

## 研究開発課題別中間評価結果

### ➤ 課題情報

**研究開発課題名** 「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」

**研究代表名** 栗栖 源嗣

### ➤ 中間評価結果：

日米欧の国際組織 wwPDB の枠組みの下、構造生物学の基盤データベースである PDB（蛋白質立体構造データバンク）と BMRB（NMR 実験情報データバンク）の日本拠点をさらに効率化して安定的に構築するとともに、AI 構造予測技術に対応するためにデータ登録システムの拡張と機械学習用選抜データセットの公開に取り組み、立体構造中の分子ポケット情報を自動アノテーションして公開する研究開発課題である。

wwPDB メンバーとして、PDB と BMRB のアーカイブの構築・データ検証・公開を行った。PDBj が分担するアジア・中東地域からの全てのデータを 100%処理し、各エントリーの論文発表に合わせて遅滞なく全世界に公開した。アジア地区におけるエントリー増加に伴い、PDBj で処理したエントリー数は、2022 年度は wwPDB に登録された全データの 29.0%（4,749 件／16,344 件）、2023 年度は 31.5%（5,376 件／17,064 件）、2024 年度は 9 月の時点で 32.1%（4,100 件／12,787 件）に上った。近年急増している AlphaFold2 や RosettaFold を併用して構造解析したエントリーを登録する際に、予測構造をどう活用したのかという情報を収集する仕組みを実装した。また wwPDB に、予測構造を併用した解析データの検証タスクフォース（Validation Task Force : VTF）を組織し、検証レポートのマスターフォーマットである PDBx/mmCIF の拡張版である IHMCIF を定義し、論文を発表した。BMRBj が主導し、NMR 制限情報の標準化を行った。2023 年度に、米国にて BMRB のグラント更新が滞り、半年間、繋ぎフاندと大学予算で暫定的な運用がされるという事態が発生した際、追加支援により BMRBj のデータ処理システムを高度化し、米国 BMRB の遅れをカバーした。ここで開発したシステムはパフォーマンスが高く、その後、米国コネチカット大の BMRB にも移植し、PDBj が開発した処理システムを用いて日米での BMRB のデータ処理が行われるようになった。BMRBj のデータ処理システムの高度化にエフォートを割く必要が生じたために、NMR の検証結果の可視化については、2024 年度以降に実施する計画変更を行った。蛋白質立体構造中におけるアミノ酸レベルの位置情報、分子ポケット情報の自動アノテーションについては、NMR 制限情報の標準化と検証結果の可視化の目処がついた段階で研究開発に取り組む予定。

2022 年度に実施した利用動向調査を踏まえ、東北大学で整備している jMorp（日本人多層オミクス参照パネル）と連携し、日本人のゲノムバリエーション情報を PDB もしくは AlphaFold の構造上にマップし、生物学・化学の視点からの検索結果を表示する PDB 統合利用ポータルβ版を完成させ、2024 年 10 月に一般公開した。化合物情報に特化した機械学習用データセットについて 2022 年度に利用動向調査を実施し、その結果を踏まえて、PDBj の化合物の検索結果に PDBj 独自の選抜基準を策定して構造情報を選抜する検証パラメータを策定した。

蛋白質研究奨励会にバックアップサーバーを整備し、阪大蛋白質研のメインサーバーと DNS を切り替えることで 365 日 24 時間安定運用できる体制を構築し、SINET の正式加盟機関として、奨励会からも 10Gbps でのバックアップサービスを開始した。wwPDB の枠組みの中で PDB China の始動を主導し、2022 年夏以降は上海で勤務する PDB China のアノテーター 4 名と共に、アジア・中東地区からのエントリーに完全に対応した。

上述の通り、進捗状況と今後の成果見込みは大変優れている。米国 BMRB が予算問題から一部機能が停止

した際は、BMRBj のデータ処理システムを高度化することでグローバルなデータ登録の遅れをカバーした。またその後、PDBj が開発した処理システムを米国 BMRB にも移植することで、大きな存在感を示した。さらに PDBc の設立支援は、近年急増する中国からのデータ登録への具体的な対応策となるばかりでなく、国際的な研究コミュニティにおける中国の孤立を避けるためにも有効な国際貢献として評価できる。研究開発計画は適切であり、今後も概ね現計画通りに推進すべきである。

以上