DBCLSにおけるデータ統合とデータベース事業のこれから

片山俊明 <ktym@dbcls.jp>
ライフサイエンス統合データベースセンター / Database Center for Life Science (DBCLS)

2023-10-05 @ 日本科学未来館

DBCLS

データベースとは

ー ー ー

一次データベース（レポジトリ）

- 求められる機能
  - 同種のデータの集約、データ形式の正規化
  - 論文発表や引用のためのID発行、永続性
  - 修正への履歴を含む対応（データは育つ）
  - 公開時期の調整と、メタデータの標準化
  - ダウンロードなど研究利用の促進
  - 国際連携 (INSDC/DDBJ, wwPDB/PDBj)
- 課題
  - オミックスなど新分野のデータへの対応 (DICP)
  - 制限公開データのアクセス管理 (DAC)
  - ID管理体系、オントロジーの共通化
  - メタデータ記載事項の標準化
  - データ量の増加、データ種類の増加

二次データベース（知識ベース）

- 求められる機能
  - データの整形・整理
  - IDや座標による関連情報の管理
  - 高度なアノテーションの付与
  - 多様な観点からのデータ検索
  - APIや計算サービスの提供
  - データ統合による知の体系化
- 課題
  - データ統合の果実を摘み取るデータ科学
  - 自然言語や画像など、非構造データからの知識抽出（ダークマター）
  - 大規模言語モデル(LLM)など新技術対応
  - データベースシステムやデータモデル刷新
  - 膨大なデータの更新・安定運用

https://www.nature.com/articles/nrg1065

データベース統合とは

Lincoln Steinさんのレビュー（20年前！）を振り返ってみる

- データが爆発的に増大している
- ゲノム・遺伝子・相互作用情報の蓄積と利用に DBは必須
- しかしデータベースの統合には課題が多い

20年も経って、さすがに状況は変わっているだろうか？

- DBごとにUIが違う… が本当の問題はさらに深いとの指摘
  - 遺伝子シンボルが生物種ごとに違うし、オーソログ情報の提供は別DBだし
    - → ここはRDFで？
  - 用語のコンセプトが生物種ごとに違う
    - → ここはオントロジーで？
- データベースをまたいで統合的に検索できたらいいのに！
  - DBごとに違うシステムのRDBが使われている
  - SQL検索インターフェイスはなくウェブページしかない
    - → ここはSPARQLで？

REVIEWS

INTEGRATING BIOLOGICAL DATABASES

Lincoln D. Stein

Recent years have seen an explosion in the amount of available biological data. More and more genomes are being sequenced and annotated, and protein and gene interaction data are accumulating. Biological databases have been invaluable for managing these data and for making them accessible. Depending on the data that they contain, the databases fulfil different functions. But, although they are architecturally similar, so far their integration has proved problematic.

Over the past two decades, databases of biological knowledge have grown from a cottage industry that was only of interest to a few specialized disciplines, to become essential resources that are used daily by biologists around the world. Examples abound, and include such diverse databases as: PubMed[1], the searchable compendium of biological literature that is maintained by the National Center for Biotechnology Information (NCBI); Ensembl[2], the database of human gene predictions that is maintained by the European Bioinformatics Institute (EBI) and the Wellcome Trust; the UCSC Genome Browser a human, mouse and rat genome browser[3] that is maintained by David Haussler's group at the University of California at Santa Cruz; FlyBase[4], the *Drosophila* research community database that is maintained by the FlyBase Consortium; WormBase[5], the *Caenorhabditis elegans* model-organism database; and the Gene Ontology (GO) database[6] of gene function, process and location terms. Many readers of this article will find it difficult to imagine conducting their work without access to one or more of these databases.

Despite having highly different functions, these databases are all architecturally similar. Each consists of three tiers of software (FIG. 1). At the bottom is a database management system (DBMS) that manages a collection of facts. At the top is the web browser that transmits requests for data to the database and renders the responses as web pages. In the middle is a software layer that mediates between the DBMS and the web browser to turn data requests into database queries, and to transform the query responses into hypertext mark-up language (HTML).

For the biological researcher, however, there are profound differences among the various biological databases. The differences begin on the first page, on which the researcher is greeted by a distinctive look and feel. For example, although Ensembl, FlyBase and the UCSC Genome Browser all provide the similar function of identifying the position of a gene of interest on the human or fly genomes, they provide distinctly different user interfaces for accessing this information. In Ensembl (FIG. 2), the user first selects the 'Human' database, which leads to a search page. Selecting 'Gene' from a pull-down search menu, and entering the name of the desired gene, leads to an intermediate page with a list of genes that have description lines containing the gene name. From here, the user selects the best match, which leads finally to a gene detail page. The position of the gene is printed at the top of this page.

In FlyBase (FIG. 3), the user selects 'Search Genes' from the list of search links on the front page, and then chooses 'Symbol/synonym/name' when prompted for the field to search from. This leads to a table of matching gene symbols, which includes the cytogenetic map position of each gene. Selecting the cytogenetic map position takes the user to a graphical display that shows the position of the gene in base-pair coordinates.

The UCSC browser (FIG. 4) requires the user to select 'Human' from a pull-down search menu and then enter the name of the gene into a search field. This leads to a page that summarizes all matches to the gene name, and, conveniently, lists the gene position directly.

Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA.
e-mail: lstein@cshl.org
doi:10.1038/nrg1065

https://www.nature.com/articles/nrg1065

データベース統合とは

データ統合のためのアプローチ

- Link integration
  - データベース間に IDの対応関係でリンクを貼る統合（いまでも安定）
- View integration
  - ひとつのページに各 DBの情報を埋め込む統合（パフォーマンスが悪かったそう）
- Data warehousing
  - 全部のデータを一箇所に集める統合（更新の手間が高くて　IGDは1年でポシャッた…）
- Web services
  - 共通のオントロジー とグローバルにユニークな IDの利用を前提とした APIによる統合

最後に提案されていた今後の方針が**ナックル&ノード**というアプローチ

- ノード
  - 各データベースごとに独立した詳細なデータモデルと適した技術を使用
- ナックル
  - 各ノードを他のノードと関連付けるために必要な情報を提供するサービス

→ それセマンティック・ウェブというやつでは？(RDF/OWL/SPARQL)

REVIEWS

INTEGRATING BIOLOGICAL DATABASES

Lincoln D. Stein

Recent years have seen an explosion in the amount of available biological data. More and more genomes are being sequenced and annotated, and protein and gene interaction data are accumulating. Biological databases have been invaluable for managing these data and for making them accessible. Depending on the data that they contain, the databases fulfil different functions. But, although they are architecturally similar, so far their integration has proved problematic.

Over the past two decades, databases of biological knowledge have grown from a cottage industry that was only of interest to a few specialized disciplines, to become essential resources that are used daily by biologists around the world. Examples abound, and include such diverse databases as: PubMed[1], the searchable compendium of biological literature that is maintained by the National Center for Biotechnology Information (NCBI); Ensembl[2], the database of human gene predictions that is maintained by the European Bioinformatics Institute (EBI) and the Wellcome Trust; the UCSC Genome Browser a human, mouse and rat genome browser[3] that is maintained by David Haussler's group at the University of California at Santa Cruz; FlyBase[4], the *Drosophila* research community database that is maintained by the FlyBase Consortium; WormBase[5], the *Caenorhabditis elegans* model-organism database; and the Gene Ontology (GO) database[6] of gene function, process and location terms. Many readers of this article will find it difficult to imagine conducting their work without access to one or more of these databases.

Despite having highly different functions, these databases are all architecturally similar. Each consists of three tiers of software (FIG. 1). At the bottom is a database management system (DBMS) that manages a collection of facts. At the top is the web browser that transmits requests for data to the database and renders the responses as web pages. In the middle is a software layer that mediates between the DBMS and the web browser to turn data requests into database queries, and to transform the query responses into hypertext mark-up language (HTML).

For the biological researcher, however, there are profound differences among the various biological databases. The differences begin on the first page, on which the researcher is greeted by a distinctive look and feel. For example, although Ensembl, FlyBase and the UCSC Genome Browser all provide the similar function of identifying the position of a gene of interest on the human or fly genomes, they provide distinctly different user interfaces for accessing this information. In Ensembl (FIG. 2), the user first selects the 'Human' database, which leads to a search page. Selecting 'Gene' from a pull-down search menu, and entering the name of the desired gene, leads to an intermediate page with a list of genes that have description lines containing the gene name. From here, the user selects the best match, which leads finally to a gene detail page. The position of the gene is printed at the top of this page.

In FlyBase (FIG. 3), the user selects 'Search Genes' from the list of search links on the front page, and then chooses 'Symbol/synonym/name' when prompted for the field to search from. This leads to a table of matching gene symbols, which includes the cytogenetic map position of each gene. Selecting the cytogenetic map position takes the user to a graphical display that shows the position of the gene in base-pair coordinates.

The UCSC browser (FIG. 4) requires the user to select 'Human' from a pull-down search menu and then enter the name of the gene into a search field. This leads to a page that summarizes all matches to the gene name, and, conveniently, lists the gene position directly.

Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA.
e-mail: lstein@cshl.org
doi:10.1038/nrg1065

https://www.nature.com/articles/nrg1065

データベース統合とは

まだまだ歴史は繰り返している気がするが…

私が全体的に楽観的であるにもかかわらず、生物学的データの統合は、考えられる将来にわたって困難な問題であり続けるだろう。データベース提供者側の協調的な努力と、 研究コミュニティの励ましと支援によってのみ、爆発的に増大する生物学的データを飼いならすことができるだろう。

REVIEWS

INTEGRATING BIOLOGICAL DATABASES

Lincoln D. Stein

Recent years have seen an explosion in the amount of available biological data. More and more genomes are being sequenced and annotated, and protein and gene interaction data are accumulating. Biological databases have been invaluable for managing these data and for making them accessible. Depending on the data that they contain, the databases fulfil different functions. But, although they are architecturally similar, so far their integration has proved problematic.

Over the past two decades, databases of biological knowledge have grown from a cottage industry that was only of interest to a few specialized disciplines, to become essential resources that are used daily by biologists around the world. Examples abound, and include such diverse databases as: PubMed[1], the searchable compendium of biological literature that is maintained by the National Center for Biotechnology Information (NCBI); Ensembl[2], the database of human gene predictions that is maintained by the European Bioinformatics Institute (EBI) and the Wellcome Trust; the UCSC Genome Browser a human, mouse and rat genome browser[3] that is maintained by David Haussler's group at the University of California at Santa Cruz; FlyBase[4], the *Drosophila* research community database that is maintained by the FlyBase Consortium; WormBase[5], the *Caenorhabditis elegans* model-organism database; and the Gene Ontology (GO) database[6] of gene function, process and location terms. Many readers of this article will find it difficult to imagine conducting their work without access to one or more of these databases.

Despite having highly different functions, these databases are all architecturally similar. Each consists of three tiers of software (FIG. 1). At the bottom is a database management system (DBMS) that manages a collection of facts. At the top is the web browser that transmits requests for data to the database and renders the responses as web pages. In the middle is a software layer that mediates between the DBMS and the web browser to turn data requests into database queries, and to transform the query responses into hypertext mark-up language (HTML).

For the biological researcher, however, there are profound differences among the various biological databases. The differences begin on the first page, on which the researcher is greeted by a distinctive look and feel. For example, although Ensembl, FlyBase and the UCSC Genome Browser all provide the similar function of identifying the position of a gene of interest on the human or fly genomes, they provide distinctly different user interfaces for accessing this information. In Ensembl (FIG. 2), the user first selects the 'Human' database, which leads to a search page. Selecting 'Gene' from a pull-down search menu, and entering the name of the desired gene, leads to an intermediate page with a list of genes that have description lines containing the gene name. From here, the user selects the best match, which leads finally to a gene detail page. The position of the gene is printed at the top of this page.

In FlyBase (FIG. 3), the user selects 'Search Genes' from the list of search links on the front page, and then chooses 'Symbol/synonym/name' when prompted for the field to search from. This leads to a table of matching gene symbols, which includes the cytogenetic map position of each gene. Selecting the cytogenetic map position takes the user to a graphical display that shows the position of the gene in base-pair coordinates.

The UCSC browser (FIG. 4) requires the user to select 'Human' from a pull-down search menu and then enter the name of the gene into a search field. This leads to a page that summarizes all matches to the gene name, and, conveniently, lists the gene position directly.

Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA.
e-mail: lstein@cshl.org
doi:10.1038/nrg1065

https://dbcls.jp/

ライフサイエンス統合データベースセンター

DBCLS

サービス（抜粋）

- TogoID – なんでもID変換サービス
- TogoDX – ヒトデータの探索的分析(EDA)
- TogoVar – 日本人バリアントDB
- TogoGenome – ゲノムDB
- TogoMedium – 培地DB
- ：
- NBDCヒトDB – （移行予定）
- RefEx – 遺伝子発現DB
- PubCaseFinder – 症状から疾患を検索
- BodyParts3D – 解剖学的人体モデル
- ：
- inMeXes – 英語論文表現の検索
- Colil – 文献引用コンテキストの検索
- Allie – 生命医科学用語の略語検索
- ：
- TogoTV – バイオインフォのYouTube

基盤技術（抜粋）

- RDF Portal – 知識グラフが全てここに
- SPARQL-proxy – エンドポイントの守護神
- Grasp – エンドポイントをGraphQL対応に
- SPARQList – 複雑なクエリをREST APIに
- SPARQL-GA – 遺伝的アルゴリズムで最適化
- RDF-config – スキーマ図やクエリを生成
- RDF-doctor – RDFデータをチェック
- D2RQ Mapper – RDBをRDFに
- Med2RDF – 医科学DBをRDFに
- ：
- TogoStanza – 可視化フレームワーク
- MetaStanza – ノーコードで可視化を実現
- ：
- PubAnnotation – 文献アノテーションを集積
- PubDictionaries – マイニング用辞書を集積
- ：
- TogoWS – 統合ウェブサービス
- TogoDB – 自分のDBを簡単に構築

知識グラフによるデータ統合と研究のサイクル

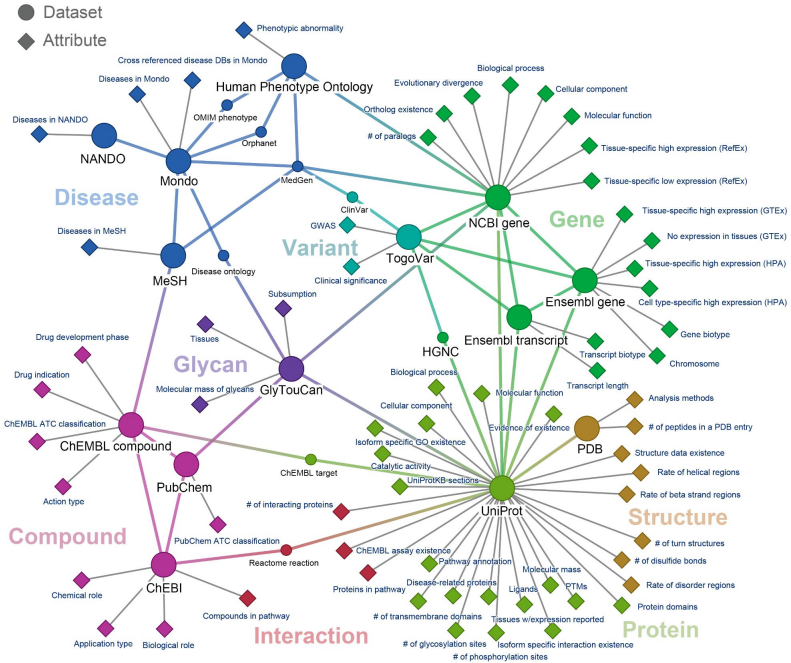
https://github.com/dbcls/sparql-proxy

Data sources
standardize
RDF
query
SPARQL
security
Proxy
performance
WS
D2RQ Mapper
DB
RDF
OPENLINK VIRTUOSO
GraphDB
Amazon Neptune
RDF-config
data transformation
curation
ID
new data
Application
end user
Component
visualize
DX
GENOME
VAR
D3SPARQL
STANZA
REST API
Robotics
experiment
Workflow
COMMON WORKFLOW LANGUAGE
reproducible
Analysis
jupyter
Studio
machine learning

https://github.com/dbcls/SPARQList

データベースの中身の統合利用の必要性

DBCLSで統合しているヒト関連データ



SPECIAL SECTION HUMAN GENOMICS

REVIEW

From variant to function in human disease genetics

Tuuli Lappalainen[1,2]* and Daniel G. MacArthur[3,4,5]*

Over the next decade, the primary challenge in human genetics will be to understand the biological mechanisms by which genetic variants influence phenotypes, including disease risk. Although the scale of this challenge is daunting, better methods for functional variant interpretation will have transformative consequences for disease diagnosis, risk prediction, and the development of new therapies. An array of new methods for characterizing variant impact at scale, using patient tissue samples as well as in vitro models, are already being applied to dissect variant mechanisms across a range of human cell types and environments. These approaches are also increasingly being deployed in clinical settings. We discuss the rationale, approaches, applications, and future outlook for characterizing the molecular and cellular effects of genetic variants.

今後10年間でヒト遺伝学の主要な課題は、遺伝子変異が疾患リスクなどの表現型に影響を与える生物学的メカニズムを理解することである

[Review article] Science (2021) 373:1464-1468

https://www.science.org/doi/10.1126/science.abi8207

知識グラフ(RDF)は既存のエコシステムがなくフルスタックの再開発が必要
データベース構築に必要な技術スタックの例

青字：汎用的なウェブ技術
赤字：RDFに求められる技術

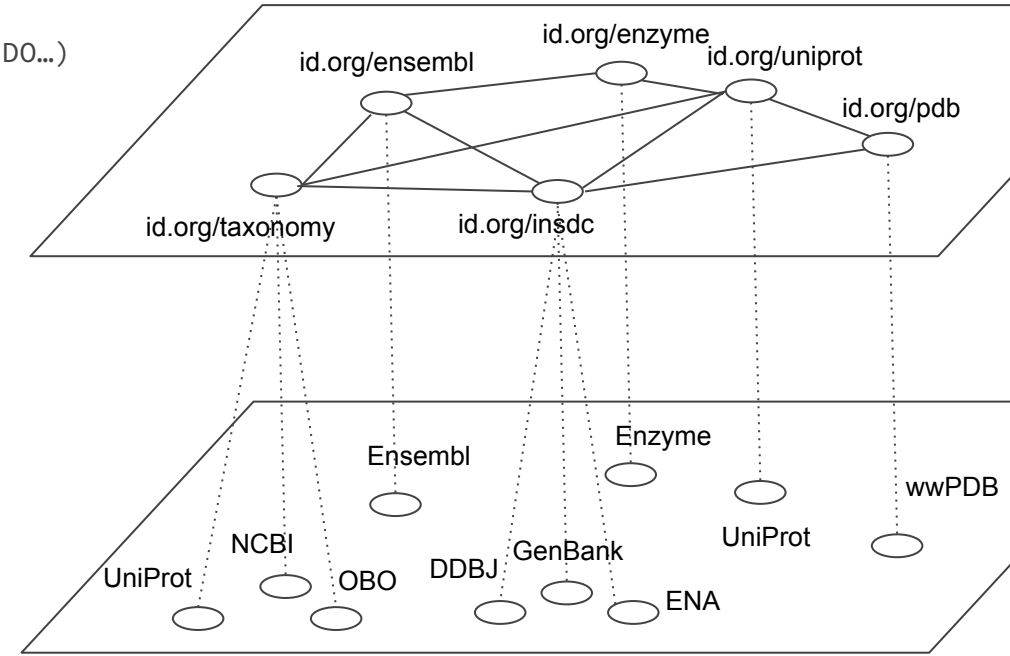
- ユーザーインターフェイス
  - HTML, CSS, SVG, JavaScript, JSON, API
  - 各種ライブラリ（D3.js, Vue.js, React.js, Three.js, …）→ d3sparql.js, MetaStanza
- データベースシステム （DBMS）
  - RDB/SQL（PostgreSQL, MySQL, SQLite, …）
  - Object store（MongoDB, …）
  - Key-value store（Redis, Memcached, …）
  - RDF/SPARQL（Virtuoso, GraphDB, …）→ 文字列検索・グラフ探索は別途開発が必要
- データの維持・管理
  - ミドルウェア（Ruby on Rails, Django, …）→ スクラッチで開発（SPARQList, TogoStanza）
  - プログラミング（Python, Node.js, Ruby, Rust, …）→ SPARQL
  - 各種パラダイム（HTTP, REST, MVC, O/Rマッピング, OWL, XSD, Git, …）
  - オンプレミスのシステム管理（Linux, Docker, ジョブ管理, cron, …）
  - クラウド（Amazon Web Service, Microsoft Azure, Google Cloud Platform, …）
  - 認証（OAuth, OpenID, …）
- RDFの場合
  - 運用ノウハウ（RDF portal, SPARQL-proxy）→ トリプルストアの性能限界、独自に開発？
  - データモデル（RDF-config）→ クエリ生成、スキーマ図生成、GraphQL対応、バリデーション、RDF生成
  - アプリケーション開発（TogoID, TogoGenome, TogoVar, …, TogoDX）→ UIとデータ解析の新パラダイム

標準化と技術開発にはコミュニティの協力が大切
→ 国際開発者会議BioHackathonを2008年から主催し国際連携を推進

- データモデル
  - 生物種 – ゲノム – 遺伝子 – 変異 – 表現型 – 疾患 – 医薬品 – マルチオミックス…
- オントロジー
  - BioPortalなど + 独自開発（FALDO, HCO, IDO…）
- グローバルにユニークな ID
  - Identifiers.orgなど
- メトリクス
  - FAIR principles, RDF化ガイドライン, …
- ワークフロー
  - WES, CWL, Galaxy, …
- 再利用モジュール
  - BioJS, TogoStanza, SPARQList, …

共通の基準に沿ったRDF化

- URIとしてIdentifiers.orgを採用
- オントロジーを共有
- SPARQListやTogoStanzaで再利用



http://biohackathon.org/

国際版 NBDC/DBCLS BioHackathon

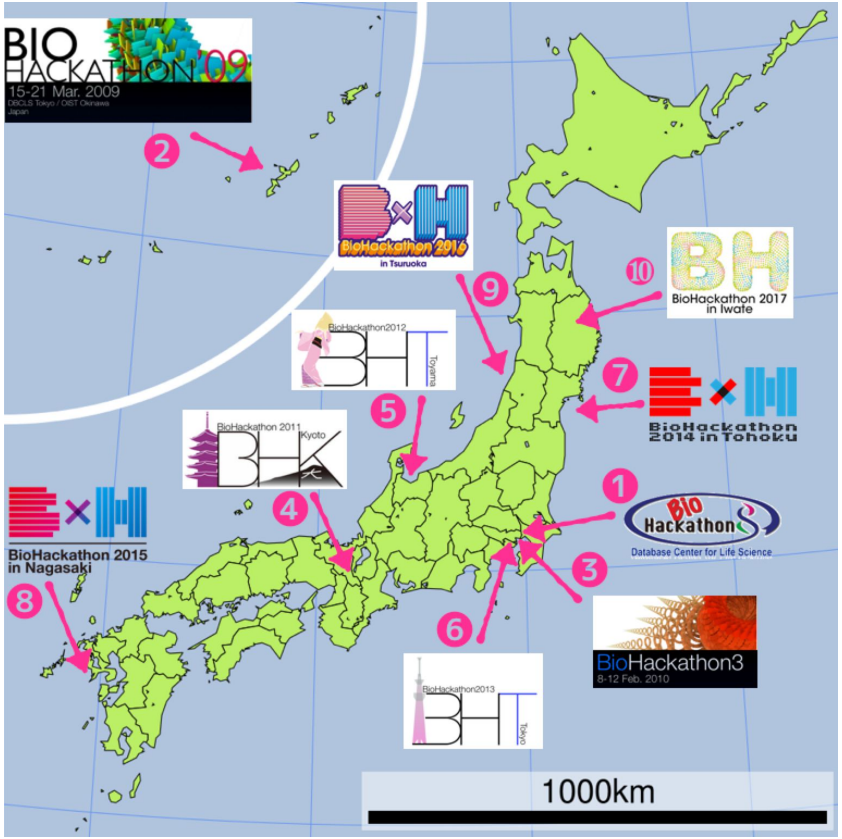
（海外へも波及: ELIXIR BioHackathon, BioHackathon-MENA, SWAT4HCLS hackathon etc.）

ーーー

ライフサイエンス統合データベースセンターでデータベース統合技術開発のため開催開始

- BioHackathon 2008-2010
  - ウェブサービス(API)による分散バイオインフォマティクスリソースの統合
- BioHackathon 2010-2019
  - セマンティック・ウェブ(RDF)による分散バイオインフォマティクスDBの統合
- BioHackathon 2023
  - 人類遺伝学・疾患研究
  - 微生物研究・有用物質生産
  - 環境・農業・食料・エネルギー問題
  - 膨大な知識グラフと文献の利活用技術開発
  - → 統合データ利用の具体的な成果を目指して

※ 国内版バイオハッカソンも毎年開催中！



https://rdfportal.org/

知識グラフ(RDF)で統合された主要な生命医科学DB

RDF: Resource Description Framework

- 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)

- 医科学 (Med2RDF)
  - ICGC, COSMIC, CIViC
  - DGIdb, OpenTG-Gates
  - ClinVar, dbSNP, dbVar
  - 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)
  - :

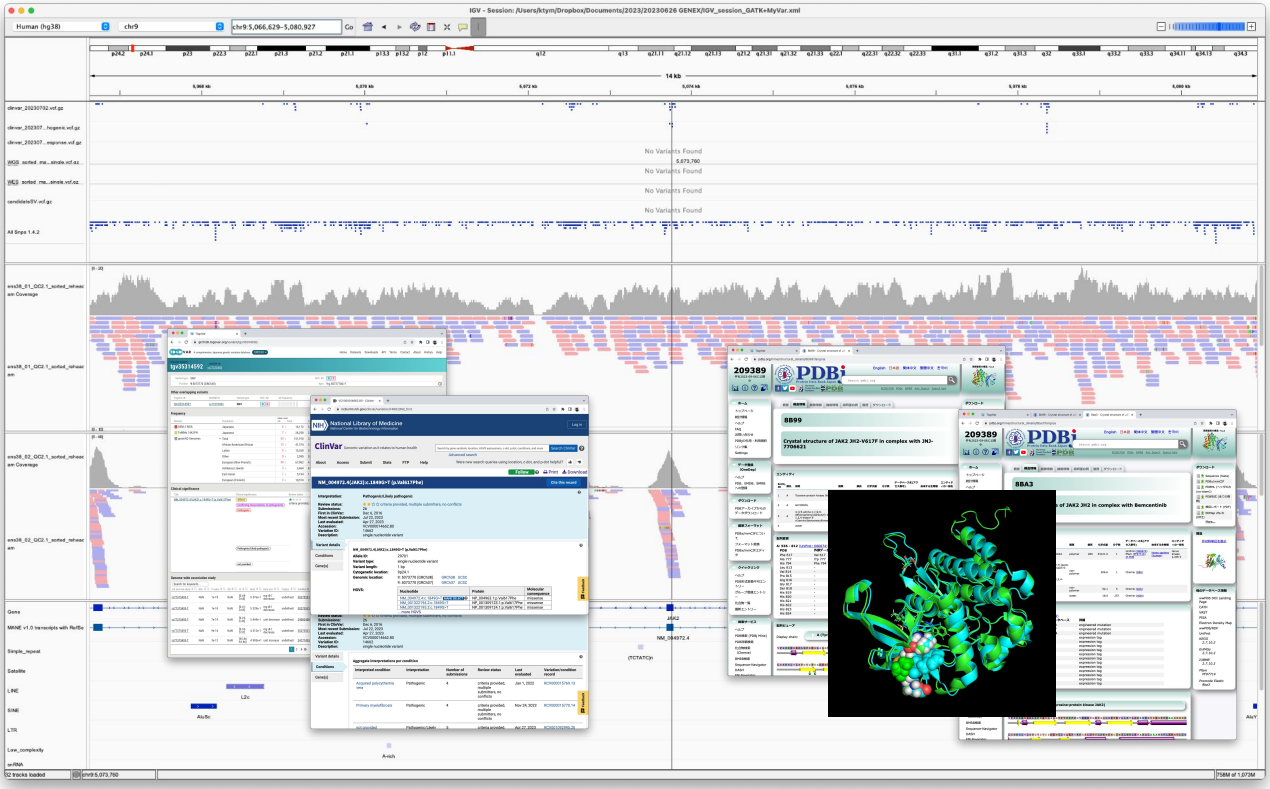
マイゲノムを眺めて見るに…

- IDの対応だけでなく座標系の対応も
- 複雑なDBスキーマの深い階層にある重要な値

まだまだ関連情報を辿るのは面倒…

- バリアント
- 遺伝子発現制御
- 機能アノテーション
- 立体構造
- 疾患情報
- 相互作用
- パスウェイ
- ：

一望できる統合データ環境がほしい！



データ統合の成果を引き出すには研究利用が必須

- 生命科学における研究DX・データ科学・オープンサイエンスの実現にむけて
  - 第5期科学技術基本計画
    - オープンサイエンスの推進
  - 第6期科学技術・イノベーション基本計画
    - 「単に研究プロセスの効率化だけではなく、研究の探索範囲の劇的な拡大、新たな仮説の発見や提示といった研究者の知的活動そのものにも踏み込んだプロセスを変革」
    - 「データを用いたインパクトの高い研究成果の創出、研究者の貴重な時間を研究ビジョンの構想や仮説の設定などより付加価値の高い知的活動へと充当」
  - DBCLSのデータベース統合や基盤技術開発の取り組みは正にこれらを実現するために行ってきた
    - 幅広いデータの統合・データの信頼性の担保・これを支えるデータインフラや計算資源の整備
    - → 見直しで**多くの研究者が統合データを利活用できるための応用研究・基盤研究を厳選**した
- DBCLSの基盤技術やデータ統合の成果を活かしていくための重点項目
  - 項目1．応用研究：日本人ゲノム研究・遺伝学研究・疾患研究に資するデータ統合
  - 項目2．応用研究：有用物質生産につながる微生物データ統合
  - 項目3．基盤研究：情報処理における新しい技術開発課題への対応
  - 項目4．基盤研究：統合データの安定的な運用と利便性の向上による利活用の促進

DBCLSの取り組む応用研究・基盤研究の概要

ーーー

- 項目1．応用研究：日本人ゲノム研究・遺伝学研究・疾患研究に資するデータ統合
  - **日本人ゲノム医療の実現には、日本のデータベースセンターが責任を持って対応する必要がある。**このため、**すでに広く使われている NBDCヒトデータベースと TogoVar**がDBCLSに移管されること、DICPでも**日本人ゲノムおよびマルチオミックス** データが集積することから、連携し日本人ゲノム研究の基盤として、より高度なデータ提供を行うための技術開発およびデータ統合を進める。
- 項目2．応用研究：有用物質生産につながる微生物データ統合
  - **近年メタゲノムをアセンブルした MAGによる微生物ゲノムが爆発的に増加** しており、このデータ解析が急務となっている。 **TogoGenomeやTogoMediumで開発してきた資産が活用できる**微生物のデータ統合に注力し、アノテーションの自動化や培養条件の推定技術開発に取り組む。
- 項目3．基盤研究：情報処理における新しい技術開発課題への対応
  - **大規模言語モデルによる情報抽出、大規模言語モデルの生成するテキストのエビデンスの担保、チャットUIを想定した新しいデータベース利用のための技術開発** に取り組む。

DBCLSと統合化推進プログラムの連携(希望)

ーーー

一次データベース（レポジトリ）

識ベース）

各DBを連携させ
データ統合による
日本独自の付加価値を

DICP

日進月歩の生命科学データを蓄積

- データ形式？
- メタデータ？
- ID体系？
- オントロジー？
- 規模・速度・運用・DBMS？
- UIの機能？

将来的に共通する基盤を共有？

標準化

効率化

DBCLS

データ統合とデータ科学への展開

- DICPデータ統合のハブに
- DB構築の技術提供
- 標準化・効率化の推進
- ベストプラクティスの確立
- LLM利用による知識抽出
- 自然言語によるDB問合せ

多様なドメインDBの理解と活用

See also 統合データベースプロジェクトから学ぶこと 高木利久 「情報処理」Vol.62 No.5 (May 2021)

日本のデータベース事業のこれから(期待)

国立のデータベースセンター維持の必要性

- 欧州バイオインフォマティクス研究所（EBI）
- アメリカ国立生物工学情報センター（NCBI）
- 中国などの台頭

データの増大は課題ではなくチャンス

- 生命科学のデータは複雑で多様
  - それぞれのデータを熟知する専門家は速成不可能
- 膨大なデータの統合運用と研究環境の整備
  - 使えるデータの周りに人と技術は育つ

データ科学の時代に取り残されないために

- 手元に参照すべき全データを整備し続ける
- 使いこなせる人材とノウハウの蓄積が国力