BioData Science Initiative

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 バイオデータ研究拠点

トーゴーの日シンポジウム AIが研究する時代のデータベース

AIに対応するデータベースの在り方とデータ統合・解析基盤の技術開発

片山俊明 (ROIS/NIG/BSI/DBCLS)

ライフサイエンス統合データベース部門 部門長・特任教授

2026-10-05

人によって「DB」のイメージは異なる

盲人象を評す - 足：柱のようだ、尾：紐のようだ、鼻：管のようだ、耳：扇のようだ、腹：壁のようだ、牙：…
幅広くオールジャパンのニーズを聞き取り、それに応えていくことが必要

大規模データを
預かってくれる



生命情報科学
研究の基盤





セキュリティと
品質管理が重要



論文や調べもの
答えを探す場所



論文投稿には
DB登録が義務



AIの材料

生命科学 × AIの根幹となるデータベース運用

生命科学の研究成果を集積し、適切にキュレーションすることで、研究者にもAIにも利用可能なデータを創出

各ドメインDB

DBCLS

生命科学研究



データ生産

論文・データ

新分野のデータ

生命科学研究成果

- 論文発表
- データベース登録



データ利用申請

DB問合せ

生命科学的な問い

- 既存のファクト知識検索
- 元データにない知見
- データ統合による推論

データベース化

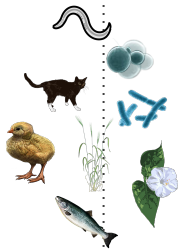
1次レポジトリ

アノテーション



キュレーション

2次データベース



ストレージ・DMP・人材の欠乏

貴重な研究データの永続的な保全

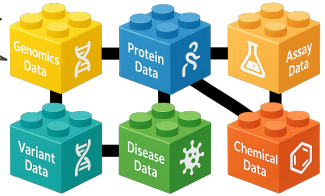
- 登録申請・利用申請業務
- 大規模ストレージ運用
- 解析環境の開発提供

データを分野別に規格化

- 専門家の整理で付加価値を創出
- 他の生物種での情報補完など

標準化・統合

2次利用に向けた専門家によるデータ整理



AIの性能は学習データの質に依存するため、「AIレディな高品質データベース」の継続的な整備

マルチモーダルなデータの標準化

- メタデータ
- データモデル・データ形式
- 標準語彙（オントロジー）
- リンク情報

基盤技術開発

- 高度なデータベース技術開発

研究利用・AI利用

データ利用の基盤を運用・開発中

科学データの取得・再利用

- 生データへのアクセス
- 統合データの機械学習
- 生物・AI由来データ選別

質問応答システム

- 自然言語問合せ
- QAタスクのデータ生成

AIエージェント

- MCPサーバ整備
- SPARQL自動生成

ナショナル DBに求められること

・国家予算で科学研究費を出す以上、得られた研究データを保全する恒久的なストレージと運用も国が責任を持つべき。
・膨大なデータは移動が困難なため計算環境がデータのそばに必要。
・生物のドメイン知識と情報技術をもつ専門家によるキュレーションの維持が重要。

DBセンターの理念とマルチモーダルな生命科学DBの統合

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（日本にはまだ無い）

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)

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

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

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

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。
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。

DDBJ+DBCLS
15億円・50人

DDBJ, DRA, JGA
BioProject, BioSample

AGD, CANNDs

GlyCosmos, Microbiome Datahub, JPOST, Mass, fanta.bio, RDF Portal, TOGOVAR, KegMol, TOGO ID, PDBj, SSBD, NanbyoData, dbTMM

NCBI
300億円・350人

GenBank, SRA, dbGaP
BioProject, BioSample

RefSeq, Taxonomy, dbSNP, dbVar, GEO, 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, IntAct, Rhea, MetaboLights, ChEMBL, ChEBI, Europe PMC, UniProt, Identifiers.org, reactome, PDBe, BioImage, EMDB, EMPIAR

ナショナル生命科学データベースセンターの必要性

生命科学研究を支える基盤が、主体的な国際連携と、AI時代の研究DXおよび生命の理解につながる

目的

生命の仕組みから疾患のメカニズムまでを理解し、安全なAI利用でQoLを向上する

研究DX

- 個別化医療
- AI創薬
- バイオものづくり

対象

- **データ提供側研究者**
  登録・倫理対応の負担を減らし公開が評価される仕組み
- **大規模データ生産拠点**
  大量データを高速に受け入れ、新技術データも迅速にDB化
- **データ利用側研究者**
  散在するDBを横断して安全に統合解析できる環境
- **大規模AI研究開発拠点**
  AI-readyなデータの継続供給とAI・DBの双方向連携

基盤

国内の複数の中核拠点に、永続的かつ十分な規模の計算機資源・ストレージを整備し、研究インフラとしてダウンタイムをゼロに、相互バックアップ体制も構築

オールジャパンで「既存・大規模・1次DB」と「新規・中小規模・2次DB」の開発を一体的に支援

理念

- **国際潮流**
  欧米に加え、アジア各国でもナショナルセンター設立が進む
- **国家による保全**
  国民のゲノム情報・ヒト研究データは国内で守り管理する
- **公共財としての集積**
  税金で得たデータを囲い込まず集積・標準化してAI利活用へ
- **オールジャパン**
  1次DB・2次DB開発者双方のニーズを満たす支援体制

トーゴー

データベースインフラのテナント化と相互運用による共有、TREを含む共通の技術開発

クリエイティブな場の提供：新しいDBを作る意欲のある人材にリソースを提供、データ統合と解析による付加価値の創造

生命科学DBの統合

- 国内外のデータをデータベースセンターに蓄積
- AIに対応するDBを実現するMCP等の技術開発
- AI-readyな知識グラフとオントロジーによる標準化
- オープン＆クローズ戦略への対応



塩基配列とアノテーション
INSDC (DDBJ/DBCLS)

アミノ酸配列とアノテーション
UniProt (SIB)

プロテオーム
jPOST
The Human Protein Atlas

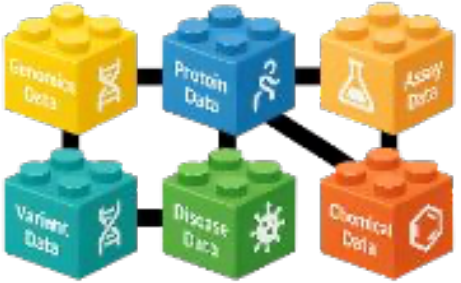
パスウェイ
Reactome (EBI)
BioModels (EBI)
BRENDA (DSMZ)
Rhea (SIB)

ゲノム情報
NBDCヒトDB・JGA (DBCLS/DDBJ)
Ensembl (EBI), HGNC (DBCLS)
RefSeq (NCBI/DBCLS), HCO (DBCLS)
GTDB (DBCLS), DBKERO (DBCLS)
OMA (SIB), HomoloGene (NCBI)

タンパク質立体構造
PDB (PDBj)
BMRB (PDBj)
FAMSBASE (Chuo U)

アッセイ
ChEMBL (EBI)

サンプル
BioSample(EBI/DDBJ)
Cellosaurus (SIB), NBRC (NITE)
JCM (RIKEN), NARO (NARO),
BacDive, MediaDive (DSMZ)
MicrobeDB.jp, MoG+, PGDBj



糖鎖
GlyCosmos
GlyTouCan
GlycoEpitope
GGDonto
PAConto
WURCS

遺伝子発現
RefEx, GTEx (DBCLS)
ExpressionAtlas (EBI), Bgee (SIB)

医科学 (Med2RDF)
PubMed, MedGen, MeSH, PubTator
ICGC, COSMIC, CIViC
DGIdb, OpenTG-Gates, HiNT, INstruct
ClinVar, dbSNP, dbVar, ExAC, gnomAD
NANDO, Nucleic Acid Drug DB

化合物
PubChem (NCBI)
ChEBI (EBI)
Nikkaji (JST)
MassBank (TUAT)
Polyinfo

バリアント
TogoVar (DBCLS), MGeND (JIHS)
ClinVar, LitVar (NCBI)
GWAS catalog (EBI)

Public Databases

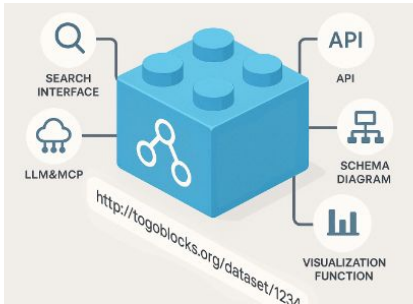
TogoMCP

SPARQL endpoint (トリプルストア)へ
自然言語により AIエージェントから MCP で自動的に問い合わせ

TogoPackage

DBCLSで開発してきたツール群を
どこでも実行可能なパッケージに

ローカルな非公開データと公共 DBを
シームレスに結合



Local Datasets



AIエージェントと生命科学

科学研究の主権はどこにあるのか

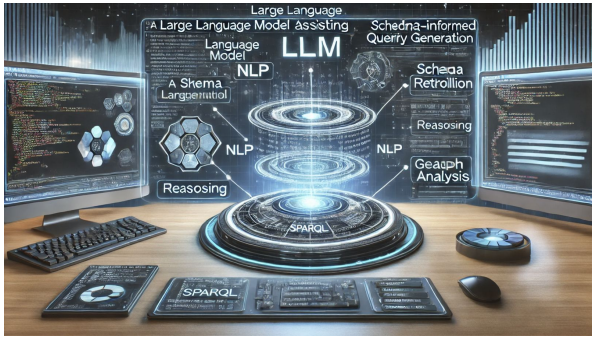
AIエージェントによる生命科学研究のDX

- スタンフォード Biomni
- ハーバード ToolUniverse
- OpenClaw ClawBio
- OpenAI GPT-Rosalind
- Google DeepMind Science Skills
- Anthropic Claude Science
- ：
- DBCLS TogoMCP, TogoPackage

ツールやDBをAPIで呼ぶ部分をラップして、というのは誰でも出来る時代、ツールやDB自体を提供できることは強みに

- **ツール**：SapporoなどWESで大規模解析も提供
- **データ**：知識グラフベースであることで差別化
- **ノウハウ**：ヒトDB, TogoVar, TogoID, RDFポータル, TogoDX/Human, PubAnnotation/Dictionaries, PubCaseFinderなど

今後はAIの安全性とデータの安全性 (TRE) の確保も課題



過去（学習）

- DBのデータを学習に利用 - AlphaFold, AlphaGenome

現在（自動化・効率化）

- DBやツールをエージェントが利用 - Skills, MCP
- 大規模実験の自動化で不足データの補完 - Physical AI

未来（無効化）

- 複雑な生命システムの理解は人間の認知能力を超える？
- ただ、真理・原理の解明は科学として続けたい