提言

持続可能な生命科学のデータ基盤の整備に向けて

令和元年（２０１９年）１１月１８日

日　本　学　術　会　議

基礎生物学委員会・統合生物学委員会・農学委員会・

基礎医学委員会・薬学委員会・情報学委員会合同

バイオインフォマティクス分科会

この提言は、日本学術会議基礎生物学委員会・統合生物学委員会・農学委員会・基礎医学委員会・薬学委員会・情報学委員会合同バイオインフォマティクス分科会の審議結果を取りまとめ公表するものである。

日本学術会議
基礎生物学委員会・統合生物学委員会・農学委員会・
基礎医学委員会・薬学委員会・情報学委員会合同
バイオインフォマティクス分科会

委員長	有田　正規	(連携会員)	大学共同利用機関法人情報・システム研究機構国立遺伝学研究所教授
副委員長	岡田　眞里子	(連携会員)	大阪大学蛋白質研究所細胞システム研究室教授
幹　事	高木　利久	(第二部会員)	富山国際大学学長
幹　事	諏訪　牧子	(連携会員)	青山学院大学理工学部化学・生命科学科教授
	岩崎　渉	(連携会員)	東京大学大学院理学系研究科准教授
	上田　泰己	(連携会員)	東京大学大学院医学系研究科機能生物学専攻教授
	久原　哲	(連携会員)	九州大学大学院農学研究院教授
	斎藤　成也	(連携会員)	大学共同利用機関法人情報・システム研究機構国立遺伝学研究所集団遺伝研究部門教授
	德永　万喜洋	(連携会員)	東京工業大学大学院生命理工学研究科生命情報専攻教授
	中村　春木	(連携会員)	国立遺伝学研究所 DDBJ センター特任教授、大阪大学名誉教授
	美宅　成樹	(連携会員)	名古屋大学名誉教授、公益財団法人豊田理化学研究所・客員フェロー

本提言の作成にあたり、以下の方々に御協力いただいた。

安達　澄子	国立研究開発法人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主査
河野　信	富山国際大学現代社会学部准教授
眞後　俊幸	国立研究開発法人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主査
木下　賢吾	国立大学法人東北大学大学院情報科学研究科教授
金久　實	京都大学化学研究所特任教授

本提言の作成にあたり、以下の職員が事務を担当した。

事務局	**高橋　雅之**	**参事官（審議第一担当）**
	酒井　謙治	**参事官（審議第一担当）付参事官補佐**
	三神　雅子	**参事官（審議第一担当）付審議専門職**

要　旨

本提言は、現在の生命科学とバイオ産業における脆弱なデータ基盤や運用体制の問題点を研究者や行政機関の関係者と共有し、今後、データ基盤を持続的に整備、発展させるための方策についてまとめている。本提言を踏まえ、持続可能なデータ基盤が整備されることが強く望まれる。

１　作成の背景

生命科学はビッグデータの網羅的な収集とその情報解析を基盤とした学問に大きく変貌しようとしている。その背景には、生体計測技術の革命的とも言える進歩、AI（人工知能）やIT技術の目覚ましい発展、公的資金で生成された研究データを広く利用可能としイノベーションにつなげようとするオープンサイエンス促進の世界的な潮流、などがある。

基礎学問にとどまらず、ゲノム医療、ゲノム創薬などの言葉に代表されるように医学や薬学の分野や、新しい育種法の開発や微生物を活用した有用物質生産など、農業やバイオマテリアル産業にも、大きな変革の波が押し寄せている。

このような動きの中で、ゲノム配列やタンパク質構造のような大量の生命情報を処理・予測するアルゴリズム研究が先行していたバイオインフォマティクスも、ゲノムに限らず様々な種類のビッグデータを整理・統合・解析し、データに潜む規則性や仮説を導くための学問へと大きく発展してきた。実験生物学から生み出されるデータを処理するためという従属的学問から、ビッグデータ時代における主役の学問になりつつある。

ビッグデータの収集とバイオインフォマティクスを用いたデータの統合・解析というアプローチを通して、生命科学におけるデータ科学の重要性は増す一方である。

２　現状及び問題点

生命科学の新たな潮流を支え更に加速するには、①データベースの構築やビッグデータ解析技術の開発、②これらのビッグデータ解析を支えるスーパーコンピュータの整備、③新たな生命科学やバイオ産業の担い手となる人材の育成、がいずれも欠かせない。本提言では、データベース、ネットワークを含めたスーパーコンピュータシステム、担い手となる人材の３点をまとめて「データ基盤」と呼ぶ。

我が国でもデータ基盤の必要性、重要性が認識されてはいるが、多くの施策がビッグデータによる経済効果や短期的なインパクトを目指して立案されるため、分野横断の基盤となる装置や人材等、インフラストラクチャーの整備が相対的に置き去りにされ、予算を削減される傾向にある。

３　提言の内容

本提言は、「データ基盤」を持続的に整備、発展させるための方策について今後議論すべき論点と方向性を明らかにすることを目的とする。現状を整理するため、１章で「データ基盤」が必要となった一般的な背景や目的を述べる。２章では世界におけるデータベース開発及びそれを担うデータベースセンターの現状を、３章ではその課題を、明らかにする。提言の中核である４章では、持続可能なデータ基盤整備のあり方を論じ、下記の５項目について今後の方策を述べる。以下にその概要を示す。

（1）データ共有政策の作成と義務化

国が省庁横断的に適用できるデータ共有政策を作成し、研究資金配分機関は政策に基づいてデータ共有の環境整備と義務付けを行うべきである。また、研究データの生産者に対してデータ共有に対する動機づけ（インセンティブ）を付与する仕組みを導入すべきである。

（2）プロジェクト立案時からのデータベース戦略策定

生命科学に必要なデータを産出し、データベースを整備する戦略が必要である。データの共有・公開・統合のために、プロジェクト立案時から、データ産出プロジェクトとデータベースセンターが密に連携すべきである。その際、新技術に基づくデータ産出や人工知能技術との連携、産業応用への貢献を充分に考慮する必要がある。

（3）データベースセンターの一元化とスーパーコンピュータの整備

複数ある国内のデータベース関連機関を統合し、国内には政策立案・研究開発ともに強力なリーダーシップを発揮しつつ、国際的には存在感と競争力を打ち出せる体制を整えるべきである。また、データ量の増加とデータ利用者の増加に対応できる、生命科学の巨大データ解析に適したネットワーク及びスーパーコンピュータの増強が必要である。

（4）人材育成と教育体制の整備

バイオインフォマティクスの人材不足の解消に向けて、高校教育・大学入試・大学（学科新設など）・民間企業のあり方、研究プロジェクトの立て方などを見直し、一過性の取組ではない中長期的な人材育成の体制整備が必要である。

（5）予算の確保、データ量やデータ種類の増加に対応した仕組みの導入

将来にわたりデータを整備・活用していくためには、国のライフサイエンス研究分野の公的資金のうち一定割合を措置するような、新たな財源モデルの構築が必要である。その際には長期安定性だけでなく、オープンサイエンスへの貢献や、費用負担の公平性にも留意する必要がある。品質管理がなされたデータベース（レポジトリ、知識ベース、統合データベース）の構築と長期にわたる管理・維持やそれらのデータを扱うスーパーコンピュータ、及び社会へ安全・高速にそれらデータを公開する仕組みをそれぞれの事情に応じて整備すべきである。

我が国においては、データ基盤整備及びその持続可能性の確保において、現時点で世界に大きな遅れを取っており、今後その差はますます開く可能性が高い。この状況を打開するためには、明確で持続可能なデータ基盤整備の具体的な推進方策が必要である。

目　次

１　背景と目的：生命科学研究における「データ基盤」の重要性

生命科学では、日々、新たな計測や観測の技術が生まれ、新しい種類のデータが日々大量に生み出されている。精密なデータが安価かつ高速に得られるようになり、情報通信技術の発展も相まって、生命科学はビッグデータを基盤とした学問へと大きく変貌しつつある。網羅的・系統的に収集された大規模データに基づく新しいタイプの研究アプローチが発展拡大しているのである。

個々人のゲノムデータに加え、生活・環境データなど、多様なデータが迅速かつ安価にビッグデータとして取得できるようになると、研究成果の社会還元が加速されるだけでなく、より社会に根ざした研究が可能になってくる。情報技術と組み合わせることで、専門家以外もデータを登録して研究に参加したり、インターネット上に公開された世界中のデータを日本語で読み解いたりすることも可能になるだろう。

バイオ産業においても、ビッグデータは、従来とは異なるアプローチによる知的財産の発掘・獲得という新たな可能性に道を拓く。従来ブラックボックスで制御不可能であったプロセスが予測・制御可能となることで、生産性が向上するだけでなく、時間や費用がかかりすぎて困難であった研究開発も実施可能となる。社会活動の多方面でビッグデータが生み出される中、創造的な組み合わせによる新規ビジネスもあろう。

生命科学の新たな潮流を更に加速し、そこから新たな研究アプローチや学問及び産業を生み出していくには、①大規模で質の高いデータベースの構築やビッグデータ解析技術の開発、②これらのビッグデータ解析を支えるスーパーコンピュータ（以下「スパコン」と略す）やネットワークの整備、③これからの新たな生命科学やバイオ産業の担い手となる人材の育成や発掘（以下「人材」と略す）、この３点がいずれも欠かせない。本提言では、「データ基盤」という言葉でデータベース、ネットワークを含むスパコン、人材の３項目を包括的に表現する。ただし、それぞれが大きなテーマであるため、データベースの構築や持続的な更新に主眼を置く。ここでは「データ基盤」の需要と現状を４つの側面から概観する。

(1)　生命科学のビッグデータ化

近年のゲノム塩基配列決定装置をはじめとした生体観測技術、生体計測技術の進歩には目覚ましいものがある。例えば、10 年を超える時間と３千億円超を費やしたヒトゲノム解読は、次世代型のゲノム配列決定装置の登場により、わずか 10 万円程度のコストで可能になった [1]。また、質量分析装置や顕微鏡技術についても急速な進展がみられる。従来に比べ、より精密なデータが安価かつ高速に得られている。これらのビッグデータを基盤とした学問として、生命科学は大きな変貌を遂げつつある。

大きな転機は、ゲノム医療、ゲノム創薬などの言葉に代表されるように、医薬学分野の変革と産業界への影響にある。個人ゲノムが安価に決められるようになり、個別化医療あ

るいは精密医療[1]と呼ばれるような新しい医療が開発され、遺伝子診断が大きなビジネスとして動きつつある。この変革はヒトゲノムにとどまらない。新しい育種法の開発や微生物を活用した有用物質生産など、農業やバイオマテリアル産業にも同じ変革の波が押し寄せている［2］。多様なデータを大量に得られるようになったことで、長年の試行錯誤や経験に基づく専門技術者の技能を客観的データによって理解し工業的に再現できるようになった。そして、「データ基盤」の整備は、こうしたビッグデータを各種の産業に役立てる前提条件に他ならない。

(2) データ駆動型科学への展開

一方、情報通信技術の発展も目覚ましい。特に近年の深層学習をはじめとした人工知能技術の発展は、科学や産業そのもののあり方を大きく変えようとしている［3］。これと上記の生命科学のビッグデータ化が相まって、生命科学及びバイオ産業は従来の仮説駆動型からデータ駆動型の科学・産業に大きく舵を切りつつある。ここで言う仮説駆動型とは、個々の研究者が持つ知識をもとに着想した仮説を検証していく、従来の研究スタイルを指す。それに対してデータ駆動型は、網羅的・系統的に収集された大規模データから（事前の仮説をあまり想定せずに）データに潜む規則性（仮説）を導こうとするタイプの研究スタイルを指す。

このような状況を背景として、医学の分野では、何万人にも及ぶ生体情報や生活習慣情報を長期間に渡って収集・解析するゲノムコホート[2]と呼ばれる大規模なプロジェクトが企画されている。日本学術会議から 10 万人規模のゲノムコホートを 10 箇所で実施する提言が出され［4］、東北では既に地域住民コホート調査（８万人）と第三世代コホート調査（７万人）が実施されている［5］。また、先に述べた各種バイオ産業の振興に際しても、バイオ研究とデジタルの融合を合言葉に、内閣府においてバイオ戦略の策定が進められ、データ駆動型科学への展開が図られている［6］。

しかし、こうした科学を実現する「データ基盤」についての議論は進んでいない。生物学の未解決問題を解き明かすには、ビッグデータを解析して新たな法則を見出し、さらなる実験をデザインしたりコンピュータの中で再現（シミュレーション）して見せるというデータ駆動型アプローチが重要であろう。

(3) オープンサイエンス[3]の潮流

---

[1] 個別化医療 (personalized medicine) とは個人に対して個別に最適な医療を提供することであり、精密医療 (precision medicine) とは、特定の傾向を持つ（例えば特定の遺伝子変異）患者集団に対して最適な医療を提供することである。厳密には両者には違いがあるが、日本においてはあまり区別せず使われている。

[2] 従来行われてきた、特定の基準で集められたヒトの集団を一定期間追跡して疾病の病因を推定する研究に、参加者のゲノム情報やオミックス情報を追加で取得して、DNA 塩基配列等と病因の関係を明らかにする研究。

[3] 誰でも研究データにアクセスできるオープンアクセスや、研究データを公開するオープンデータという概念を包括する大きな概念で、第五期科学技術基本計画（2016 年）にも言及されている。

生命科学・産業におけるこのような動きと並行して、広く学術の分野でオープンサイエンスという動きが生まれつつある。これは主に公的資金を用いて生成された研究データを専門家間にとどめず、研究成果の利用を一般に広く普及させ、イノベーションの創出につなげようとするサイエンスの進め方を意味する。2013 年の G８科学大臣会合をきっかけに、学術の幅広い分野でオープン化が重要であるという概念が急速に広まった［7］。この動きを受けて、データジャーナル[4]の創刊、データアーカイブサイト[5]の設置も相次いでいる。今後、生命科学・産業においても、ビッグデータ化、データ駆動型科学への展開にますます拍車がかかる状況が生まれるが、相応の「データ基盤」が必要なことは言うまでもない。

(4) バイオインフォマティクス分野の発展と変容

バイオインフォマティクスという学問は、ヒトゲノム計画を契機として、大量のゲノム配列データを処理し、その意味付けを行う研究に発展し、さらにその後オミックスと呼ばれる様々な種類のビッグデータを整理・統合・解析し、データに潜む規則性や仮説を導くための研究へと発展してきた。つまり、実験生物学から生み出される大量データを処理するためという従属的、脇役としての学問から、ビッグデータ時代における主役の学問に登りつめてきた。その重要性・必要性は、基礎生物学にとどまらず、医学農学薬学といった応用学問や、それを基盤とした幅広いバイオ産業でも広く認識されてきている。

バイオインフォマティクスの発展にはこれらを支える「データ基盤」が不可欠である。知恵と意欲のある若手研究者が自分のパソコンからビッグデータにアクセスし、世の中を変えるような大きな発見ができる時代がいま現実になりつつある。本分科会が 2014 年９月に表出した報告「大容量情報時代の次世代生物学」［8］において取り上げている、生物学の未解決問題の解決にも、「データ基盤」が本質的な役割を果たすものと期待される。

以上のように、データベース、ネットワークを含むスパコン、人材の３つの柱からなる「データ基盤」は、我が国においても必要性、重要性が認識されてはいる。ただ、多くの施策がビッグデータによる経済効果や短期的なインパクトを目指して立案されるため、分野横断の基盤となる装置や人材等、インフラストラクチャーの整備が相対的に置き去りにされる傾向にある。また、昨今の経済状況は、必ずしもこれらの整備に充分な予算を国が措置する状況にはなく、現状ではむしろ予算を削減する方向にある。

「データ基盤」は持続的に維持発展させてこそ意味がある。にもかかわらず、多くの科学政策や我が国の予算の仕組みはそうはなっていない。欧米の優れたデータベースには、

---

[4] 従来の学術誌（ジャーナル）は、研究データを公共のリポジトリに登録した上で、そのデータから導かれる結論や考察を記載した論文を掲載する。データジャーナルは、データそのものを公共リポジトリに登録する点では論文と変わらないが、研究に使用した試料（生物種や細胞等）やデータの測定手法など、データに関する詳細な説明を掲載する。

[5] 時限プロジェクト等で産出されたデータは、プロジェクト終了後に維持費用が捻出できずデータが消失してしまう可能性がある。これを防ぐために、データを受け入れて中長期的にデータを提供できるようにするためのウェブサイト。

決まって維持管理のための組織がある。それらの歴史を見ると、何十年にもわたって継続的に開発維持されてきたことが一目瞭然である。時限的、競争的、一過的な仕組みだけでは「データ基盤」の整備には不充分なのである。

財政が苦しい状況であるときにこそ、研究資金配分機関は「データ基盤」を整備し、それを持続可能なものにする努力が必要である。「データ基盤」は、コホート研究などの大規模プロジェクトを効率よく進めるという実用面（データ駆動型科学の実現）から重要なだけではない。生命そのものの仕組みを解き明かす面（仮説駆動型科学の推進）でも大変重要である。データの共有化、公共財化を図ることにより、研究の効率化、研究資源の有効活用を実現しなければならない。

2　世界におけるデータベース・データベースセンターの現状

(1) データベース構築の歴史とその意義

生命科学分野では、現在のオープンサイエンスを何十年も前から先取りする形でデータの共有（データベース構築）が行われてきた。文献データについては、現在 MEDLINE あるいは PubMed の名で知られる文献の書誌情報データベースの前身である Medical Literature Analysis and Retrieval System (MEDLARS) が 1960 年代より構築が進められた。研究データについては、アミノ酸配列のデータベースである UniProt の前身が 1960 年代より、タンパク質立体構造のデータベースである PDB (Protein Data Bank) が 1970 年代より、DNA 塩基配列のデータベースである GenBank/EMBL/DDBJ が 1980 年代より構築された。ゲノムについては、後述する米国の NCBI や欧州の EBI といったデータベースセンター及びカリフォルニア大学サンタクルーズ校などがデータとブラウザをあわせて提供している。現在は、生命科学の非常に幅広い分野において様々なデータベースが公開、共有されている。生命科学分野のデータベースの数はおおよそ２万にも及ぶ［9］。

このように古くからデータ共有及びそのためのデータベース構築が行われてきた理由として、以下の要因が挙げられる。

１） データから導かれる知識は少数の数式やルールで記載することが困難なため、データそのものが重要であり、価値を持つ。よって、その保存と共有が大切である。
２） データが取得方法や実験条件など様々な文脈依存性を持ち、研究の再現性や検証には元のデータが必要である。
３） 研究資金配分機関や出版社がデータ共有（データベース登録）を義務付けてきた。
４) その受け皿としてのデータベースセンターが、特に欧米において整備されてきた。
５) 近年では、重複の排除、研究資金・資源の効率化、研究不正への対応の面でもデータ共有が重要視されるようになった。

研究データだけでなく文献やオントロジー（知識や概念の明示的な記述）[6]に関するデータベースも数多く作られている。現在の生命科学はこれら膨大なデータと知識の共有によって推進され、データベースの存在無しには研究が進められない。それは基礎生物学にとどまらず、応用面（医学、薬学、農学、環境学など）でも同様である。データベースは、まさに研究のインフラでありフロンティアであると言えよう。

(2) 多様化するデータベース開発の現状

現在では非常に幅広い分野で様々なタイプのデータベースが開発され、研究のみならず産業界で大いに活用されている。データベースには主に、生データを格納するための一次データベース（リポジトリ）と解析したデータや文献からの知識を格納するための二次データベース（知識ベースや統合データベース）がある。

---

[6] オントロジーとは、知識や概念を表現する際に、それらの関係性までをも含めて、（コンピュータが扱えるような形で）整理・体系化したもの。オントロジーを利用することで、コンピュータを用いた推論が可能となる。例えば、”果物”を”リンゴ”の上位概念、”食物”の下位概念とすることで、コンピュータはリンゴが食物であることを認識できる。

国際的な一次データベースとしては、DNA 塩基配列データを格納する International Nucleotide Sequence Database Collaboration (INSDC：前述の GenBank/EMBL/DDBJ を含む枠組み)、タンパク質立体構造データの Worldwide Protein Data Bank (wwPDB) がある。二次データベースとしては、ヒトの変異情報を収集した ClinGen、生物のシステム情報をまとめた Kyoto Encyclopedia of Genes and Genomes (KEGG) などがある。さらに近年では、個人情報保護の観点からアクセス制限が必要な個人ゲノム情報のためのデータベースとして、Database of Genotypes and Phenotypes (dbGaP)、European Genome-phenome Archive (EGA)、Japanese Genotype-phenotype Archive (JGA)なども開発されている。

主要な国内外のデータベースについて参考資料２に挙げたが、これ以外にも多種多様なデータベースが開発されている。その数は我が国だけでも 1,000 を超え、世界では２万にも及ぶと考えられている［9］。また、これらのデータの容量は既に数ペタから数十ペタバイトに達しているが、現在もムーアの法則[7]を上回る勢いでデータが生産されており［10］、今後ますますデータを格納するためのディスク容量が必要になる。

さらに論文に関しても、前述の PubMed に登録されている論文数だけで 2,800 万件を超え、毎年出版される論文数も増加の一途をたどっている。PubMed は書誌情報のデータベースであるが、PubMed Central と呼ばれるセクションで論文全体（フルペーパー）の無償公開も進めている。オープンアクセスジャーナル[8]や、bioRxiv のようなプレプリントサーバ[9]の普及もあり、現在 500 万件を超える数の論文が無料で読めるようになっている。

また、論文の根拠となるデータをデータベースに登録するこれまでの動きに加え（前述の一次データベースの活用や出版社によるアーカイブサイトの設置）、論文には直接は結びつかないが、重要と思われるデータを論文とは別に出版することを目的としたデータジャーナルも登場し、データを共有する動きはますます活発化してきている。

このように多種多様なデータベースが生み出されている要因として、生命科学分野の多様性が挙げられる。それぞれの研究分野、生物種、分子種、測定手法などに応じて個別のデータベースが必要である。また、歴史的にそれぞれの分野が独自の方法でデータベースを開発してきた経緯もあり、用語やフォーマットの不整合などにより統合化が進んでいないことも多様性を生み出している一因であろう。

(3) 国内外のデータベースセンターの状況

米国では National Institutes of Health (NIH) 傘下の National Center for Biotechnology Information (NCBI) が、欧州では European Molecular Biology Laboratory (EMBL) 傘下の European Bioinformatics Institute (EBI) が中心となってデータベースの整備を進めている（表１）。この二つのセンターは後述する日本の生命情報・DDBJ センターと連携して

---

[7] 米インテル社のゴードン・ムーアによって提唱された、18 ヶ月でコンピュータの性能が２倍になるとされる経験則。

[8] 掲載された論文を、誰でも無料で全文を読むことができる学術誌。従来のように読む側が購読料を支払う形ではなく、主に論文の著者が掲載料を支払う形で運営されている。

[9] 論文を学術誌に投稿して査読を受ける前に、論文の原稿をアップロードするためのサーバ。研究成果をいち早く共有する目的で利用され、誰でも無料で全文を読むことができる。

DNA 塩基配列データベースの構築を進めている。中国では中国科学院 Beijing Institute of Genomics (BIG)が、これらの国際的な動向とは独立に、ゲノムデータなどの収集を始めている。

日本では科学技術振興機構（JST）のバイオサイエンスデータベースセンター（NBDC）、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）、国立遺伝学研究所の生命情報・DDBJ センター、大阪大学蛋白質研究所の Protein Data Bank of Japan (PDBj)などが連携してデータベースの整備を進めている。この中で、NBDC と DBCLS は、データの共有、統合によりデータの価値を最大化することを目的とした「統合データベースプロジェクト」の遂行を目的に設立された組織である。これらについては参考資料３に設立経緯を示しておく。他のデータベースセンターの設置やその目的などについては、参考文献の URL を参照されたい［11，12］。

(4) データベース開発とデータベースセンターの意義

データベースの整備とその利活用は、生命科学やバイオ産業を推進するために不可欠である。別の言い方をすれば、データベース作りは研究そのものであり、これからのバイオ産業の根幹をなす。日々新たな計測技術が生まれ、それに伴い日々新たなデータが生み出されている状況を踏まえれば、これまでのデータを整理統合して保管しておくだけでなく、新たなデータと過去のデータとを統合し、データの意味付け、注釈付けを最新のものに更新することが欠かせない。つまり、データベースを長期的永続的に構築・更新・維持していくことが、何よりも欠かせない。

このことから、最も重要な研究資源であるデータベースの整備を欧米に頼るわけにはいかない。データベースの構築・維持・管理を放棄すれば、これからのオープンサイエンス化の流れを考えても、我が国で産出したデータが海外へ一方的に流出するだけである。国内においてこれらデータを最大限に活用するためには、データ及びデータベースをハンドリングする技術・人材を自前で持たないと、我が国の生命科学・産業そのものが成り立たない。逆に、データベースを整備し最大限に活用することができれば、多くの関係者（データ生産者、データ利用者、公的研究機関、民間企業、など）に大きな利益をもたらすことは明らかであろう。基礎研究はもちろん、医療・創薬から育種農業、有用物質生産にいたるまで多くの分野に利益をもたらす。また、データベースを整備し、俯瞰することで初めて、どの分野の、どの種類のデータが不足しているかが明らかとなり、今後のプロジェクトの方向性が見えてくる。つまりデータベースは政策立案にとっても大きな意味を持つ。

しかしながら、生命科学のすべてのデータベースを我が国だけで整備することは予算的にも人的にも不可能である。そこで、整備すべきデータベースを戦略的に絞り込む必要がある。例えば、我が国において整備すべきデータベースとして、日本固有のもの、日本の強みをいかすものが考えられる。これらの例として、日本人ゲノムのデータベースや日本の気候・土地環境・害虫環境などに適した産業上有用な植物や微生物のゲノムやオミッ

クスのデータベースが挙げられる。海外との協調と競争を踏まえた国際連携も必要で、その観点からも日本からの貢献は不可欠である。

表１　各国の主要なデータベースセンターとそれらの比較

センター (国・地域)	役割	主な データベース	予算※	人員※ (人)
NCBI （米国）	NIH 傘下である National Library of Medicine (NLM)の付属機関。NLM は生物医学情報の集積をミッションとし、その中で NCBI は分子生物学に関するデータ及び生物医学文献に特化した部門。	Genome, GenBank, GEO, dbGaP, SRA, OMIM, PubChem, Pubmed	約 181.2 億円 1 億 7580 万ドル	287
EBI （欧州）	EMBL 傘下の非営利学術機関。バイオインフォマティクスの研究とサービスの中心機関。	Ensembl, ENA, EGA, ArrayExpress, ChEMBL, UniProt	約 77.1 億円 6,720 万ユーロ	513
NBDC /DBCLS (日本)	NBDC はライフサイエンス分野のデータベース統合を目的として発足。DBCLS は文部科学省ライフサイエンス分野の統合データベース整備事業の中核機関として設立。 両センターはライフサイエンスデータベース統合推進事業において共同研究開発を実施。	NBDC ヒトデータベース, NBDC RDF ポータル	約 14.1 億円	64
DDBJ (日本)	国立遺伝学研究所の付属施設。「生命情報学」の我が国における研究拠点。 INSDC を運営。	JGA, DDBJ	約 14.6 億円	39

※予算、人員は平成 27（2015）年度時点

出典：NBDC、「ライフサイエンスデータベース統合推進事業　事業報告書」［13］

３　我が国における課題

これまで述べてきたように、生命科学のデータと知識は爆発的に増えている。それらは単に量的に増えるだけでなく、種類の多様性も増している。このような膨大で多種多様なデータや知識を格納し、活用するための「データ基盤」をどのように構築、更新、維持していくかは世界的にも難しい課題である。

「データ基盤」整備の問題は、多額の公的資金を投じて得られた成果を社会に還元する、説明責任（アカウンタビリティ）を果たす、研究の重複を排除し税金を効率よく使う、と言った観点からも重要である。先に述べたように、生命科学においてはデータそのものが成果であるという側面があり、これを公共財化し維持することは、研究の重複を防ぐ意味からも重要である。また、研究成果（データ）が死蔵される、消失する、充分に活用されない場合は我が国にとって大きな損失になる。

「データ基盤」の整備に関する課題、特に我が国における問題として以下が挙げられる。

(1)　データ共有政策の欠如

欧米においては、研究資金配分機関が、強力なデータ共有政策を設けている。また、研究予算の申請時にデータ管理計画 [10] の提出が義務付けられており、プロジェクトで取得するデータの取扱いについても審査を受けるのが一般的である。そして、ヒトゲノムのような機微情報データに関するセキュリティも堅固である。サーバ上のデータは Advanced Encryption Standard(AES)暗号化され、建物内には警備員が配置され、敷地には入構規制がある。

一方、日本では、内閣府などを中心にオープンサイエンスへの機運が高まっている［7］が、データ共有政策の制定、義務化すら不充分である。日本医療研究開発機構（AMED）のデータ共有政策［14］など、政策を掲げる機関は出てきているものの、研究データの公開・共有状況が一元管理されておらず、再利用や検証ができない。データ管理計画の提出義務化も一部の研究資金配分機関で始まったばかり（JST が 2017 年から［15］、経済産業省、AMED が 2018 年から［16，17］）で、データ共有に対する意識が研究資金配分機関や研究者間に浸透しているとは言い難い。

また、ヒトゲノムのような機微情報のデータベースによる公開と共有に関しても、研究プロジェクト間共通のルール制定が進んでいない。NBDC は公的資金を用いて産生されたヒトに関するデータ一般に適用することを目的として、ヒトデータ共有ガイドライン［18］を作成した。しかし、本ガイドラインは強制力を持たず、現状では、各研究プロジェクトが独自に指針を策定し、運用している。結果として、どのようなデータが取得され、どこまで公開され、セキュリティの状態すら分からぬ形で個別の研究が進んでいる。また、データベースによる公開と共有にあたっては試料提供者からの同意が不可欠であるが、これに関する試料提供者への説明方法・文章について共通の形式がない。このため、例え

[10] 研究の進行にともなって生産されるデータをどのように取り扱い、いつ、どこで、どのような方法でデータを共有するのかについての計画。一般に、欧米では予算申請時のデータ管理計画提出が義務化されている。

ば、研究実施者が情報の公開と共有を行おうとした際、試料提供者に充分な説明を行っていないことが判明し、資料提供者への再説明が必要になることも少なくない。データセキュリティに関しては改正個人情報保護法により暗号化が推奨されるものの、暗号化方法や建物の物理的セキュリティに関しては各機関が試行錯誤する状態である。

(2) データベース戦略の欠如

DBCLS や NBDC の設立以前、我が国におけるデータベース構築は、研究プロジェクトごとに独自に行われ、データベースの開発やプロジェクト終了後の維持・管理に関する明確な戦略は存在しなかった。DBCLS や NBDC の設立によって、我が国として研究分野や研究プロジェクトを俯瞰したデータベース整備戦略や、プロジェクト終了後の維持、統合や活用の促進といった問題は改善の兆しがある。しかし先に述べたように、データ産出プロジェクトがどうあるべきかを議論する枠組み、産出されたデータを一元的に管理するデータベースセンター、既存のデータを最大限活用して新たな知識発見につなげる戦略的な研究事業のいずれも存在しない。このように、データの生産から利活用まで、多くの課題を抱えている。

(3) データベースセンター及びスパコン連携の欠如

DBCLS や NBDC 等、国内のデータベースセンターは、欧米のそれらと比べ、予算的にも人員的にも圧倒的に規模が小さい。多岐にわたる生命科学やバイオ産業のニーズに充分には対応できず、整備できるデータベースやデータが限られているのが現状である。また、法人化の影響でセンターの運用方針が異なっているため、非効率でもある。

データベースから安全に高品質のデータが社会へ発信されるためには、データの品質管理が一定のルールの下でなされ、利用者が個々のデータの品質を確認できる仕組みが必要である。またデータの保管（アーカイビング）も常に安全に運用される必要がある。品質管理には、専門分野の知見を持つキュレータ（データや知識の編集者）[11]が、それぞれの分野の研究者集団の支援を受けつつ作業をする必要がある。アーカイビングには、情報科学・情報工学の技術者がデータのバックアップ及び最新のハードウェアや基盤ソフトウェアのアップデートを常に実施する必要がある。それには、データベースセンター間の連携による効率化も必要となる。

生命科学ビッグデータを活用するには、スパコンが必須である。我が国には、生命科学の用に供する目的で、国立遺伝学研究所、東京大学医科学研究所、東北大学東北メディカル・メガバンク機構など、大量データの近くにスパコンが導入されている。これらのスパコンは、単に膨大な解析量に対処する計算機資源の面からだけでなく、データベースセン

[11] キュレータは、目的に応じて必要な情報を取捨選択し、提示する専門家。文献から必要な知識を取り出してデータベースを構築するような仕事に従事する。分野横断的な知識や取り扱うデータへの深い理解が必要となり、高度な知的作業が要求される。

ターからのデータ転送時間を短縮し、ヒト個人ゲノムなど機密性の高いデータ等を安全に解析するためにも必須である。

しかし、上記のスパコンの運営においては、データの伸び、解析の需要に追いつくため、個別の予算確保に苦労しているのが現状である。例を挙げると、国立遺伝学研究所はストレージ 43.5 ペタバイト(PB)（高速ディスク 13.5 PB、アーカイブ用ディスク 15 PB、テープ 15PB）、総コア数 1 万、総メモリ 90 TB というスパコンを 2019 年に導入したが、登録するログインユーザー数は 900 を超えており、データの格納面でも、解析する計算パワーの面でも、需要に応えられる見通しはない［19］。

(4) 人材育成と教育体制の不備

バイオインフォマティクス人材の不足、人材育成の必要性は、これまでも叫ばれてきた［20, 21］が改善されていない。これまで約 20 年に渡り、人材不足解消のための試み（例えば、平成 13 年度からの科学技術振興調整費による新興分野人材育成など）がなされてきたが、需要の伸びがそれを上回っている。上に述べたように、データベースの構築や解析の必要性は基礎だけでなく、生命科学及びバイオ産業全般に及んできており（例えば、病院などでゲノム情報解析の人材が必要になってきたなど）、人材の需要は拡大の一途を辿っている。

バイオインフォマティクスの人材と一口に言っても、世界最先端のアルゴリズムを開発する研究者、データベースの構築をするためのキュレータ、データベースの管理をするシステムエンジニア、データベースを使って実験研究者の解析を支援する人材、など多岐にわたる。それぞれ必要とされる人材によって教育内容も異なる。人材育成の問題は、キャリアパス、評価の方法、大学教育のあり方まで含めて検討せねばならない。

(5) 予算の不足、縮小

我が国におけるバイオデータ、バイオインフォマティクス分野への予算配分は、科学技術振興機構のバイオインフォマティクス推進センター事業（平成 13 年から平成 24 年にかけて実施）における予算額に比べても、減少の一途を辿っている。その一例として、ライフサイエンスデータベース統合推進事業とその前身となったライフサイエンス分野の統合データベース整備事業（統合データベースプロジェクト）における予算の推移を表２に示す。表中の予算額は文部科学省管轄のものだけの推移を示したものであるが、統合データベースプロジェクト（参考資料３）［22］は、内閣府総合科学技術会議（当時）の推進のもと、農林水産省、厚生労働省、経済産業省がそれぞれ予算を確保していた。それが文部科学省以外は継続されなかったことを考えると予算の減少幅が大幅であることがわかる。このままではデータ量の伸びやデータ種類の増加に対応できず、国の予算で産出された重要なデータが充分に保全、整理、利用されないどころか、雲散霧消してしまう。

表２　バイオサイエンスデータベースセンター（NBDC）関連予算の推移

年度	事業	年額予算
平成 19	文部科学省ライフサイエンス分野の統合データベース整備事業（統合データベースプロジェクト） ＋バイオインフォマティクス推進センター事業	約 33 億円
平成 23	ライフサイエンスデータベース統合推進事業※	約 17 億円
平成 28	ライフサイエンスデータベース統合推進事業	約 14 億円

※平成 23 年度に文部科学省 統合データベースプロジェクトとバイオインフォマティクス推進センター事業を融合する形でライフサイエンスデータベース統合推進事業が開始し、合わせて発足した NBDC が事業実施することとなった。

出典：NBDC、「ライフサイエンスデータベース統合推進事業　事業報告書」［13］

予算の縮小問題は我が国に限った話ではない。欧米でもデータベース予算を減らす動きがある。 米国政府の資金配分機関もデータベースのプロジェクトが終了すれば支援は打ち切っている。モデル生物別のデータベースに対しても 2020 年までに新たな財源を確保するよう要請している［23］。しかしながら、表 1 に示したように、欧米とは元々の予算に大きな違いがあり、一概には論じられない。

４　提言：持続可能な「データ基盤」整備のあり方

これまで、爆発的に増えるデータの視点から、生命科学・産業を取り巻く状況、「データ基盤」整備の現状と問題点を述べてきた。我が国においては、「データ基盤」の整備及びその持続可能性の確保において、多くの問題がある。現時点で世界に大きな遅れを取っており、今後その差がますます開く可能性が高い。このような危機感を研究資金配分機関や研究者の間で共有することが本提言の目的の一つである。ここでは、３章で述べた五つの課題それぞれについて、どのような解消策、推進策を取るべきか、その方向性や検討項目について以下のとおり提言する。

(1)　データ共有政策の作成と義務化

国が省庁横断的に適用できるデータ共有政策を作成し、各研究資金配分機関は政策に基づいてデータ共有の環境整備と義務付けを行うべきである。さらに各研究資金配分機関は、研究プロジェクトの申請者にデータ管理計画の提出を義務付け、それの遂行状況を（中間）評価の際に用いるべきである。そして、研究データの生産者に対して、データ共有に対するインセンティブ付与の仕組みを導入すべきである。

インセンティブの例として、高品質なデータを提供した研究者に研究課題や研究者個人の業績の評価に加点する仕組みが考えられる。適切な評価のために、データやデータベースの利用に対して適切な引用や謝辞を義務付けるのも一つの方法である。また、データ共有に関する費用（例えば、データを再利用可能に整理する経費）や、データ登録・公開に関する費用を研究資金配分機関が負担することも考えられる。

データ共有政策には、データやデータベースのライセンスのあり方に関しても、方針が提示されていることが望ましい。なお、データ共有の義務化に際しては、ヒトゲノムデータなどの機微情報について、個人情報の保護や倫理面、セキュリティにも充分配慮したルール作りが必要である。

データ共有の範囲は我が国だけではない。オープンサイエンスの潮流は国境を超える話であり、国内のルール作りに際しては国際的なデータ共有、オープンサイエンスの視点が欠かせない。また、データ共有の促進、データの公共財化は、社会全体にとっても重要である。

(2)　プロジェクト立案時からのデータベース戦略策定

これまでは先に研究プロジェクトの立案実行があり、それに伴って出てくるデータを後付けで（プロジェクト終了後に）、共有したり使いやすく統合したりされてきた。しかしながら、公的研究資金による研究成果を収集・保全するだけでは、データ駆動型科学を推進するには不充分である。生命現象は多様であり、それを担う分子ネットワークは複雑であり、それらの反映として、データには文脈依存性や曖昧性がある。個々の研究プロジェクトから出てくるデータを寄せ集めただけでは、データの突き合わせが出来なかったり、探索空間が広すぎたりして、イノベーションに繋がるような規則性は見いだせない。

生命科学の発展のためには、必要なデータは何かをまず俯瞰的に検討し、それに従いデータを産出・整備する戦略が必要である。そして、共有・公開・統合に適したデータが産出されるよう、プロジェクト立案時からデータ整備の方法に関して、データ産出プロジェクトとデータベースセンターが密に連携すべきである。その意味において、データベースセンターはプロジェクト立案時、データ生産時、データ解析・解釈時を通じて、すなわちデータの公開前の段階から、研究を支援する体制を構築すべきである。

データ産出拠点を立案するに際しては、新技術に基づくデータ産出や人工知能技術との連携、産業応用への貢献が充分に考慮される必要がある。生命科学においては日々新たな計測技術、観測技術が生まれ、新たな種類のデータがもたらされている。現在、バイオイメージングやメタゲノム解析、大規模ゲノム解析などの新技術に基づくデータ生産拠点が構想されている［24］。こうしたデータ及び知識は、データ解釈の多様性、曖昧性、文脈依存性、などの問題を抱えている。そのため、人工知能研究者など生命科学と異なる分野の研究者と連携して、データベース統合、オントロジー整備などをより一層進める必要があろう。これは、後述するデータサイエンティスト育成にもつながる事業である。

(3) データベースセンターの一元化とスーパーコンピュータの整備

複数ある国内のデータベース関連機関を統合し、国内には政策立案・研究開発ともに強力なリーダーシップを発揮でき、国際的には存在感と競争力を打ち出せる体制を整えるべきである。残念ながら、我が国では依然として AMED を含め省庁や組織の壁があり、データベース、データベースセンター、人材育成の統合が難しい。これら縦割りの解消が不可欠である。更に、各省の公的資金でそれぞれにプロジェクトの立案とデータ産出を行うのは非効率である。上記（2）に述べたように、データベースセンターがデータ産出プロジェクトを束ね、生命科学に必要なデータが産出・整備・利活用されるようにすべきである。

２章で述べたように、現在では計算機の性能の伸びよりもデータ生産の伸びが大きく、研究者が扱うデータ量も増加している。その結果、これまでのように研究室内の小規模なサーバではデータを処理できない。このため、巨大データの解析に適した高速のネットワークと、増大する大規模データに対応できるストレージ機能を備えた専用スパコンを増強すべきである。

これらの専用スパコンにはデータ保存と利活用の戦略が必要で、それを検討するのはデータベースセンターである。また、この問題を解決する技術（データ圧縮、分散処理、など）の開発も望まれる。スパコンを研究のインフラとして安定的に運用するためには、大規模な解析を行う利用者に対する課金も必要であろう。

さらに、スパコンどうしの連携やクラウドの活用も効率化のために検討すべきであるが、これらに対しては、ヒトゲノムなどの機微情報をどう安全に管理するか、我が国の個人情報保護法に配慮しながら進める必要がある。生命科学の研究には特有のデータベースや解析ソフトの導入・管理も必要なため、スパコンどうしの連携やクラウドの活用には充分な配慮が必要である。

(4) 人材育成と教育体制の整備

バイオインフォマティクス分野の人材不足解消には、高校教育のあり方、大学入試のあり方、大学（学科新設など）のあり方、民間企業のあり方、研究プロジェクトの立て方などの見直しの議論が必要である。また、バイオインフォマティクスという生命科学分野に限らず、より広くデータサイエンティストを育成すべく、短期的ではない中長期的な人材育成の整備が必要である。以下に人材育成に必要な検討項目を列挙する。

- バイオインフォマティクス人材を必要とする（あるいは将来的に必要とする）研究分野や産業を明確にし、それに必要な人材像と人数を見積もること。それに基づき養成のロードマップを明らかにすること。
- 養成すべき人材の具体的なイメージ（研究者、研究支援者、キュレータ、システムエンジニアなど）を明らかにし、それらの分類をするとともに、レベルに応じて身に付けるべきスキルを明らかにすること。［25］
- 上記の分類に従って、それぞれのカリキュラムを作成し、また、そのための教科書や教材を揃えること。
- 上記のカリキュラムに従って、それらを教えるべき時期や教育体制（高校生レベル、学部レベル、大学院レベル、社会人再教育）のあるべき姿を明確化にし、その整備を図ること。
- 個々の機関での取組の充実だけでなく組織間連携（例えば、大学、理化学研究所、JST、産業技術総合研究所など）による教育システムの構築を図ること。
- 研究支援者やキュレータなどの評価やインセンティブ付与の仕組みを設計すること。また、それに基づいた研究機関や民間企業でのキャリアパスを明確にし、ポジションを確保すること。
- サイエンスカフェなどの活用を図り、この分野の魅力を広く発信すること。

(5) 予算の確保、データ量やデータ種類の増加に対応した仕組みの導入

将来にわたりデータを整備・活用していくためには、新たな財源モデルの構築が必要である。国際的に様々な財源モデルが検討されているが［26，27，28］、我が国のオープンサイエンスへの取組状況や公的研究資金制度なども考慮に入れつつ、新たな財源を導入していく時期にきている。財源モデルの検討に当たっては、長期安定性だけでなく、オープンサイエンスへの貢献や、費用負担の公平性にも留意する必要がある。また、一次データベース（レポジトリ）と二次データベース（知識ベースや統合データベース）、それらのデータを扱うスパコンが必要であり、それぞれの事情に応じた基盤整備が必要である。特に、ジャーナル等でデータ登録が義務付けられているような一次データベースについては、データ保全の観点から特に持続性が重要であり、安定運営への国際的な責任や公平な負担という観点から、運営機関に配分される競争的資金や運営費交付金ではない、安定

的な財源によって運営されるべきである。逆にスパコンの場合は受益者負担が原則となり、課金システムが妥当だろう。

長期安定的な財源モデルの一つとして、国のライフサイエンス研究分野の公的資金のうち一定割合を措置するという考え方があり（インフラストラクチャーモデル）、長期安定性のほかオープンサイエンスへの貢献や費用負担の公平性にも優れるとされている［26］。一方で、費用負担の按分には一定割合を設定することが容易ではなく、一度割合を設定すると変更が難しいことから、導入には慎重であるべきとの考え方もある［28］。データ量に対応させるという観点からは、データ産出者からデータ登録料を徴収するという方法（データ登録料モデル）も考えられるが、安易な導入はデータの公開・共有を阻害しかねない。本章の提言**(1)**で述べたようなデータ公開・共有を義務付けるデータ共有政策を適用した上で、個々のデータ産出者ではなく研究資金配分機関が一括で契約する等の工夫が必要である。

二次データベースについてはモデル生物別やオミックス別など多数のデータベースが国内外で構築されており、研究開発要素が強く、国際競争の観点から戦略的な研究開発支援が必要である。日本でも KEGG のようにオープンアクセスを基本としながら一部を有料化するデータベースも出てきており、国の支援と自己収入の共存を前提としたデータベース構築・運用支援の仕組みが必要である。

5　おわりに

生命科学・ バイオ産業におけるビッグデータの出現、オープンサイエンス化の流れ、これらがもたらす研究開発環境及び研究開発アプローチの変化について概観したうえで、このような新たな潮流に対応するためには、持続的な「データ基盤」（データベース、スパコン、人材）の整備が不可欠であることを述べてきた。また、欧米と比較し、我が国の「データ基盤」は整備が遅れていること、今後ますますその整備における格差が生じる可能性があり、これらが我が国の生命科学・バイオ産業の国際競争力に大きな障害、足枷となることを論じた。

本提言を契機に、上記のような危機意識が研究資金配分機関や研究者の間で共有され、我が国において中長期的な視点に立った骨太の「データ基盤」整備戦略が立てられることを切に望む。

＜参考文献＞

［1］ National Human Genome Research Institute (NHGRI) 、「DNA Sequencing Costs: Data」、2018 年 4 月 25 日 https://www.genome.gov/27541954/dna-sequencing-costs-data/

［2］ 産業競争力懇談会、「デジタルを融合したバイオ産業戦略」、2018 年 2 月 21 日 http://www.cocn.jp/theme104-L.pdf

［3］ 総務省情報通信審議会情報通信技術分科会技術戦略委員会、「次世代人工知能推進戦略」 http://www.soumu.go.jp/main_content/000424360.pdf

［4］ 日本学術会議第二部ゲノムコホート研究体制検討分科会、提言「100 万人ゲノムコホート研究の実施に向けて」、2013 年 7 月 26 日 http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-22-t176-1.pdf

［5］ 東北メディカル・メガバンク機構 http://www.megabank.tohoku.ac.jp/result

［6］ 内閣府総合科学技術・イノベーション会議バイオ戦略検討ワーキンググループ http://www8.cao.go.jp/cstp/tyousakai/bio/

［7］ 内閣府総合科学技術・イノベーション会議国際的動向を踏まえたオープンサイエンスに関する検討会、「我が国におけるオープンサイエンス推進のあり方について～サイエンスの新たな飛躍の時代の幕開け」、2015 年 3 月 http://www8.cao.go.jp/cstp/sonota/openscience/

［8］ 日本学術会議基礎生物学委員会・統合生物学委員会・農学委員会・基礎医学委員会・薬学委員会・情報学委員会合同 バイオインフォマティクス分科会、報告「大容量情報時代の次世代生物学」、2014 年 9 月 17 日 http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-22-h140917-1.pdf

［9］ 高木 利久、「研究リソースとしてのバイオデータとその活用」、実験医学、34、79-83(2016)

［10］C.E. Cook et. al, “The European Bioinformatics Institute in 2017: data coordination and integration” Nucleic Acids Res., 46, D21-D29 (2018) doi: 10.1093/nar/gkx1154

［11］国立遺伝学研究所 DDBJ センター https://www.ddbj.nig.ac.jp/index.html

［12］大阪大学蛋白質研究所 Protein Data Bank of Japan (PDBj) https://pdbj.org/

［13］国立研究開発法人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、「ライフサイエンスデータベース統合推進事業 事業報告書」、2016 年 11 月 https://biosciencedbc.jp/gadget/unei/jigyou_houkoku.pdf

［14］国立研究開発法人日本医療研究開発機構、「疾病克服に向けたゲノム医療実現化プロジェクト ゲノム医療実現のためのデータシェアリングポリシー」、2016 年 4 月 https://www.amed.go.jp/content/000004858.pdf

［15］ 国立研究開発法人科学技術振興機構、「オープンサイエンス促進に向けた研究成果の取扱いに関する JST の基本方針」、2017 年 4 月 https://www.jst.go.jp/pr/intro/openscience/policy_openscience.pdf

［16］経済産業省、「委託研究開発におけるデータマネジメントに関する運用ガイドライン」、2018 年 4 月
http://www.meti.go.jp/policy/innovation_policy/data_manegement.html
［17］国立研究開発法人日本医療研究開発機構、「データマネジメントプランの提出の義務化について」、2018 年 3 月
https://www.amed.go.jp/koubo/datamanagement.html
［18］国立研究開発法人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 ヒトデータ共有ガイドライン ver. 3.0」、2016 年 2 月
https://humandbs.biosciencedbc.jp/guidelines/data-sharing-guidelines
［19］情報・システム研究機構国立遺伝学研究所 DNA データ利用委員会、「DDBJ 事業報告 2016」、2017 年 2 月 27 日
https://www.ddbj.nig.ac.jp/activities/annualreport.html
［20］J. Chang, “Core services: Reward bioinformaticians”, Nature, 520, 151-152 (2015) doi:10.1038/520151a
［21］国立研究開発法人科学技術振興機構、「研究開発の俯瞰報告書　ライフサイエンス・臨床医学分野（2017 年）」、2017 年 6 月
https://www.jst.go.jp/crds/pdf/2016/FR/CRDS-FY2016-FR-06.pdf
［22］白木澤佳子、高木利久、「ライフサイエンス分野のデータベース統合を目指してバイオサイエンスデータベースセンター（NBDC）の発足」、情報管理、54(3)、144-151(2011)
https://doi.org/10.1241/johokanri.54.144
［23］J. Kaiser, “Funding for key data resources in jeopardy”, Science, 351, 14 (2016) DOI: 10.1126/science.351.6268.14
［24］日本学術会議 基礎生物学委員会・統合生物学委員会合同 生物物理学分科会、提言「生命科学の発展を加速する次世代統合バイオイメージング科学の研究推進」、2017 年 9 月 20 日
http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-23-t250-5.pdf
［25］国立研究開発法人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、「バイオインフォマティクス人材に関するアンケート調査」、2013 年 6 月
https://biosciencedbc.jp/gadget/chousa/Questionnaire.pdf
［26］C. Ember and R. Hanisch, “Sustaining Domain Repositories for Digital Data: A White Paper” (2013)
DOIs: 10.3886/SustainingDomainRepositoriesDigitalData
［27］W. Anderson et. al, “Towards coordinated international support of core data resources for the life sciences” (2017) doi: https://doi.org/10.1101/110825
［28］OECD, "Business models for sustainable research data repositories", OECD Science, Technology and Industry Policy Papers, No. 47, OECD Publishing, Paris, (2017) http://dx.doi.org/10.1787/302b12bb-en

＜参考資料１＞分科会審議経過

第 23 期

平成 27 年８月６日	第１回分科会 活動方針について
平成 28 年 11 月 15 日	第２回分科会 活動計画について
平成 29 年３月 28 日	第３回分科会 提言の骨子案について
平成 29 年９月 26 日	第４回分科会 提言の骨子の改定案について

第 24 期

平成 30 年４月 24 日	第１回分科会 提言の作成について
令和元年　６月 27 日	日本学術会議幹事会（第２７９回） 提言「持続可能な生命科学のデータ基盤の整備に向けて」について承認

＜参考資料２＞国内外の主要な生命科学データベース

カテゴリ	タイプ	データベース
DNA 塩基配列	一次	Sequence Read Archive (SRA), European Nucleotide Archive (ENA), **DDBJ Read Archive (DRA)**, Genome Sequence Archive (GSA)
遺伝子	一次	*International Nucleotide Sequence Database Collaboration (INSDC)* [GenBank, European Nucleotide Archive (ENA), **DNA Data Bank of Japan (DDBJ)**]
タンパク質立体構造	一次	*Worldwide Protein Data Bank (wwPDB)* [Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), Protein Data Bank in Europe (PDBe), **Protein Data Bank Japan (PDBj)**, Biological Magnetic Resonance Data Bank (BMRB)]
遺伝子発現	一次	Gene Expression Omnibus (GEO), ArrayExpress, **Genomic Expression Archive (GEA)**
プロテオーム	一次	*ProteomeXchange* [PRIDE, PASSEL, MassIVE, **jPOSTrepo**, iProX, Panorama Public]
メタボローム	一次	*MetabolomeXchange* [MetaboLights, Metabolomic Repository Bordeaux, Metabolomics Workbench, **Metabolonote**]
化合物	一次	PubChem, ChEMBL, ChemSpider, 日化辞
文献	一次	PubMed, PubMed Central (PMC), Europe PMC
ヒトデータ（制限公開）	一次	Database of Genotypes and Phenotypes (dbGaP), European Genome-phenome Archive (EGA), **Genotype-phenotype Archive (JGA)**、Genomic Data Commons (GDC)
ヒト変異	一次・二次	dbSNP, European Variation Archive (EVA) ClinVar, ClinGen, **TogoVar**, Exome Aggregation Consortium (ExAC), Catalogue of Somatic Mutations in Cancer (COSMIC)
ゲノム情報	二次	Ensembl, UCSC Genome Browser, **Functional Annotation of the Mammalian Genome (FANTOM)**

エピゲノム	二次	*International Human Epigenome Consortium (IHEC)* [Encyclopedia of DNA Elements at UCSC (ENCODE), Blueprint, **CREST/IHEC**], **ChIP-Atlas**
マルチオミクス	二次	**Database of Kashiwa Encyclopedia for human genome mutations in Regulatory regions and their Omics contexts (DBKERO)**
遺伝病	二次	Online Mendelian Inheritance in Man (OMIM)
タンパク質アノテーション	二次	UniProtKB, neXtprot, Pfam, Integrated resource of protein families, domains and functional sites (InterPro)
タンパク質相互作用	二次	IntAct, Molecular Interaction Database (MINT), Database of Interacting Proteins (DIP)
酵素反応	二次	Rhea, BRENDA, The Carbohydrate-Active enZYmes Database (CAZy)
パスウェイ	二次	**KEGG PATHWAY**, Reactome, BioCyc, WikiPathways
システム情報	二次	**Kyoto Encyclopedia of Genes and Genomes (KEGG)**
生命動態情報	二次	**Systems Science of Biological Dynamics (SSBD)**
生物種別	二次	Mouse Genome Database (MGD), Rat Genome Database (RGD), Zebrafish Information Network (ZFIN), WormBase, FlyBase, The Arabidopsis Information Resource (TAIR), **Plant Genome Database Japan (PGDBj),** Saccharomyces Genome Database (SGD), **Cyanobase, MicrobeDB.jp**
糖鎖情報	一次・二次	**GlyTouCan, Japan Consortium for Glycobiology and Glycotechnology DataBase (JCGGDB)**, MonosaccharideDB, GlycomeDB

国際的な協力体制で構築されているデータベースは、コンソーシアム名を斜体で記述し、加盟しているデータベース名を [ ] 内に記述した。日本で開発されたデータベースは**太字**で示した。

＜参考資料３＞統合データベースプロジェクトの沿革

平成 12 年 11 月	科学技術会議 ライフサイエンス部会 ゲノム科学委員会「ゲノム情報科学におけるわが国の戦略について」 ※人材養成、データベース構築、情報解析技術開発の３つの観点から推進戦略を提案
平成 13 年４月	科学技術振興機構（JST）にバイオインフォマティクス推進センター（BIRD）を設置
平成 17 年８月	科学技術・学術審議会　研究計画・評価分科会　ライフサイエンス委員会　データベース整備戦略作業部会「我が国におけるライフサイエンス分野のデータベース整備戦略のあり方について」 ※戦略委員会の設置、ポータルサイトの構築、統合データベースのための技術開発、人材養成を緊急に取り組むべき課題として提言
平成 18 年４月	農林水産省、経済産業省で統合データベースのプロジェクトが開始
平成 18 年９月	情報・システム研究機構（ROIS）を中核機関とした「ライフサイエンス分野の統合データベース整備事業（統合データベースプロジェクト）」が開始
平成 20 年 12 月	科学技術・学術審議会　研究計画・評価分科会　ライフサイエンス委員会ライフサイエンス情報基盤整備作業部会「ライフサイエンスデータベースの統合・維持・運用の在り方」 ※DBCLS と BIRD の一体的な運用を、JST に設置する新たな組織で行うことを提言
平成 21 年５月	総合科学技術会議　ライフサイエンス PT　統合データベース　タスクフォース「統合データベース　タスクフォース報告書」 ※統合データベース構築のための体制整備、ロードマップ等について検討
平成 23 年４月	JST にバイオサイエンスデータベースセンター（NBDC）を設置 ライフサイエンスデータベース統合推進事業が開始

出典：国立研究開発法人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 パンフレット」、https://biosciencedbc.jp/gadget/pamphlet/NBDC_pamphlet.pdf