【報告様式Ａ】

作成／更新日：令和４年５月２６日
課題管理番号：21am0109006j005

日本医療研究開発機構　創薬等ライフサイエンス研究支援基盤事業
事後評価報告書

公開

Ⅰ 基本情報

補助事業課題名：　（日本語）創薬等ライフサイエンス研究支援基盤事業
（プログラム名）　（英　語）Platform Project for Supporting Drug Discovery and Life Science Research

実施期間：平成 29 年４月１日～令和４年３月 31 日

補助事業担当者　氏名：（日本語）栗栖源嗣
（英　語）Kurisu, Genji

補助事業担当者　所属機関・部署・役職：
（日本語）国立大学法人大阪大学・蛋白質研究所・教授
（英　語）Professor, Institute for Protein Research, Osaka University

Ⅱ 補助事業の概要

我々は，各プラットフォームで生まれる様々なデータを，収集・管理・公開することで，成果を最大限活用することを目的として事業展開した。この目的を実現するため，大阪大学では事業で整備されてきたタンパク質関連の有用なデータベースを継承・維持・管理・運用し，本プロジェクトに必要とする情報を提供した。これらデータベースをインターネットに公開することで，外部研究者と事業推進者の研究活動をサポートした。また，事業内で得られる各種データを新しく収集・蓄積・管理することも担当した。支援と高度化，さらに他グループと連携して行った事業の成果と意義について記す。

（１）支援

1-1. タンパク質構造データベースの継承と運用

我々の研究グループでは，３つの国際的なデータベース，(i)蛋白質構造データバンク（Protein Data Bank: PDB），(ii)生体系核磁気共鳴データバンク（Biological Magnetic Resonance Data Bank: BMRB）と (iii)電子顕微鏡マップデータバンク（Electron Microscopy Data Bank: EMDB）を運用・公開している。このうち EMDB データベースは，本 BINDS 事業によりサポートされている。大阪大学では，日本を含むアジア地区から EMDB へ登録されたエントリーを 100%処理し，1 件の漏れなくデータを公開している

（https://pdbj.org/emnavi）。2017 年 4 月の事業開始から現在まで月毎の登録数も右肩あがりで増加し，大阪大学で登録した EMDB の総エントリー数は 2000 件を超えている。これは，この期間に構造決定された全世界の電子顕微鏡マップデータの約 15%に相当する。また PDB に関連して，これまでの事業で整備されたデータベースおよびツール（HOMCOS，hGtoP，Omokage 検索）を更新・維持・管理し，総合的な構造データベースとしてアノニマスな利用に供することで本事業推進者をサポートした。

我が国を含むアジア地域からのデータ登録を遅滞なく遂行しており，構造生物学コミュニティーに貢献できたと考えている。

1-2．タンパク質構造データベースの公開と活用

PDB や EMDB データベースを基に，ホモロジーモデリング，分子動力学シミュレーション，および AI を活用した迅速な NMR 構造ソフト群を活用して，個別構造解析支援を行った。PDBj で開発された電子顕微鏡による構造検索サービス（EM Navigator），分子の形について EMDB を中心に横断的に検索するサービス（Omokage search），高度化事業で整備を進めているクライオ電子顕微鏡画像データベース（EMPIAR）を含めて総合的な電子顕微鏡による構造データベースとして公開・発信した。特に，新型コロナウイルス（SARS-CoV-2）の構造情報に関しては、できるだけ早い構造データの蓄積と公開が期待された。通常，登録済み PDB データは，該当の情報を含む研究が論文として発表されるまで非公開とされ，実験のオリジナリティーが担保される仕組みになっている。しかし，蛋白質の構造情報は，立体構造に基づいた創薬研究に積極的に活用される基盤情報であり，SARS-CoV-2 の構造情報も，できるだけ早い構造データの蓄積と公開が期待された。そこで，大阪大学では，世界で最初の SARS-CoV-2 の構造情報が PDBj に登録された際に，我々が個別に登録者（研究者）に連絡をとり、論文発表を待たずに即時公開することを強く勧める体制を構築した。この体制により，SARS-CoV-2 に直接関係あるエントリーは日米欧の各拠点の区別なく，論文発表を待たずに即座にデータ公開できるようになった。特に，COVID-19 特集ページを開設し，多くの利用者が SARS-CoV-2 のデータを早く・正確に利用できるよう整備しサポートした。既に，我々がデータ登録し国際基準で早期公開処理を行った構造データを用いて，スーパーコンピューターを用いた創薬研究や分子設計研究などに積極的に活用されている。

（2）高度化

現在の創薬研究および基礎生命科学研究等のライフサイエンス研究で生み出されるデータはビッグデータであり，これらをとりあつかう技術はまだ完成していない。我々のグループでは，これらのデータを扱うデータベースの高度化を目的として活動した。大阪大学では、[a] 本事業で得られるさまざまなデータを蓄積し，新しいデータベースに格納する技術を開発した、そして [b] すでに存在する多くのデータベースを接続し，新しい情報を取得する技術の開発に貢献した。

2-1．クライオ電子顕微鏡画像データベースの開発

EMBL-EBI との間に学術交流協定を締結して，クライオ電子顕微鏡画像データベースである EMPIAR（Electron Microscopy Public Image Archive）の日本ミラーサイトを開設し，2018 年 12 月にデータ公開を開始した（https://empira.pdbj.org）。クライオ電子顕微鏡の 2 次元画像データをデータベース化するには，セキュアな蓄積技術の開発，10 Gb 以上の高速ネットワークの活用が欠かせない。事業開始当初の想定を上回る猛烈な勢いで公開データ量が増加した。そこで，サーバーを冗長化することによるバックアップの方針を切り替えて，LTO テープと最適化ユニットに導入された Cold Storage システム（早稲田大学）を活用したバックアップシステムを構築した。2018 年に追加支援していただき導入した Aspera と Globus を用いた高速通信システムのセットアップも行い，効率的に大容量データを転送，バックアップする仕組みを確立した。本事業終了時点（2022 年 3 月末）では，大阪大学の EMPIAR-PDBj のデータベースに 1.7 PB のデータを蓄積し，一般に公開している。米国に同様のデータベースが存在しないため，アジアだけでなく米国（特に西海岸）からのアクセス数が多い。

2-2. 計算結果データベースの開発

分子動力学およびドッキング等の計算結果のデータベース化においても，クライオ電子顕微鏡画像データと同様に巨大なデータを取り扱う。PDB でこれまでに開発した技術を援用し，計算データのアップロードシステムと公開サーバーを構築した。データベースの運用スタイルは，データ寄託方式とした。このデータベースは Biological Structural Model Archive と名付け，BSM-Arch と称し 2018 年 12 月よりサーバーを外部公開して運用している（https://bsma.pdbj.or.jp)。PDBx/mmJSON フォーマットを基本とし，論文にひも付けされていないトラジェクトリーまでカバーできる圧縮フォーマットを開発して公開している。独自に各エントリーに対して DOI を付す機能を実装し，インシリコユニットを初めとする研究者が解析・論文発表したデータの収集している。

（3）連携

3-1. 構造解析ユニット（電子顕微鏡関係者）との連携 I

初年次に，構造解析ユニットの阪大・岩崎グループ（当時)，東大・吉川グループと共同で，Cold Storage に未公開データを保存する方策を検討し，2 年次以降に大容量の Cold Storage のメタデータの効率的作成方法や，データの転送方法（rsync の活用）などを実装して，大容量バックアップの仕組みを構築した。その他に，構造解析ユニットのメンバーから，個別に講習会やチュートリアル利用のための EMPIAR の画像データダウンロードについて問い合わせがあったため，サーバー上に「Workshop」用のリソースとしてファイルサイズの小さな講習会用エントリーを公開している（https://empiar.pdbj.org/workshop)。

3-2. 構造解析ユニット（電子顕微鏡関係者）との連携 II

現在の EMDB には，メタデータとして顕微鏡の種類を記載できるが，どの大学のどの顕微鏡であるかを記載する仕組みとはなっていない。そこで，構造解析ユニット（電子顕微鏡関係）からの要請により，大阪大学が独自に EMDB に構造解析ユニットの電子顕微鏡設置場所を追加でアノテーションする仕組みを構築することにした。新規データベース EMJP（https://emjp.pdb.org）として整備した。

3-3. インシリコユニット（計算科学関係者）との連携

計算結果データベースの開発では，インシリコユニットとの連携が必須である。我々は，初年次にインシリコ・ユニットのユニット会議にオブザーバー出席し，協力を要請すると共に，自動 login システムの構築や入力フォーマットの簡素化など，要望を聞く機会をもった。2 年次以降，インシリコユニットからの要望に応える形で，スターフォーマットによる半自動入力システムを開発し，ORCiD ID でログインするシステムを実装した。新型コロナウイルに関するエントリーについては，PDB と同様に論文公開を待たずに計算結果を公開する仕組みを導入した)。

In this project, we poerated the project to maximize the utilization of the results or the various raw data generated by each platform. My team maintained, managed, and operated the useful PDB-related databases that had been developed in the project, and provided the information publicly necessary for this project. By making these databases available on the Internet, we supported the research activities of external researchers and BINDS project members. The following is a summary of the results and significance of the support and achievement of this project, as well as the project conducted in collaboration with other groups.

(1) Support

1-1 Stable Data-in of the protein structure databases

Our research group has been operating the three international databases, (i) Protein Data Bank (PDB), (ii) Biological Magnetic Resonance Data Bank (BMRB), and (iii) Electron Microscopy Data Bank (EMDB), which are jointly managed by the worldwide Protein Data Bank. The EMDB database in Asia is financially supported by the BINDS project. My team has processed 100% of the entries registered in the EMDB from Japan and other Asian countries and has released the whole data without any delay (https://pdbj.org/emnavi). Since the start of the project in April 2017, the number of monthly registrations has been increasing steadily, and the total number of EMDB entries registered at Osaka University has reached 2000. This is about 15% of the global EMDB data in the world determined during the same period.

1-2. Data-out activity and utilization of protein structure databases

We supported individual structure analysis based on PDB and EMDB databases by using homology modeling, molecular dynamics simulation, and rapid NMR structure determination utilizing AI.PDB-related databases and tools (HOMCOS, hGtoP, Omokage search) developed in previous PDIS projects have been updated, maintained, and managed. We started that the urgent release of SARS-CoV-2 related entries before publication. Normally, registered PDB data are kept private until a study containing the relevant information is published as a paper. However, protein structural information is fundamental information that is actively used in drug discovery, and it was expected that the structural data of SARS-CoV-2 would be public as soon as possible. Therefore, we established a new system in which we individually contacted the authors and strongly recommended the immediate release of the entries without waiting for publication of a paper. The structural data that we have processed and immediately released are already being actively used in drug discovery and structure-based drug-design.

(2) Development

The data generated in the state-of-the-art Life Science research are big, and the technology for handling them is not yet fully developed. In our group at Osaka University, [a] we developed a technology to archive various big data obtained in this project and store them in a relational database, and [b] contributed to develop a new technology to link them to other existing databases.

2-1. development of cryo-EM image database

We started a Japanese mirror site of EMPIAR (Electron Microscopy Public Image Archive), a cryo-electron microscopy image database (https:// empira.pdbj.org). The data size of EMPIAR archive increased faster than expected at the starting time of the project. Therefore, we changed our policy of backing up by redundant disk space and built a cold backup system utilizing LTO tapes and the blue-ray disks (Waseda University). At the end of this project (end of March 2022), 1.7 PB of data has been archived in the EMPIAR-PDBj database at Osaka University and is available to the public. Since there is no similar database in the U.S., the database is accessed not only from Asia but also from the U.S. (especially from the West Coast).

2-2. development of computational model archive

In the field of computational structural biology, such as molecular dynamics and ligand docking, etc., huge data are handled. We constructed an upload system and a public database for these computational models by using the technology already developed in PDBj. The database is operated as a data repository. The database is named Biological Structural Model Archive and is called BSM-

Arch, and the server has been open to the public since December 2018 (https://bsma.pdbj.or.jp). We have developed and published a compressed format that can cover even trajectories from the molecular dynamic calculations. BSM-Arc entries have a unique DOI.

(3) Collaboration with other research groups

3-1. Collaboration with the Electron Microscopy groups I

When staring this project, we collaborated with Iwasaki Group of IPR (Osaka Univ.) and Kikkawa Group (Univ. Tokyo) to find the ways to store unpublished big data in Cold Storage. Later in the project, we constructed a large-scale backup system using the Cold Storage at Waseda Univ. In addition, we have received inquiries from members of the Structural Analysis Unit about downloading EMPIAR image data for individual workshops and tutorials, so we have made a small file size entry for workshops available on the server as a resource for "Workshops".
(https://empiar.pdbj.org/workshop).

3-2. collaboration with the Electron Microscope groups II

The current EMDB includes the entity describing the type of microscope as metadata, but it does not contain which microscope is used at which institution. Therefore, at the request of the Structural Analysis Unit (Electron Microscopy group), we established a system to annotate additional information of which microscope at which institution in the EMDB. The database was named as EMJP (https://emjp.pdb.org) and already published.

3-3. collaboration with the computational biology group

In the first year of the project, we attended the unit meeting of the computational biology group as an observer and asked for their collaboration and had an opportunity to hear their requests for data-archiving. Based on their requests, we developed a semi-automatic input system in the START format and implemented a login system using ORCiD IDs. For entries related to the COVID-19, we introduced a system to release the entries prior for the publication of the paper, similarly to the case of PDB.