資料 2
AI for Science を支える研究データの
管理・利活用と流通の在り方
ワーキンググループ(第５回)
令和８年５月22日

ライフサイエンス分野のデータ基盤

情報・システム研究機構　国立遺伝学研究所
バイオデータ研究拠点（BSI）ライフサイエンス統合データベース部門（DBCLS）

五斗進・片山俊明

ライフサイエンス研究データとDBの特徴

データの多様性

- **多様な研究分野**：生物学、生化学、分子生物学、医学、薬学、生態学、農学、…
- **多様な研究対象**：ゲノム、遺伝子、タンパク質、遺伝子発現、低分子化合物、細胞、生物、環境、疾患、…
- **多様な計測装置**：DNAシーケンサー、質量分析、画像解析、…

大規模なデータ

- **大規模な計測データ**
  - 塩基配列データベース：100PB クラスのストレージが必要
  - プロテオームデータベース：1回の実験で数10TBのデータ
- **大規模な文献データ**
  - PubMed に 4,000 万件以上の論文

多種多様なDB

- **大小様々なデータベースが既に存在**

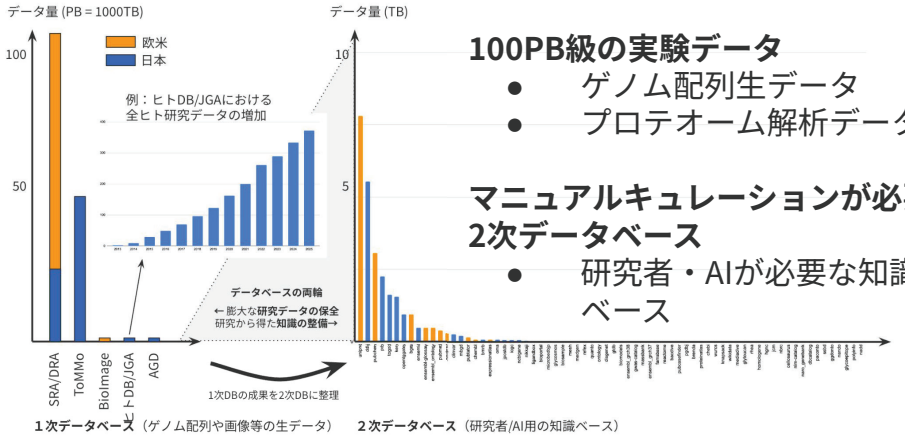
世界全体では**>7000DB**
- Integbio DBカタログ
- FairSharing
- Database Commons

そのうち1329個が日本のDB
（819DBは横断検索に対応）

Cross search

>1000 オントロジー
- Ontology Lookup Service
- BioPortal

OLS ONTOLOGY SEARCH　BioPortal



100PB級の実験データ
- ゲノム配列生データ
- プロテオーム解析データ

マニュアルキュレーションが必要な2次データベース
- 研究者・AIが必要な知識ベース

ライフサイエンス統合DBプロジェクトの立ち上げとその経緯

(2001-2010 BIRD) 2006- 統合DBプロジェクト　2007-DBCLS　2011-NBDC　2025-NLDP

目的：世界的に分散され、フォーマット、オントロジー、 IFがバラバラなDBを連携・統合して格段に使い易くすることにより、生命研究・バイオ産業の大幅な効率化を図る

・その手段として、以下を推進
- ・オープンサイエンスの推進（データの流通、データの保全、ヒトデータ共有ポリシー）
- ・統合に資する様々な技術開発
- ・カタログ、横断検索、アーカイブ、など統合の成果や各種DBのポータルサービス事業
  4省連携（文科省、厚労省、経産省、農水省）の枠組み
- ・ファンディングによる統合利用を意識した多様なDB構築（本格型、育成型）
- ・国際標準化、国際連携、国内連携、技術普及活動
- ・DB人材、バイオインフォ人材の育成

・ターゲット層
- ・ライフサイエンス研究者
  （基礎・応用生命研究者、医薬科学研究者、バイオインフォマティシャンなど）

FAIR原則をベースとしたデータの利活用促進

＊科学データ共有の基準としてのFAIR原則。Wilkinson, et al. (2016) Scientific Data.
DBCLSが主催している BioHackathon での議論が元になっている

Findable: データの所在を明らかにする

- Integbioデータベースカタログ（NBDC）：約2500件のDB

Accessible: データをアクセス可能にする

- データベース横断検索（NBDC）：819DB
- データベースアーカイブ（NBDC）：150DB
- 世界各地のデータベースセンター・リポジトリ
  - 米国NCBI、欧州EBI、スイスSIB、遺伝研DDBJ、京大GenomeNet/KEGG など
  - タンパク質立体構造、プロテオーム、グライコーム、メタボローム、ヒトデータベース

Interoperable: データを相互参照可能にする

Reusable: データを再利用可能にする

- データベースをダウンロードすると使える形になっている。
- データベース間で使用する用語やフォーマットを統一して使えるようにする。
- DBCLS での基盤技術開発。

統合プロジェクト、DBセンターの成果（一部）

・ライフサイエンス研究者向け AIツールなどの開発・ポータル事業

アプリケーション開発：ヒトDB(提供申請約1300, 利用申請約600 – 半数以上が海外から)、
TogoVar(20DB, 9.3億バリアント, 25.8万人)、
PubCaseFinder(15DB, 200万データ)、TogoTV(2311動画, 300万回視聴, 1.1万人登録)、
DBカタログ(2,572DB)、DB横断検索(819DB)、アーカイブ(157DB)、など

データセット整備：TogoID(114DB, 52億IDペア)、RDFポータル(70DB, 1600億トリプル)、
PubAnnotation(1,700万文献, 650注釈プロジェクト)、TogoDX(20DB, 65属性)、など

ツール開発：SPARQL-proxy、SPARQList、Grasp、RDF-config、TogoStanza、TogoWS、など

データセット整備やツール群開発は当初バイオインフォマティシャン用を想定していたが、その後 AI用の基盤になることが判明。現在AI利活用に不可欠の存在となった

・その他の活動

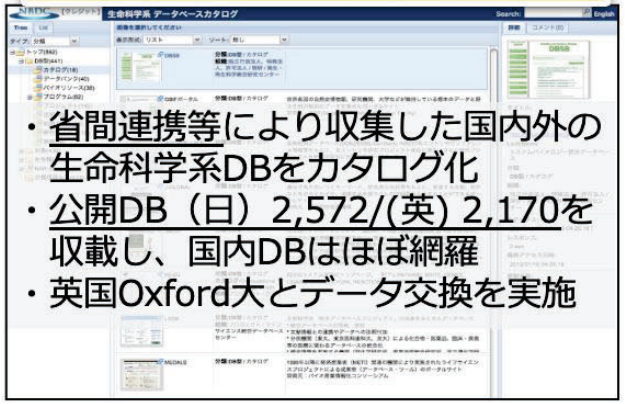
国際連携(BioHackathon15回100人, BLAH9回50人, グラフサミット6回30人, 20ヶ国)、
国内連携(バイオハッカソン16回80人, Togothon157回60人)、
トーゴーの日シンポ(500名)、AJCAS(5回, 1,700名)、など

・ファンディングにより構築された DB群

（本格型）PDBj, KEGG MEDICUS, jPOST, GlyCosmos, Shin-MassBank, SSBD, INTRARED, MicrobiomeDatahub
（育成型）ATTED-II, JoGo, DeepspaceDB, MIIB-AI, Cell IO, integMet, SSCV DB, PHi-C DB, Cura Toxii

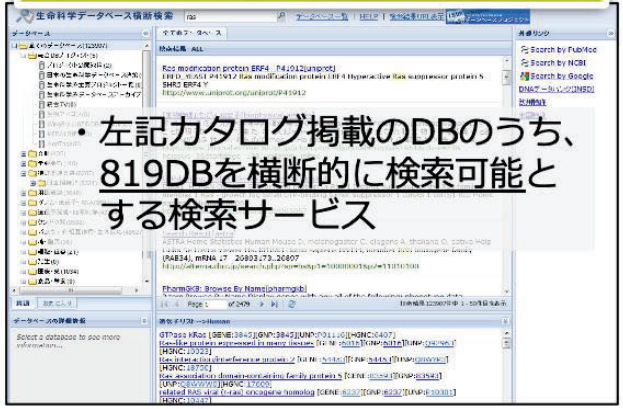
DBサービスの一例

生命科学系データベースカタログ



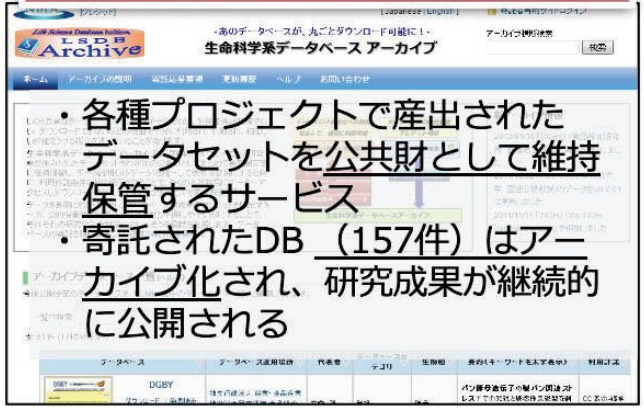
・省間連携等により収集した国内外の生命科学系DBをカタログ化
・公開DB（日）2,572/(英) 2,170を収載し、国内DBはほぼ網羅
・英国Oxford大とデータ交換を実施

生命科学データベース横断検索



・左記カタログ掲載のDBのうち、819DBを横断的に検索可能とする検索サービス

生命科学系データベースアーカイブ

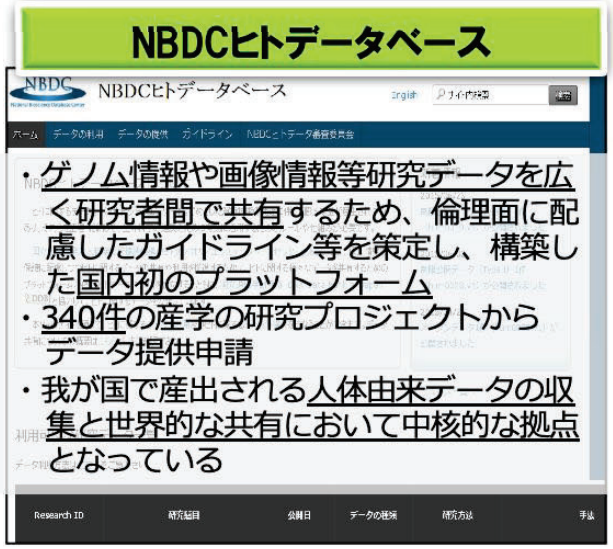


・各種プロジェクトで産出されたデータセットを公共財として維持保管するサービス
・寄託されたDB（157件）はアーカイブ化され、研究成果が継続的に公開される

RDFポータル

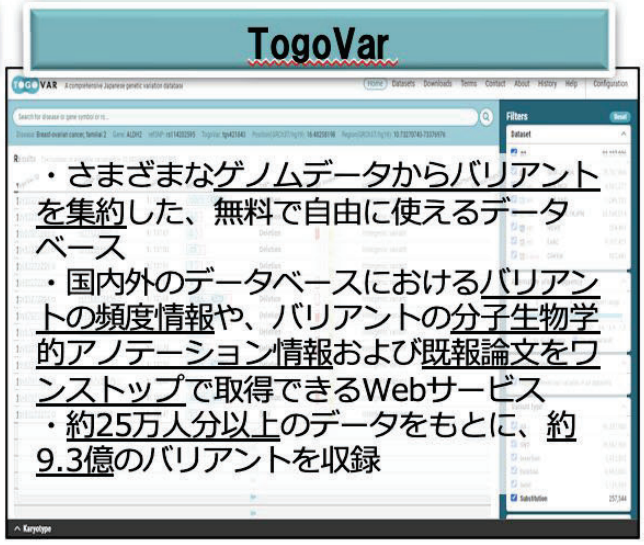
・分野横断的な研究促進に貢献するため、連携が容易で機械可読なRDF形式で統一したデータベースを集積したポータルサイト
・70件のRDF形式の生命科学データベースを用意

NBDCヒトデータベース



・ゲノム情報や画像情報等研究データを広く研究者間で共有するため、倫理面に配慮したガイドライン等を策定し、構築した国内初のプラットフォーム
・340件の産学の研究プロジェクトからデータ提供申請
・我が国で産出される人体由来データの収集と世界的な共有において中核的な拠点となっている

TogoVar



・さまざまなゲノムデータからバリアントを集約した、無料で自由に使えるデータベース
・国内外のデータベースにおけるバリアントの頻度情報や、バリアントの分子生物学的アノテーション情報および既報論文をワンストップで取得できるWebサービス
・約25万人分以上のデータをもとに、約9.3億のバリアントを収録

https://rdfportal.org/

2025年度追加分データ、全2200億トリプル

知識グラフ(RDF)で統合された主要な生命医科学DB

共通のオントロジーやID対応関係によってAIが解釈可能なデータセット

- 塩基配列とアノテーション
  - INSDC (DDBJ/DBCLS)
- ゲノム情報
  - Ensembl (EBI)
  - RefSeq (TogoGenome)
- アミノ酸配列とアノテーション
  - UniProt (SIB)
- タンパク質立体構造
  - PDB (PDBj)
  - BMRB (PDBj)
  - FAMSBASE (Chuo U)
- 化合物
  - PubChem (NCBI)
  - ChEMBL (EBI)
  - Nikkaji (JST)
- 遺伝子発現
  - RefEx,GTEx (DBCLS)
  - ExpressionAtlas (EBI)
- サンプル
  - BioSamples (EBI/DDBJ)
  - JCM (RIKEN)

モデル生物
- NBRP(一部)
酵素
- BRENDA

- 医科学 (Med2RDF)
  - ICGC, COSMIC, CIViC
  - DGIdb, OpenTG-Gates
  - ClinVar, dbSNP, dbVa
  - ExAC, gnomAD
  - HiNT, INstruct
- 糖鎖
  - GlyTouCan, GlycoEpitope, WURCS, GGDonto, PAConto
- プロテオーム
  - jPOST
  - The Human Protein Atlas
- パスウェイ
  - Reactome (EBI)
- その他
  - MeSH (NCBI)
  - BioModels (EBI)
  - MBGD (NIBB/DBCLS)
  - Quanto (DBCLS)
  - :

バリアント
- dbNSFP
- dbscSNV
がん
- TCGA

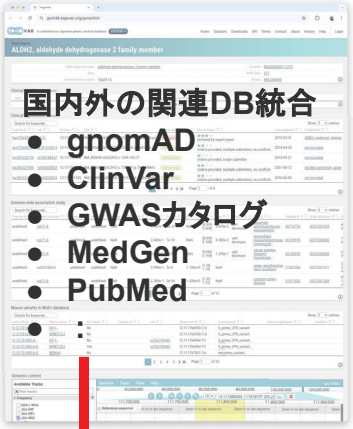
メタボローム
- MassBank

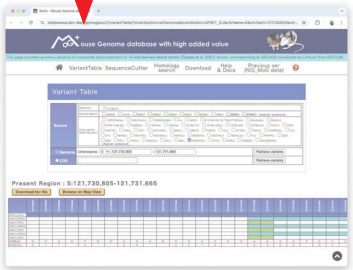
微生物・培地
- BacDive
- MediaDive
- AMR

統合DBとそれに基づく外部DBとの連携の一例

TogoVar: 日本人ゲノムバリアント頻度の統合DB

JoGo: ハプロタイプDB
(九州大学)



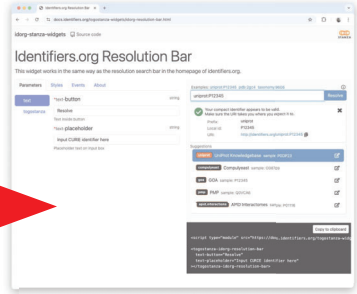


実験用モデルマウス検索
- ヒトと同じ変異を持つ

MoG+: モデルマウスDB
(RIKEN)

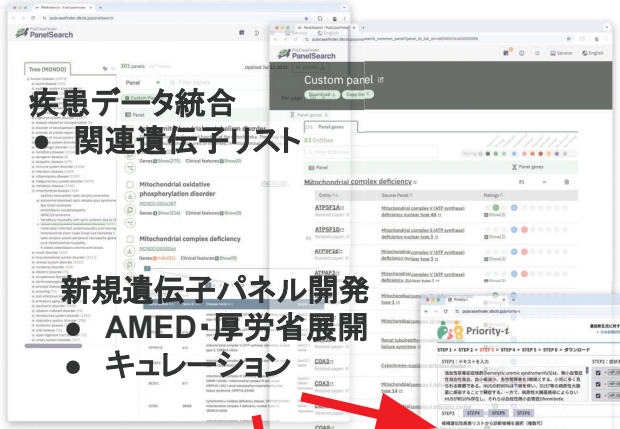
TogoID: DB間のID対応サービス

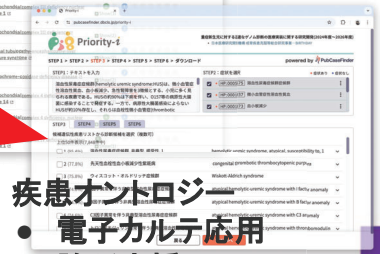
Id.org に変換機能提供
(欧州EBI)



DBCLS技術提供
(欧州EBI)

PubCaseFinder: 疾患検索統合DB





NanbyoData: 難病DB
Priority-i: 重症新生児ゲノム診断
(厚労省)

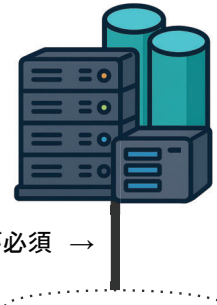
マルチモーダルな生命科学DBの統合と国際情勢

海外ではナショナルセンターが統合的なデータベースの整備とサービス開発を戦略的に実施

アジアでも、インド IBDC・中国 CNCB・韓国KOBICなどナショナルセンターの設立が進む。

国家生物信息中心 China National Center for Bioinformation
kobic Korea Bioinformation Center
IBDC Indian Biological Data Centre

ナショナル生命科学データベースセンター



データとスパコンは直結が必須 →

ゲノム・DNAデータ

NCBI・EBI
(>100PB)

日本の生命科学データを永続的に保全するDMPと提供申請・利用申請業務

ストレージと運用を維持するための国策が必要

全生命科学データの大部分を占める

欧米では100PB超のストレージを確保

DDBJ (25PB)
ヒトDB・JGA・AGD他は散在

ストレージ枯渇
中国等の台頭

マルチオミクスデータ

NBDC (1PB)

マルチモーダル・画像データ

RIKEN・PDBj

医科学・医療データ

AMED・ToMMo・BBJ・NCCなど

知識・文献データ

DBCLS (0.5PB)

この一部をAIが利用

AIの成果をDB化

AlphaFold

実験由来データとAI由来データの識別

DBの管理機関や省庁が分散し、データが散在していることと、永続的な運用が課題。予算人員に比して対応分野は広い。

欧米はナショナルセンターで国家レベルのデータマネージメントを実施、主要なデータベースが集約され相乗効果で付加価値創出や、製薬などの産業応用にも繋がっている。
欧州では Elixir 支援のもと EOSC-Life とも連携して進められている。

DDBJ+DBCLS
15億円・50人

DDBJ, DRA, JGA
BioProject, BioSample

AGD, CANNDs

GlyCosmos
Microbiome Datahub
JPOST
Mass
fantasia.bio
RDF Portal
TOGOVAR
KEGG
TOGO ID
PDBj Protein Data Bank Japan
SSBD
NanbyoData
dbTMM

NCBI
300億円・350人

GenBank, SRA, dbGaP
BioProject, BioSample

RefSeq
Taxonomy
dbSNP
dbVar
GEO Gene Expression Omnibus
Protein
PubChem
CDD
PubMed
MeSH
ClinicalTrials.gov
ClinVar
MedGen
GTR

EMBL-EBI
180億円・650人

ENA, (Federated) EGA
BioProject, BioSam

Ensembl
EVA
GWAS Catalog
MGnify
Rfam
ArrayExpress
PRIDE Proteomics IDEnt
IntAct
Rhea
MetaboLights
ChEMBL
ChEBI
Europe PMC
UniProt
Identifiers.org
reactome
PDBe Protein Data Bank in Europe
BioImage
EMDB Electron Microscopy Data Bank
EMPIAR

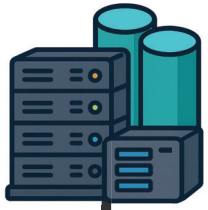
マルチモーダルな生命科学DBの統合と国際情勢

海外ではナショナルセンターが統合的なデータベースの整備とサービス開発を戦略的に実施

アジアでも、インド IBDC・中国 CNCB・韓国KOBICなどナショナルセンターの設立が進む。

国家生物信息中心 China National Center for Bioinformation
kobic Korea Bioinformation Center
IBDC Indian Biological Data Centre

ナショナル生命科学データベースセンター



データとスパコンは直結が必須 →

塩基配列など生データのリポジトリは国際連携で開発・運用

DBの管理機関や省庁が分散し、データが散在していることと、永続的な運用が課題。予算人員に比して対応分野は広い。

欧米はナショナルセンターで国家レベルのデータマネージメントを実施、主要なデータベースが集約され相乗効果で付加価値創出や、製薬などの産業応用にも繋がっている。
欧州では Elixir 支援のもと EOSC-Life とも連携して進められている。

ゲノム・DNAデータ

NCBI・EBI
(>100PB)

全生命科学データの大部分を占める

欧米では100PB超のストレージを確保

日本の生命科学データを永続的に保全する DMPと提供申請・利用申請業務

ストレージと運用を維持するための国策が必要

マルチオミクスデータ

DDBJ (25PB)
ヒトDB・JGA・AGD他は散在
NBDC (1PB)

ストレージ枯渇
中国等の台頭

マルチモーダル・画像データ

RIKEN・PDBj

この一部をAIが利用

医科学・医療データ

AMED・ToMMo・BBJ・NCCなど

AIの成果をDB化

知識・文献データ

DBCLS (0.5PB)

AlphaFold

実験由来データとAI由来データの識別

DDBJ+DBCLS
15億円・50人

DDBJ, DRA, JGA
BioProject, BioSample

AGD, CANNDs

GlyCosmos, Microbiome Datahub, JPOST, Mass, fantasia.bio, RDF Portal, TOGOVAR, KEGG, TOGO ID, PDBj Protein Data Bank Japan, SSBD, NanbyoData, dbTMM

NCBI
300億円・350人

GenBank, SRA, dbGaP
BioProject, BioSample

RefSeq, Taxonomy, dbSNP, dbVar, GEO Gene Expression Omnibus, Protein, PubChem, CDD, PubMed, MeSH, ClinicalTrials.gov, ClinVar, MedGen, GTR

EMBL-EBI
180億円・650人

ENA, (Federated) EGA
BioProject, BioSam

Ensembl, EVA, GWAS Catalog, MGnify, Rfam, ArrayExpress, PRIDE Proteomics IDEnt, IntAct, Rhea, MetaboLights, ChEMBL, ChEBI, Europe PMC, UniProt, Identifiers.org, reactome, PDBe Protein Data Bank in Europe, BioImage, EMDB Electron Microscopy Data Bank, EMPIAR

NII知識基盤（AI基盤モデル）との関係と期待

- **ライフサイエンス分野データ基盤と共通の基本理念**
  - オープンサイエンス
  - 研究透明性
  - データ利活用促進

- **ライフサイエンス分野データ基盤にある背景と現状**
  - データの多様性＆大規模性
  - 専門家によるアノテーション、キュレーション、オントロジーの必要性
  - 国際連携による各研究対象に対する生データ登録用リポジトリの運用
  - 大規模データのストレージと分野に特有の解析基盤の一体的提供

- **NII知識基盤への期待**
  - データベース運用の基盤となる標準ツール（認証、トレーサビリティ、セキュリティなど）や部品の提供
    - 共通認証機構の提供：Gakunin 以上の機能。
      - 海外の研究者への提供、研究者の要件確認など
    - 分野間データ連携検索の提供：メタデータの共通化、Data Sharing Policy のトップダウン的な適用
    - 共通大規模ストレージの提供：分野ごとの事情を反映できるような環境の必要性