炭素数さえ決まれば、組成式の推定は非常に楽になります。この工夫により、生体を構成する有機元素（C,H,N,O,P,S）で構成される組成式なら、ほぼ間違いなく決定できるようになりました。

さらに、既存のデータベースから化合物の構造とMS/MSスペクトルの関係性を見出し、測定された物質に対して「フラボノイド、ステロイド」といったおおまかな分類を推定できるようにしました。さらに、既知の物質に糖や脂肪酸が付加した構造かどうかを、スペクトルの関連性から見出す手法も開発しました。

こうした一連の手法を用いることで、6 科 12 種の植物から合計 3604 個の代謝物の炭素数を決定し、そのうち 1133 個について組成式も決めることができました。そこには、最終的に 69 個の新規構造が含まれていることも突き止めました。また、今回の測定では 500 以上の代謝物が特定の科（アブラナ科、イネ科など）のみで作られていること、シロイヌナズナから開花や乾燥耐性に関わる代謝物群も検出できることなども明らかになりました。

スペクトルの重複を分離するデコンボリューション

液体クロマトグラフィー - タンデム質量分析（LC-MS/MS）で得られるシグナル（マススペクトル）を分離し、個別の物質候補に帰属させる技術（デコンボリューション）の開発も進めています。質量分析計から得られるシグナルには、複数の物質情報が重複した状態のものが多くあります。大半の研究者は、重複した状態のままデータベースで検索して物質名を推定していますが、これでは正確な解析はできません。私たちは、クロマトグラフィーの溶出時間が物質ごとに異なることを利用して波形情報からシグナルを分離するアルゴリズムを開発し、2015 年に MS-DIAL というソフトウェアを完成させました。これはカリフォルニア大学デイビス校のオリバー・フィーン教授との共同研究による成果です。

MS-DIAL は主要な質量分析装置メーカーの生データを直接読み込み、解析結果をグラフィカルに描出できます。2015 年の解析では、ミドリムシ（ユーグレナ）をはじめとする 9 種の微細藻類を対象に分析を行い、1023種もの脂質を検出しました。その内訳から藻類の系統関係を再現できただけでなく、2 つの種（クロレラ種、ナンノクロロプシス種）のいずれに属するかが不明だった藻類が、クロレラ種に属するとの示唆を得ることもできました。

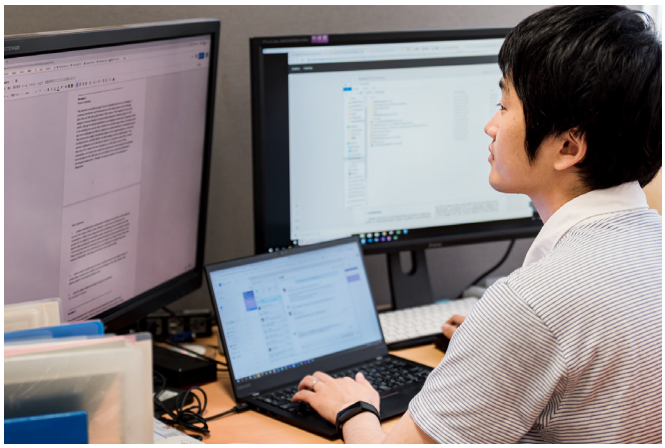
フィーン教授とはその後も共同研究を続け、メタボローム情報のリポジトリを使って、がん組織などに含まれる未知代謝物の構造を同定する技術も開発しました。データベースにない構造を見つけるのは至難の技で、多くの研究者にとっては「1 つでも 2 つでもみつけられれば嬉しい」という状況にあります。私たちのソ

フトウェアやメタボローム情報のリポジトリを使えば、再利用データの中からでも「これまで知られていなかった物質」を発見できるのです。

多くのプロジェクトを同時並行で進める日々

植物や藻類を対象にした例をあげましたが、私は生物の種類には全くこだわっていません。また、解析対象も、質量分析で得られるスペクトルだけでなく、物質や構造式、ゲノム、遺伝子、アミノ酸など、生物に関するあらゆるデータに対応したいと考えています。このような考え方は、私の最大の興味が「物質が自然界でどのようにめぐり、代謝されるのか」という点にあることによります。物質を作る生物、利用する生物、分解する生物が、何をどう代謝し、何に作り変えているかを知りたいのです。メタボロミクスもゲノミクスもプロテオミクスも、そのための手法に過ぎないと考えています。

現在は、ピロリ菌のゲノムリアレンジメント、ビフィズス菌の糖関連酵素、メタボローム解析のデコンボリューション技術、質量イメージング用画像処理、生薬のメタボロームなど、多岐にわたるプロジェクトを同時並行で進めています。私の頭のなかでは、それぞれを区別することはなく、自然に共存している感じです。もちろん私一人では不可能で、いずれもウェットな実験をする研究者に協力いただいていますし、研究室の学生には一人ひとり異なるテーマに取り組んでもらいます。学生には、自分の興味に従って積極的に参画してもらっています。



データや研究は、誰のもの？　誰のため？

私のモチベーション、方向性、研究アプローチについてお話してきましたが、私の考え方はなかなか理解されず、研究を進めるのにも多くの困難があります。その要因の一つは、データの扱いや競争的資金、大学のポジション等に絡む既得権の問題です。メタボローム解析領域に限らず、生命科学分野全体において、研究者は得られた情報を共有・公開したがりません。時間がない、細かいデータとして拾っていないなど、言い訳はさまざまですが、平たく言えば、「他の研究者が自分のデータを精査したり、再利用することを良しとしない風潮」があるのです。私はこうした状況は、競争的資金を得たい、特許や知的財産権を独占したいといった理由からくる、研究者や所属機関のわがままな姿勢だと解釈しています。研究予算など、極論を言えば、等分で配ればよいのではないかと思います。また、学術界やメディアが、研究者を発表論文誌の種類、論文の本数、論文の被引用数ばかりで評価する姿勢もおかしいと思いますね。

欧米では、資金提供機関が、発表したものかどうかに関わらず、研究者にデータを公開するよう義務付けるのが慣例です。他人が広く使える状態にしないと、次の研究資金が獲得できないしくみになっているのです。ところが日本においては、研究者が得た情報をデータベース等で共有するという文化がありませんし、行政側にもその意図は見えません。私はなんとかして研究者の意識を改革したいと考えています。とくに若手には、世の中への成果還元を見据えてほしいと願います。幸い、遺伝研には私と同じように考える研究者もいるので、少しずつ前進できたらと思います。

情報の公開が進めば、間違った情報に振り回されることが少なくなるのではないでしょうか。研究者に限らず、知りたいと思った人が誰でもデータベースや論文にアクセスし、自分の知りたい情報を得られることが望ましいと思います。そうなれば、偽の健康食品やサプリメント、怪しい遺伝子検査サービスなどを判断する際などにも、先行研究の成果が役立つことでしょう。

コンピュータを作れるほどの人材を生物学に

インフォマティクスの教育においても、日本はうまくいっているとはいえません。最大の問題は、バイオ領域のインフォマティクスを学べる専門の学部がなく、「真のバイオインフォマティシャン」が多く育たないことです。今の大学で教える「〇〇情報学部」などの名前がついた教育では、目標が曖昧だったり内容が古すぎたりして実践的ではないと感じます。早い段階からコンピュータの歴史や、プログラミング言語の思想といった基礎概念を習得

させ、大学において自力でコンピュータが作れるような人材まで育て、そのうえで生物学の分野に入ってもらわないと現状は変わらないと思います。

私の研究室は、バイオインフォマティクスに関連することなら何でもできます。本人の希望を重要視しますが、「これだけをしたい」というよりは、広い視野をもった学生に入ってほしいですね。私自身、高校時代には生物に全く興味がありませんでした。生物学は単なる暗記科目と誤解していましたので。大学時代に数学、物理、化学などで挫折感を味わい、計算機科学に方向を変え、最後に生き物の情報解析にたどり着きました。若い頃に自分の道を見極めるのは難しいと思いますので、まずはいろいろやってみたらいいのではないでしょうか。

聞　き　手：サイエンスライター　西村 尚子
写 真 撮 影：リサーチ・アドミニストレーター室　来栖 光彦

2019年7月