



生命科学分野における データ共有の取り組み

応
専

■ 高祖歩美 (科学技術振興機構 バイオサイエンスデータベースセンター)

生命科学分野におけるデータ共有

生命科学分野では、公表されている研究結果を追試するため、再現するために研究者コミュニティあるいは、広く社会にその研究結果の元となったデータを公開し、再利用できるようにすることを促してきた。この取り組みは、データ共有 (data sharing) と呼ばれる。また、データの利用条件を緩和して誰もが自由にデータを利用できるように公開するオープンデータの活動もデータ共有の一環として、ヒトのゲノムが解読された 2000 年はじめ頃から盛んになっている。本稿では生命科学分野におけるデータ共有がいかに発展してきたかを概説しつつ、国内外の生命科学分野のオープンなデータも紹介する。

公的なデータベース

生命科学分野におけるデータ共有に、公的なデータベースが果たしてきた役割は大きい。欧米を中心に 1970 年代ごろから現在に至るまで、登録されたデータを実質的に誰もが無償で自由に利用できる公的なデータベースが作られ、公開されてきた。公的なデータベースの構築は、ある研究結果の追試や再現性を確認するためには、その結果の元となったデータが必要である、という学問上の要請によるところが大きい。たとえば、DNA の塩基配列については、日本・米国・欧州の 3 極で運営する国際塩基配列データベース (International Nucleotide Sequence Database^{☆1}, 図-1) が存在し^{☆2}, 2013 年現在、世界中の研究者が解読した 26 万種以上の生物種の塩基配列を入手できる¹⁾。そのほかにも、遺伝子発現



図-1 日本・米国・欧州の3極で運営する国際塩基配列データベース DDBJ の Web サイト (<http://www.ddbj.nig.ac.jp/intro-j.html>) より CC-BY により提供されている画像を利用した。

データやタンパク質のアミノ酸配列のデータ、タンパク質を含む高分子の構造のデータ、化合物のデータなど多様な種類のデータを収録した公的なデータベースがある。このような公的なデータベースを米国では、米国国立生物工学情報センター (National Center for Biotechnology Information : NCBI^{☆3}) が、欧州では、欧州バイオインフォマティクス研究所 (European Bioinformatics Institute : EMBL^{☆4}) が中心となって整備し、維持している。多岐にわたるデータベースの一覧は、それぞれの機関の Web サイト

☆1 <http://www.insdc.org/>

☆2 国際塩基配列データベースは、国立遺伝学研究所で運営される DNA Databank of Japan (DDBJ), 米国国立生物工学情報センターで運営される GenBank と欧州バイオインフォマティクス研究所で運営される European Nucleotide Archive (ENA) によって構成されている。3 極はそれぞれ異なった名前のデータベースを運営しているが、登録されている DNA または RNA の塩基配列は同じである。

☆3 <http://www.ncbi.nlm.nih.gov/>

☆4 <http://www.ebi.ac.uk/>

トから把握できる。

公的なデータベースが発展した背景には、多くの学術雑誌が論文内で引用するデータをあらかじめ公的なデータベースに登録することを義務づけるようになったことが影響している。たとえば、DNAの塩基配列については、日本・米国・欧州の塩基配列データベースのいずれかにあらかじめデータを登録し、登録によって得られるデータの登録番号（アクセス番号）を論文内に記載しなければならない。同様の論文投稿の規定は、遺伝子発現データやタンパク質のアミノ酸配列のデータ、タンパク質を含む高分子の構造などのデータにも課されている。それまでは、データベースを作成し、維持する側が塩基配列のデータを論文内から1つ1つ拾ってデータベースに登録していた。しかし、解読される塩基配列の量が増えるにつれ、このような方法では追いつかないこと、塩基配列の詳細は論文中にそれを引用した執筆者、すなわちデータの産出者自身がよく知っていることなどを理由に、1980年代の終わり頃から執筆者も登録に協力する仕組みへと変わった。この仕組みは学術論文の投稿規定が変わることにより実現され、新しい規定の下では、論文の執筆者が塩基配列をデータベースに登録することとなった。この学術雑誌の義務づけによって、公的なデータベースにおけるデータの集積が加速され、生命科学分野におけるデータの共有は促進されてきた。

一方で、公的データベースの中に登録されているデータはすべてオープンデータであるか、という点について曖昧さが残る。それは、データの利用条件に由来している。ここでのオープンデータとは、英国の非営利団体 Open Knowledge Foundation が掲げる定義に基づいて^{☆5}、“自由に使えて再利用もでき、かつ誰でも再配布できるようなデータのことだ。従うべき決まりは、せいぜい「作者のクレジットを残す」あるいは「同じ条件で配布する」”^{☆6}としよう。たとえば、NCBI では、上述の塩基配列データベースに登録されている塩基配列の利用や配布を何ら制限しないし、その利用や配布に制限を設けるような塩基配列の登録は認めないとしている。一方で、

登録されている塩基配列の中には、他人が特許、著作権などの知的財産権を主張し得るもの、あるいは取得しているようなものが含まれている可能性があり、登録されている塩基配列に関する知的財産権を NCBI が譲り受けてデータベースから公開しているわけではない。そのため、NCBI は個々のデータの利用、複製、再配布について無制限に許可することはできないし、NCBI としての意見を表明することもできないとしている^{☆7}。つまり、登録されているデータの利用にあたっては、利用者の責任で他人の知的財産権を侵害しないようにする必要がある。実質上、日々の研究を行う生命科学分野の研究者にとって公的なデータベースは、インターネットにさえ接続できれば、そこで公開されているデータを自由に無償で利用できる。しかし昨今のオープンデータの定義に照らし合わせると公的なデータベースに収録されているデータの位置づけは難しい。

ヒトゲノムプロジェクト

生命科学分野におけるデータ共有の歴史においてヒトゲノムプロジェクトは、1つの転換点と言える。ヒトゲノムプロジェクトとは、欧米や日本を始めとする多数の国が参加してヒトのゲノムの全塩基配列を解読した国際的なプロジェクトである。上述のようにそれまでは学術雑誌によって、データベースへの塩基配列の登録が促されていた。一方で、創薬や医療への応用、特許の取得などの観点からヒトの塩基配列については、特別な取り扱いが必要との見解もあった。これを体現したのがヒトゲノムプロジェクトと競ってゲノム解読を進めた Craig Venter 率いる Celera Genomics 社であった。同社は独自に解読した塩基配列を特許化しようとして世界的に大議論となった。しかし、ヒトゲノムプロジェクトでは、解読された塩基配列はパブリックドメインに帰するものとされ、各国の関係者はそのデータを解析

☆5 <http://opendefinition.org/>

☆6 <http://opendatahandbook.org/ja/what-is-open-data/index.html> より

☆7 <http://www.ncbi.nlm.nih.gov/About/disclaimer.html#disclaimer>

後 24 時間以内に公開して、誰でも自由に利用できるよう取り決めた（バミューダ原則）。このバミューダ原則に基づいて、塩基配列がパブリックドメインに置かれて、その配列の特許や著作権などの他人の知的財産権を気にすることなく、ヒトゲノムプロジェクトを通して得られたデータを利用できることとなった。

このような流れに沿って、ヒトのゲノムデータ以外にもその利用条件の緩和と明示がな

されているデータが出てきている。知的財産権のうち著作権の利用条件を緩和するもので、クリエイティブ・コモンズ・ライセンス（CC ライセンス）を使用するものが多く見られる^{☆8}。CC ライセンスとは、インターネット上に公開した自分の作品の利用条件をマークで表せるツールである。2002 年に米国の NPO 法人クリエイティブ・コモンズが著作物の利用と流通を促進するために提供をはじめたライセンスで、2013 年現在では 70 もの国や地域で使用されている。CC ライセンスには、4 つの利用条件（表示、継承、非営利、改変禁止）を組み合わせた 6 つのライセンスとそれぞれのライセンスに対応するマークがあり、マークを作品に表示して利用条件を明らかにできる。たとえば、欧州バイオインフォマティクス研究所では、医薬品や医薬品の候補となるような低分子の化合物の情報を集めたデータベース、ChEMBL^{☆9}（図-2）を公開し、ChEMBL に収録されているすべてのデータを CC ライセンス表示 - 継承 3.0 非移植^{☆10}で提供している。CC 表示 - 継承は、その利用にあたって、1) 作品の原作者の氏名、作品のタイトルと URL を表示し、2) CC 表示 -

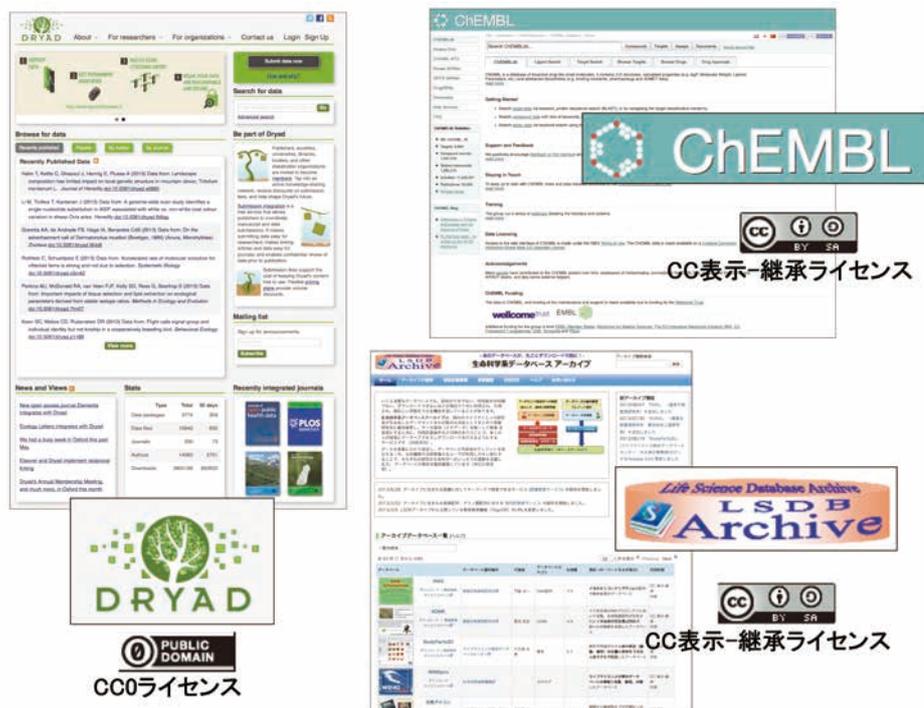


図-2 データの著作権について利用条件が緩和されているデータベース（Dryad, ChEMBL, 生命科学系データベースアーカイブ）

継承のライセンスで改変した作品を公開し、提供することを条件に、たとえば、作品を複製したり、改変したり、商用目的で利用できる。さらに利用条件が緩和されている例として、米国立進化統合センター（National Evolutionary Synthesis Center^{☆11}）とノースカロライナ大学チャペルヒル校によって運営されている Dryad^{☆12}がある。もともと環境学や生態学に関連するデータの保管場所として立ち上げられたものだが、現在では広く生命科学分野のデータも収録対象としている。Dryad に収録されているデータは、パブリックドメインに置かれており、CC0（ゼロ）と呼ばれるライセンスを明示している。CC0 は著作権および著作権に関する諸権利の行使

☆8 データ共有は科学の発展に資するため、データの一部の知的財産権について利用条件を緩和するだけでなく、データをパブリックドメインに置くべきとの見解もある。たとえば、パントン原則（<http://pantonprinciples.org/>）など。

☆9 <https://www.ebi.ac.uk/chembl/>

☆10 通常、CC ライセンスは世界のどの国や地域でも同じ効果を得るために、また簡単に理解され利用されるようにするために、各国や地域の法律に合わせて作成されている（移植という）。一方で非移植とは、特定の国や地域の法律に合わせて作成されているのではなく、6 つの著作権に関する国際的な条約に基づいて作成されていることを指す。

☆11 <http://www.nescent.org/>

☆12 <http://datadryad.org/>

を、法律で認められる限り、放棄、または差し控えることを表明するライセンスである。Dryad に収録されているデータのほとんどは査読された学術論文と紐づいているという特徴があり^{☆13}、2013年8月現在、10,000件以上のデータファイルと4,000件近くのデータパッケージ（1つの論文と紐づいた複数のデータファイルの集合体）が収録されている。ひるがえって国内では、科学技術振興機構 バイオサイエンスデータベースセンター（NBDC）が我が国の生命科学分野におけるデータベースの再利用と保存を推進するために「生命科学系データベースアーカイブ」^{☆14}を提供している。「生命科学系データベースアーカイブ」もCCライセンス表示 - 継承日本2.1を標準的なライセンスとして採用して、60件以上のデータセット（Dryadのデータパッケージに相当する概念）を収録している。こちらの特徴は、前2つと異なり、元々データベースとして独自に公開されていたデータセットが多く収録されている点、英語のみならず日本語でのサービス提供がある点にある。

データを共有する方針

これまで見てきたようなデータベースは、各国の政府やファンディング機関が示しているデータを共有する方針に支えられている。データ共有の方針は米国国立衛生研究所（National Institute of Health : NIH）でいち早く2003年頃から導入されたもので、研究費の申請者にその研究費を用いて産出したデータを広く、研究者間や社会一般に公開して、共有することを求める。米国の大学における科学技術の基礎研究を支援する主要な助成機関の1つ、米国国立科学財団（National Science Foundation : NSF）は、NSFの研究費を申請する際にその研究費を使って取得するデータをどのように共有し、管理するのかを記した計画書を提出するよう求める。この計画書が添付されていない申請書は、不採択あるいは審査されることなく返却される。また、計画の実施は次の研究費の採択や不採択にも影響する。同じように英

国のウェルカム・トラスト（Wellcome Trust）や英国がん研究所（Cancer Research UK : CRUK）、英国バイオテクノロジー・生物科学研究会議（Biotechnology and Biological Sciences Research Council : BBSRC）などの英国の主要な助成機関でもデータ共有を申請者の義務として位置づけている（表-1）。また、国際的にも生命科学分野に限らず、公的な資金を投じて得られた研究データのアクセスと共有を促進するために経済協力開発機構（OECD）の加盟国により2006年に「OECD Principles and Guidelines for Access to Research Data from Public Funding（公的資金による研究データへのアクセスに関する原則及びガイドライン）」が採択されている^{☆15}。

我が国でも欧米よりも少し遅れて2008年頃から文部科学省や厚生労働省、科学技術振興機構の助成を受けてなされた一部の研究について、公募の段階からデータ共有に協力するよう呼びかけが始まっている。具体的には、公募要領にて論文発表等で公表された成果にかかわる生データの複製物、または構築した公開用データベースがNBDCへ提供されるよう呼びかけられている。この呼びかけによって提供されたデータは、たとえば、上述の「生命科学系データベースアーカイブ」に収録されるという流れになる。一方で、この呼びかけには欧米のような拘束力や義務づけがない点が、今後の課題として認識されている²⁾。

今後の展開

ここまで読み進んでいただいた読者は、生命科学分野におけるオープンデータには取り組むべき課題はもうないと思われるかもしれない。しかし、そんなことはない。まず、研究者のデータ共有に対する意識の改善に継続して取り組んでいく必要がある。データ共有に対する意識は、分野ごとに違いがある

☆13 このほかに Dryad には査読されていない学術論文と紐づいたデータも含まれている。たとえば、博士論文で引用されているデータなどがこれに該当する。

☆14 <http://dbarchive.biosciencedbc.jp/>

☆15 <http://www.oecd.org/sti/sci-tech/38500813.pdf>

組織名	事業や補助金の名称	データを共有する方針の導入時期*	備考
米国国立衛生研究所 (NIH)	-	2003 年～	年間 50 万ドル以上の直接費の助成を受ける研究に限る
米国国立科学財団 (NSF)	-	2011 年～	
英国バイオテクノロジー・生物科学研究会議 (BBSRC)	-	2007 年～	
英国ウェルカム・トラスト	-	2007 年～	
英国がん研究所 (CRUK)	-	2007 年～	
文部科学省	文部科学省委託事業	2008 年度～	
文部科学省	科学研究費助成事業 特別推進研究 基盤研究 (S・A・B・C) 挑戦的萌芽研究 若手研究 (A・B)	2012 年度～	
厚生労働省	厚生労働科学研究費補助金	2011 年度～	「人体に由来するデータを取り扱う研究は除く」とのただし書きあり
科学技術振興機構	戦略的創造研究推進事業 さきがけ, CREST	2011 年度～	
科学技術振興機構	戦略的創造研究推進事業 先端的低炭素化技術開発 (ALCA)	2013 年度～	

* 我が国では、公募要領に協力依頼がはじめて掲載された時期を指す

表-1 各国のデータを共有する方針の導入時期

とされているものの、2011年にScience誌が自身の論文の査読者を対象にアンケート調査を行ったところ、およそ1,700人の回答者のうちの約8%がデータをほかの研究者もアクセスすることができる場所に登録しており、約半数が研究室内に保存している、つまり関係者以外はアクセスできない、と答えている。一方で、約半数を超える回答者は論文で引用されているデータの提供を依頼したことがあると答えており³⁾、日々の研究活動の中でデータを共有する必要性を感じつつも、なかなか自分のデータを広く共有するまでには至っていない様子が見えてくる。

次に、計測技術や情報処理技術の向上により、個人のデータを共有することができるようになり、ヒトに関連するデータの共有についてさらなる議論が必要となっている。ゲノム情報を例にとると、ヒトゲノムプロジェクトが実施された1990年代ごろは、ヒト一人の全塩基配列を解読するために10数年もの歳月と複数の国にまたがる国際的な研究グループの動員を要した。しかし現在ではヒト一人の全塩基配列を1つの研究者グループがたった数日で、しかも安価に解読できる。塩基配列などを含む個人のゲノム情報を個人の健康情報など、さまざま

な情報と組み合わせれば、複数の要因が関与している疾患の解明や予防、診断、治療に生命科学研究はいっそう貢献できると考えられている。一方で、個人のゲノムの塩基配列は、個人ごとに異なり、塩基配列に基づいて将来どのような病気にかかりやすいかなどを把握することができるため、雇用や保険の差別につながる可能性があり、倫理的な配慮が必要な情報である。加えて、個人のゲノム情報とその他の情報を組み合わせることによって、個人が特定できる場合もある(たとえば、珍しい病気等を発症している患者のゲノム情報とその患者の居住地の情報とを組み合わせる)ため、個人情報としての側面も有する。これは、ゲノム情報に限らず、健康情報や脳画像などにも当てはまる。したがって、ヒトに関連するデータを共有する場合には、データの特성에応じて個人情報の保護や倫理的な観点にも配慮し、適切なアクセス制限や情報セキュリティ対策を講じる必要がある。我が国では生命科学分野においてNBDCがヒトに関連するデータを共有する場合のガイドラインを策定して公開している^{☆16)}。

☆16 <http://humandbs.biosciencedbc.jp/guidelines/>

また国際的には OECD 加盟国は、2009 年に「OECD Guidelines on Human Biobanks and Genetic Research Databases (ヒトのバイオバンクおよび遺伝学研究用のデータベースに関する OECD ガイドライン)」を採択して、世界的に調和のとれた政策のもとでヒトに関するデータを共有することが望ましいとしている^{☆17}。このほかにも、2013 年 6 月に 40 カ国以上にわたる約 70 の代表的な医療機関、研究所や関連団体によって、ヒトのゲノム情報や診療情報を世界的に共有するための協定が締結された^{☆18}。この協定には我が国の医療機関も参画しており、今後はこの協定のもとでヒトのゲノム情報や医療情報を共有するための共通の情報基盤の開発や、プライバシーや倫理に配慮した共通のガイドラインの策定、関連する手続の調和などが進められていくこととなっている。

このようなヒトに関連するデータを共有することの有用性は、米国で 2004 年頃から始まったアルツハイマー病の予防、治療と根治を目指すプロジェクト (Alzheimer's Disease Neuroimaging Initiative, ADNI^{☆19}) の活動において見られる。ADNI では、プロジェクトの一環として取得された 800 名を上回る実験参加者の健康情報や脳画像、ゲノム情報などを世界中の研究者に無償で提供している^{☆20}。これによって、これまでにアルツハイマー病に関連する 350 以上の論文が発表され、データは全世界的

に約 2,500 名の研究者に利用され、アルツハイマー病の研究に役立てられている。また、今後は新たに個々の実験参加者の全ゲノムの塩基配列が共有される予定であり、アルツハイマー病の研究を加速させるだけでなく、ヒトに関連するデータの共有をも加速されると予想される。

まとめと今後への期待

本稿では、生命科学分野におけるデータ共有の進展を国内外の状況を踏まえて紹介した。オープンデータの取り組みは、オープン化されたデータが社会で活用されて、新しいサービスや産業、価値の創出につながってはじめてその目標が達成されたと言える。そのような意味で本稿が生命科学分野においてオープン化されているデータを利活用するきっかけとなれば幸いである。

参考文献

- 1) Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Sayers, E. W. : GenBank, Nucleic Acids Research, Vol.41, pp.D36-D42 (2013).
- 2) 高祖歩美 : 生命科学分野におけるデータの共有の現状と課題, 情報管理, Vol.56, No.5, pp.294-301(2013).
- 3) Science Staff. : Challenges and Opportunities, Science, Vol.331, No.6018, pp.692-693 (2011).

(2013 年 9 月 1 日受付)

謝 辞 本稿を執筆するにあたり、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の川本祥子氏に貴重な意見やコメントをいただいた。

☆17 <http://www.oecd.org/sti/biotech/44054609.pdf>
☆18 International partners describe global alliance to enable secure sharing of genomic and clinical data, Broad Communications, June 4th, 2013.
<http://www.broadinstitute.org/news/globalalliance>
☆19 <http://www.adni-info.org/>
☆20 <http://adni.loni.ucla.edu/>

■ 高祖歩美 a2koso@jst.go.jp

東京大学教養学部卒業、首都大学東京大学院博士課程単位取得満期退学。博士(言語学)。現在、(独)科学技術振興機構バイオサイエンスデータベースセンター所属。