# 蛋白質立体構造データベースの高度化

大阪大学蛋白質研究所
中村　春木

# Advancement and Standardization of Protein Structure Database

Haruki Nakamura
Institute for Protein Research,
Osaka University

The three-dimensional (3D) structural database for biological macromolecules, Protein Data Bank (PDB), has been developed and managed in collaboration among USA, Europe, and Japan, founding the new international organization, world-wide PDB (wwPDB). We organize the PDB japan (PDBj), which curates, edits and provides the structural data., and construct a new data browser using an XML-DB with the SOAP service. The errors in the conventional format describing the 3D macromolecular structures have been completely repaired with the aid of a new and canonical XML description, PDBML. In addition, we have developed several tools and services: a new molecular graphics viewer, *j*V, which directly parses the PDBML, and search services for the analogous queries of the backbone folds and the molecular surface shapes.

## 1. はじめに

欧米・日本を中心とする国際的な構造ゲノム／構造プロテオミクス・プロジェクトの進展によって、多くの蛋白質立体構造が従来に比べて迅速に決定される時代を迎えつつある。蛋白質立体構造データベース（PDB: Protein Data Bank）に登録されている立体構造データは1990年代から急増し、2006年2月の時点では総計35,000件以上の実験によって決定された構造が登録されている。PDB は地球規模で利用されてきたものの、構造生物学研究の成果をまとめあげただけのデータベースに留まってきた。そこで本研究開発では、蛋白質の立体構造とゲノム情報との結びつきを強める一方、XML などの最新情報技術を利用し、国際的な連携のもとに世界標準としての新しいデータ記述（PDBML）と解析ツールや二次データベースを開発して付加価値を付け加え、構造生物学者だけを対象とする専門的なデータベースから、広く生命科学の研究者、産業界、さらには一般の人にも役立つデータベースに高度化することを目的とした。

## 2. 研究開発の成果概要

### 2.1 PDB データベース業務の日本の分担作業の実施

PDBj では、専門のキュレータを育成し、wwPDB[1]の一員として、日本国内はもとより、アジア・オセアニア地区からの登録を原則的に担当している。PDBj における登録処理件数は、2001年356件、2002年648件、2003年935件、2004年1,586件、2005年2,101件と急増している。世界全体では2005年は6,043件であり、日本の寄与は34.8%にのぼった。

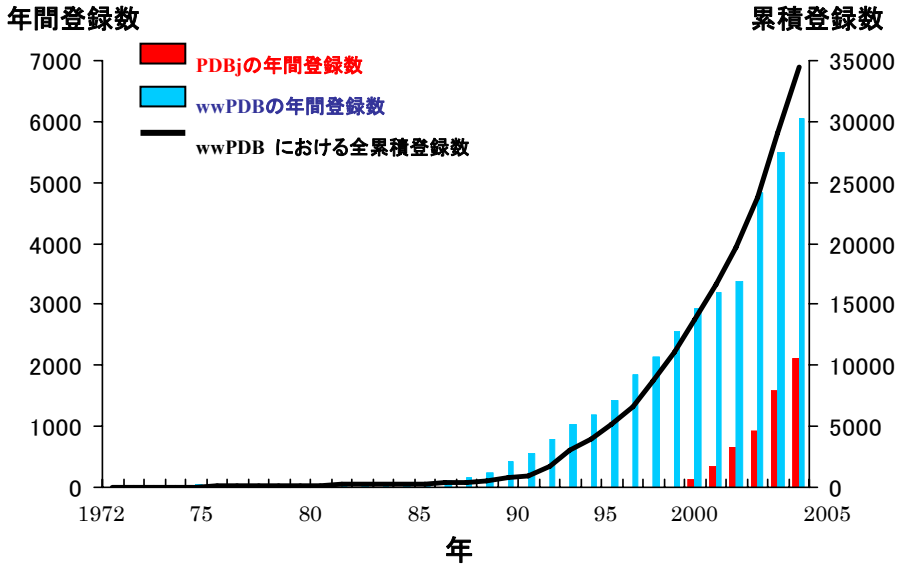


図 1. 日本蛋白質構造データバンク(PDBj)と国際蛋白質構造データバンク(wwPDB)の年間登録数の変遷。

PDBj において生体高分子の立体構造を登録する際には、インターネット上の登録サーバ ADIT を利用する。日本語による案内ページを作成・公開し電子メールによる PDB データ登録に関する質問を受付け、登録者に対する便宜を図っている。(図2-a, b)

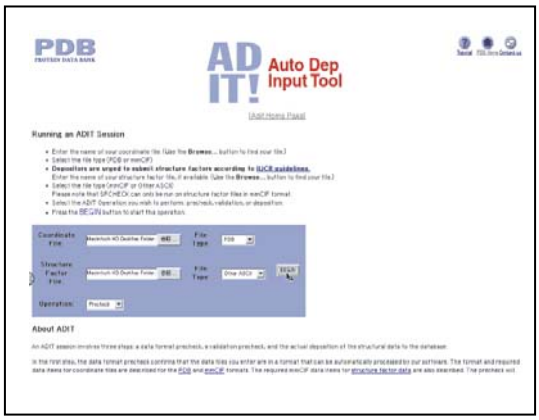


図 2. a) 登録用 ADIT システム。


b) 日本語による登録案内ページ

一方、生体分子に対する核磁気共鳴(NMR)実験データのデータベースである BMRB (BioMagResBank:米国 Wisconsin 大学が開発) のミラーサイトを共同して立ち上げ、また、日本での登録作業も2005年から開始した。これまでに、31件の登録が PDBj においてなされている。

## 2.2 蛋白質立体構造データベースの標準化記述

PDB では、Brookhaven National Laboratory による伝統的な PDB フォーマットと呼ばれるフラットファイル・フォーマットが現在でも標準的に用いられているが、いろいろな歴史的経緯から、記述法が必ずしも統一的でなく、特に古いエントリーに関しては正しくデータがフォーマットどおりに記述されているかどうかの検証(以下、このことを validation と記す)も十分とは言えない状態にあった。そのため、データベースにアクセスする際には、利用者はその多様な記述法と例外処理に悩まされてきた。また、30年前

に作られた全くフラットなデータ記述法のため、現在の高度なデータベース技術の応用には障害となっていた。米国の RCSB では、国際結晶学会が低分子のデータ記述として確立している Crystallographic Information Format (CIF)を、蛋白質や核酸の高分子用に拡張した Macromolecular Crystallographic Information Format (mmCIF) を以前から提案し、厳密かつ確立された定義辞書を構築していたが、mmCIF 形式のデータを読むツールとしてのパーサや validator 等のソフトの開発が遅れたため、バイオインフォマティクスの研究者による利用がほとんどなされていない。また、結晶学者は従来の伝統的な PDB フォーマットをむしろ好んでいたため、結果として、この mmCIF は利用頻度が低いままになっていた。

このため、PDB データそのものの品質管理を将来にわたり保つだけでなく、他のゲノム配列やプロテオームデータベースとのデータグリッドなどによる統合化のためには、validation のためのツールがそろっており、他のデータベースにおいても最近利用が始まった eXtensible Markup Language (XML) を用いた記述が強く望まれていた。我々PDBj のグループは米国の RCSB-PDB と共同し、国際標準としての正規 XML 記述を PDBML という名称で2004年に確立した[2-4](スキーマは http:// pdbml.pdb.org/schema/pdbx.xsd、データは ftp サイトの ftp://beta.rcsb.org/pub/pdb/uniformity/data/XML/all/ に置かれている)。

PDBML の特徴は、辞書が確立している mmCIF との互換性を基本的に保証していることである。mmCIF における name と value とで表現される data item は、XML における element 中の tag と content に対応する。また、従来の DTD を用いた書式定義でなく、浮動小数データ等のテキスト以外のデータ形式も扱える XML Schema が用いられている。XML による一般的な問題として、増大するファイルサイズの問題がある。しかし、多くの場合、検索は residue レベルまでの情報に対して行われるため、データの9割以上を占める座標や温度因子等の原子情報は検索の対象としない別ファイルに格納し、やはり XML 化されたアレイ(多次元)情報としてコンパクトに記述する方式を、extatom 形式という名称で、PDBj からの提案として採用されている(スキーマは http://pdbml.pdb.org/schema/pdbx-v1.005-ext.xsd、データは ftp サイトの ftp://beta.rcsb.org/pub/pdb/uniformity/ data/XML/all-extatom および all-noatom に置いてある)。この方式により、検索が高速となり、またファイルサイズも2～3倍程度に納まっている。

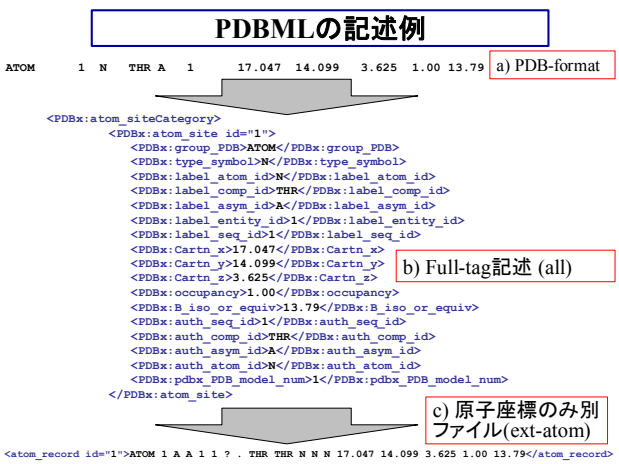


**図 3. a) 従来の PDB フォーマット、b)フルタグの PDBML 書式、c)PDBML のうちの ext－atom 形式による記述例。**

PDBML による記述例として、図3に1つの原子の座標行(atom 行)の記述例を示す。従来の PDB フォーマットでは1行(80文字)以下で簡単に表現できたもの(図3a)が、フルタグ形式の PDBML ではサイズが10倍ほどに膨れ上がっている(図3b)。ただし、図3c の ext－atom 形式ではファイルサイズが爆発せずにすんでいる。この正規 XML 記述である PDBML が完成したおかげで、PDB の全データに対して全く validation エラーをなくすことができた。

さらに、ブックキーピングを行いつつデータ内容

の拡張が容易に行えるというXML書式の特長を活かして、オリジナルのPDBデータに多く欠損している分子機能や実験条件等の情報を文献から抽出して追加する作業を継続して行っている。その数は15,979件(平成18年1月16日付け)に達しており、同時に関連する13,808件の文献をスキャンし、PDFおよび WORD の書式による電子化を行って蓄積しアーカイブも作成している。これらの文献から抽出された情報は XML 化されオリジナルの PDBML データに加え、内部的に PDBMLplus と称する我々独自のデータベースを構築している。

この拡張された PDBMLplus データを基に、native XML-DB による蛋白質構造データ検索システム(xPSSS:xml-based Protein Structure Search Service, http://www.pdbj.org/ xpsss/)を構築し、公開して利用者の便宜を図っている(図4a, b)。この検索システムは、native XML-DB の特徴を生かして、XPath による検索が行える Web サービスも実現し、公開している。この時、XML データの伝送プロトコルである SOAP (Simple Object Access Protocol) を用いることで、xPSSS のブラウザを介した作業を人手で行うことなく、コンピュータ同士で直接情報をやり取りし、網羅的な解析等に適した利用ができる。PDBj では、XPath による検索クエリをパラメータとする SOAP サービスも運用している。

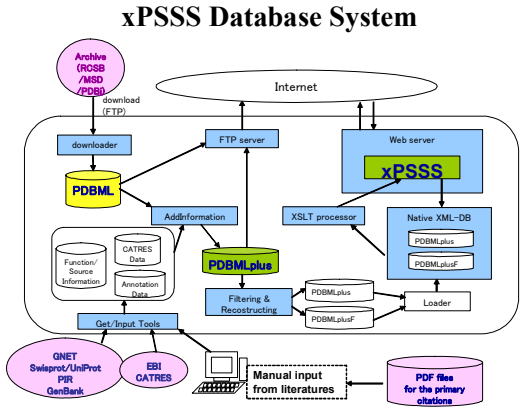


図 4. a) PDBMLplus データベースと xPSSS 検索システム。


b) xPSSS の Top 画面

さらに、これらテキスト・ベースでの情報検索に加えて、2006年3月には、蛋白質の類似構造(フォールド)の検索[5](Structure Navigator)と、eF-site[6]を拡張した蛋白質類似分子表面の検索[7,8]のサービスを開始し、蛋白質構造・形状というアナログデータが直接アナログ情報のままで検索できるようになる。

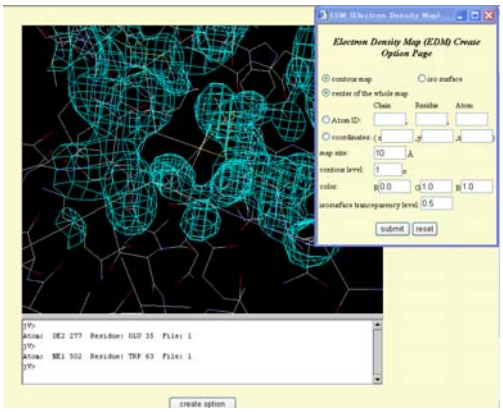


図 5. jV3 ｱﾌﾟﾚｯﾄによる電子密度マップ表示例

一方、PDB にこれまで蓄積されてきた実験データである構造因子から電子密度マップ(2Fo-Fc)を作成して DB 化し、その interactive 画像表示を行うサービスを、*jV*3[6]を利用して運用を開始した。ユーザは、会話的にいくつかのパラメーターを選び、等電子密度線と等値表面による電子密度マップを作成し、可視化される仕組みとなっている(図5)。

### 2.3 蛋白質構造を基にした解析システムの開発と二次データベースの構築

スタンドアローンとしてもアプレットとしても利用できるJava ベースの新しい3Dビューア(PDBjViewer (*jV*))[6]をJOGLを採用して実行速度を確保できるものを開発し、パブリックドメインとしてソースおよびバイナリプログラムの公開を行った。また、このアプレットを用いて、xPSSS、eProtS, eF-site などの PDBj が開発した様々のサービスのビューアに応用している。

一方、蛋白質表面形状と物性に関するデータベース(*e*F-site)[6]、蛋白質ダイナミクス・データベース(ProMode)[9]、マルチプル・アラインメントによる蛋白質立体構造の比較サービス(ASH/GASH)[10]の開発を行い、いずれもWeb ページ上で公開している。

高度化された PDB データベースを基にして、立体構造解析を専門としない一般の生物学・生化学の研究者や大学生および高校生を対象とした、教育用の蛋白質構造データベース(eProtS: encyclopedia of Protein Structures)の英語版と日本語版の公開を継続的に進めた。これまでに210件の蛋白質についての解説を実施している。

### 2.4 wwPDB の活動と諮問委員会の開催、その他の広報・学会活動

国際蛋白質構造データバンク(wwPDB)[1] の第1回諮問委員会 (wwPDBAC:world wide Protein Data Bank Advisory Committee) が、米国ワシントン DC において2004年11月21日に開催した。Structural GenomiX 社の Stephen K. Burley 博士が議長となり、その他に Wayne A. Hendrickson 教授(RCSB, コロンビア大学)を始めとするPDB 各サイトの代表諮問者、国際学会の代表者、wwPDB 各サイトのリーダー、運営開発基金の配分機関代表者ら、総勢16名が参加した。日本からの参加者は、西村 善文 教授(横浜市立大学), 若槻 壮市 教授(フォトンファクトリー)、小池俊行 氏 (JST), 高橋秀貴 氏 (JST)、中村 春木 教授(PDBj, 大阪大学蛋白質研究所)の5名であった。会議は、Helen Berman 教授によるPDBとwwPDBに関するレビュー、そしてwwPDBの3つのサイトの代表者(RCSB, MSD-EBI, PDBj)からそれぞれのサイトの活動状況についての報告をおこなった。次に、参加者全員による wwPDB 諮問委員会のミッションについて合意・確認をした。そして、PDB ファイル・フォーマット、PDB エントリーの改善、X 線結晶解析・NMR スペクトル解析・電子顕微鏡解析の構造生物学実験データの登録問題、実験データの品質管理問題、wwPDB および wwPDB 諮問委員会メンバー間の今後のコミュニケーション等、様々な問題が議論された。

平成17年8月30日には、イタリア・フィレンツェにおいて、wwPDB の第2回助言者委員会が開催された。第1回と同様、Dr. Stephen K. Burley が議長となり、PDB 各サイトの代表諮問者、国際学会の代表者、wwPDB 各サイトのリーダー、運営開発基金の配分機関代表者ら、総勢12名が参加した。日本からは、若槻 壮市 教授(フォトンファクトリー), 由良 敬 研究副主幹(原研量子生命), 中村 春木 教授(PDBj, 大阪大学), 小池俊行 氏 (JST)の4名が参加した。今回の会議では、PDB データ記述の誤記述の修正と統合化、wwPDB と BMRB との関係の強化、電子顕微鏡データベースとの協力、理論モデルの取り扱い、登録数の急増への対応、wwPDB 諮問委員会メンバーによる運営開発基金の配分機関への働きかけの必要性等が議論され、wwPDB のメンバーに具体的な課題が課された。次回第3回国際諮問委員会は、平成18年10月に、東京で開催される International Conference on Structural Genomics のサテライト会議に合わせて、東京近辺にて PDBj が主催して開催する予定となった。PDB

データ記述の誤記述の修正と統合化における日本の分担は、Primary Citation に対しての誤記述や記述の欠落を修正・追加することであり、平成17年2月から開始し、人手によって、多くの誤記述の修正と欠落している文献情報を追加した。

その他の広報・教育関連の活動として、ニュースレターを年間2回発行する他、国際学会（2004年4月第1回環太平洋蛋白質科学国際会議（横浜）、2005年8月第20回国際結晶学会（フィレンツェ））・国内学会（2004年12月日本生物物理学会第42回年会（京都））へのブース出展やシンポジウムの共催による開催を行った。また、国立遺伝学研究所の DDBJ と協力し、大阪大学中之島センター・キャンパス・イノベーションセンターを利用して、DDBJing & PDBjing －講習会 in 大阪－を、2004年3月2日、2006年2月2～3日の2回にわたって開催し、一般社会人や関西、中国地方の大学生・大学院生を対象として、データベース利用法についての講習会を無料で行った。2回とも約25名程度の参加者があり、講演と、自らの PC を持ち込んで行う実習が実施され、好評であった。

**図 6. DDBJing & PDBjing 会場の様子。**

### 3. まとめ

本研究開発を通して、国内では、「日本蛋白質構造データバンク（PDBj:Protein Data Bank japan）」を設立し、複数の国内の大学の研究室をサテライト・サイトとする研究開発体制を構築する一方、国際的には、米国 RCSB や欧州 EBI と共同で「国際蛋白質構造データバンク（worldwide Protein Data Bank:wwPDB）」を設立して国際的な連携を強化させ、研究開発を推進した。当初の計画目標である、（1）PDB データベース業務の日本の分担作業の実施、（2）生体分子磁気共鳴データバンク（BioMagResBank:BMRB）の日本の分担作業の実施、（3）蛋白質立体構造データベースの標準化記述、（4）蛋白質構造を基にした解析システムの開発と二次データベースの構築、（5）日本国内での教育用データベースの作成と公開、はいずれの項目も実施することができた。今後、ますます急増すると思われる生体高分子の立体構造データを精度の高い品質管理を行ってデータベースへ登録するとともに、高度なデータ検索サービスや種々の利用しやすいツールの開発を進め、利用しやすく信頼性の高い蛋白質の構造と機能についての情報源として国内外へ発信し、普及・教育活動等も積極的に行って、社会へ貢献することを目指す。

### 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）

（1）新規蛋白質立体構造データベース構築グループ

グループリーダー　中村　春木（大阪大学蛋白質研究所）

（2）解析システム開発と二次データベースグループ

グループリーダー　中村　春木（大阪大学蛋白質研究所）

（3）PDB データベース管理運営グループ
グループリーダー　楠木　正己（大阪大学蛋白質研究所）
（4）BioMagResBank (BMRB)データベース管理運営グループ
グループリーダー　阿久津　秀雄（大阪大学蛋白質研究所）
（5）教育用蛋白質データベースの作成と公開グループ
グループリーダー　中村　春木（大阪大学蛋白質研究所）

## 5. 参考文献


[1] H. Berman, K. Henrick, H. Nakamura, Announcing the worldwide Protein Data Bank. *Nature Struct. Biol.* **10** (12), 980.(2003).

[2] J. Westbrook, N. Ito, H. Nakamura, K. Henrick, H. M. Berman, PDBML: The representation of archival macromolecular structure data in XML. *Biooinformatics* **21** (7), 988-992 (2005).

[3] 中村春木, PDBjの国際協力による登録作業と品質管理, ユーザサービス. *日本結晶学会誌*, **47** (5), 334-340 (2005).

[4] 中村春木, 特集「バイオデータベースの今」バックボーンデータベースの標準化: PDBj. *情報処理学会会誌*, **47** (3), (2006) 印刷中.

[5] D. M. Standley, H. Toh, H. Nakamura, Detecting local structural similarity in proteins by maximizing number of equivalent residues. *PROTEINS*, **57** (2), .381-391 (2004).

[6] K. Kinoshita, H. Nakamura, eF-site and PDBjViewer: database and viewer for protein functional sites. *Bioinformatics* **20** (8), 1329-1330 (2004).

[7] K. Kinoshita, H. Nakamura, Protein informatics towards function identification. *Curr. Opin. Struct. Biol.* **13** (3), 396-400 (2003).

[8] K. Kinoshita, H. Nakamura, Identification of the ligand binding sites on the molecular surface of proteins. *Protein Science* **14** (3), 711-718 (2005).

[9] H. Wako, M. Kato, and S. Endo, ProMode: a database of normal mode analyses on protein molecules with a full-atom model. *Bioinformatics* **20** (13), 2035 (2004).

[10] D. M. Standley, H. toh, H. Nakamura, GASH: An improved algorithm for maximizing the number of equivalent residues between two protein structures. *BMC Bioinformatics* **6**, 221 (2005).