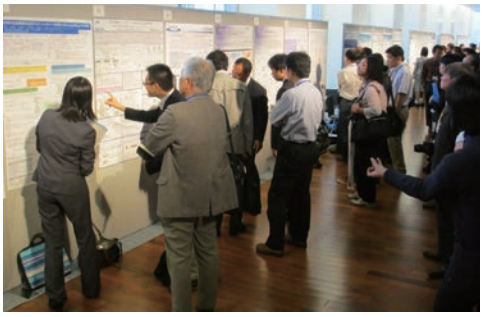
生命科学データベースを未来の産業革命に

JSTバイオサイエンスデータベースセンター（NBDC）の取り組み

生命科学のデータが爆発的に増え続けている。せっかく大量に蓄積したデータベースも、各地に散在しているのが現状だ。ビッグデータとして活用するには、これらの統合が不可欠となる。JSTのバイオサイエンスデータベースセンター（NBDC）は、日本で生命科学系のデータベース統合化を進める中核的組織で、そのための研究開発費の支援もしている。NBDCは2011年度から3年間の第1段階を終え、14年度からは第2段階に入った。将来的には、医療や住宅、農業、エネルギーなどの産業にも大きな革新を起こすと期待されている。NBDCの現在と将来像、さらに微生物統合データベースの開発について、高木利久NBDCセンター長と東京工業大学の黒川顕教授にそれぞれ話を聞いた。



ライフサイエンスデータベース統合推進事業によるサービス提供や研究活動についての発表や、統合にまつわる問題について議論を深めるため、10月5日を「トーゴーの日」としてシンポジウムを開催している。

●分野を超えたデータベースの統合を目指す

蓄積データを、使えるデータに整える

「データベースは生命科学にとって、他の分野以上に重要な意味を持つようになっています」。NBDCセンター長の高木さんは特殊性をこう語る。「生命科学はある意味で“記述の学問”なのです。何らかの遺伝子変異が病気の原因になることがありますが、病気の発症には生活習慣や環境などのさまざまな要因が関わってきます。これは簡単な法則で表すことができません」。遺伝子の変異と病気との関係をさまざまな遺伝子や病気に関して記述する。そのようなことが基本になる分野だけに、生命科学ではデータベースの重要性が高くなるのだと強調する。

各地の大学や研究所に散在するデータベースの統合も重要になっている。他の学問分野、例えば素粒子物理学の実験では、巨大加速器は建設費や運用費が巨額なため、多くの物理学者が同じ加速器で研究を進めており、データは自然に集中してたまる。

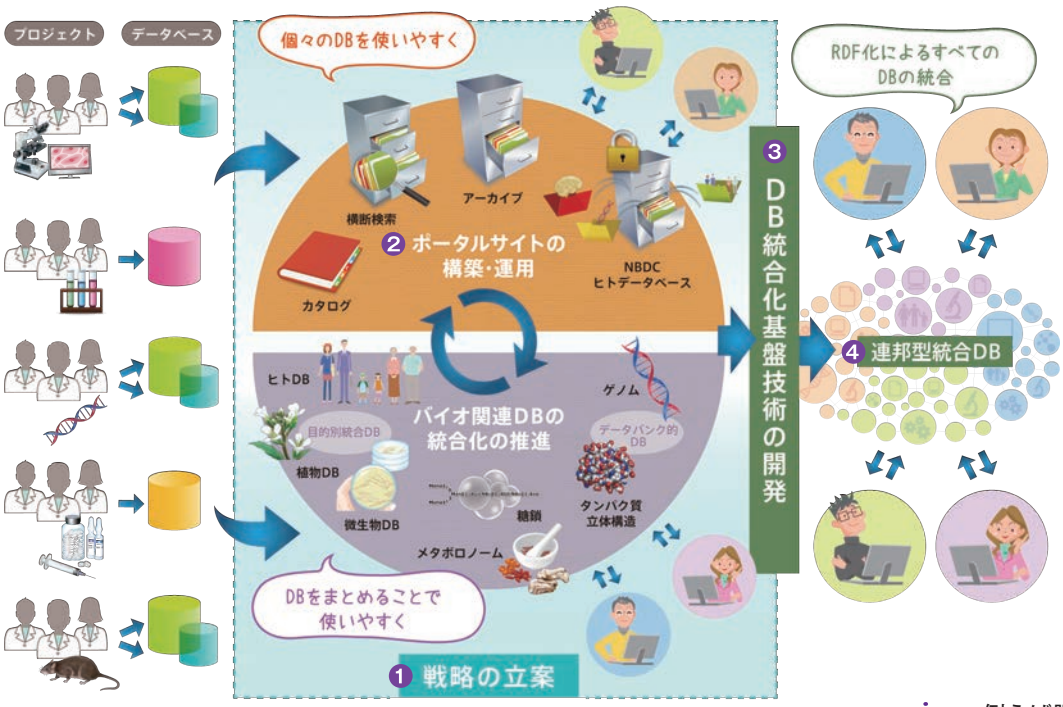
ところが生命科学の分野では、最先端のシークエンサー（DNA解析装置）でも比較的入手しやすく、1つの大学に何台もの装置があることも珍しくない。またテーマが異なれば観点も異なり、装置が異なると同じ試料でも得られるデータが微妙に違ってくることもある。

「それらのデータをうまくまとめないと、ビッグデータとしては使えません。ビッグデータ解析は、統計解析によってデータの中にある規則性を見いだすものです。そのためには大量のデータを使いやすい形で整えることが必要なのです」と話す。

また、物理学などとは異なり、生命科学のデータの場合は文脈依存性や多義性、曖昧性が生じることがある。例えば顕微鏡で観察した細胞の形を、ある研究者は「丸い」と記述し、同じ細胞でも別の研究者は「楕円」と記述することがある。

高木 利久 たかぎ・としひさ
JSTバイオサイエンスデータベースセンター長
東京大学大学院理学系研究科・国立遺伝学研究所教授

1976年、東京大学工学部卒業 。86年、工学博士（九州大学）。九州大学情報処理教育センター助教授、東京大学医科学研究所助教授、同教授、東京大学大学院新領域創成科学研究科教授などを経て、2014年より現職。また、07年、情報・システム研究機構ライフサイエンス統合データベースセンター長（10年まで）。09年より国立遺伝学研究所生命情報研究センター教授、12年よりDDBJセンター長。



NBDCの事業推進の構成

NBDCは、生命科学分野のデータベースを統合し、データの価値を最大化することにより、日本のユーザー、さらには世界のユーザーに貢献できるデータベースセンターとなることをミッションとし、①戦略の立案、②ポータルサイトの構築・運用、③データベース統合化基盤技術の開発④バイオ関連データベース統合化の推進、を4つの柱として事業を推進している。

事業の成果により、利用者はさまざまな生命科学分野データベースの一元的な利用が可能となり、日本の生命科学に関する研究成果の広範な共有を図ることが可能となる。

「こうした表記の揺れについても、同じものなのか違うものなのかを整理していかないと使えるデータにはなりません」。

活用と保護のルール作りを急ぐ

NBDCの第1段階の成果は、4省連携の体制が整ったことがまず挙げられる。文部科学省、厚生労働省、農林水産省、経済産業省の4省のデータベースを束ねて、1つのポータルサイト（ネットの玄関口）を作った。「integbio.jp」では、分野横断的に収集された情報を一括利用することが可能になった。

もう1つは、ヒトのデータを扱うためのガイドラインを作ったことだ。個人のゲノム（全遺伝情報）データには、病気のなりやすさや親子関係など、さまざまな個人情報が含まれている。「それだけにデータの管理が重要になり、情報の取り扱い方を厳格に決めておく必要があります。まずはNBDCのガイドラインを作り、承認された人だけがアクセスできるNBDCヒトデータベースとして運用を開始しました。今後は環境情報や健康情報、より臨床に近い情報も組み込んで解析ができるような仕組みを作りたいと考えています」。

また国民の税金を使う研究については、できるだけ「すべてのデータを誰でも利用できるようなルールを作ろうと政府に働きかけをしています」。アメリカでは税金を使った研究データは明確に国の共有財産としてデータの共有化が進んでいる。日本ではそのようなルールはない。研究機関や研究者単位でバラバラにデータを抱え込んでしまう傾向があり、ビッグデータにつながらないのが現状だ。

さらに著作権などデータの権利関係の整理も進めてきた。古い映画などで、権利関係が不明瞭なために流通できないケースがある。「サイエンスではそれは許されません。しかし法律的にしっかりしていないと、そのつもりがなくても結果的にデータを抱え込んでしまいます。これらを整理するのもNBDCの活動の１つです」と、活用から保護まで幅広く捉えている。

伝統はなくとも独自の強みがある

第2段階での最大のテーマは、分野を超えたデータベースの統合だ。「これまでは植物や微生物、動物、人間というように分野ごとにデータベースを作ってきました。今後はすべてのデータベースをつなぐ方向で進めたい。そうすることで、分野を超えた解析が可能になります」。

例えば腸内細菌と病気との関連では、微生物とヒトの研究とが関係する。また植物の生育は、土壌中の微生物や土壌の栄養状態などと密接に関連しており、それらのデータをつなげることも重要になってくる。

さまざまなデータベースを統合するために、NBDCでは情報を意味づけする「セマンティック・ウェブ」の技術を使っている。「情報の意味をコンピューターでも扱えるようにする記述方式であるRDF（p.11上図）の生命科学での利用は私たちが世界をリードしています。今後もイニシアチブをとり続けたい」と高木さん。

実は、データベースに対する日本の取り組みは、欧米に比べて歴史が浅い。しかし、生命科学の分野では、過去20年間もかけて作られたゲノムのデータが、最近では1年間で得られるような状況だ。「極論すれば過去20年間のデータは捨てたり新しい形に作り直してもいい。伝統がないことを逆に強みにすることが可能で、その1つの側面が、RDFを使ったデータベース化なのです」。

「私たちが作っているデータベースが研究に直接役立つ事例がようやく少しずつ出始めています。これまではルールを定め、理解を得るなど基礎的な準備をしてきました。これからは収穫の時期に入るのでぐんと飛躍したいと考えています」と、将来への期待感を語った。

●環境情報も含めた微生物データでメタゲノム解析を

門外漢でも使えるデータベースを

NBDCでは、統合化推進プログラムとして、生命科学系の複数の分野のデータベースの統合化のために研究開発費を支援している。東京工業大学地球生命研究所教授の黒川顕さんが中心となり、国立遺伝学研究所の中村保一さん、基礎生物学研究所の内山郁夫さんらとともに進める「MicrobeDB.jp」は、その研究課題の1つだ。

MicrobeDB.jpは、「ゲノム情報」と「メタゲノム情報」のデータベースを統合した微生物データベースである。「メタゲノム」とはさまざまな生物の遺伝情報の集合体のことで、メタゲノム解析は、環境中から得た多様な生物のゲノムをまとめて解析する手法だ。例えるなら、ゲノム解析が1本の木の解析とすると、メタゲノム解析は森全体の解析ともいえる。

「ゲノム解析では、ある微生物の遺伝子を解析し、それを基にさらに詳細な研究へと深掘りしていきます。これに対してメタゲノム解析では、微生物群集全体の遺伝子配列の情報だけでなく、微生物の群集が存在している環境情報も付随してきます」と黒川さん。ヒトの腸内でみると、年齢や性別、体重や身長、食生活、病歴などが環境情報となる。メタゲノム解析で出てくる情報は、その環境の中でどのような微生物群集が存在し、その微生物群集にはどのような遺伝子があるのかというデータの集合体である。

黒川さんらは「散在している微生物のデータをすべて集めて、ゲノムを中心にしてすべてのデータをつないでいく」ことを目標にしている。第1期に続き、14年度からの第2期でも引き続き統合化推進プログラムの一環としてプロジェクトを進めている。

実はMicrobeDB.jp作成に取りかかる前、微生物に関するデータベースはすでにたくさんできていた。だがいずれも微生物の専門家向けのものだった。「微生物の専門家以外の人にも使いやすいデータベースにすることが目標です」と黒川さん。それは微生物が地球上の至るところに存在しており、微生物研究は農業や食品、材料、環境、健康などさまざまな分野の研究と結び付けられるからだ。「微生物はそれができる数少ない分野の1つです。そのためには、ゲノム情報や環境情報など、さまざまな情報を統合したデータベースを作らないといけません」。

手作業で言葉の関係性を整理

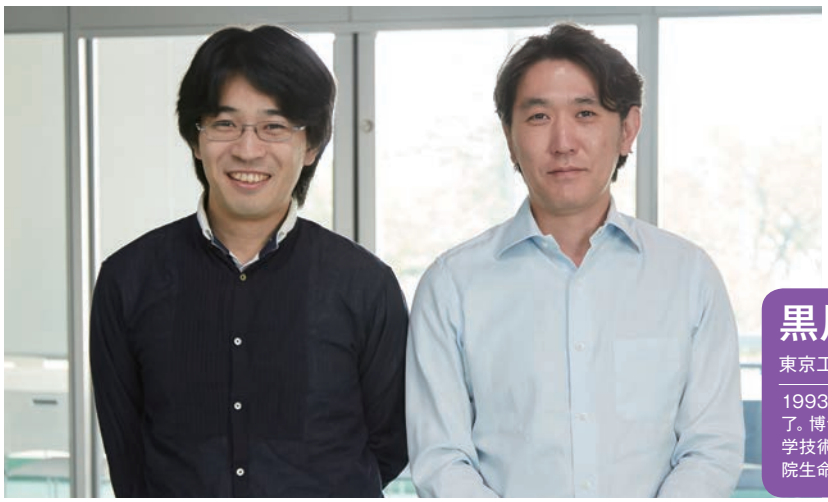
黒川さんらはNBDCと共同研究を行っている情報・システム研究機構（ROIS）のライフサイエンス統合データベースセンター（DBCLS）と密接な関係を持ちながら開発を進めてきた。

MicrobeDB.jpデモサイトのトップページ
MicrobeDB.jpは黒川さんらが統合化推進プログラムで開発した微生物統合データベース。どうやって求める情報にたどり着いたらよいのかわからない門外漢でも使いやすいよう、シンプルに検索できるようになっている。森さんは統合化推進プログラムで統合されたデータベースを対象としてデータを解析するツール等を開発し、それを用いて新たな知識発見を目指す、「統合データ解析トライアル」において、MicrobeDB.jpを利用して、メタゲノム配列データの解析ツールを開発した。

「DBCLSで作られた基盤技術は、東工大のMicrobeDB.jpの中核になっています。協力関係がなければ、第1期の3年間でここまで進めることは無理だったでしょう」と語るのは、MicrobeDB.jpの技術的な面で中心的な役割を担っている黒川研究室助教の森宙史さんだ。

「MicrobeDB.jpのバックグラウンドになっている技術は、セマンティック・ウェブの技術です。情報を意味づけするときに重要なのは、RDF形式の記述方式と『オントロジー』という単語間の関係性を整理する技術です」と続ける。簡単にいえば、RDFは文法で、オントロジーは辞書に相当する。

「ヒトの腸内の研究は、『ヒューマン・インテスティン』とか『ヒューマン・ガット』、『ヒューマン・フィーシーズ』などバラバラに呼ばれています。コンピューターにはそれらの単語がほぼ同じ意味だということを教えておく必要があります。しかしガット（内臓）とフィーシーズ（排泄物）とは辞書的にはまったく意味が違うので、言葉のつながりを考えながら関係性を定義


東京工業大学大学院生命理工学研究科助教の森宙史さん（左）と黒川さん（右）

黒川 顕 くろかわ・けん
東京工業大学地球生命研究所教授・副所長
1993年東北大学理学部卒業。98年大阪大学薬学研究科修了。博士（薬学）。大阪大学微生物病研究所助手、奈良先端科学技術大学院大学情報科学研究科助教授、東京工業大学大学院生命理工学研究科教授などを経て、2013年より現職。

します。これがオントロジーで、情報を正確に記述するためにはまずそれをあらかじめ作っておく必要があります」と黒川さん。

さらに、例えば論文に書いてあるデータにどの語彙を当てはめれば良いかを決める「オントロジーマッピング」という作業も必要になる。それらの作業はコンピューターのみでは困難なため、分担しながら手作りしたという。

「RDFの形でデータを書いておけば、データベースの統合が簡単にできるようになります。またデータが分散していても、散在したデータベースから情報を集めてきて結果を表示することができます。最初に手間がかかっても基本構造を作っておけば、比較的簡単に実現できることが、セマンティック・ウェブの素晴らしいところです」という。

地質、海洋、気象など異分野との統合も

「最初は大変でした。まずデータを集め、データの関係性を整理することで、1期目はほぼ終わってしまいました」と森さん。ノウハウが蓄積できたので、2期目はRDF化や語彙のマッピングなどを自動化し、省力化で持続可能なデータベースにしようと考えている。さらに、使い勝手も良くする。検索結果を羅列するだけでなく、視覚化などの手法を組み合わせた高度な解析結果を提供することも目指しているという。

微生物はヒトの活動とは密接な関わりがある。そのため医療関係や農業分野だけでなく住宅や便器などのメーカーからもアプローチがある。今後は、微生物以外のデータと結び付けることが目標だと黒川さんはいう。

統合化推進プログラムで進めているほかの研究開発課題との連携も考えている。例えば、かずさDNA研究所（千葉県）の田畑哲之さんが進める植物ゲノムのデータベースや、国立遺伝学研究所の有田正規さんが進めているメタボロームのデータベースとの連携などだ。

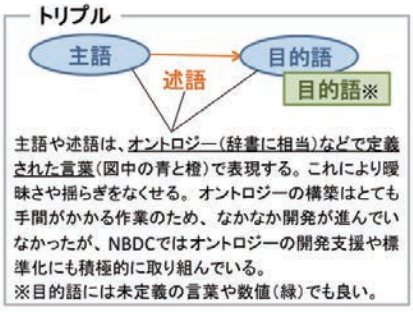
さらに黒川さんは、異分野のデータベース研究にも目を向けている。地質学関連や海洋、気象関連、さらには人工衛星によるリモートセンシングのデータベースなどとの統合だ。「すでに異分野の研究者とも話し合いを始めています」と話す。

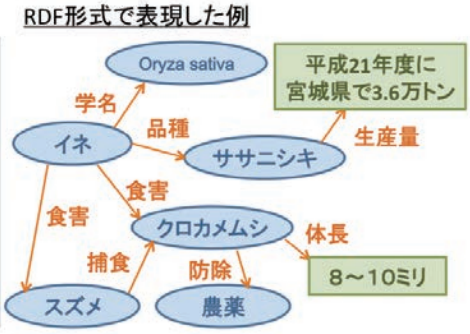
頭の中には、データベース開発が起こす将来の画期的な産業革命が描かれているようだ。

RDF（Resource Description Framework）とは

あらゆる“ものごと”を、コンピューターが理解（処理）できるように記述するための形式（文法）。RDF形式の基本文は「トリプル（3つ組）」と呼ばれ、「主語」と「目的語」を、それらの関係を表す「述語」でつなぐ（左下）。右下の例のように、「イネの学名はOryza sativaです」などの文章（自然言語）をRDF形式で記述することで、さまざまな概念がつながる。

RDFを使うと、他分野のデータベースとの統合解析が可能になることで発見の幅が広がる。また、適切なオントロジーを実装することでコンピューターによる推論が可能になり、新しい知識の発見につながる。





オールジャパンでのデータベース整備が重要

長洲 毅志 ながす・たけし
JSTバイオサイエンスデータベースセンター統合化推進プログラム研究総括
エーザイ株式会社アドバイザー

NBDCの第1段階の3年間は、組織作りと新しい枠組みの活動であっという間に過ぎました。日本でメインのデータベースセンターとしての稼働が実質的に始まったといえます。利用面では、先行しているアメリカのデータに向かいがちですが、これからはNBDCに豊富なデータがあるから皆さん使いましょうといえる時代になってきたと思います。もちろん、ヒトのデータなど、量的に不足していますが、企業に対してもデータが公開されていますので、将来が楽しみな状況になってきました。

今後の展望として、成果を出すためには、ある程度の量のヒトゲノムのデータが集まって、オープンにされることが必要です。これは国の法や制度の整備を含めると数年かかると思いますが、そのころには、ヒトのデータを皆さんの研究に活用できるようにしたいと考えています。「ヒトのゲノム配列やプロテオーム（たんぱく質に関する情報）やメタボローム（代謝に関する情報）」などといった「オミックス情報」のほか、ヒトの行動パターンなど、遺伝子とは少し離れたところにある情報をもう少し豊富にしていくべきでしょう。

個別の企業や大学が、使いたい情報をすべて自分のところだけで集めて囲っているだけではいけません。オールジャパンで情報を集めてデータベースを作り、それをもとに皆で研究を進める体制が必要です。関連分野の情報をお互いに使えるようになることで、日本の生命科学研究は発展が加速されるでしょう。NBDCのデータベースがその中核を担うべきと考えています。

また生命科学と情報分野の両輪を扱える人材も必要です。両分野ともに精通している人が日本にもっといて欲しいと思います。アメリカなどにはすでに一流の人材がたくさんいます。人材を養成する体制を作らないと海外に太刀打ちできなくなります。そうなる前に何とか手を打たないといけません。

データベースは過去の研究や知的活動の蓄積であり、未来を豊かにする資源でもあるのです。