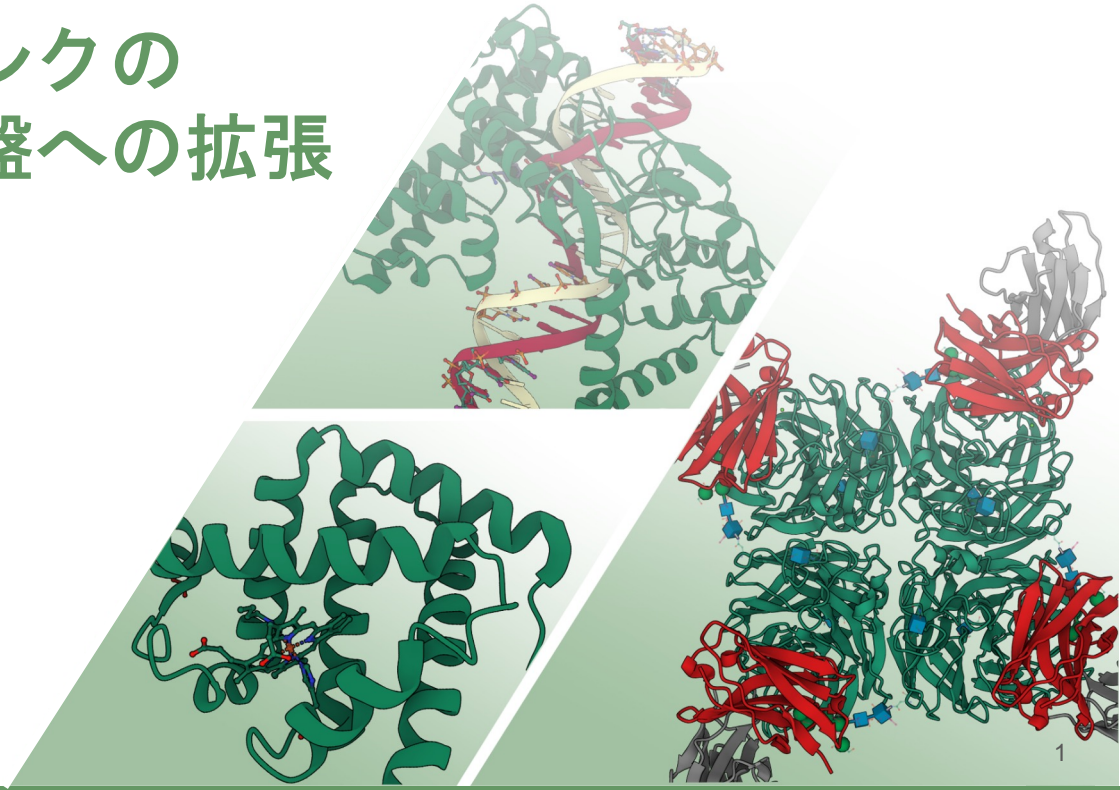
蛋白質構造データバンクの
データ駆動型研究基盤への拡張

栗栖源嗣

大阪大学蛋白質研究所
（財）蛋白質研究奨励会



統合化推進プログラムで整備するデータベース

生体高分子の3次元構造に関する包括的な情報

本プログラムではPDBとBMRBという2つの基盤データベースを対象とします。主たるデータベースであるPDBは1971年からデータが集積され，情報は無償で利用できます。今後ますますデータ蓄積数の増加が見込まれています。

- **Protein Data Bank (PDB)**
  - X線，NMR，EMで決定した座標データ
  - X線の構造因子
  - 中性子の構造因子
- **Biologial Magnetic Resonance Data Bank (BMRB)**
  - 化学シフト
  - 距離/角度制限情報
  - その他

研究開発項目の概要

①国際連携による国際基準のデータ基盤構築

②国際的なプレゼンス

③知識発見・課題解決を支援する機能の開発

④分野・領域を超えたデータ統合とDB連携

⑤研究ニーズや実験技術の新しい動向への対応

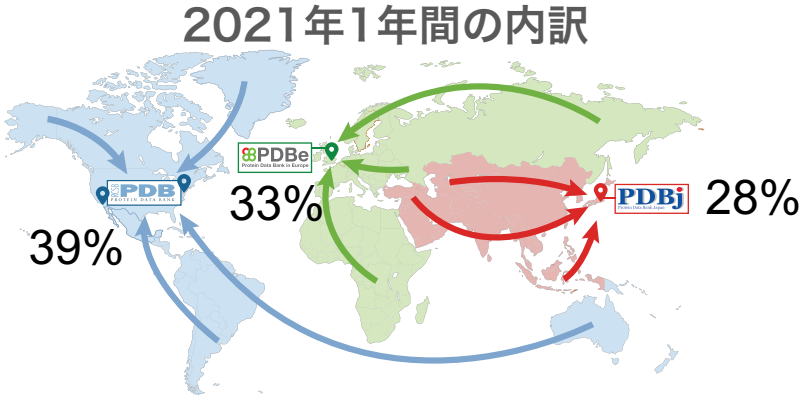
⑥研究コミュニティと連携

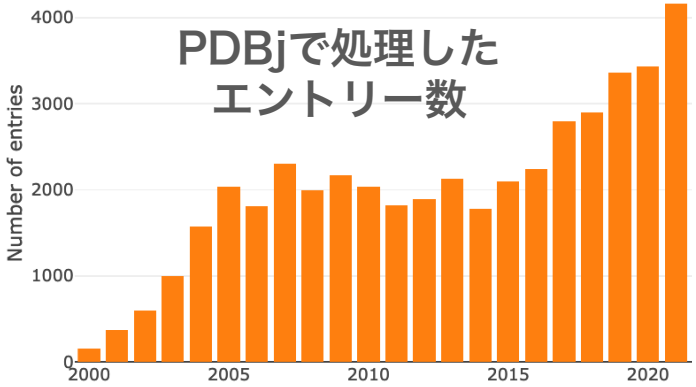
① 国際連携による国際基準のデータ基盤構築

・wwPDB全体のエントリーに占めるPDBj20年間の登録・編集の割合は**約22%**
(45,080/198,464 = 0.227)

・2021年1年間に限ると，アジア地区のデータ量増加率が他地域よりも多いので，PDBjの登録・編集の割合は**約28%**に増加
(4,160/14,570 = 0.286)

・2022年2月から**PDB China**が准メンバーとして始動。8月よりPDBjがアノテーショントレーニングを開始予定。准メンバー期間は中国を含む全アジア・中東地区のデータに対しPDBjが最終チェックする。

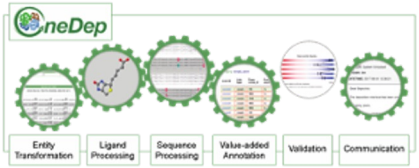




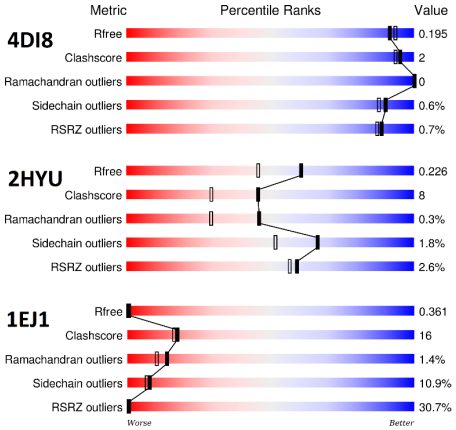
・データ処理にかかる１エントリーあたりの時間を短縮しデータの品質を維持する目的で，**電子顕微鏡構造の検証レポートにおける検証項目を追加**する。（先行例：2022年2月25日にhalf mapsの登録必須化された）

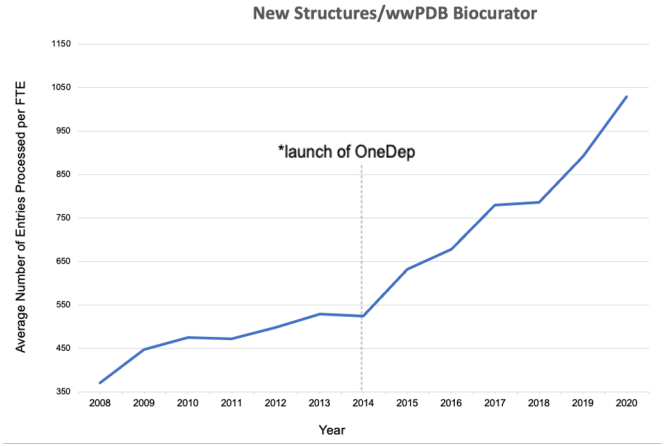
・OneDep登録システムの高度化を進め，**１エントリーを処理するのにかかる時間を短縮**する。

・それらにより，アノテータの人数を増やすことなく増加する一途を辿る構造データに対し，遅滞することなく信頼度の高いデータベース構築を達成する。

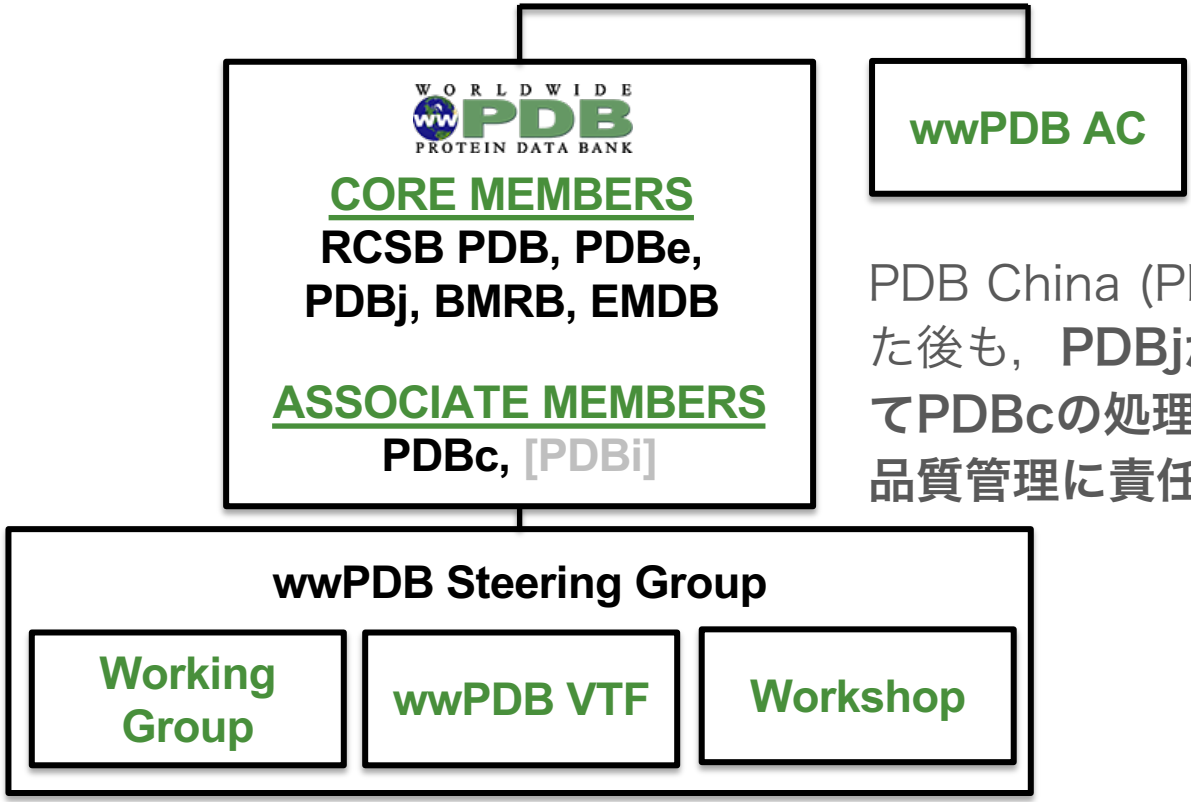


Overall Quality





② 国際的なプレゼンス



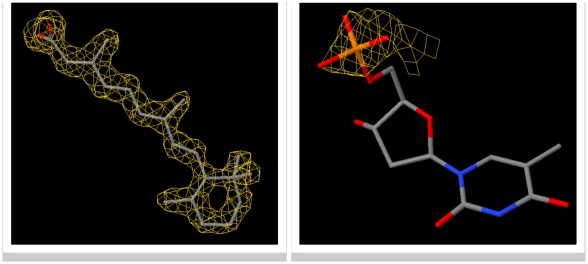
PDB China (PDBc) が立ち上がった後も，**PDBjがメンター機関としてPDBcの処理したエントリーの品質管理に責任**を持ちます。

③ 知識発見・課題解決を支援する機能の開発

・これまでの統合化プロジェクトで整備したRDF対応の検証レポートを活用し，**化合物情報に特化した機械学習用データセットを毎週公開**する。

・同じ分解能でも右図のように，実験データとの異なる一致度を示すエントリーが存在するため，蛋白研の実験研究者と協力して，複数の判断基準を導入し**機械的にフィルタリングしたデータセットとして公開**。

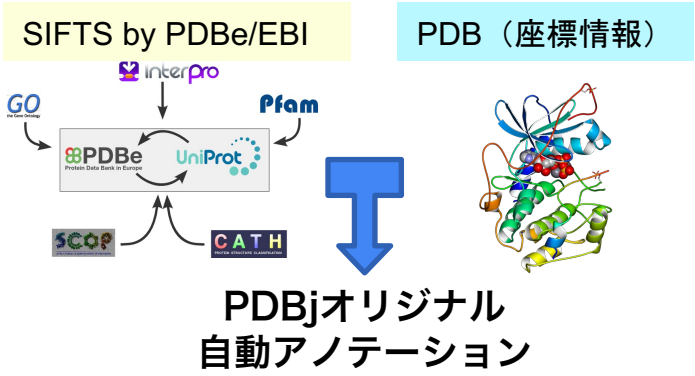
・1次データベースを提供するwwPDBの地域拠点から化合物情報に特化した機械学習用データセットを公開することはAI開発を推進する。



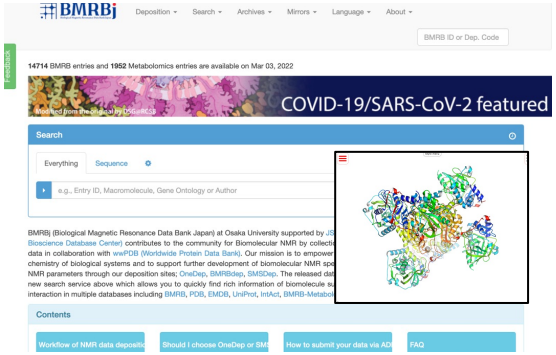
同分解能でも異なる実験データとの一致度の例

左：RSR=0.10, CC=0.95
右：RSR=0.41, CC=0.70

・前課題で外部リソースと座標情報の統合により各残基の立体構造中での役割（分子内・分子間コンタクト）と機能を自動的にアノテーションする仕組みを構築した。同じ枠組みで，**分子ポケット情報を自動アノテーションし，医薬品候補化合物との結合予測を含むリソースと統合**できるようにする。



・統一されたNMR-STARフォーマットに未対応のBMRBデータを一元的に標準化し，**NMR実験データを分子ビューアで可視化**できるようにして，BMRBjのウェブからサービスを提供する。



④ 分野・領域を超えたデータ統合とDB連携

・これまでの統合化プロジェクトにおいて，PDB, BMRB, および各検証レポートのRDF化が完了している。その豊富な情報を活用して，複数のデータベースを横断的に利用する**データ駆動型研究に資する統合利用ポータルを構築**する。

・バイオインフォマティックス学会，分子生物学会や生物物理学会で等でWS/セミナーを実施して意向調査を実施します。それにより，データ駆動型研究を実施する生命情報科学研究者のDB統合利用を促進するとともに，各生命現象を追求する生命科学研究者の個別利用も支援します。

⑤ 研究ニーズや実験技術の新しい動向への対応

個別の構造情報をWeb上のビューアで検証するような利用スタイルだけではなく，PDBjの提供するサービスにより構造生物学の専門知識を必要とせず**機械学習をベースとした研究者の研究活動を大きく推進する**ことを期待している。 急増するAIによる構造予測手法（AlphaFold2等）を併用した**Integrated/Hybrid構造解析に対応した検証レポート**も検討します。

⑥ 研究コミュニティと連携

DB側主導のTask Forceをコミュニティーレベルで組織し，データ処理にかかる１エントリーあたりの時間を短縮するとともに，データの品質を維持する目的で**電子顕微鏡構造の検証レポートにおける検証項目を追加**します。
学会等での意向調査をベースにしたサービスの最適実装を目指します。

PDBjメンバー

- 統括責任者

  栗栖源嗣　（大阪大学蛋白質研究所）

PDB/EMDBデータベース構築グループ

中川敦史　（大阪大学蛋白質研究所・教授）
于　健　（大阪大学蛋白質研究所・特任准教授）
張　羽澄　（大阪大学蛋白質研究所・特任研究員）
池川恭代　（大阪大学蛋白質研究所・特任研究員）
佐藤純子　（大阪大学蛋白質研究所・特任研究員）
金　宙妍　（大阪大学蛋白質研究所・特任研究員）
丹羽智美　（大阪大学蛋白質研究所）

PDB/EMDBデータベース高度化グループ

水口賢司　（大阪大学蛋白質研究所・教授）
Bekker, Gert-Jan　（大阪大学蛋白質研究所・特任助教）
長尾知生子　（大阪大学蛋白質研究所・助教）
山下鈴子　（大阪大学蛋白質研究所・技術専門職員）
工藤高裕　（大阪大学蛋白質研究所・特任研究員）

BMRBデータベース管理運営グループ

藤原敏道　（大阪大学蛋白質研究所・教授）
児嶋長次郎　（横浜国立大学工学部・教授）
宮ノ入洋平　（大阪大学蛋白質研究所・准教授）
岩田武史　（大阪大学蛋白質研究所・特任研究員）

EMPIARグループ

中根崇智　（大阪大学蛋白質研究所・特任准教授）
常住規代　（大阪大学蛋白質研究所・特任研究員）

事務職員

佐久間量子　（大阪大学蛋白質研究所・特任事務職員）

PRF分室

栗栖源嗣（兼）　（（財）蛋白質研究奨励会・客員研究員）
磯山正治　（（財）蛋白質研究奨励会・室長）
横地政志　（（財）蛋白質研究奨励会・研究員）
見学有美子　（（財）蛋白質研究奨励会・研究員）

WORLDWIDE PDB
PROTEIN DATA BANK

The worldwide Protein Data Bank

www.wwPDB.org • info@wwPDB.org

RCSB PDB PROTEIN DATA BANK

BMRB Biological Magnetic Resonance Data Bank

PDBe Protein Data Bank in Europe

EMDB Electron Microscopy Data Bank

PDBj Protein Data Bank Japan