ライフサイエンスDBの在り方について①　ー経緯ー

第112回
ライフサイエンス委員会
令和6年2月16日
資料3
文部科学省

1．ライフサイエンスデータベース統合推進事業の経緯と取組状況

（1）ライフサイエンスデータベース統合推進事業の経緯

ライフサイエンスPT統合ＤＢタスクフォース報告書（平成21年５月・総合科学技術会議）において、『我が国における恒久的な統合データベース整備に向けてのロードマップとして、平成23年度以降の現実的体制を第一段階として、「統合データベースセンター（仮称）」（センター）の整備を行い、その後、第二段階として、整備した体制の強化を図りつつ、我が国として目指すべき統合データベースに相応しいセンター機能が発揮できる体制を構築していくこととする』とされた。

これを受けて同23年、JSTにNational Bioscience Database Center（NBDC）を設置することし、ライフサイエンスデータベース統合推進事業を開始した。その後、第二段階においても、引き続きNBDCを中心とした現行の体制で推進していくことが、平成25年１月に「総合科学技術会議ライフイノベーション戦略協議懇談会」で了承された。

【平成21年度当時、センターに求められた機能】
データベース統合に必要な調査、データベースの統合に必要な標準化、システムの構築・維持・管理、ポータルサイトの構築、データベースの受入れ・管理・更新、データベースの品質管理、各省等のデータベースとのネットワークの構築、海外との連携、データベースの統合化や高度な検索等、統合的利用のための技術開発

JSTの令和２年度における業務実績評価において、「ライフサイエンスデータベース統合推進事業については、NBDC発足から10年を迎えたことから、ポータルサイト運用、データベース統合、基盤技術開発の各取組に関する今後の進め方について、これまでの成果や課題を踏まえて検討することを求めたい。」とされた。
また、ライフサイエンス委員会における今後のライフサイエンス研究の在り方に関する議論、バイオ戦略の見直しや健康・医療戦略等の政府戦略の検討を見据え、ライフサイエンスDBの在り方について、今後の方向性を検討することとしたい。

ライフサイエンスDBの在り方について② －主な成果１－

文部科学省

（２）ライフサイエンスデータベース統合推進事業の主な成果

NBDCにおいて、上記の方針に基づき、本事業で取り組んできた主な成果は以下のとおり。

① 統合化推進プログラム：ファンディングによる統合データベース（DB）の整備

	PDBj	ChIP-Atlas	KEGG MEDICUS	GlyCosmos Portal	MicrobeDB.jp	Plant GARDEN	jPOST
概要	日米欧の三極で共同運営するタンパク質立体構造のDB	エピゲノム(ChIP-Seq)データを再解析・整理したDB	ゲノムと疾患・医薬品とその作用を関連付けたDB	糖鎖構造や糖鎖合成遺伝子等の糖鎖関連情報を統合したDB	微生物の遺伝子・分類・生育環境等の情報を統合したDB	さまざまな植物のゲノムや遺伝子情報をまとめたDB	生物種･翻訳後修飾･絶対発現量の横断的統合プロテオームDB
データ登録件数	約21万4千	約37万6千	約1万9千	約43万8千	約190万	約1167万	約2千※プロジェクト件数
利用状況（月平均アクセス数）	約18万	約1万2千	約600万	約9千8百	約3千4百	約2千9百	約1万8千
論文数	35	7	3	7	-	3	7
論文被引用数※	4,772	550	6,023	120	-	51	1,216

※登録データの論文引用に加え、研究構想の段階においても数多くのライフサイエンス研究で活用されていることが想定される。

・製薬会社がPDBjのタンパク質構造情報を用いて、新型コロナウイルス感染症の新規経口治療薬を開発。
・ChIP-Atlasのデータを使って解析し、心房細動に強く関与していると思われる因子（ESRRG）を発見。

② ポータルサイトの構築・各種DBサービスの提供

	DBカタログ	DB横断検索	DBアーカイブ	RDFポータル	TogoVar	ヒトDB
概要	生命科学系DBのカタログ化	生命科学系DB790件の横断検索サービス	寄託DB153件のアーカイブ化し公開するサービス	RDF形式で統一したDB40件を集積したポータル	ゲノムデータからバリアントを集積したDB	ゲノム情報等研究データのプラットフォーム
利用状況（月平均ユーザー数）	約8千	約2万	約2万	約3千	約1千	236※累計利用申請数

・ヒトDBに掲載されているJ-ADNI研究のデータセットは多数の利用申請があり、アルツハイマー病の研究で広く活用され、論文が多数発表されている。

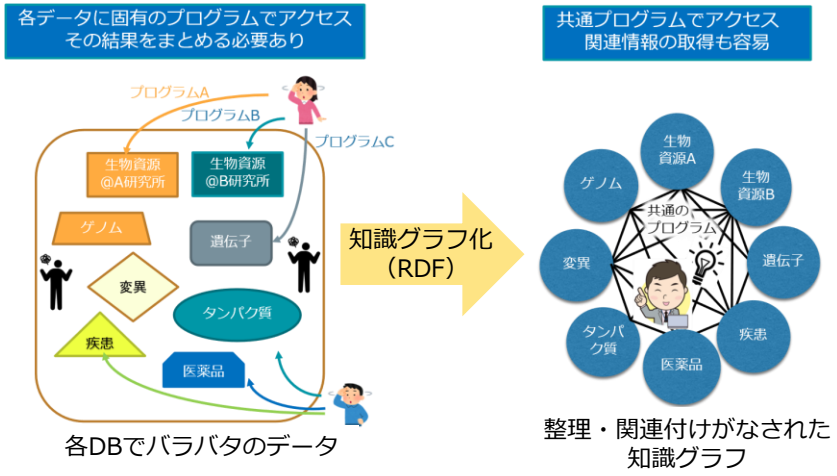
ライフサイエンスDBの在り方について②　ー主な成果２ー

文部科学省

③DB統合化のための各種基盤技術の開発

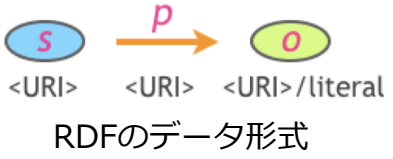
○Resource Description　Framework（RDF）によるDB間データの統合化

- 国内外で取得される多種多様なデータを入手・利用するために必要なデータ基盤整備として、DB登録データの記述にRDFを採用し、統合データの整備を推進。
- RDFは、関連データを検索するために共通化できるデータ形式で、W3Cによって国際標準化されている。そのため、フォーマットの共通化やデータ統合に適している。
- PDBj等のファンディングしたDBやDDBJを含め100件を超えるDBのデータをRDF化。
- RDFによるデータ統合で個別データ（遺伝子、タンパク質、化合物等）と他のオミクスデータが下図のように知識グラフ化されており、国内外のDBの効率的な構築や製薬企業における解析環境の構築に貢献。

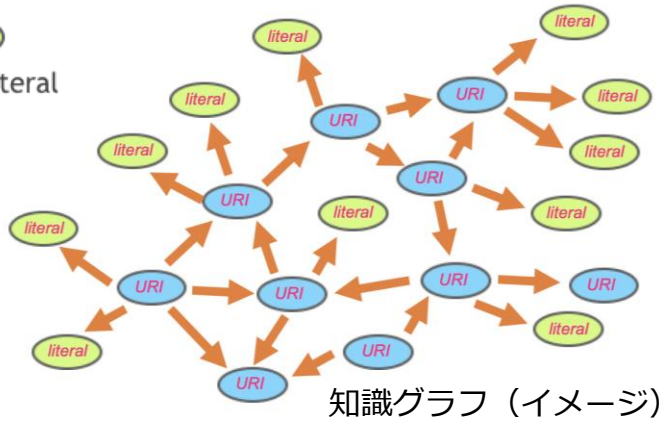


各DBでバラバタのデータ

整理・関連付けがなされた
知識グラフ



RDFのデータ形式



知識グラフ（イメージ）

○TogoVar（日本人ゲノム情報解析など）等のアプリケーション開発

- プロジェクト横断的なヒトゲノムデータ活用のため、様々なゲノムデータからバリアントを集約した「TogoVar」を開発。約8.6億のバリアントを収録。新規に22のバリアントを発見し、日本人PCD患者特徴的にDRC1遺伝子のコピー数多型が多いことの解明に貢献した。
- 統合データ活用インターフェースとして「TogoDX」を開発。統合された多種多様なデータを利用者の発想で柔軟に組み合わせることができる。
- そのほか、統合データ利活用促進に向けたアプリケーションを３件開発。

ライフサイエンスDBの在り方について③　ー課題ー

文部科学省

2．ライフサイエンスDBにおける課題

事業開始から10年が経過し、ライフサイエンスDBの現状を改めて振り返ると、以下の課題がある。

（1） DBの安定的な維持・管理

大学等による組織的な支援が確立されていない現状において、JSTのNBDC事業推進部のファンディングが途絶えると、国際的な地位を確立している多くのDBを財源不足等により運営できなくなる可能性が高く、安定的に運営できるDB整備からは程遠い現状。特に長期的に実績のあるDBを維持・管理できなくなると、日本の国際的なプレゼンス低下にも繋がる可能性があり、研究基盤として維持・管理できる体制や方策が必要。一方で新しい分野のDBの開発は、研究コミュニティのニーズに応じて発展するものなので、競争的な支援が望ましいと考えられる。

（2）DBの統合化

多様な関連データを整理し統合するためにRDF化を推進してきたが、開発段階の途上にあり、いくつかの課題が存在している。特に、複雑に繋がったデータを利用するインターフェースにはまだ課題が残っており、ライフサイエンス研究者では活用しにくい現状を改善する必要がある。

急速に発展するAI技術や大規模言語モデルを活用するなど、新たなデータ統合検索技術を開発し、ライフサイエンス研究者がより統合的にデータを活用できる仕組みを構築することが急務。また、RDF化されたデータをどのように利活用していくかの視点も重要。

今後のライフサイエンスの研究動向を考慮すると、核酸、タンパク質、代謝、細胞、組織、疾患等の各階層データを整理・関連付けて（統合的に）検索・解析できるツールが益々重要。

（3）人材育成

研究基盤としてDBの活用が必須の中で、アカデミアにおけるキャリアパスの未確立や任期付き雇用等により、民間企業へ人材が流出しており、DBの開発・維持・管理やキュレーションを担う研究者の人材が不足。

DBを利活用・発展させ、ライフサイエンス研究における国際競争力を維持するためには、生命科学と情報科学の両分野における知識が必要なバイオインフォマティクス研究者の人材育成が急務。

ライフサイエンスDBの在り方について④　ー今後の方向性(案)ー

文部科学省

3．今後の方向性（案）

（1）統合化推進プログラムによるDBの継続的支援

現在、DBの開発・維持・管理に特化したファンディングは他に存在しないため、当該プログラムによる継続的支援は必須。支援に当たっては、DBの種類（一次DB、二次DB）、分野（ゲノム、タンパク質、代謝等）、開発段階（萌芽期、定常期、拡大期）に応じたファンディングの継続が必要ではないか。その際、以下（３）と連携した統合利用に向けた仕組みが必要ではないか。

（2）DBの一元的管理

論文投稿時に必要となるデータを公平に受理・登録・整理し、国際連携のもと再利用可能な形で維持しており、データのアーカイブ機能を有する、一次DBは継続的かつ安定的に維持・管理することが必須。このため、運用コストや人件費等の観点で、国によるDBの一元的管理や財政的支援の在り方をライフサイエンス研究基盤として検討することが必要ではないか。

（3）DB高度化のための基盤技術開発

① 検索インターフェースに係る技術開発

大規模言語モデル等のAI技術を活用したDBの統合的検索技術、大規模データの利用技術などのDB高度化のための先端的技術開発に対する支援を開始し、開発した技術の実用性やそれによる研究成果を発信、普及することが必要ではないか。

② ①と連携した統合データの充実化

DBの統合化には、標準化やID・用語の統一、メタデータ整備、データ形式の統一などが必要。基盤的な検索・利用技術と組み合わせ、これらを整理・関連付けさせた知識グラフ（学習データ）の充実化を引き続き実施。論文情報や研究データ取得から知識グラフ構築までを自動化する技術も開発し、データ統合の効率化も必要ではないか。

（4）バイオインフォマティクス人材の育成

新しいバイオインフォマティクス人材として、AIを含む情報科学系のデータ技術を応用できるライフサイエンス研究者を戦略的に育成。生命科学のみならず人文・社会学にも通じた分野横断的な人材もあわせて育成。

ライフサイエンスデータベース統合推進事業

令和６年度予算額（案）　：1,311百万円
（前年度予算額　　：1,311百万円）
※運営費交付金中の推計額

参考

背景・課題

◯ 個々のプロジェクトのデータベース作成や終了後の運用・管理の難しさ等、基盤としてのデータ整備に課題がある中で、国として利活用を促進する方針。

◯ 爆発的に増加するデータや知識の利活用を推進するためには研究データや利活用に係るニーズの多様化への対応が喫緊の課題であり、個別に作成されたデータベースを利用者の新たな知識発見や課題解決に資するよう、連携させたデータ基盤を整備することが必要。

【成長戦略等における記載】「統合イノベーション戦略2023」、第2章、2、(2)新たな研究システムの構築（オープンサイエンスとデータ駆動型研究等の推進）「最先端のデータ駆動型研究、AI駆動型研究の実施を促進するとともに、これらの新たな研究手法を支える情報科学技術の研究を進める。」

事業概要

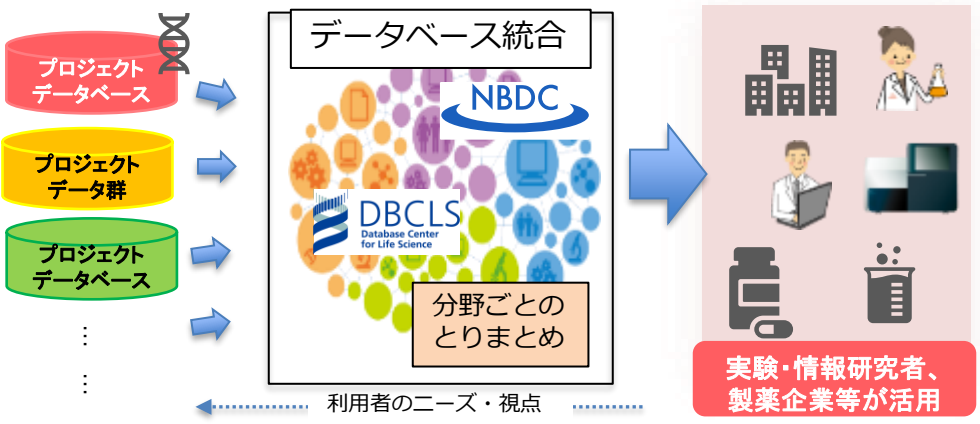
【事業の目的・目標】

我が国におけるライフサイエンス研究の成果が、広く研究者コミュニティに共有かつ活用されることにより、基礎研究や産業応用研究につながる研究開発を含むライフサイエンス研究全体が活性化されることを目的とする。

【事業概要・イメージ】

・我が国のライフサイエンス分野のデータベース統合にかかる実務や研究開発をNBDC事業推進部が推進。

・産出されたデータを利用者の視点に立って統合化し、効率よく研究者、産業界、さらには国民に還元していくための統合的なデータベースの構築・利活用促進と、それに関連したバイオインフォマティクス研究の推進。



【事業スキーム】

- ✓ 支援対象機関：大学、国立研究開発法人等（11課題（令和5年度））
- ✓ 事業規模：13.1億円（令和5年度）
- ✓ 事業期間：平成23年度～

国内の大学等の関係機関と連携してデータベースの統合を推進

国 →（運営費交付金）→ JST →（委託）→ 大学・国立研究開発法人 等

① 統合化推進プログラム：ライフサイエンスに関する**国内外のデータを統合的に扱うためのデータベースの開発**を推進

② ポータルサイト運用・事業運営：**データベース検索・継続的公開**等、データ利活用促進

③ 基盤技術開発：**データベース統合に向けた基盤的な技術の開発**

【これまでの成果】

・日本人ゲノム多様性統合データベース「TogoVar」を構築・運用
平成30年6月にヒトゲノムの多様性（バリアント）頻度情報を集計し、公開。令和4年度末までの全データ収録件数は約8.6億件（令和3年度の約4倍）に増加。

・プロテオームデータベース（jPOST）の国際的知名度向上
統合化推進プログラムで開発したjPOSTは、欧米を含む43ヶ国からのアクセスが8割を超過、令和4年度にはアクセス数が前年度比約2倍に増加。

【活用事例】

・製薬企業が、PDBjを用いて既知タンパク質の構造データから未知の受容体タンパク質の構造を予想→**新薬候補の発見**

・ChIP-Atlasを用いることで、利用者が自ら実験を行うことなく多種多様な細胞のデータを比較解析→**診断研究への貢献**

統合化推進プログラムの成果例

参考

文部科学省

PDBj

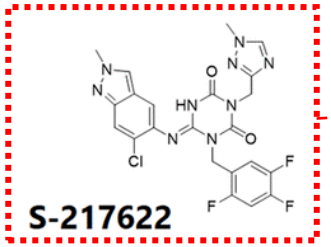
- PDBjは、米国や欧州の研究機関と連携し、タンパク質立体構造の世界標準データレポジトリ（PDB）を運営している。そして、PDBjはアジアオセアニア地区の担当として、世界全体で約20万件あるデータのうち、24%を登録・公開してきた。
- 製薬企業が、**PDBjに登録されたタンパク質の構造情報を用いて、新型コロナウイルス感染症の新規経口治療薬を創製することに成功**し、厚生労働省に承認された。

＜研究開発のポイント＞

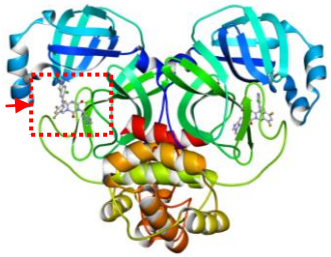
酵素は、タンパク質の立体構造に特異的な化合物等が結合することで活性が変化する。Structure-Based Drug Design（SBDD）は、タンパク質立体構造情報を活用し、タンパク質の不活性化に必要な特異的化合物を設計する技術、新たな医薬品の候補となる化合物を効率的にスクリーニングできる。

PDBにはSBDDに必須の正確なタンパク質立体構造情報が膨大に登録されており、創薬研究への活用が劇的に進んでいる。

製薬企業において、新型コロナウイルスが自己複製する際に重要となる、酵素の活性部位に結合する化合物の創製を行う前段階として、SBDDによるスクリーニングを実施し、早期の新規化合物創製に成功した。



PDBを活用して新たに創製した、新型コロナウイルスの経口治療薬：化合物S-217622



新型コロナウイルスのタンパク質複合体

Yamamoto, S.,Yamane, J.,Tachibana, Y. (2021) doi: 10.2210/pdb7vu6/pdb
Unoh,Y et al. (2022) *J Med Chem.* 65(9):6499-6512 doi: 10.1021/acs.jmedchem.2c00117.

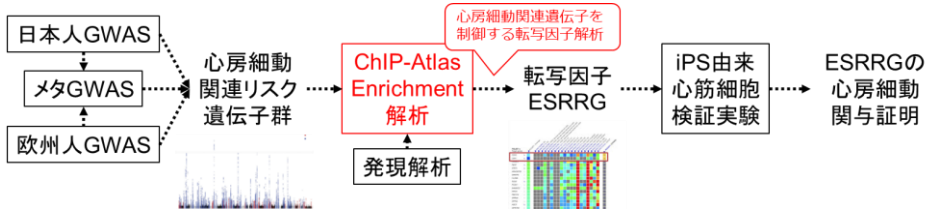
ChIP-Atlas

- ChIP-Atlasは、ヒト・マウス・ラット・ハエ・線虫・酵母の公開されているほぼ全てのエピゲノミクスデータ（※1）を再解析し、比較解析可能な形で公開しているデータベース。
  ※1 エピゲノムとは、遺伝子に付加されたメチル化等の情報を指し、エピゲノミクスとは、遺伝子配列を変えることなくエピゲノムにより遺伝子の働きを制御する仕組みを研究することをいう。
- **ChIP-Atlasを使ってヒト疾患や生体機能に関係する遺伝子の発現制御メカニズムの解明が可能**になり、様々な研究へ貢献している。
- 京都大学と理研は、ChIP-Atlasのデータを使って解析し、心房細動に強く関与していると思われる因子（ESRRG）を発見した。この仮説に基づき、iPS細胞から作製した心筋細胞を使って検証実験を行い、心筋細胞においてその因子が心房細動の発症に決定的に関与していることを証明した。

＜研究開発のポイント＞

GWASデータ（※2）とChIP-Atlas解析により疾患の発症メカニズムを予測し、その仮説を検証実験で証明した。

※2 特定の疾患や体質などと関連する遺伝的な特徴を見つけ出すために行われる手法。多数の人々のゲノム情報を集めて網羅的に解析することで、一塩基多型（SNP）という遺伝情報のわずかな違いを探し出し、疾患や体質に関連する因子を特定できる。



Oki, S., Ohta, T. (2015): ChIP-Atlas. https://chip-atlas.org
Miyazawa, K. et al. (2023) *Nat Genet.* 55(2):187-197. doi: 10.1038/s41588-022-01284-9.