蛋白質構造データバンクの国際的な構築と高度化（PDBj）

大阪大学蛋白質研究所
中村 春木

International Development and Advancement of Protein Structure Databank (PDBj)

Haruki Nakamura
Institute for Protein Research, Osaka University

The PDB Japan (PDBj) has developed and managed the three-dimensional structural database for biological macromolecules, as one of the members of the worldwide Protein Data Bank (wwPDB) in collaboration with RCSB-PDB and BMRB in USA, and PDBe in Europe. We have so far processed about a quarter of the all deposited structural data to the wwPDB, and provides them from a newly constructed data management system, PDBj Mine, based on the canonical XML data, PDBML. In addition, we have developed many tools and services to search the similar folds, similar local structures, and similar molecular surfaces, to understand molecular functions.

1. はじめに

本申請の代表研究者である中村春木は、2001 年度から JST-BIRD の支援を受けて日本蛋白質構造データバンク（PDBj）を大阪大学蛋白質研究所内に組織し、米国・欧州と協力して wwPDB (worldwide PDB：国際蛋白質構造データバンク)を設立し、国際協力による PDB データベースの維持・管理とサービス・システムの開発・高度化を進めてきた。本開発研究において PDBj は、wwPDB 設立メンバーの一員として、米国構造バイオインフォマティクス共同機構（Research Collaboratory for Structural Bioinformatics: RCSB-PDB）および欧州分子生物学研究所の PDBe－欧州バイオインフォマティクス研究所(European Bioinformatics Institute: EBI）と、NMR データベースとしての BMRB（BioMagResBank）との協力作業により、蛋白質構造データバンクおよびNMR実験データバンクの維持・管理を継続して進めた。さらに、種々の検索や二次的データベースの開発・公開を行って構造データベースの高度化と標準化をはかり、高度な Web サービス環境を提供する一方、利用者が使いやすいシステムとした。

2. 研究開発の成果

2.1 蛋白質構造データバンクの国際的な構築

PDBj は、wwPDB 設立メンバーの一員として、RCSB-PDB、PDBe、および BMRB との協力作業により、蛋白質構造データバンクと NMR 実験データバンクの構築を行っている[1][2][3]。特にアジア、オセアニア地区からの登録を wwPDB 内では分担しており、その地域

の研究者（登録者とデータ閲覧等の利用者）のため、英語以外に日本語、中国語（簡体字と繁体字)、韓国語の解説ページを設け、利用者の便宜を図っている。

PDBj では、2010 年 12 月末までに 16,345 件の構造データに対する検証・編集・登録の処理をしている。世界全体では同期間の総登録処理数は 66,490 件であり、約 25%を PDBj にて分担処理してきたことになる（図 1）。ちなみに、1970 年代からの世界全体での総累積は 2011 年初頭には 70,303 件に達している。

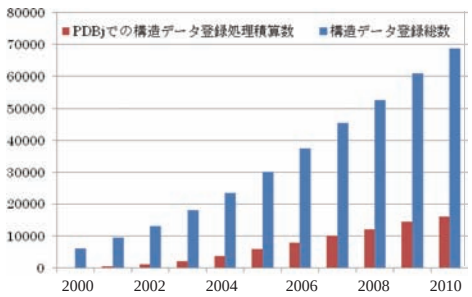

図 1：wwPDB (青)全体と PDBj (赤)の登録数の推移

一方、PDBj-BMRB グループでは、NMR 実験データバンクの構築を行っている米国ウィスコンシン大学マディソン校（PI: John L. Markley 教授）と協力し、生体高分子 NMR データバンクを協力して維持・管理し、NMR 実験データの登録処理を行っている[4]。PDBj-BMRB がデータ登録を 2005 年に始めて以来、2010 年 12 月末までに 760 件の登録処理を実施しており、これは、これまで BMRB 全体に蓄積されたデータの約 12%（総数 6,424 件）に達している。2010 年 12 月からは、NMR によって決定された原子座標の登録には実験データの登録も同時に行うことが必須となり、アノテーションを含めて、wwPDB のメンバーとして対応をしている。さらに、NMR によって構造解析が行われたペプチドや核酸、糖以外の低分子量生体分子構造の登録処理システム（SMSDep: Small Molecule Structure Deposition）を米国 BMRB との協力により、PDBj-BMRB で構築し運営を開始した。

PDBj からの立体構造データファイルのダウンロード数は 2010 年 4 月から 12 月末までの間 15,952,441 件であり、同時期の PDBj Mine によるデータ検索数は 1,913,913 件に達している。2009 年 10 月から、毎週アップデートされる立体構造データの公開を、グリニッジ時刻で毎週水曜日の午前 0 時（日本時間では水曜日の午前 9 時）と定め、PDBj、RCSB-PDB、PDBe-EBI の三ヶ所で同時に行われる新たな仕組みを PDBj が提案し、実施されている。

wwPDB では、毎年、持ち回りで wwPDBAC (wwPDB Advisory Committee: 国際諮問委員会)を開催しているが、2009 年 11 月には、我々PDBj が wwPDBAC 会議を大阪大学蛋白質研究所にて主催し、現状における問題点と将来のための施策を議論した（図 2）。現在、wwPDB では、将来のデータ増加とさらに高精度のデータ検証を実現させるため、共通の仕様による自動的な登録システム（Common Deposition & Annotation）のパイプライン構築を実施中であり、PDBj では特にマスター・フォーマットの検証ツールの開発を実施している。


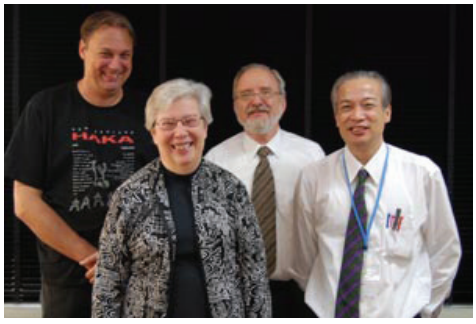
図 2：2009 年 11 月 6 日に大阪大学蛋白質研究所で行われた wwPDBAC 会議のメンバー(左）と wwPDB heads (右）

2.2 PDBj（日本蛋白質構造データバンク）の高度化

我々PDBj におけるデータベース検索・閲覧は、データ記述とその検証がスキーマに沿って極めて厳密に行える XML を用いた PDBML[5]に基づいている。しかし、native XML-DB では階層が深くなると探索に時間がかかることに加え、必ずしも性能が合致するフリーソフトが開発されなかった。そのため、PDBML に基づきつつフリーの relational DB (PostgreSQL) の管理によって高速に処理できる仕組み（PDBj Mine）を独自に開発し、以前の市販の native XML-DB システムに置き換え、新たなデータ管理システムの公開と運営を実施している[6]。この新システムでは、検索数の増加に対応して PC クラスターの増設を行っても、データベースのライセンス費用が高額化しないため、少ない費用で容易に対処できるという利点がある。また、利用者が自由にミラーサイトを構築することも可能となり、既に、Web ページからの PDBj Mine システムのダウンロードが行われている。一方、PDBj の Web ページを、研究者だけでなく一般社会人や学生にもより広く利用していただくため、日本語での利用者インターフェースを充実させた。具体的には、Web ページや利用の手引きの日本語化に加えて、日本語によるキーワード検索とその結果の表示ページの日本語化を行うとともに、ビデオでの日本語による利用法ガイドも作成し配布している。さらに、利用者からの意見を、学会のランチョンセミナー等におけるアンケートで収集し、利用しやすく、かつ SOAP や REST サービスなどによって Web サービスをさらに強化し、より多くの実験情報や機能情報を閲覧できる仕組みとしている。

ところで、構造データがゲノムなどの配列データや文献データと異なる点は、データそのものが本質的にアナログ値であり、検索する場合には原子座標の数値を探すのではなく、「こんな形や、分子表面、動き（ダイナミクスの特徴）をする蛋白質はないか？」という「アナログ検索」が必要とされ、そこから蛋白質の新たな生物学的機能等の考察が可能となる。これらのアナログ検索は、必ずしも世の中には揃っておらず、これまで PDBj では独自の開発を行ってきた。実際、類似フォールドの検索サービス（Structure Navigator）[7][8][9]、蛋白質フォールドの俯瞰的ビューア（Protein Globe）[9]、構造に基づく蛋白質ファミリーの推定（SeSAW）[10]、類似した蛋白質リガンド結合部位の検索（GIRAF）[11]、蛋白質分子表面データベース(*e*F-site)の構築[12]とそれに基づく類似表面の検索（*e*F-seek）[13]、基準振動解析によって計算される蛋白質動的性質のデータベース（*Pro*Mode）[14]を開発した。特に *Pro*Mode については、Elastic Network Model を用いた新たなアルゴリズムにより、原理的に PDB に登録されている全ての蛋白質、核酸、糖についての基準振動解析を可能とする *Pro*Mode-Elastic を開発し、データベースの拡充を進めている。さらに、上記アナログ検索の結果である構造や分子表面の画像を表示するため、アプレットとして、またスタンドアローンとしても利用できる JAVA ベースのフリーの graphic viewer (jV3)も開発した[12]。

一方、阪大蛋白研と九大生医研との協力によって、進化情報を加味して配列・構造から機能を推定するシステムを開発し、公開した。具体的には、(a)フォールドの分類とそれを利用した高速検索システムの構築と公開、(b) 構造・配列統合マルチプル・アラインメント・システム（MAFFTash）の構築と公開、(c) 進化トレース解析の自動化とデータベース化（SEALA: SEquence ALignment Analyzer）の開発と公開、を行った。特に最新の SEALA では、アミノ酸配列アラインメントの各サイトについて保存度、変異度を計算する手法、また進化トレース

法に代表されるアラインメントをいくつかのグループに分割して、そのグループ間での機能的差異を生じるサイトを計算する手法について、これまで報告されている手法を収集し、その計算サービスを実施する。計算結果は、グラフ、テキスト、また立体構造上の表示のいずれかを選択して出力される[15]。さらに、本研究開発のスタート時に計画していたアミノ酸配列のアラインメント情報から、蛋白質のホモロジー・モデルを作成するサービスとして、新たにテンプレート法とフラグメント法を融合させた Spanner を開発し、公開をした。

以上のサービスを実際に利用して、タンパク 3000 などの構造ゲノム科学プロジェクトで構造が解かれた蛋白質の機能推定[16][17]や、マイクロアレイ解析実験で発見された新たな蛋白質の機能推定[18]にも利用され、またリガンド分子認識や蛋白質間相互作用のための共通な部分構造モチーフの同定の研究が実施され（図 3）[19][20]、新たな知見として論文に発表されている。

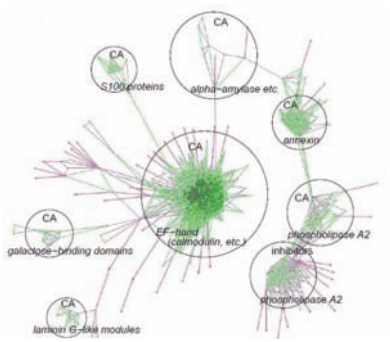

図 3：GIRAF によって見出された、Ca^{2+}イオン結合部位の共通な部分構造の類似性 [19]

図 4：電子顕微鏡構造を PDB データにリンクさせて表示する Yorodumi のページ例。左側の画面はムービー表示となっている。

EBI で収集している 3 次元電子顕微鏡構造データベース（EMDB）のデータ閲覧サービスである EM Navigator を開発し、Yorodumi ポータルからデータ公開を開始した。2010 年 12 月末までに、EMDB の全データ 1,349 件を登録してムービーも作成し、対応する PDB の原子モデルとともに表示・公開した（図 4）。

教育的なサイトとしての英語と日本語による蛋白質構造事典：*e*ProtS (encyclopedia of Protein Structures)では、350 項目の蛋白質とその構造について平易な解説を加え、さらに RCSB-PDB が作成している教育サイトである MOM (Molecule of the Month)については、RCSB-PDB との協力により事前に日本語訳を作成して同時公開している。

2.3 PDBj-BMRB（NMR 化学シフト・データバンク）の高度化

蛋白質、核酸、脂質等の生体高分子の NMR 解析は、構造情報以外にもリガンド相互作用、動的構造転移等の機能解明に寄与する空間、時間分解能の高い情報を与える。この NMR 実験情報を利用価値の高いデータベースとして実現するためには、多様な試料状態と実験法、取得データに応じた高度に組織化したデータ構造を必要とする。PDBj-BMRB では、この多様な情報のために複雑化していた登録・アノテーション作業をシームレスに実現するプログラム BESS を 2009 年度に開発し、年間 300 件以上の理化学研究所からの NMR 化学シフトデータの登録を少人数で実現できた。さらに、複雑な NMR データ構造をオントロジー工学で容易に内部構築しフォーマットの変換を行うコアプログラム MagRO を開発し、標準的な

NMR 解析ソフトウエア Sparky にプラグインとして実装し公開した。また NMR 立体構造アンサンブルから収束領域を自動検出し分離するプログラム fit_robot、および NMR データと立体構造との矛盾点を表示する分子モデルビュアーMagMol も開発、公開した。これらプログラムは、PDBj-BMRB の web サイトからダウンロードされているが、登録データの妥当性評価の点で、登録者の負担軽減だけでなくデータベースの質的向上に大きく貢献すると期待される。

3. まとめ

データの品質を維持しつつ最小限のアノテータによる人手で登録業務を推進し、高度なデータベースとして利用価値を高めるため、PDBj および PDBj-BMRB では、wwPDB における国際連携により、コスト・パフォーマンスが良く、かつ恒久的に持続可能なデータバンク構築を行う努力を続けてきた。ライフサイエンスのデータベースにおいては、その Stakeholders として、データを生産し登録・寄託する研究者、その情報をデータベース化して整理・管理しインターネットで配布する管理者、そしてデータベースを活用する利用者、の 3 者があり、それぞれの役割りと寄与とを相互に理解しあい尊重しあって研究し活動することにより、ライフサイエンスの研究自体に飛躍的な発展が望まれると考えている。国内において多数の構造生物学研究者が価値の高い新奇な構造データを解明しつづけており、今後ともそれらデータを登録・管理する基盤的構造データバンクの役割は極めて重要である。将来にわたり、持続可能な国際的構造データバンクの維持管理に努めていきたい。

最後に、これまで、PDBj および PDBj-BMRB の国際的構造データバンク活動に対するご支援をいただいた JST-BIRD に感謝する。

4. 研究開発実施体制

代表研究者　中村　春木（大阪大学蛋白質研究所）

研究開発題目

(1)　PDBj の国際的運営と高度化

グループリーダー：中村　春木（大阪大学蛋白質研究所）

岩崎　憲治（大阪大学蛋白質研究所）

金城　玲（大阪大学蛋白質研究所）

Daron M. Standley（大阪大学免疫学フロンティア研究センター）

鈴木　博文（大阪大学蛋白質研究所）

輪湖　博（早稲田大学社会科学部）

木下　賢吾（東北大学大学院情報科学研究科）

藤　博幸（CBRC-AIST）

加藤　和貴（CBRC-AIST）

(2)　PDB データ登録・品質管理

グループリーダー：中川　敦史（大阪大学蛋白質研究所）

松浦　孝範（大阪大学蛋白質研究所）

(3) BioMagResBank(BMRB)データベースの構築と高度化

グループリーダー：藤原　敏道（大阪大学蛋白質研究所）

阿久津　秀雄（大阪大学蛋白質研究所）

小林　直宏（大阪大学蛋白質研究所）

5. 参考文献 （下線は、本研究開発プロジェクト参加者）

[1] H. Berman, K. Henrick, H. Nakamura, J. L. Markley, The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucl. Acids Res.* **35**, D301-D303 (2007)

[2] K. Henrick, Z. Feng, W. F. Bluhm, D. Dimitropoulos, J. F. Doreleijers, S. Dutta, J. L. Flippen-Anderson, J. Ionides, C. Kamada, E. Krissinel, C. L. Lawson, J. L. Markley, H. Nakamura, R. Newman, Y. Shimizu, J. Swaminathan, S. Velankar, J. Ory, E. L. Ulrich, W. Vranken, J. Westbrook, R. Yamashita, H. Yang, J. Young, M. Yousufuddin, H. M. Berman, Remediation of the protein data bank archive. *Nucl. Acids Res.* **36**, D426-D433 (2008)

[3] 中村 春木,「蛋白質構造情報の高度化と統合データベース」 *蛋白質核酸酵素* **52** (14), 1897-1905 (2007)

[4] J. L. Markley, E. L. Ulrich, H. M. Berman, K. Henrick, H. Nakamura, H. Akutsu, BioMagResBank (BMRB) as a partner in the Worldwide Protein Data Bank (wwPDB): new policies affecting biomolecular NMR depositions. *J. Biomol. NMR* **40**, 153-155 (2008)

[5] J. Westbrook, N. Ito, H. Nakamura, K. Henrick, H. M. Berman, PDBML: The representation of archival macromolecular structure data in XML. *Bioinformatics* **21**, 988-992 (2005)

[6] A. R. Kinjo, R. Yamashita, H. Nakamura, PDBj Mine: design and implementation of relational database interface for Protein Data Bank Japan, *Database* **2010**, ID baq021 (2010)

[7] D. M. Standley, H. Toh, H. Nakamura, Detecting local structural similarity in proteins by maximizing number of equivalent residues. *PROTEINS* **57**, 381-391 (2004)

[8] D. M. Standley, H. Toh, H. Nakamura, ASH structure alignment package: Sensitivity and selectivity in domain classification. *BMC Bioinformatics* **8**, 116 (2007)

[9] D. M. Standley, A. R. Kinjo, K. Kinoshita, H. Nakamura, Protein structure databases with new Web services for structural biology and biomedical research. *Brief. Bioinfo.* **9**, 276-285 (2008)

[10] D. M. Standley, R. Yamashita, A. R. Kinjo, H. Toh, H. Nakamura, SeSAW: balancing sequence and structural information in protein function mapping. *Bioinformatics* **26**, 1258-1259 (2010)

[11] A. R. Kinjo, H. Nakamura, Similarity search for local protein structures at atomic resolution by exploiting a database management system. *Biophysics* **3**, 75-84 (2007)

[12] K. Kinoshita, H. Nakamura, *e*F-site and PDBjViewer: Database and viewer for protein functional sites. *Bioinformatics* **20**, 1329-1330 (2004)

[13] K. Kinoshita, Y. Murakami, H. Nakamura, *e*F-seek: Prediction of the functional sites of proteins by searching for similar electrostatic potential and molecular surface shape. *Nucl. Acids Res.* **35**, W398-W402 (2007)

[14] H. Wako, M. Kato, S. Endo, *Pro*Mode: a database of normal mode analyses on protein molecules with a full-atom model. *Bioinformatics* **20**, 2035-2043 (2004)

[15] F. Johansson, H. Toh, A comparative study of conservation and variation scores, *BMC Bioinformatics* **11**, 388 (2010)

[16] D. M. Standley, H. Toh, H. Nakamura, Functional annotation by sequence-weighted structure alignments: Statistical analysis and case studies from the Protein 3000 structural genomics project in Japan. *PROTEINS* **72**, 1333-1351 (2008)

[17] D. M. Standley, 中村 春木, 「蛋白質の構造から機能推定へ：構造バイオインフォマティクスによる解析」 *蛋白質核酸酵素* **53** (5), 638-644 (2008)

[18] K. Matsushita, O. Takeuchi, D. M. Standley, Y. Kumagai, T. Kawagoe, T. Miyake, T. Sato, H. Kato, T. Tsujimura, H. Nakamura, S. Akira, The CCCH-type zinc finger protein, Zc3h12a, is a ribonuclease essential for controlling immune responses by regulating mRNA decay. *Nature* **458**, 1185-1190 (2009)

[19] A. R. Kinjo, H. Nakamura, Comprehensive structural classification of ligand binding motifs in proteins. *Structure* **17**, 234-246 (2009)

[20] A. R. Kinjo, H. Nakamura, Geometric similarities of protein-protein interfaces at atomic resolution are only observed within homologous families: An exhaustive structural classification study. *J. Mol. Biol.* **399**, 526-540 (2010)