特集

データの共有と統合で生命科学に変革を

生命科学分野のデータベース統合を目指すバイオサイエンスデータベースセンター（NBDC）が誕生して10年。これまでの成果や研究を取り巻く環境の変化、今後の展望について、高木利久センター長らに話を聞いた。

データが価値を生む ビッグデータ時代の到来

生命科学では50年に一度の周期で大きな変革が起きている、とNBDC前運営委員長である堀田凱樹博士は言う。その言葉通り、1900年にはメンデルの法則の再発見、53年にはDNA二重らせん構造の発見が生命科学に大きなパラダイムシフトをもたらした。そして2003年のヒトゲノムの解読は生命科学の研究スタイルを大きく変え、情報生物学の時代が幕を開けた。これを先取りする形で、01年JSTはバイオインフォマティクス推進センター（BIRD）を立ち上げ、生物学の情報技術開発、データベース構築、人材育成を目指した。

その後の10年で、ゲノム解読装置の性能が1万倍になるなど計測技術は革命的といえるほど進歩し、データ量は指数関数的に増加した。人工知能やスーパーコンピューターも驚くほど進歩した。これにより、さまざまな種類のデータの組み合わせから人工知能を使って新たな仮説を生み出すという研究アプローチが生まれた。これを「データ駆動型科学」と呼ぶ。データそのものが価値を生むビッグデータ時代が訪れたのである。

研究者は自分の実験で得たデータだけでなく、他の研究者のデータもより積極的に必要とし始めた。しかし当時は多くのデータが研究室や研究プロジェクトの中に囲い込まれ、他人のデータを活用するのは容易でなかった。

データを世の中に解き放ち、誰でも自由に使えるようにすることが必要だった。複数のデータベースの組み合わせから仮説を得るには、公開するだけでなく各データベースのフォーマットや専門用語を統一しなければならない。これがデータベース統合プロジェクトの始まりである。このプロジェクトは06年頃から文部科学省を中心に生まれ、その推進センターとして11年にNBDCが設立された。

データを使いやすく提示 研究者の利便性向上

「世界には2万以上の生命科学データベースがあるといわれますが、フォーマット、専門用語、精度などがばらばらで実験研究者はもとより、バイオインフォマティクスの専門家でも多様で膨大なデータを使いこなすには高いスキルが必要です。NBDCではデータの共有と統合を通じて研究者の負担を大幅に減らし、データベースの利便性を高める努力をしてきました」とNBDCの高木利久センター長は言う。

例として高木センター長は13年に開始したNBDCヒトデータベースの構築を挙げる。これは疾患研究のために主に日本人のゲノムデータなどを集めたものである。データの登録数は年々増え続け、疾患研究に欠かせないデータベースになった（図1）。

NBDCで企画運営に携わる眞後俊幸主査は、データ統合の具体例を挙げる。これまでに報告された遺伝子のオンオフに関わる文献に用いられた膨大なデータは、公共のデータベースに登録され世界中の誰でも見ることができるという。そこでNBDCの支援で京都大学の沖真弥特定准教授はそれらのデータを整理統合し、見やすく表示した「ChIP-Atlas」を開発したところ、公開からわずか5年で150以上の研究論文で引用されたという。「データがどこかにあって使えるというだけでは駄目なのです。それらを整理統合して初めて価値が出ます」と眞後主査は成果を強調する。

これと並行して、世界中の主要なデータベースの「RDF化」を進めてきた。RDFはデータとその意味を主語、述語、目的語の3つ組で統一的に表現するもので、これにより多数のデータベースが網の目のようにつながり、2次的、3次的データベースを素早く柔軟に作ったり、さまざまな視点で考察したり、人工知能を含む多様なツールを適用できるため、容易に仮説を生成できる（図2）。

例えば、「TogoVar」というデータベースはRDF化されたデータベースを用いたことにより非常に効率良く作ることができた。開発コストが低減した結果、実験研究者が求めてきたサービスを短期間で提供できるようになったのだ。

RDFデータベースは現時点ではまだ検索用語が実験研究者向けではないので、文献データベースのように研究者が日常的に使うまでには至っていない。高木センター長は「はからずも新型コロナウイルスの影響でオンライン会議が一気に普及したように、資材不足などで実験がしにくくなっている今、データベースの活用が研究推進力として重要度を増してくれば、RDFデータベースを縦横無尽に使ってデータ駆動型科学を実践する段階が訪れると思います。それまでは地道にデータベース基盤を整備していきます」と言う。

伝道と人材育成が鍵 データ駆動型科学の実践

「統合データベースができ、格段に使いやすくなったとはいえ、研究者がストレスなく使えるまでにはまだまだです。これからはデータベース作りだけでなくデータ駆動型科学を実践し、RDFデータベースも含めデータベースがこんなに研究に役立つんだよということを示していきたい」と高木センター長。

NBDCの守備範囲は医学、創薬、農学、有用物質生産など幅広いが、「まずは医学の分野で研究者や臨床医が日常的に使えるものを目指します」と続ける。TogoVarだけでなく、入力した塩基配列に対してゲノム編集で重要となるCRISPR-Cas9システムのガイドRNAを設計できる「CRISPRdirect」や、患者の症状を入力すると関連する希少・難治性疾患の候補を可能性が高い順に自動的にリストアップする医療者向け検索システム「PubCaseFinder」などのサービスを通じて、その目標は確実に達成されつつある。

誰でも簡単に使えるデータベースを作るには高度な専門知識が求められる。「NBDCの主な役割ではありませんがデータベース作りの面白さ、重要性を広く伝え、この分野への人材の参入、育成にも貢献したい」と話す高木センター長自身も、データ駆動型科学の伝道者だ。

新技術との融合 30年後の生命科学とは

50年周期なら、次の変革は2050年だ。その時には何が起きるのか。「30年先のことは予想もつきませんが、10年後くらいには実験ロボットを使って網羅的で均質なビッグデータが簡単に得られ、それを人工知能で統合解析し、自動的に仮説を出すという時代になっているのではないでしょうか」と高木センター長。

ロボットと人工知能がさらに活躍する未来に、生命科学を取り巻く状況はどうなっているのか興味は尽きない。データベース開発を通じて、生命科学研究に新潮流を生み出し続けるというNBDCの夢はこれからも続いていく。

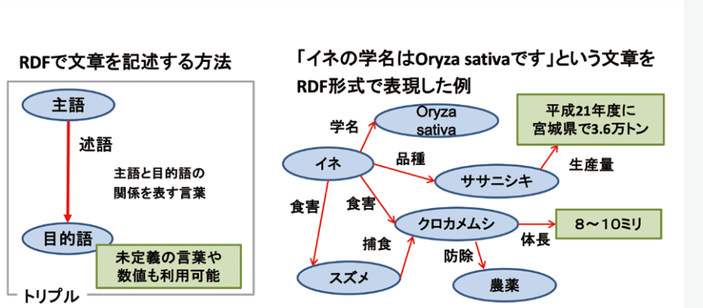

図2　RDF（Resource Description Framework）は、コンピューターが処理できるよう、さまざまな事象の「主語」と「目的語」、それらの関係を表す「述語」でつないだ「3つ組（トリプル）」表記する文法形式だ（左）。例えば「イネの学名はOryza sativaです」という文章をRDF形式で記述すると、さまざまな概念をつなげることができる（右）。RDF化によって、他分野のデータベースとの統合解析の他、コンピューターによる推論も可能になり、新しい発見につながる。

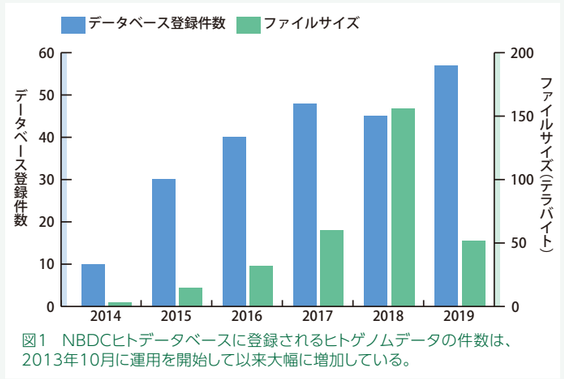

図1　NBDCヒトデータベースに登録されるヒトゲノムデータの件数は、2013年10月に運用を開始して以来大幅に増加している。

日本人ゲノム多様性統合データベース「TogoVar」疾患の解明に期待

日本人のバリアント情報を公開 論文情報もワンストップで提供

ヒトゲノム配列の解読は1990年代より欧米を中心に進められ、2003年には標準的なゲノム配列「参照配列」として公開された。同じヒトという種でも、ゲノム配列は人種や個人ごとに異なる。

NBDCは、日本人のゲノム配列の多様性に関する情報を集めたデータベースとしてTogoVarを18年6月に公開した。「参照配列との違いを表すバリアントの出現頻度を日本人集団と他集団で比較するだけでなく、生物学的意味や関連する論文もワンストップで検索できます」とTogoVarの開発に携わる豊岡理人研究員は説明する（図3）。

希少疾患の原因を絞り込み 国内外がデータに注目

20年7月には、全ゲノム解析した日本人集団のサンプル数としては最大規模となる、7609人分のゲノム配列データを統合・再解析して得られたバリアント情報の総合的なリストを日本の複数機関が連携して作成し、TogoVarから公開した。このデータは国内外の大学や研究所、民間企業など、公開からわずか2カ月で100近い研究グループに利用されているそうだ。

子供だけに発症した希少疾患がある場合、その両親にはなく子供だけにあるバリアントは疾患と関係する可能性がある。しかし、親子のゲノムを比較しただけでは子供に固有のバリアントの数がまだ多く、どれが疾患の原因であるかの特定は困難だ。「日本人集団の中に存在するバリアントも子供のバリアントから取り除ければ、疾患の原因となるバリアントをより早く特定できます」と豊岡研究員（図4）。これまで未知・未解明だった日本人特異的な希少疾患や単一遺伝子疾患の病因解明に役立つ基礎的なデータとして期待される。

「今回のデータでは日本人固有のバリアントがたくさん見つかっています。日本人に特異的なデータを蓄積、公開することは日本の研究者だけでなくアジアを中心に海外の研究者へも貢献できると考えています」とも豊岡研究員は話す。実際、TogoVarは50カ国からのアクセスがあり、利用されている。

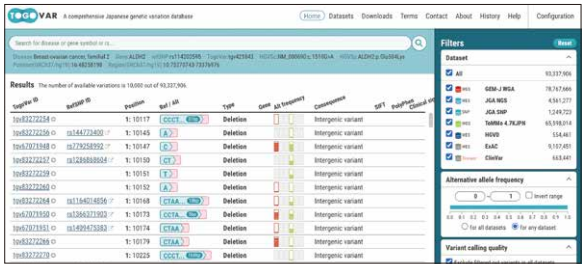

図3　TogoVarは、トップページから個々のバリアント情報、データベースごとの頻度、疾患との関与などをワンストップで検索できる。

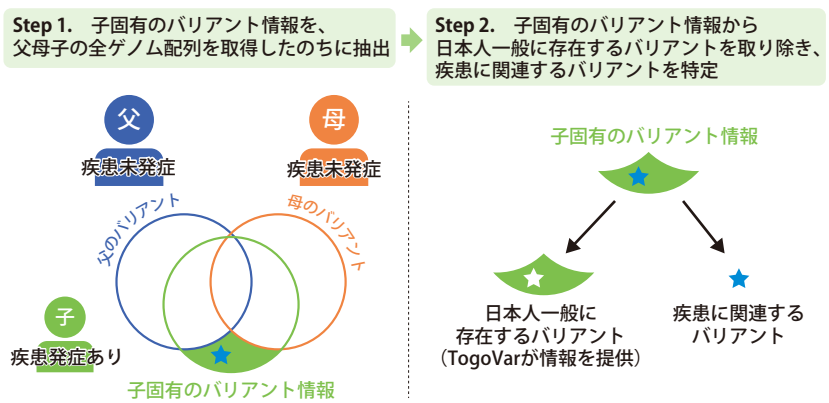

図4　TogoVarで提供する日本人のバリアント情報を使い、日本人特異的な希少疾患や単一遺伝子疾患の病因解明を促進する。

研究者が使いたくなる ユーザー目線の開発

豊岡研究員はシステムエンジニアなどを経て、遺伝子解析の道に進み、TogoVarの開発に関わることになった。

「タイで結核の疾患関連遺伝子研究のためにゲノム解析をしていたことがありますが、今取り組んでいる仕事は、当時の自分が欲しかったデータベースを作り出すことです。TogoVarの開発では、使う側の立場に立って、インターフェースも使いやすく工夫しました」と振り返る。

今後はゲノムのバリアント情報だけでなく、遺伝子や疾患をまとめた情報や遺伝子発現に関係する情報も追加し、さらなる利便性の向上を目指すという。