

ライフサイエンス統合データベースの 推進方策について（報告）

平成24年4月

大学共同利用機関法人情報・システム研究機構

ライフサイエンス統合データベースの将来構想検討会議

目次

前文	1
第1章 データベースの統合の意義とその重要性について	2
第2章 国内外におけるデータベース整備と統合化への取組みと達成状況	3
2.1 統合化以前の状況	3
2.1.1 諸外国の取組み	3
2.1.2 我が国の対応	3
2.2 統合化に向けた国内の取組み	4
2.2.1 文部科学省	4
2.2.2 経済産業省	4
2.2.3 農林水産省	5
2.2.4 厚生労働省	5
2.2.5 内閣府が果たしてきた役割	6
2.3 統合化の現状と動向	6
2.3.1 米国での取組み	6
2.3.2 欧州での取組み	7
2.3.3 中国の挑戦	7
2.3.4 諸外国の人材養成の状況	8
2.3.5 我が国の進捗状況	8
第3章 我が国が今後取り組むべき課題と方策	9
3.1 ライフサイエンス分野の動向を踏まえたデータベースの今後の課題	9
3.2 新たに取り組むべき具体的なテーマとその進め方	10
3.3 研究技術開発にかかる人材とその養成のあり方	10
3.4 持続可能な予算の仕組みと組織運営等	11
3.5 国内4省連携及び国際連携のあり方	12
3.6 アカデミックセクターが特に果たすべき役割	13
3.7 第二段階のセンターが持つべき役割	13
第4章 DBCLS の将来像とそれを実現するための方策	14
結 び	18
資料編	
1. 委員名簿	19
2. 会議開催録	19
3. これまでの統合化プロジェクトの経緯	20
4. これまでの取組み（4省）	22
5. 日本と欧米や中国との比較	25
6. 諸外国の人材養成の状況	28

前 文

近年、ライフサイエンス分野の著しい発展により、膨大な量のデータが産出され、その結果数多くのデータベースが出現することとなった。これにともない、第3期「科学技術基本計画」に基づき、総合科学技術会議において、ライフサイエンス分野の推進戦略「世界最高水準のライフサイエンス基盤整備」が策定され、文部科学省では平成18(2006)年9月に「統合データベースプロジェクト」の開始を決定した。これを受けて、データ中心科学の推進を標榜する大学共同利用機関法人である情報・システム研究機構（以下「ROIS」又は「機構」という）では、このプロジェクトの中核的な推進機関として、その担い手となるべく平成19(2007)年4月に機構本部に直轄組織としてライフサイエンス統合データベースセンター（以下「DBCLS」という）を設置した。

上記プロジェクトは平成22(2010)年度末に終了することから、平成21(2009)年4月に内閣府ライフサイエンスPT統合DBタスクフォースが取りまとめた「統合データベースタスクフォース報告書」に基づき、上記プロジェクトと平成13(2001)年度から実施している科学技術振興機構（以下「JST」という）のバイオインフォマティクス推進センター事業（以下「BIRD」という）とが一体化され、平成23(2011)年4月にJSTの下にバイオサイエンスデータベースセンター（以下「NBDC」という）が発足した。これにともない、それまで同プロジェクトの中核を担っていたDBCLSは、NBDCの公募事業である基盤技術開発プログラムに参画し、主に統合化のための研究開発と人材養成を担うこととなった。すなわち、統合データベースプロジェクトは、事業予算上NBDCの下に形式的に一体化されたものの、実態はNBDCとDBCLSの二つの組織を中心に展開されることになったのである。

しかしながら、内閣府のロードマップによれば、現行の体制は平成25(2013)年度末までの暫定的なものであり、それ以降のいわゆる第二段階のあり方については方向性しか記載されていない。近い将来、内閣府等においても審議が行われるものと思われるが、今後の検討や予算スケジュールを考えると早急な対応が望まれるところである。学術研究の推進・振興を負託された機構としても、内閣府等によって描かれるライフサイエンス分野におけるデータベースの構想を踏まえつつも、大学や研究所が行う科学研究の観点から、自らの役割について検討することが求められている。このような事情により、このたび機構の下にライフサイエンス統合データベース将来構想検討会議が設置された。本報告書は、同機構の長の諮問に答えて、同会議の審議をとりまとめたものである。

第1章 データベースの統合の意義とその重要性について

21世紀はライフサイエンスの世紀といわれて久しい。我が国においても、ミレニアムプロジェクトの立ち上げ以降、特に平成15(2003)年のヒトゲノムの塩基配列の解読完成と前後して、ライフサイエンス分野にかかる研究予算は急速に増大し、次々と巨大なプロジェクトや研究事業が立ち上がった。その成果のアウトカムとして、時間の経過とともに膨大なデータ量を抱えた多様なデータベースが数多く出現した。これにともない、これらの貴重な研究データを適切な形で統一的に収集、保存、整理して、国のレベルで有機的に統合し、有効活用を図るべきであるとの声が学术界だけでなく産業界からも出された。

個々の研究においても、投下される研究費用の巨大化や日進月歩の技術革新により、網羅的にデータが産出されるが、研究論文に用立てられるデータの占める割合は必ずしも高くはなく、その多くは顧みられなくなる傾向にある。しかし、ユニークな手法や視点に富むデータをほかの研究において十分に活用できる可能性は高い。統合化によって有効活用が図られることになれば、科学研究の発展に大きく寄与することは明白である。整備が進む欧米の先進諸国における成果事例や、新興国における国を挙げた取組みは、その重要性の証左であるといえよう。

殊に未曾有の財政危機に直面する我が国においては、今後は従来のような研究投資は望めそうもない。そのような中で、統合化されたデータベースの提供は、将来において、国家的な規模での効果的な研究費・資源の利用に大いに資するだけでなく、また新たな研究課題に挑もうとする後進の若手研究者に対して有力な支援ともなるであろう。

実際に研究の現場においては、ゲノム計画やポストゲノム計画などによる網羅的データの産出、これらのデータを基盤としたデータ駆動型研究の展開など、データベースの重要性は飛躍的に高まっており、今後の科学研究を推進していくための必要不可欠な知的基盤であるとの認識は深まりつつある。

それに加えて、ここ数年の新型シーケンサーの登場やゲノムコホート研究、がんゲノム研究などの進展により、基礎研究だけでなく、臨床、創薬、育種などへの応用という面でもデータベースの重要性は格段に高まっている。

さらに最近の傾向として注意すべきは、質的にも量的にもこれまでとは異なるタイプのデータが出現し始めていることである。具体的には、データの

- 1 大規模化 ペタバイト、エクサバイト級のデータ
- 2 多様化 脳画像、がん細胞画像などをはじめとする新たな種類のデータ
- 3 いわゆる
 パーソナル化 パーソナルゲノムや臨床データ
- 4 定量化 1分子や1細胞に関する種々の定量データ

などが挙げられるが、これらに加えてインターネット、データベース、並列処理に関するコンピュータ技術も、日々、目覚ましい進展を遂げている。

このような新たな局面を迎え、データベースの整備や統合化に関しては、従来とは別次元の国家レベル戦略が必要とされてきており、特に技術、人材、制度、体制、予算については、これまでの枠組みにとらわれず、今後のあるべき姿を全体的に改めて検証・検討することが重要である。

本報告書では、統合化に関していくつもの報告書が出ていることから、それらの報告書等が世に出て以降の、ここ数年の間に進捗した事業の顕著な成果や到達点、統合化を取り巻く技術的・社会的な環境の変化、国際的な潮流や世界水準等を俯瞰する。それら

を踏まえて、大学や研究機関等の研究者や研究の現場からの視点で、具体的に取り組まねばならない課題や事項を洗い出し、DBCLS、それは取りも直さず、我が国及び機構におけるデータベースのあり方でもあるが、その将来構想・展開に資することとしたい。

第2章 国内外におけるデータベース整備と統合化への取組みと達成状況

2.1 統合化以前の状況ーデータベース及びセンターの整備ー

2.1.1 諸外国の取組み

米国では、昭和 39(1964) 年に国立衛生研究所 (NIH) の下部組織である国立医学図書館 (NLM) において、現在の文献データベースサービス (PubMed) につながる活動が始まった。その後、昭和 44(1969) 年には遺伝子変異疾患データベース (MIM) が、昭和 46(1971) 年にはタンパク質立体構造データベース (PDB) が、昭和 57(1982) 年には核酸配列データベース (GenBank) が、というように、現在のライフサイエンス研究になくってはならないデータベースが続々と誕生した。さらに、昭和 63(1988) 年には NLM の下に国立バイオテクノロジー情報センター (NCBI : National Center for Biotechnology Information) が設立され、ライフサイエンス分野のデータベースの開発及び維持更新を専門に担い、バイオインフォマティクスの研究とデータベースや情報解析のサービスの中核拠点として機能することとなった。NCBI の予算規模は 72 億円、人員規模は約 600 名である。そのサービスとしては、GenBank を中心とする核酸配列データを中核に、標準配列である RefSeq データベースの提供、Entrez システムによる統合データベース環境の構築及び世界標準的な相同 (ホモロジー) 解析ソフトウェアである BLAST を中心とする各種解析ソフトウェアなどにその特徴がある。さらに、NLM の下部機関としての特徴を活かした PubMed の提供は、ほかにはない大きな特徴となっている。

欧州では、昭和 55(1980) 年に欧州分子生物学研究所 (EMBL) において核酸配列データベース EMBL-Bank の前身が形づくられた。平成 4(1992) 年には NCBI と同様の目的で EMBL の下部組織として、欧州バイオインフォマティクス研究所 (EBI : European Bioinformatics Institute) が設立された。EBI の予算規模は 43 億円、人員規模は約 500 名である。サービスの特徴として、UniProt や InterPro といったデータベースに代表されるタンパク質配列を対象とした機能情報の提供と、Ensembl と呼ばれるデータベースにおける真核生物のゲノムを対象とした詳細なアノテーション (データに生物学的医学的な解釈を加えること) 情報の提供などがある。資金的には、英国のウェルカム財団や米国の NIH から得て活動している。そのため、他機関との共同開発も多いのが特色である。

NCBI、EBI とともに人員の 2 割から 3 割程度が研究要員として働いている。それらの組織はただ単にデータベースの整備に限定されたものではなく、データベースの整備やサービスの提供とそれに関連する研究開発とが対になった形で構成されている。また、それらの組織においては、データベースの開発を明確な目標としたプロジェクト制で実施されている。

2.1.2 我が国の対応

我が国における日本 DNA データバンク (DDBJ : DNA Data Bank of Japan) は、欧州の EMBL-Bank/EBI 及び米国の GenBank/NCBI と並び、International Nucleotide

Sequence Database (INSD) を構築・維持する世界の三極の 1 つとして、昭和 61(1986)年に国立遺伝学研究所の遺伝情報研究センター内に設置された。この INSD データは目的や国籍に拘わらず閲覧転用できる科学の共有財であり、世界中の研究者は 3 つの機関のどれかを通じて INSD にデータを登録できることとなっている。ちなみに我が国からの登録の 99%以上が、DDBJ を通じて行われている。DDBJ は文部科学省予算により、同研究所の生命情報・DDBJ 研究センターの下で運営されている。

平成 12(2000)年 11 月、当時の科学技術会議ライフサイエンス部会ゲノム科学委員会から、バイオインフォマティクスの人材養成、研究開発の振興、データベース整備戦略の 3 つの課題に関する推進方策の提言が出され、平成 13(2001)年度より JST に BIRD が設立された。この BIRD 事業 のファンディングにより、ライフサイエンス分野の基盤的データベースや新たなデータベースの開発を推進するための支援が行われることとなった。これによりバイオインフォマティクス研究の推進や、大阪大学の Protein Data Bank Japan (PDBj)、京都大学の Kyoto Encyclopedia of Genes and Genomes (KEGG) といった世界的に定評のあるデータベースの構築等が進められた。

2.2 統合化に向けた国内の取組み

2.2.1 文部科学省

我が国のライフサイエンス分野のデータベースは国際的に高い評価を受けているものがある一方で、その他のデータベースの多くは、各機関や各プロジェクトが個々に作った状態のままになっており、これらに関連付けて使おうにも往々にして使い勝手が悪く、基本的サービスの多くが海外に依存していたり、継続的に維持されない状況にあり、国内の主要なデータベースの統合化と継続的な維持方策が必要であるとの指摘を受けることとなった。そこで我が国のライフサイエンス関係のデータベースの利便性の向上を図るため、当該関係データベースの整備戦略の立案や評価支援、データベース統合化及び利活用のための基盤技術開発、ポータルサイトの整備等を行い、統合化を推進することを目的として、平成 18(2006)年度より統合化のための委託事業が開始され、このプロジェクトの受け皿として、平成 19(2007)年度に機構の下に DBCLS が設置されたのは前述のとおりである。

平成 23(2011)年度からは、NBDC が JST に設立され、戦略立案、ポータルサービス、基盤技術開発、分野別の統合化推進からなる 4 つの事業を展開している。基盤技術開発については DBCLS が担当しており、各分野の統合化推進は 10 大学等が連携して行っている。

これまでの成果としては、DBCLS、NBDC 合わせて、現在約 50 のサービスを提供しており、利用実績としては月間約 9 万ユーザー、230 万のアクセスがある。

2.2.2 経済産業省

産業技術総合研究所及びバイオ産業情報化コンソーシアム (JBiC) が中心となり、これまでいくつかのデータベース統合化に向けた取組みが実施されてきた。その主な内容としては、平成 12(2000)~16(2004)年度に実施された「バイオインフォマティクス関連データベース整備 (ミレニアムプロジェクト・統合データベース)」では、ヒト遺伝子アノテーションデータベース H-InvDB が構築され、平成 17(2005)~19(2007)年度の「ゲノム情報統合プロジェクト」では、H-InvDB の更新及び選択的スプライシング、DNA 多型、PPI、比較ゲノム等の追加データベースの構築・公開が行われた。平成

20(2008)~22(2010)年度の「経済産業省統合データベースプロジェクト」では、経済産業省関連の公的資金研究から得られる研究成果を、産業上の有用性を評価した上で、統合データベース化することを目的としてポータルサイト MEDALS の公開、横断検索サービスの公開、ウェブサービスを用いたヒト分子関連情報の統合化が実施された。

平成 23(2011)年度から開始された「経済産業省ライフサイエンスデータベースプロジェクト」では、NBDC を中核とした政府全体の統合データベースの構築に協力するため、経済産業省関連の生命科学のデータベースに関して、ポータル構築連携、横断検索連携、アーカイブ構築連携が実施されている。

現在 MEDALS では、73 件のデータベースと 70 件の解析ツールを公開しており、53 プロジェクトをカバーしている。

2.2.3 農林水産省

平成 18(2006)~22(2010)年度において、農林水産生物のゲノムや遺伝子情報を効率的に利活用するシステムを構築することにより、新産業の創出に資することを目的として、農業生物資源研究所が中心となり、イネ、カイコ、家畜などの農林水産生物のゲノムや遺伝子情報を統合した農林水産生物ゲノム情報統合データベース (AgriTOGO) を開発し、運用を行ってきた。

平成 23(2011)年度からは、「食料・農業・農村基本計画」に掲げられている新品種や革新的な生産技術の開発、新需要を創出する付加価値の高い農産物・食品、農林水産生物の機能を利用した新素材・医薬品の開発等を計画的・効率的に推進するために、「画期的な農畜産物作出のためのゲノム情報データベースの整備」を行うこととなった。引き続き、同研究所が中核機関となり、AgriTOGO のデータを引き継ぐだけでなく、今後の超大量シーケンシング時代への対応として、大容量ストレージと高速ネットワークの整備に取り組んでいるほか、育種のためのマーカー作成や表現型と配列の関連を大規模解析するシステムの開発を目指している。また、NBDC と連携してデータアーカイブの提供や横断検索機能の開発も行っている。

現在 AgriTOGO では 47 件のデータベースを公開している。

2.2.4 厚生労働省

平成 12(2000)年度からの 4 年間に、厚生省ミレニアムプロジェクトとして「遺伝子解析による疾病対策・創薬推進事業」が行われた。高齢者の主要な疾患の遺伝子を解明することで、予防、治療法などの確立、画期的な新薬の開発などの推進を目指して、国立がんセンターが中心となり、アルツハイマー、がん、糖尿病、高血圧、喘息及び薬理遺伝学 (ファーマコジェネティクス) に関する疾患ゲノムデータベース (GeMDBJ) が公開された。

平成 23(2011)年度からは、医薬基盤研究所により、同研究所が所有するトキシコゲノミクスデータベース (Open TG-GATEs) や GeMDBJ、同研究所が開発した創薬ターゲットの絞り込みを支援するシステム (TargetMine) をはじめとするヒト由来のデータベース 8 件に加えて、NBDC との相互運用により、医学・生命科学関連の多数のデータベースを併せて横断検索できるシステム (Sagace) を構築し、公開している。ただし、同プロジェクトはあくまでボトムアップ型であり、ほかの 3 省のように省全体としての統合化とまでは至っていないが、政府全体の統合データベースへ、トキシコゲノミクスデータのアーカイブを提供したり、4 省合同ポータルサイト (integbio.jp) へ

の連携も行われている。また、データ提供の面においては、厚生労働科学研究費補助金によって構築されたヒト以外のデータベースは、公募要領にて統合データベースへの提供を義務づけられている。

現在 Sagace では、300 以上の医学・生命科学関連のデータベースを横断検索することができる。

2.2.5 内閣府が果たしてきた役割

平成 17(2005)~19(2007)年度において、世界最高水準の研究開発環境を提供するライフサイエンスデータベースの構築を中心としたライフサイエンス研究における国際的優位性の確保を目標として、科学技術連携施策群「生命科学基礎・基盤」に着手した。その補完的研究課題として、国内主要データベースを中心として統合化に関するフィージビリティ・スタディを通じた実現性検証を目的とする「生命科学データベース統合に関する調査研究」を科学技術振興調整費により実施した。

また、平成 20(2008)年度より総合科学技術会議ライフサイエンス PT に統合 DB クフォースを設置し、さらに平成 22(2010)年度より統合データベース推進タスクフォースを設置した。これにより今後のデータベース整備のロードマップを提示するなど、省庁の枠組みを超えた政府全体の統合データベースの構築の推進に努めている。

2.3 統合化の現状と動向

2.3.1 米国での取り組み

前述の NCBI のそもそもの目的は核酸配列データベースの開発提供であり、いわゆる統合データベースを目的としているわけではない。しかしながら、多彩な核酸配列関連データやそれらを解析するツールやサービスの提供、NLM 傘下であることを活かした文献情報とのリンクなどにより、膨大なデータの集合体をベースにそれらの利便を高めるための多様な機能の拡充を進めた結果、分子生物学の研究のための統合データベースとして、国際的なスタンダードとなっている。

米国ではさらにバイオメディカル研究におけるデータ共有とネットを介した共同研究を促進するための仕掛けとして、National Center for Research Resources (NCRR) が資金提供した Biomedical Informatics Research Network (BIRN) プロジェクトが、平成 13(2001)年から 5 年間、24 百万ドルのプロジェクトとしてスタートし現在も継続中である。これは 28 大学及び 37 研究グループが保有する装置、ソフトウェア、技術、データなどを共同利用化するものである。当初は脳のイメージング研究に集中されているが、将来はほかの分野にも応用することとなっている。平成 23(2011)年の NCRR の改組にともない、現在は National Institute of General Medical Sciences (NIGMS) の監督下にある。

また、平成 16(2004)年からは、cancer Biomedical Informatics Grid (caBIG) イニシアティブを開始している。これはグリッドコンピューティングによる、がん関連のデータ、研究ツール及び科学者を結ぶ仮想インフォマティクスインフラストラクチャの構築を目的としている。National Cancer Institute (NCI) から資金を受けている全米のがんセンターの参加など、かなり大規模なものが構想されており、国際的な協力関係も含まれるが日本は未参加である。また、サービス関連では民間企業の参加もできるが、オープンソース、オープンアクセス、自由参加が原則となっている。

さらに、ライフサイエンス分野には限定していないが、2 億ドル以上の資金を投じた

ビッグデータ研究開発イニシアティブが最近発表されるなど、大規模データの収集、格納、保存、管理、分析及び共有に関する技術開発が今後活発に進められる予定である。

2.3.2 欧州での取組み

EBI も NCBI と同様に核酸配列データベースの開発提供を主目的とし、こちらも必ずしも統合のための組織ではないが、Ensembl 等核酸配列データベースを基にした独自の集積型のデータベースを開発し、膨大な関連データベースとともに、解析ツールなど複合的・統合的なサービスを提供している。欧州全体の各国や財団のファンド、NIH からの資金によって維持されている一方で、殊に EMBL-EBI へは産業界からのサポートもあり、産業界からのニーズへの対応も考慮されている。

英国ではウェルカム財団が出資する Wellcome Trust Sanger Institute (旧 Sanger Centre) は平成 4(1992)年にヒトゲノムプロジェクトの 1 大解析センターとして発足した。平成 12(2000)年以降は健康と疾患に関わるゲノムの理解をミッションとして、ヒトゲノムをはじめとしたいくつかの研究分野に注力し、現在では約 900 名の研究者を擁して数多くのプロジェクトやコンソーシアムとコラボレーションしデータベースの整備・統合に貢献している。

このようなセンターとしての活動のほかに、欧州においては、ライフサイエンス研究とその医学、環境、バイオ産業への応用を支援する生物学的情報の持続可能な基盤として、European Life Sciences Infrastructure for Biological Information (ELIXIR) が設置されている。EBI がコアとなり、13 か国 32 機関が参加するコンソーシアム体制により、ハブとなる機関と各国に設置される 20 程度のノードからなるネットワークを構築する予定で、現在は準備フェーズ (平成 20(2008)~23(2011)年末) の終了時期を迎えている。平成 24(2012)年 4 月現在、ELIXIR 実施協力の覚書を締結した国は 11 か国となった。予算の流れはハブとノードでは大きく異なっており、ハブに対しては EU 全体のコンソーシアムにより予算を確保し、ノードに対しては各国の国家予算によって賄われることを想定している。平成 23(2011)年 12 月には英国より 7,500 万ポンドの資金を調達し、現在ハブとなる建物の建設を開始 (完成は平成 25(2013)年秋)、今後計算機の調達などにも使われる。

2.3.3 中国の挑戦一次世代シーケンスデータを産出する大型センターの設立一

中国の BGI (旧 Beijing Genomics Institute) は、平成 22(2010)年には 500 名以上 (内バイオインフォマティクス関連約 200 名) のスタッフを抱え、シーケンス能力：5 TBase/日、計算機能力：コア数 50,000、総メモリ 200 TB、ディスク容量 1,000 PB で解析を行うと発表した。これは、米国でも最大級のワシントン大学ゲノムセンター (平成 23(2011)年 5 月にゲノム研究所に改名) のシーケンス能力：70 GBase/日、計算機能力：コア数 3,000、ディスク容量 3 PB、スタッフ数：約 300 名 (内バイオインフォマティクス関連約 50 名) と比較しても格段に強力な解析パワーを持つことが分かる。中国のこの急激な台頭の背景には国家的な支援があり、核酸配列データの集積量において INSD と肩を並べるまでに成長することが見込まれ、アジアのハブとしての機能を今後主張してくることが容易に想像される。

参考：東京大学医科学研究所ヒトゲノム解析センターのマシンパワーはコア数 6,464、総メモリ 27 TB、ディスク容量約 1PB であり、国立遺伝学研究所 DDBJ のスーパーコンピュータは、コア数 6,560、総メモリ 36TB、ディスク容量 5PB である。

2.3.4 諸外国の人材養成の状況

バイオインフォマティクスや計算生物学、又はゲノム科学の専門分野の1コースとして、大学や専門教育機関における教育プログラムが多数組まれている。国際計算生物学会の調べによれば、専攻としてあるいは副専攻として組まれているプログラムの数は全部で延べ232（学士：38、修士：112、博士：54、その他：28）件であり、そのうち8割以上が欧米先進国に集中している。詳細については、資料編6を参照。

2.3.5 我が国の進捗状況

データベースがライフサイエンス分野で不可欠の存在であるにもかかわらず、統合化とそのために必要となる戦略立案、技術開発、人材養成、データ共有のルールづくり、行政内での相互連携などそれらへの取組みに対して、以前より我が国は総合的で戦略的な視点が欠落している旨長らく指弾されてきた。しかし、平成17(2005)年頃からの内閣府の調査研究、それに基づく平成18(2006)年度からの関係府省の統合化プロジェクトの下で繰り広げられる研究現場での努力により、人材養成を除いて、状況は大きく改善されつつある。その概略は以下のとおりである。

① 戦略立案

内閣府は統合データベース推進タスクフォースにて、文部科学省はDBCLS及び平成23(2011)年度からはNBDCにおいて、またそれぞれの府省の統合化プロジェクト推進機関においては、各々の戦略が立案され、実行に移されてきたが、それらの取組みはいわれるほどに4省がばらばらではなく、シンポジウムの開催などを通じて何らかの連携を行う形での実現が図られてきている。

② 技術開発

文部科学省ではDBCLSを中心に、経済産業省は産業技術総合研究所バイオメディシナル情報研究センター、農林水産省は農業生物資源研究所、厚生労働省は医薬基盤研究所等において、鋭意取組みが進んでいる。

③ 人材養成

データベースにかかわる人材の養成はBIRDやDBCLSを中心にオン・ザ・ジョブ・トレーニング(OJT)の形で実施されてきた。また、文部科学省の統合DBプロジェクトの一環として、東京大学、お茶の水女子大学、長浜バイオ大学においても人材養成が図られてきた。しかしながら、それらの人材養成の受け皿の問題もあり、欧米に比して大きく遅れをとっている。また、データベースを活用して新たな知識発見を行う、いわゆるバイオインフォマティクスの人材に関しても、それらの教育機関の数は欧米に比べるべくもなく、深刻な状況である。

④ データ共有のルールづくり

DBCLSから文部科学省その他への働きかけが行われ、現在は、文部科学省ライフサイエンス課による委託プロジェクト、JSTのCREST・さきがけ、厚生労働科学研究費補助金については公募要領に統合化プロジェクトへのデータ提供義務に関する条項が記載されるまでになった。

⑤ 4省連携

カタログ、横断検索、アーカイブの3大サービスについては、4省が連携し、これらの相互乗り入れや一体化が進行中である。これを背景に平成23(2011)年12月には合同ポータルサイトの開設も行われ、殊に現場レベルから始まった連携は大きく前進

しつつある。

第3章 我が国が今後取り組むべき課題と方策

このように国内で統合化への努力が続けられる一方で、周囲の理解や資金の不足等により、また統合化プロジェクトが実施されてから日が浅いこともあり、データベースの整備・統合化及びその利活用に関しては、諸外国との比較では未だに遅れをとっている。さらには新興国の台頭も相まって、これまで以上の取組みをしないと世界の中で取り残される恐れが出てきた。また、ライフサイエンスの領域での著しい技術革新、バイオロジー分野のみならず医学・医療への関心の拡大、停滞する研究資金の影響にともなう出口志向の高まり、科学的成果の社会への還元等の政治的社会的要請への対応に支障を来す可能性が出てきた。政治行政の混迷もあり、現状では中央レベルでの調整・取りまとめは遅々として進まず、このままでは将来的な展望が開けない状況にある。本章では前章の論述を踏まえて、今後我が国が取り組むべき課題を洗い出し、まずは目指すべきその全体像を描くこととしたい。

3.1 ライフサイエンス分野の動向を踏まえたデータベースの今後の課題

アカデミックの観点からデータベースの重要性が説かれてから12年が経過した。一方、産業界からデータベースの統合に強い期待が寄せられてから7年も経った。これらを受ける形でこれまで10年以上にわたり、内閣府や関連府省におけるいくつもの委員会や作業部会において、データベースの意義や重要性、そのためのナショナルセンターや人材養成の必要性が説かれ、推進方策が策定されてきた。そこで描かれた方策の必要性は現在も基本的に変わっていない。今後は、それをいかに実行に移すかということと、データの大規模化、多様化、いわゆるパーソナル化、定量化などや臨床や育種への応用が現実化してきたという近年の状況の変化に対応して、設計の一部を見直すとともに、これまでの取組みを大幅に拡充することが必要である。例えば、患者個人由来のゲノムデータを診療に活用する可能性が大きく開かれつつある。患者データの匿名性を確保しつつ、得られる情報を共有することは今後の医生物学研究の発展を図る上で不可欠であり、これに対応することが必要である。

ここ10年ほどで想定されるライフサイエンス分野の動向をデータベースとのかかわりで眺めると以下のようなキーワードが浮かぶ。

網羅的、データ駆動型、データ中心科学、パーソナルゲノム、エピゲノム、ゲノムコホート、メタ解析、画像データ、育種、創薬、生命動態、細胞シミュレーション、表現型、ペタバイト級データ、システムの理解、構造生命科学

これらのキーワードに代表されるサイエンスには、データベース化とその統合や解析のために新たな技術や人材が必要とされる。特にデータの共有、データベースの整備や統合化、解析ツールの開発とそのための人材養成はますます重要な課題となっており、これがうまく実現されなければ、我が国のライフサイエンス研究は早晚立ち行かなくなる。また患者のデータを扱うヒトゲノムに関わる倫理的課題を理解できるバイオインフォマティクスの専門家の養成も必要となる。

またライフサイエンス分野と環境分野、地球観測分野などとの連携も視野に入ってくる。これらの分野とのデータ統合も進むであろうし、それらの分野と共通する課題（データ共有）や共通する技術（統合化技術）も出てくる。基礎研究にとどまらず、応用研究、すなわち臨床、育種、バイオテクノロジー全般にも当然当てはまり、小手先ではない、

骨太の計画立案とその速やかな実行が望まれる。

ここ数年以内には個人の全ゲノム配列が極めて安価に解析できる時代が到来することが予想される。そうなれば、個人の DNA 配列情報に基づく病気の診断、治療及び予防などの新しい医学が一気に加速するであろう。医療統合データベースの果たす役割は計り知れないが、そこでは大きく 2 つのタイプのデータベースを考慮する必要がある。一つ目は、個々人のゲノム情報や医療画像情報、臨床検査のデータベースである。個人情報保護の観点及び圧倒的なデータ量のため、ローカルにデータベース化される可能性が高いが、今回の DBCLS の統合データベースの将来計画の中にも含めるかどうかについては、議論の余地がある。二つ目は、ゲノムコホート研究等からのデータ、またそれに付随する医療関連情報などのデータベース化である。これらは個人情報の保護、極めて大規模なデータ量を考えると、その扱いには従来と異なる対応が求められる。例えば、医療でのレファレンスデータベースでは、医師や看護師、検査技師などに利用しやすくするために、データベースの統合化やユーザインターフェースの整備を急ぐべきであり、それはまさに DBCLS が取り組むべき課題である。

今後の統合データベースとそのための制度設計に際しては、これらについても考慮に入れた計画の策定が必要になる。

3.2 新たに取り組むべき具体的なテーマとその進め方

これからますます増大するデータベースやバイオインフォマティクスの需要を満足するためには、戦略立案、ポータルサービス、3大サービス（カタログ、横断検索、アーカイブ）、分野別データベース統合、4省連携、技術開発のこれまでの取組みに加えて、データの大規模化、多様化、いわゆるパーソナル化、定量化に対応する、基礎だけでなく応用・診療にも使える技術を開発すること、また我が国のバイオインフォマティクス人材の不足とそれへの取組みの貧弱さを考慮しつつそのための人材養成を、総合的な観点から取り組むことが重要である。

より具体的には、例えば、以下のようなテーマや課題に取り組むべきである。

- ① 大規模データの格納、比較、検索、表示、転送など処理技術開発とそのための持続可能な計算機資源の確保と活用術
- ② メディカルデータ、クリニカルデータ、育種データ、医薬品データなど新たなデータの処理と統合化技術
- ③ 画像データの検索、正規化、比較、意味抽出、メタデータ付与
- ④ パーソナルゲノムや臨床データの秘匿や暗号化の技術、それによる予測や診断のための知識発見
- ⑤ 定量データの標準化技術やそれらのデータからのモデル構築と細胞・臓器・人体のシミュレーション、モデルと計測データとの融合技術
- ⑥ ゲノムやタンパク質構造の情報のみならず、国内で産出される基礎生命科学研究の結果に関する膨大な情報を基盤的なデータベースとして確立

これらの技術開発とそれによるデータベースの統合化、データベースサービス、そのための人材養成、それらの方針や国内外との連携のあり方に関する戦略立案は、一体的に推進されるべきである。

3.3 研究技術開発にかかる人材とその養成のあり方

質的にも量的にも従来とは異なるデータが出現しており、また、医療などへの応用が現実化してきたことにより、それらのデータベース化や統合化には新たな技術開発が不可欠である。いうまでもないが、それには最新鋭の情報技術も効果的かつ迅速に取り入れる必要がある。

これには、これまでの人員（現在の DBCLS プロパー研究員は 12 名、NBDC の研究員と合わせても 20 名程度）では対応は不可能であり、少なくとも研究員を 100 名ほどの規模に増員する必要がある。しかし、これでも欧米や中国と比較しても数分の 1 程度にしかならない。

一方、データベースの統合化には、研究員以外にもオントロジーの構築やデータの意味付けなどを支援する研究支援員（アノテータ、キュレータ）やデータベースの管理を行うシステムエンジニアなどが不可欠である。これらの人員（現在は 10 名弱）についても、大幅な人員の確保を図る必要がある。

研究員や支援員の増員を図るためには、恒久ポジションの確保はもとより、高度に知的な活動をするがいわゆる従来の研究者とは活動形態が異なる研究者を評価する仕組みの構築や、研究員、研究支援員それぞれに応じたキャリアパスの設計などが必要不可欠であり、これらの問題が解決されなければ良い優れた人材の確保は難しいといえる。

また、既存の人材を集めるだけでなく、人材養成の仕組みの構築が必要である。これはデータベースの統合化のみならず、個々のデータベース構築やデータベースからの知識発見にも不可欠であり、そのための人材養成の機能を強化する必要がある。大学等との連携による人材養成ネットワークの構築などを強力に進めるべきである。

そうした中において、総合研究大学院大学を含めた大学のさらなる活用が検討されるべきである。総合研究大学院大学においては、関連科目の開設、全学教育事業への応募やリサーチアシスタントの積極的活用等が、また他大学においては連携大学院制度の活用や特別共同利用研究員の受入れによる人材養成が考えられる。

なお、国内だけでなく海外機関との協定等の締結に基づいた研究者や留学生の受入れを推進するのも一方策と考える。

先に述べたように、データベースの構築や統合化に関する人材だけでなく、それを解析して新たな知識発見を行うバイオインフォマティクスや、臨床データやパーソナルゲノムデータから疾患の診断や予防に役立つ知識を発見するメディカルインフォマティクスに関する人材、さらには創薬などのためのケミカルバイオロジストなども、我が国は大きく不足している。これらの人材を養成するためには、データベースの構築や統合化とは必要とされるスキルが若干異なり、大学などの教育機関で果たすべき役割ともいえるが、現状では取組みが不十分な面もあり、これらの点も十分に考慮した総合的な人材養成の設計が必要である。

以上に述べたように、バイオインフォマティクス人材の養成の必要性はこれまで 10 年以上にわたって叫ばれ続け、解消に向けてさまざまな努力が続けられてきた。しかしながら、それらは成功したとはいえない。すでに述べたような取組みだけでなく、例えば産業界との連携体制など新たなシステムを導入する必要がある。これについては、現在文部科学省のライフサイエンス委員会で議論がなされていることもあり、そちらに委ねたい。

3.4 持続可能な予算の仕組みと組織運営等

今まで述べた課題を解決するためには、現在の体制では実現はほとんど不可能に近い。

現行の統合化プロジェクトの予算規模は全体で約 17 億円であり、この運営費では、新たな動きや人員増への対応は到底不可能であり、客観的な見地からも研究員だけでも 100 名規模のセンターを実現するためには少なくとも 30 億円程度の予算規模が必要である。この金額は、3 千億円とも 4 千億円ともいわれる我が国のライフサイエンス予算の 1 %ほどに相当するが、ライフサイエンス予算のかなりの部分がデータ産出に費やされている現状を考慮すると、それらのデータを有効に活用ならしめるために同予算の 1 %を使用するというのは、合理的かつ妥当であるといえる。

ただし、仮に 30 億円程度の予算を確保できたとしても、データの指数関数的な増大には必ずしも対応できるとはいえない。国のプロジェクトの成果として構築されるデータベースの増大に応じて、自動的に予算が増額される仕組み、例えばデータベースの構築が必要、あるいは指向するプロジェクトからは、予め一定予算を統合化プロジェクトに対して供出させ、サービスが提供できるような仕組みが必要である。

なお、上記金額では大規模データの格納や解析のためのスーパーコンピュータの費用まで賄うことはできない。機構傘下の国立遺伝学研究所で使用されているスーパーコンピュータを活用するのはもちろんのこと、「京」を中心とした我が国のハイパフォーマンス・コンピューティング・インフラストラクチャ（HPCI）を活用する仕組みを構築すべきである。

現段階では、既存のデータベースの統合化が精一杯であり、個々のデータベースの構築の支援までは困難である。しかし、個々のデータベースの高品質化、高度化がなければ統合化も意味が薄れることとなるため、これらについても支援策の強化も重要である。特にパーソナルゲノムデータの構築支援は重要な課題である。

今後新たなプロジェクトが立ち上がる際には、データベース構築を必ず担保するなどの制約を設けることはもちろん、統合化プロジェクトへのデータ提供を義務付ける仕組みもこれまで以上に広げる必要がある。

3.5 国内 4 省連携及び国際連携のあり方

総合科学技術会議のライフサイエンス PT 統合 DB タスクフォースが取りまとめた「統合データベースタスクフォース報告書」におけるロードマップを実現すべく、内閣府は、「指令機能」を担う組織によって、平成 26(2014)年度以降のいわゆる第二段階のデータベースの統合化の具体策を急ぎとりまとめる義務がある。その意味で総合科学技術会議の責任は重い。

関係者の努力も相まって、4 省それぞれで行われてきた統合化プロジェクトの連携という点においてここ数年で大きな進展をもたらした。カタログ、横断検索、アーカイブの 3 大サービスについては、4 省の実務者の合意の下に、相互乗り入れや合体の方向で着々と成果を挙げつつある。平成 23(2011)年 12 月には、4 省合同のポータルサイトが実現するまでに至り、連携からさらに進めてサービスの一体化までもが視野に入ってきた。今後はこの動きをさらに加速し、さらなるステップである、目的別、分野別のデータベース再編成に向けて進むべきである。

しかしながら、ゲノムコホートプロジェクト等で構築されようとしている臨床データなどの連携については、まだまだこれからであり、さらなる議論を重ねる必要があるほか、そのための技術開発も必要と考える。臨床データの中でも、ゲノム情報の持つ臨床的有用性は特筆に値する。個人情報と分離した上でゲノム情報を公的なデータベースに蓄積し、研究者に公開するための標準的なスキームの確立が不可欠である。

なお、データベースは府省ごとに統合化しても効果はそれほど大きくない。4省が協力しないと成り立たないとするならば、その際は4省のいずれが担当しても難しく、基礎データは文部科学省、臨床データは厚生労働省というように担当を明確化し、それらをつなぐ役割を統合化プロジェクトを推進するセンターが担うこととして位置づけるべきである。

国際連携については、現在はホームページや一部のサービスの英語化のレベルにとどまっている。それらを加速させるだけでなく、NCBIやEBIあるいはBGI等との連携、すみ分けを検討し、必要に応じて国際交流協定の締結も進めるべきであるが、データベースが我が国の研究の最重要基盤であることを考えると、戦略的に国際連携を行う必要がある。例えば、DNA塩基配列などは、DDBJでは常時NCBIやEBIと相互にデータを交換し合っており、国際レベルでの労力の重複が避けられている。一方で、海外では進められていない日本独自のデータベース構築の検討も必要であろう。

3.6 アカデミックセクターが特に果たすべき役割

我が国のライフサイエンス予算を有効活用するには、その予算の大部分がデータに化している現状では、データの共有化、統合化の意義は大きい。ELIXIRなど欧米においても、まさにそのような方向でプロジェクトの立案が図られている。

アカデミックセクターとしては、データの共有、統合化によるオープンイノベーションを牽引すべき使命がある。この使命を果たすには、

- ① データ共有やオープンアクセスの成功例を積み上げること
 - ② データ共有やオープンアクセスのルールづくり、基盤づくりに貢献すること
 - ③ ヒト由来データなどについてはプライバシーの保護との両立をどう図るべきかについて技術的、制度的な仕組みづくりに際して、データベース運営者・利用者の立場から積極的に提言すること
 - ④ 共有、統合化、解析のための種々の技術開発を行うこと、またそのための情報技術の基礎的な研究開発を行うこと
 - ⑤ それらを担う人材（研究者、技術者、アナテータ、キュレータ、システムエンジニア）を養成するための教育システムを確立し、また、それらの多様な人材の評価法の開発とキャリアパスの設計について取り組むこと
 - ⑥ どういうものを作るのかという明確なイメージをつくり上げること、中心となるデータベースを設定しNCBIなどの世界標準との連続性を確保すること
- などが挙げられる。

上述の技術や人材を背景に、個々のプロジェクトや研究で構築されるデータベースに対して、統合化に向けたガイドラインを提供するとともに、場合によっては、プロジェクトの企画時又は実施開始時からデータベース構築に協力することなどが期待される。

3.7 第二段階のセンターが持つべき役割

ナショナルセンターとして、第二段階の統合データベースセンターが持つべき機能については、戦略の立案、ポータルサービス、目的別・分野別データベース開発（臨床、創薬など応用分野のデータを含む）、統合技術開発すなわち情報技術開発及びオントロジーなどのコンテンツ開発、人材養成（研究者、技術者、アナテータ、キュレータ、システムエンジニアなどのほかデータベース構築だけでなく解析研究者を含む）、各プロジェクトに対するデータベース構築への支援、これらすべてを実現するための計算機資

源の確保運用及び広報普及活動、国際連携の窓口、国内のデータベース活動の連絡調整などが挙げられる。

研究機能については、若い研究者にインセンティブを与えるものであるが、その研究の対象はバイオインフォマティクス全般とするかデータベースの整備統合に限るかは難しい問題である。一般的なバイオインフォマティクスの研究開発はほかでも行われていることを勘案すると、ここでの対象は統合化技術とデータ解析技術に特化した研究を展開するのが適切であると考えられる。

個人情報を含む公開が難しいデータについては、例えば厚生労働省関係の臨床データの一部等は、内容によって別扱いとするが、その場合でも匿名化や統計情報への変換作業について協力するなどの連携関係を保ちつつ、公開可能な情報をできるだけ国民に公開できるように努める。特に、公的研究費により実施されている日本人の大規模シーケンスデータにより蓄積されつつある日本人の遺伝子多型の頻度情報を速やかに医学研究者に公開することは、我が国の医学研究の発展につながる。

上記の取組みは一体的に設計、運用されるべきである。

第4章 DBCLSの将来像とそれを実現するための方策

前章において、平成26(2014)年度以降の第二段階におけるナショナルセンターの持つべき役割について述べた。将来構想を考えると、前文に記したようにNBDCとDBCLSが相互に分担しそれぞれの事業を展開している現状において、DBCLSが今後どのような役割を果たしていかなければならないのかという問題は避けて通れない。オールジャパンの視点で俯瞰したとき、総合科学技術会議のライフサイエンスPTは、「①引き続き、JST内の組織としてさらに充実強化を図り組織を編成していくこと、あるいは②分野的に近い機関に移行すること、③新たな機関を設置すること」と提言しているが、いずれの方法が採られるかは、政府レベルで検討・判断されるべき課題である。今後半世紀以上の長期にわたり、我が国のライフサイエンスの基盤となる統合データベースの観点から総合的に取り組まねばならない。

1 第二段階の体制として、例えば、

①DBCLS/ROISとNBDC/JSTとを、どちらの組織・体制を中心にするかは今後の課題として、一つの組織にまとめあげ、ナショナルセンターとして機能させる。

②現状の体制(NBDC、DBCLS、DDBJ、その他各府省傘下の研究機関)は存続させることとしても、実質的に機能面でより一体的な運営を可能とする仕組みを作る。などが考えられるが、しかしいずれの形態になろうとも、データベースの統合化には高度な専門性を備えた人材が必要であり、現時点で唯一その条件を満たすことができるDBCLSは、大学共同利用機関法人傘下の組織として、大学や研究機関と連携しながら、今後も研究開発、人材養成、データベースの開発支援などを担わなければならない。また、それを可能ならしめるためにも、機構もその責務及び負託の重大性を十分に認識し、機構内での確固たる地位を確立できるよう積極的に取り組むべきである。こうした観点に立てば、DBCLSとNBDCとが一体的に運営ができる環境は自ずと醸成されていくだろう。

2 そのためには、DBCLSが独自のロードマップを作成し、新たなマニフェストを設け、抜本的な改革に取り組む必要がある。それは研究において国際競争に打ち勝つためにも

世界水準を満たすものでなければならず、今後のスケジュールからも大至急具備する必要がある。ロードマップはアカデミーの需要に応え、自由で豊かな発想を重視し、最新の科学を追究、反映されなければならず、なるべく多くの有識者や研究現場の声に耳を傾けるべきである。DBCLS を擁する機構は、政府の行政機関やその傘下にある独立行政法人ではなく、学術研究を推進する組織であり、アカデミックセクターに所属しており、その長所を十分に活かすべきである。

なお、オープンでかつ効率的な運営を構築するためには、外部の有識者を中心とした運営委員会による経営など、ややもすると画一的でセクショナリズムに陥りやすい行政や親組織の影響を受けないような仕組みを担保することも肝要である。

ただし、人材養成などをはじめとして、DBCLS 単独あるいは ROIS 単独のロードマップだけでは国の取組みとして不十分であり、他府省や産業界も巻き込んだ大きな動きを展開する必要がある。

- 3 NBDC からの委託業務にとどまらず、関連コミュニティの要請に対して自発的・積極的に応えつつ、活動を展開すべきである。これは、技術革新の目覚ましい時代において、大学共同利用機関法人である機構が果たすべき機能であり、責務でもある。広くアカデミーから期待される新たな分野の開拓、基礎的な生物学の領域のみならず、大学や研究所で行われている医学や薬学のデータベースの作成への協力、応用開発研究にかかわる領域への意欲的な取組みが期待される。
- 4 上記の新分野に加えて、臨床、創薬といった医学系データベースなど統合化が困難であるとされるデータベースとの連携強化、より高度な検索への取組み、さらには知識発見技術、シミュレーション技術の開発など、情報爆発時代にライフサイエンスのイノベーションの基盤として、持続可能なデータベース統合化の技術を開発することが肝要である。
- 5 独自の研究技術開発拠点として恒久的な研究の場を設けることは喫緊の課題である。レンタルスペースにおいて活動する現在の状況は、効果的な研究の遂行、データベースの永続性やそうした事業に対する信頼性、優秀な人材の確保などの側面から、決して好ましい状態とはいえない。機構は立川や三島地区等に新たな施設設備のための十分なスペースを保有していることから、資産の有効活用を図り DBCLS の施設を整備し、大学と同様な研究環境、全国の学術研究機関等との交流の場を提供すべきである。

前章の「持続可能な予算の仕組み」で述べたが、我が国のライフサイエンス予算の大部分がデータ産出側に費やされている現状を訴えつつ、相応の経費がデータベースの利活用にあてられるべきであるとアピールすることが重要である。政府に対する概算要求においては、高エネルギー物理学や天文学などにかかる巨大科学の予算とは比較にならないほどの金額ではあるが、機構は日本学術会議などの応援を得て、積極的にチャレンジする姿勢が大切である。特に、インフラの整備と人材の確保・養成には、相当額の初期投資が欠かせない。
- 6 優秀なスタッフを確保するためには、外部資金の獲得という組織独自の努力だけでなく、「情報・システム研究」を志向する機関の名に相応しい機構の全面的・計画的な協力支援が不可欠である。複数の関連する研究分野の大学共同利用機関により法人を形成

したという法人化のメリット、またデータセントリックな科学手法を推進するという情報・システム研究機構そのものの設立趣旨を活かすことが期待される。

研究スタッフの増員には、キャリアパスの問題の解決が必要であり、現時点ではそれに対応できる人材がすぐに確保できるわけではない。幸いなことに機構傘下の各研究所は、生物学や情報学関連の研究者を多数擁している。新たな事業や研究分野への取組みや新技術への対応には、それらの関連研究者の参画は不可欠であり、研究所を越えた研究者間の連携の強化、相互交流の促進は、結局は各研究所の発展にも寄与するものと信じる。

そのような連携強化、相互交流により、例えば、比較ゲノムインフォマティクスの手法によるゲノム機能解析研究、ゲノム情報学での大規模計算の高速化アルゴリズムの開発やシステム生物学における知識発見といった生物学関係の情報処理、オントロジー工学の研究やデータの効率的格納、検索のための完結データ構造といったデータベース技術の研究、セマンティックウェブにおけるコミュニケーションとインタラクションに関する研究、データベースにおける画像解析研究、遺伝子過程モデルの植物個体群への応用、ゲノムスケール遺伝子発現・プロテオームデータの統計解析、ゲノム多様性の解析、遺伝子発現データの解析といった遺伝子・ゲノム関係解析研究、ゲノム構造の比較、ゲノム構造の翻訳産物としての蛋白質多様性、生命情報の数理的研究などが大きく加速する。これらは、世界的にもユニークであり、遺伝学、情報学、統計数理、極域科学の融合的発展を実現できる我が国で唯一の研究組織として、機構が一体となって牽引し、長期的に支援、維持しなければならない。

7 我が国の先を歩んでいる欧米の機関の事例を見るに、並べて最初からデータベースの統合化を標榜して構築されたのではなく、核となる重要なデータベースを有する強力な研究機関が、研究や時代の流れに沿って発展し牽引してきたことが分かる。国立遺伝学研究所においては、ゲノムデータに関して国際的な拠点である DDBJ を擁している。DBCLS と同じ機構内の DDBJ とは連携協働し、一体的な運用が行われることにより、DBCLS は DDBJ が有するスーパーコンピュータ、人員などの研究資源を共有し安定的な活動を続ける、一方で DDBJ はこれにより一層の機能強化を図ることがともに可能になり、シナジー効果を高めることが期待できる。また、国立遺伝学研究所や機構傘下のほかの研究機関との協力により、総合研究大学院大学等をはじめとして全国の大学や研究機関との共同研究を通じて、研究・技術開発のための人材養成を大幅に強化する。これに関しては現在文部科学省ライフサイエンス委員会で進められている人材養成に関する新たな仕組みづくりなどの議論を踏まえつつ、それと呼応した形でさらなる方策を検討すべきである。

8 DBCLS は、運営費の大部分を JST の研究委託費に依存する現状に満足せず、研究者自らが自由な発想の下に研究テーマを設定する科学研究費補助金の獲得、ほかの研究プロジェクトとの連携・分担、共同研究契約による受託研究など積極的に外部資金の獲得に努めるべきである。

運用経費の確保という点においては、外部資金の確保のほかに自己収入の獲得という方法も考えられる。代表的な手段として、機構傘下の国立情報学研究所が運用する NII 学術コンテンツ・ポータル (GeNii) のように受益者負担の考え方のもとに、運営に係る経費分のみをユーザーから情報提供料として実費徴収することも可能である。実際に

他機関のデータベースの統合化自体や統合化支援、医療情報等特定のユーザーに限定されるサービスの提供、あるいは維持できなくなったデータベースの運用管理等の請負いに関しては、検討の余地はある。ただし、世界的にライフサイエンス分野のデータベースは無償で提供されているほか、国のプロジェクトで構築したデータベースに課金することの是非も問われることとなる。今後、慎重にその是非を検討する必要がある。

- 9 データ共有に関しては、これまでの働きかけをさらに進めるとともに、オープンイノベーションの成功例を示す取組みを強化することが必要である。未だ離陸段階にある統合化の事業にとっては積極的な広報普及活動も重要な業務であることを忘れてはならない。
- 10 将来構想の検討にあたっては、JST の科学技術情報発信・流通総合システム (J-STAGE) や国立情報学研究所の NII 論文情報ナビゲータ (CiNii) 等の文献事業などのオープンアクセス化の動きとの連携も十分に視野に入れておくことも重要である。
- 11 個人情報がかかわるヒトゲノムや臨床データの倫理問題、データベース作成にあたってのルール化など、アカデミーだけでは解決できない課題については、行政府や JST などの外部の機関に協力を仰ぎ、また、現在進行中の大規模コホート研究などとも連携し、調整を図りつつ、データベースの運営者としての立場と、潜在的な利用者の立場を代表するものとして、倫理問題に関するコンセンサスの形成についてイニシアティブを取ることが求められる。

結 び

本報告書では、ライフサイエンス分野の急激な進展に対応した統合データベースの意義と今後のあり方、取り組むべき課題、それを実施するナショナルセンターの機能や体制、その中で DBCLS の果たすべき役割、それを支援するための機構の責任、などについて述べてきた。これを受け、以下の取組みが実行に移されることを期待して結びとする。

- 1 機構は本報告書の指摘や助言を真摯に受け止め、平成 26(2014)年度以降の体制を睨みつつ、DBCLS が作成するロードマップに基づき、将来体制や事業について、予算規模をはじめとして具体的な年次計画を早急に取りまとめるべきである。
- 2 機構は DBCLS あるいはその発展型を大学共同利用機関法人の主要な機能として、全面的に位置づけ、必要な予算、人員、施設設備について、関係各所に働きかけを行い、実現に向けた努力を行うべきである。これらと並行して、機構はその設立趣旨にのっとり、DBCLS と DDBJ との一体的な運用、総合研究大学院大学の活用等、機構を構成するあるいは関係する研究機関や組織と DBCLS との連携のあり方の具体案を作成し、平成 26(2014)年度を待たず始められるものから速やかに実現する必要がある。

資料編

1. 委員名簿 (50 音順)

小原 雄治	情報・システム研究機構・理事 (国立遺伝学研究所・所長、ライフサイエンス統合データベースセンター・センター長)
郷 通子	情報・システム研究機構・理事
五條堀 孝	情報・システム研究機構国立遺伝学研究所・副所長
主査 榑 佳之	豊橋技術科学大学・学長
篠崎 一雄	理化学研究所植物科学研究センター・センター長
末松 誠	慶應義塾大学・医学部長
高木 利久	東京大学大学院新領域創成科学研究科・教授
中村 春木	大阪大学蛋白質研究所プロテオミクス総合研究センター・センター長
永井 良三	自治医科大学・学長
長洲 毅志	エーザイ株式会社・理事
米澤 明憲	理化学研究所計算科学研究機構・副機構長

2. 会議開催録

第1回	平成23年10月19日(水)	18:00～20:00
第2回	平成23年11月7日(月)	18:00～20:00
第3回	平成23年12月26日(月)	15:00～17:00
第4回	平成24年2月16日(木)	15:00～17:00
第5回	平成24年3月29日(木)	16:00～18:00

3. これまでの統合化プロジェクトの経緯

年度	ROIS/DBCLS	内閣府/総合科学技術会議	文部科学省	JST
平成 12 (2000)		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」		
平成 13 (2001)				4月 バイオインフォマティクス推進 センター事業 (BIRD) 開始
平成 14 (2002)		12月 BT 戦略会議「バイオテクノロジー 戦略大綱」	6月 科学技術・学術審議会研究計画・評 価分科会「ライフサイエンスに関す る研究開発の推進方策について」	
平成 17 (2005)		12月 連携施策群「ポストゲノム」 「生命科学データベース統合に関 する調査研究」(-平成19年度) 3月 第 3 期科学技術基本計画 (平成 18-22 年度) 閣議決定		
平成 18 (2006)	6月 概算要求 9月 統合 DB プロジェクトを受託 (- 平成18年度) 12月 統合データベースセンター (仮 称) 設立準備委員会設置		5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 DB 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」 9月 統合 DB プロジェクト開始 (-平成 18年度)	
平成 19 (2007)	4月 DBCLS 設置、第 1 期開始 (-平成 23 年度) 統合 DB プロジェクト受託(-平成 22 年度)		4月 統合 DB プロジェクト開始 (中核機関、分担機関により実施) 10月 統合 DB プロジェクト補完課題実施 機関決定	
平成 20 (2008)		5月 ライフサイエンス PT「統合 DB タ スクフォース」設置	7月 統合 DB プロジェクト中間評価 1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	

年度	ROIS/DBCLS	内閣府/総合科学技術会議	文部科学省	JST
			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」	
平成 21 (2009)		4月 統合 DB タスクフォース報告書		9月 ライフサイエンス分野統合データベースセンター設置準備委員会設置
平成 22 (2010)	12月 ライフサイエンスデータベース統合推進事業応募	6月 ライフサイエンス PT「統合 DB 推進タスクフォース」設置	3月 統合 DB プロジェクト事後評価 統合 DB プロジェクト終了	6月 ライフサイエンス分野統合データベースセンター設置準備の検討とりまとめ 10月 NBDC 準備室設置 11月 ライフサイエンスデータベース統合推進事業公募開始
平成 23 (2011)	4月 ライフサイエンスデータベース統合推進事業「基盤技術開発プログラム」受託（-平成25年度） 10月 ライフサイエンス統合データベースの将来構想検討会議設置	4月 <ロードマップによる第一段階 開始> 8月 第 4 期科学技術基本計画（平成 23-27 年度）閣議決定		4月 NBDC 設置 ライフサイエンスデータベース統合推進事業開始 3月 BIRD 終了
平成 24 (2012)	4月 DBCLS 第 2 期開始（-平成 27 年度） ライフサイエンス統合データベースの将来構想検討会議「ライフサイエンス統合データベースの推進方策について（報告）」			
平成 25 (2013)		3月 <ロードマップによる第一段階 終了>		
平成 26 (2014)		4月 <ロードマップによる第二段階 開始>		

4. これまでの取組み（4省）

(1) 文部科学省

年度	事業名	事業主体	予算	概要	成果
平成13(2001)- 平成23(2011)	バイオインフォマティクス推進センター事業 (BIRD)	JST/BIRD ※平成23年度は NBDC	186.3億円	<ul style="list-style-type: none"> 生命情報データベースの高度化、標準化 バイオインフォマティクスの創造的研究開発 研究成果情報の発信 	<ul style="list-style-type: none"> PDBj、KEGG など国内の主要なデータベースが国際標準へ発展 大学等による研究開発22件 ゲノムリテラシー講座45件開催、データベース（各機関提供）23件、ツール20件を公開 (http://www-bird.jst.go.jp/database/)
平成18(2006)- 平成22(2010)	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DBプロジェクト）	文部科学省 ※平成21年度より一部JST	50.4億円	<ul style="list-style-type: none"> 中核機関：ROIS/DBCLS を代表機関とする参画機関8機関、分担機関：3グループ7機関、補完課題：4機関 データベース整備戦略の立案、評価支援 データベース統合化基盤技術開発 ポータルサイト整備 	<ul style="list-style-type: none"> ポータルサイト (http://lifesciencedb.jp/) 開設、横断検索など50サービス提供、月間約96万ページビュー 経済産業省との連携により、250種類のデータベースに対する相互検索サービスを提供
平成23(2011)-	ライフサイエンスデータベース統合推進事業	JST/NBDC	17億円 (平成23年度)	<ul style="list-style-type: none"> 戦略立案 ポータルサービス 基盤技術開発（ファンディング） 分野別の統合化推進（ファンディング） 	<ul style="list-style-type: none"> 基盤技術開発プログラム1課題、統合化推進プログラム10課題（平成23年度）を実施 文部科学省統合DBプロジェクトから引き継いだサービス等を公開 (http://biosciencedbc.jp/) 4省合同ポータルサイト「integbio.jp」（インテグバイオ、http://integbio.jp/）共同開設

(2) 経済産業省

年度	事業名	事業主体	予算	概要	成果
平成12(2000)- 平成16(2004)	バイオインフォマティクス関連データベース整備 ※ミレニアムPJ	経済産業省	30.26億円	<ul style="list-style-type: none"> 中核機関：生物情報解析研究センター（バイオ産業情報化コンソーシアム (JBIC) と産業技術総合研究所の共同センター） ヒトゲノム関連データベースと解析用ソフトウェアを統合利用できるシステムの整備 	<ul style="list-style-type: none"> ヒト完全長cDNA統合データベース (H-InvDB) の構築、公開 (http://h-invitational.jp/hinv/ahg-db/index_ja.jsp) JBIC バイオDBシステムにて20以上の公共データベースの一括検索・比較・解析サービスを公開

年度	事業名	事業主体	予算	概要	成果
平成17(2005)- 平成19(2007)	ゲノム情報統合プロジェクト	経済産業省	15.81 億円	中核機関：生物情報解析研究センター ・ヒトの全遺伝子を対象とした統合データベース作成	<ul style="list-style-type: none"> ・H-InvDB の更新及び選択的スプライシングの情報整備 ・DNA 多型、タンパク質間相互作用 (PPI)、比較ゲノム等のデータベース公開
平成20(2008)- 平成22(2010)	統合データベースプロジェクト	経済産業省	1.9 億円	実施機関：JBiC、BIRC <ul style="list-style-type: none"> ・成果物情報を提供するポータルサイト構築 ・H-InvDB を基礎とした関連研究成果を連携利用できるシステムの構築 	<ul style="list-style-type: none"> ・ポータルサイト MEDALS (http://medals.jp/) 公開、横断検索サービス提供、成果物 129 件掲載、年間約 11 万超アクセス (平成 22 年 1 -12 月) ・文部科学省統合 DB プロジェクトと連携し、250 種類のデータベースに対する相互検索サービスを提供 ・H-InvDB に LEGENDA (PubMed から遺伝子・疾患・化合物の関係についての知識を集めたデータベース) を統合 ・データベース間のリンクを自動で更新するリンク自動管理システムの開発
平成23(2011)-	ライフサイエンスデータベースプロジェクト	経済産業省	0.33 億円 (平成 23 年度)	実施機関：BIRC <ul style="list-style-type: none"> ・MEDALS の更新、維持 ・NBDC との連携業務 (ポータル構築、横断検索、アーカイブ構築) 	<ul style="list-style-type: none"> ・4 省合同ポータルサイト「integbio.jp」(インテグバイオ、http://integbio.jp/) 共同開設 ・現在 MEDALS は、データベース 73 件、解析ツール 70 件を公開、53 件のプロジェクトをカバー

(3) 農林水産省

年度	事業名	事業主体	予算	概要	成果
平成18(2006)- 平成22(2010)	農林水産生物ゲノム情報統合データベースの構築	農林水産省	27.67 億円	実施機関：農業生物資源研究所 <ul style="list-style-type: none"> ・イネ、カイコ、家畜などの農林水産生物のゲノムや遺伝子情報を統合 	<ul style="list-style-type: none"> ・農林水産生物ゲノム情報統合データベース (AgriTOGO、http://togo.dna.affrc.go.jp/) を公開、イネ 26 種類、カイコ 8 種類、ブタ 6 種類の生物ゲノム情報を提供、年間 5,900 万アクセス(平成 21 年度)
平成23(2011)-	画期的な農畜産物作出のためのゲノム情報データベースの整備	農林水産省	3.43 億円 (平成 23 年度)	実施機関：農業生物資源研究所 <ul style="list-style-type: none"> ・大容量ストレージと高速ネットワークの整備 ・表現型と配列の関連を大規模解析するシステムの開発 ・NBDC との連携業務 (横断検索、アーカイブ構築) 	<ul style="list-style-type: none"> ・4 省合同ポータルサイト「integbio.jp」(インテグバイオ、http://integbio.jp/) 共同開設

(4) 厚生労働省

年度	事業名	事業主体	予算	概要	成果
平成12(2000)- 平成15(2003)	遺伝子解析による疾病対策・創薬推進事業 ※ミレニアムPJ	厚生省	138億円(平成13年度-平成14年度)	実施機関：国立がんセンターほか6機関 ・アルツハイマー、がん、糖尿病、高血圧、喘息及び薬理遺伝学に関する遺伝子情報を統合	・疾患ゲノムデータベース (GeMDBJ) の公開 (https://gemdbj.nibio.go.jp/dgdb/)
平成14(2002)- 平成23(2011)	・トキシコゲノミクスプロジェクト (平成14年度-平成18年度) ・トキシコゲノミクス・インフォマティクスプロジェクト (平成19年度-平成23年度)	医薬基盤研究所、国立医薬品食品衛生研究所、製薬企業	厚生労働科学研究費補助金27.95億円+企業からの共同研究費25.83億円(平成14年度-平成18年度)	・遺伝子発現解析によるゲノムレベルでの毒性発現機構解明 ・化合物の遺伝子への影響に関するデータベースの構築による副作用の少ない医薬品開発の支援	・世界最大規模の毒性データベース OpenTG-GATEs (http://toxico.nibio.go.jp/) にて170化合物に係るデータと病理所見等のフェノタイプデータを公開
平成23(2011)-	医薬基盤研究所データベースの統合化と外部連携	医薬基盤研究所	NBDCからの共同研究費0.1億円(平成23年度)	・医薬基盤研究所で公開している8つのデータベースに加えて、NBDCの横断検索との相互運用	・300以上の創薬・疾患関連データベースを対象とした横断検索システム Sagace を公開 (http://sagace.nibio.go.jp/) ・4省合同ポータルサイト「integbio.jp」(インテグバイオ、 http://integbio.jp/) 共同開設

5. 日本と欧米や中国との比較

	日本		米国	欧州	中国	
	ROIS		NBDC/JST	NCBI	EBI	BGI
	DBCLS	DDBJ				
組織形態	ライフサイエンス分野におけるデータベースの利便性や付加価値の向上に関する研究開発を担う我が国唯一の機関	機構傘下の国立遺伝学研究所の附属施設 「生命情報学」の我が国における研究拠点 我が国を代表するDNAデータベースを運営	DB基盤技術と分野別統合化の委託機関を公募し、ライフサイエンスデータベース統合推進事業を推進 研究部門と事務局で構成	NIH傘下のNLMの付属機関 分子生物学分野を支援するソフトの提供と計算機を利用した基礎研究機関	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	ヒトや動植物、微生物のゲノム解析研究を手がけるDNA解析研究機関
組織の持続性	予算の9割近くをNBDCからの時限付委託費により運営	国立遺伝学研究所の運営費交付金により運営	JSTの運営費交付金(ライフサイエンスデータベース統合推進事業)により運営	根拠法：Public Law 100-607	費用の半分は20か国の公的研究資金で運営される EMBLから提供 残りは、ウェルカム財団、NIH、UK Research Councilsの資金等	中国科学院より施設及び設立資金を提供
予算	4.4億円 (平成23(2011)年度)	12億円 (平成23(2011)年度)	17億円(委託費含む) (平成23(2011)年度)	72億円(\$87.3M、平成22(2010)年、前年度比4.5%増)	>43億円(>€40M、平成22(2010)年)	\$450K(平成11(1999)年)、\$9M(平成13(2001)年RiceGenomeの解読時)、\$1.5M(平成21(2009)年深圳移転時)、\$3M/年等の助成金 年\$1.5B/10年の貸付 \$45M(平成21(2009)年)、\$180M(平成22(2010)年予想)の収入 (いずれも人民元による)

	日本		米国	欧州	中国	
	ROIS		NBDC/JST	NCBI	EBI	BGI
	DBCLS	DDBJ				
人員	28名（事務部門含） 任期付雇用職員のみ	62名（事務部門含）（サ ービス/研究/その他 =79/12/9%）	19名（NBDCスタッフと して、兼務含む）	約600名（平成22(2010) 年推定、うち正規職員約 250名）（サービス/研究 /その他=55/30/15%）	約500名（平成22(2010) 年、原則有期雇用、英国 が40%を占めるが、多岐 にわたる国から参加）（サ ービス/研究/その他 =66/22/12%）	500名以上（内バイオイン フォマティクススタッフ 数：約200名）
サービ スの 概要	統合データベースの構築 及び高度な検索技術を実 現するための基盤技術開 発	国際塩基配列データベー スの共同構築と運営	我が国を代表するナショ ナルセンターに向けたデ ータベースの統合を推進	配列情報データの標準配 列(RefSeq)の提供や Entrezによる統合デー タベース、各種解析ソフ ト提供の世界的な中心 アクセス数：4,000万/日	タンパク質配列を基礎 とした機能情報 (UniProt、InterPro)や、 真核生物のゲノム情報に GWAS等をアノテートし た統合サービス (Ensembl)及び ChEMBLなどの低分子 薬物情報も含む	大規模ゲノムシーケン シング、効率的バイオイ ンフォマティクス解析、 革新的遺伝子医療を基盤 としたプラットフォームの 構築
特色	RDF技術を用いた分散 型の高度なデータベース 統合、大規模データの利 用技術開発、良質な日本 語コンテンツの作成、国 内外関係機関との連携	塩基配列バンク構築、国 プロDB構築の受け皿 タンパク構造(GTOP)や 遺伝子発現(CIBEX)など 独自データベース開発	BIRD、文部科学省統合 DBプロジェクトの成果 を継承 基幹DB支援、新規DB開 発支援、バイオインフォ マティクス研究支援	塩基配列バンク構築、 PubMed等文献情報と配 列等の分子情報の提供と Entrezによる統合、 BLAST等解析ツールの 充実、データベース間の 連携	塩基配列バンク構築、タ ンパク配列を基礎とした 機能情報、ゲノム情報の 統合サービス、ウェブサ ービス等へのイニシアテ ィブ、タンパク情報の充 実	中国国内5拠点（深圳 香 港 北京 杭州、武漢）、国 際4拠点（BGI Americas、 BGI Europe、BGI Japan、BGI Hong Kong、 いずれも2010から）で事 業を展開
提供 デー タ ベース	種々の事業成果DBを横 断検索等により一括でき るように提供	20種類	種々の事業成果DBを横 断検索等により一括利 用できるように提供	35種類 （ほとんどが基幹DB相 当）	60種類 （うち15種類が基幹DB）	—

	日本		米国	欧州	中国	
	ROIS		NBDC/JST	NCBI	EBI	BGI
	DBCLS	DDBJ				
国内プロジェクトとの連携	NBDC及びNBDC委託機関との連携	ゲノムネットワークなど他プロジェクトに参画	統合化推進プログラムにより国内の代表的データベースの構築・高度化を支援	スタッフは、塩基配列解析、遺伝子同定、遺伝子発現に関する実験的解析において、NIH の他機関と協力	EU、NIH などの研究資金を得てプロジェクトに参加 英国のウェルカム財団、サンガー研究所と連携	ヒトゲノムプロジェクト、国際HapMap計画に参画
自前の研究機能	研究課題ごとに7つのチームに分かれ、それぞれが連携しながら基盤技術研究開発を中心に実施	生命情報・DDBJ 研究センターの4研究室で実施	戦略立案やデータの公開、国際連携に係る研究を一部実施	基礎研究グループは、Computational Biology Branch の中にあり、70名のsenior scientist、staff scientist、research fellow、postdoctoral fellow からなる	バイオインフォマティクスの研究グループが16ある EBI 独自のデータベースや機能サービスを担うグループを含む	1,000動植物参照ゲノム、10,000微生物ゲノム、単一性遺伝子疾患・複雑疾患の原因となる一般・希少変異の発見などを含めたデジタルライブラリーの構築
人材養成・教育機関との連携	NBDCとの共催による講習会等の実施 リサーチアシスタントを積極的に採用し研究開発に参加させることによる人材養成	総合研究大学院大学傘下の研究機関として博士課程教育を実施 実習付の講習会の実施	BIRDでの人材養成講座及び、DBCLSとの共催による講習会等の開催	"NCBI Discovery Workshops" (2日間の実習つき講習会) を年7回 (NIHで3回、他地区で4回) 開催 その他テーマごとのミニコース開催、教材提供等	学位 (PhD) 取得を目指す学生から独立した研究者に対する種々のコースを提供 海外でも講習会を実施	—
その他特記事項	文部科学省統合DBプロジェクトでの成果の一部をNBDCに移管し、一体運用に取り組む	NCBIのGenbank、EBIのEMBL-bankとの三極で国際塩基配列DBを構築	平成23(2011)年度に発足CSTP/ライフPT/統合DBタスクフォース報告書による統合DB整備のロードマップに沿った施策 他省との連携も主な課題として取り組む	NLMには外部への研究資金配布機能がある NIHの研究資金により実施された研究に由来する論文やデータについて、受け皿をNCBI やNLMが用意し、公開・共有を一部義務化	企業へ最先端の技術を普及することや企業からの寄付を得るための仕組みを整備 ELIXIRをはじめ、EUが資金提供しているいくつかのプロジェクトのコア機関	シーケンシング能力： 5 TBase/日、計算機能力：コア数50,000、総メモリ200TB、ディスク容量1,000 PB

6. 諸外国の人材養成の状況

大学や専門教育機関における教育プログラム数

(単位: 件)

エリア	取得できる学位等				主な教育機関等
	学士	修士	博士	その他	
米国及びカナダ	20	40	43	18	Universite de Montreal(カナダ)、Indiana University、Ohio University、Oregon Health & Science University、University of California 各校、University of Delaware
欧州	3	54	4	7	英国が最も多く、ついでドイツ、スペインが多い 全体として大学が多いが、ドイツでは Max Planck もコースを提供
アジア	6	12	3	2	ほとんどがインドの大学及び研究機関 その他では韓国の研究機関が 1 件
その他 (アフリカ、中東、豪州、南米)	9	6	4	1	オーストラリアでは各大学、南アフリカでは South African National Bioinformatics Institute、イスラエルでは Weizmann Institute of Science が中心