# データジャーナルとの連携で
# データ解析の自動化へ

## －開発 10 年目を迎えたプロテオーム統合データベース jPOST での世界初の試み－

新潟大学大学院医歯学総合研究科バイオインフォーマティクス分野の奥田修二郎教授、京都大学大学院薬学研究科生体分子計測学分野の石濱泰教授らの研究グループは、公開されているプロテオーム（注 1）のデータをデータベースに収録するに当たって不可欠な「詳細なメタデータ（注 2）」を収集するため、日本プロテオーム学会との協力の下、データ論文を掲載するデータジャーナル（注 3）「Journal of Proteome Data and Methods」を創刊しました。これによってメタデータを提供する研究者にもインセンティブが生まれ、また提出されたメタデータから半自動的に再解析を進める仕組みを整備したことによって、今後のデータベースへの大量データ収録の見通しが立ちました。これは世界的に問題視されている「メタデータ収集問題」への、世界初の直接的な対策です。

**【本研究成果のポイント】**

- プロテオームのデータを再利用してデータベース化する際には詳細なメタデータが必要である。しかしこの収集は困難で、海外でも大きな問題として手付かずに近い。
- 世界初の試みとして、データベースと直接連動してメタデータを収集するデータジャーナルを創刊した。これによって、研究者側にも詳細なメタデータを投稿するインセンティブが生じる。
- 大量のデータ処理に対応できるよう、メタデータを（半）自動で解析するシステムも開発し、今後の大規模データベース構築を目指す。

### Ⅰ．研究の背景

ヒトのドラフトゲノムマップの完成が宣言されてからおよそ四半世紀、「ポストゲノム時代」と謳われる現在の科学研究では、「オープンサイエンス」（注 4）が強く主張されています。「オープンサイエンス」の語には様々な意味合いを含みますが、その一つが研究で使用した測定データ（いわゆる生データ）を公開し、誰もが容易に再利用できるようにする「オープンデータ」の考え方です。プロテオーム分野では、この「オープンデータ」の考え方に基づき、ProteomeXchange コンソーシアムを中心に世界中に複数のデータリポジトリ（注 5）が設けられています。日本からは新潟大学の他、京都大学、熊本大学、北里大学、および情報・システ

ム研究機構(データサイエンス共同利用基盤施設ライフサイエンス統合データベースセンター)の 5 施設が共同して開発と運営を進める jPOST（Japan Proteome Standard Repository/Database）が、この ProteomeXchange コンソーシアムに加盟しています。

jPOST プロジェクトは、国立研究開発法人科学技術振興機構（JST)・ライフサイエンスデータベース統合推進事業の一環として 2015 年 4 月に開始され、今年はプロジェクト開始から 10 年目に当たります。jPOST は、(i) 各研究者の測定した生データ（多くは質量分析のデータ）を寄託し公開するための**リポジトリ**・(ii) データを統一的な基準で独自に解析し直すための**再解析プロトコル**・(iii) 再解析した結果を格納し、一般的なデータベースと同様、様々なキーワードで解析結果を検索・表示できる**データベース**　の 3 要素から構成されます。

このうち**リポジトリ**は 2016 年 5 月に公開され、順調に寄託プロジェクト（データ）数が増加しています。当初はアジア・オセアニア地域からの寄託を想定していたのに対し、相当数の寄託が欧州及び北米からも寄せられていて、この 10 年間に、プロテオミクス（注 1）のための質量分析生データリポジトリとして世界的に完全に認知されたと言えます。

## Ⅱ．研究の概要･成果

jPOST では、リポジトリに登録された生データの再解析のために**新しい信頼性評価基準 UniScore** を開発しています。この再解析を正確に実施するためには、データの属性について説明する**メタデータ**が重要です。jPOST 開設時に比べてデータが大型化･より複雑化しており、再解析にはより詳細なメタデータが必要になったことなど、多くの問題が生じています。そこで、プロジェクト毎に詳細なメタデータをどのように準備するかが最大の問題となっていました。この「メタデータの整備が最大の問題である」という認識は、2017 年 9 月に英国・ヨーロッパ生命情報学研究所（European Bioinformatics Institute; EBI）にて、世界最大規模のプロテオーム･データリポジトリ PRIDE (Proteome Identification Database）の開発メンバーと情報交換した際に指摘されたことでもあります。

この問題に対応するため jPOST では、データの寄託者自身が詳細なメタデータを投稿するための仕組みを考えました。現在はオープンデータの思想に基づいて各研究分野でデータの再利用が進められていますが、プロテオミクスを含む他の分野も同様に、データの再利用（再解析）にはメタデータが必要です。このため現在は、「データ論文」を扱うデータジャーナルが創刊されるようになっています。この認識に基づいて、日本プロテオーム学会では、jPOST と連携し、プロテオームのデータ、特に jPOST での再解析を念頭に置いた「メタデータの専門論文」を電子出版するデータジャーナル「Journal of Proteome Data and Methods (JPDM)」を 2019 年に創刊しました（図 1)。

これによって、

- 研究者はデータを寄託すると共に、*JPDM* に詳細なデータ論文を投稿することで、業績が論文として 1 本増加する
- 研究論文の精査によってメタデータを抽出する場合に相当するメタデータ記入フォーマットである *JPDM* フォーマットが利用できる
- *JPDM* フォーマットで記述されたメタデータは、対応する生データと合わせて自動（少な

くとも半自動）で再解析を実行できる

という、個々の研究者と jPOST の双方にメリットがある、少なくとも**プロテオーム分野では世界初のシステム**が完成しつつあり、今までボトルネックになってきた再解析をようやく加速できるようになりました。

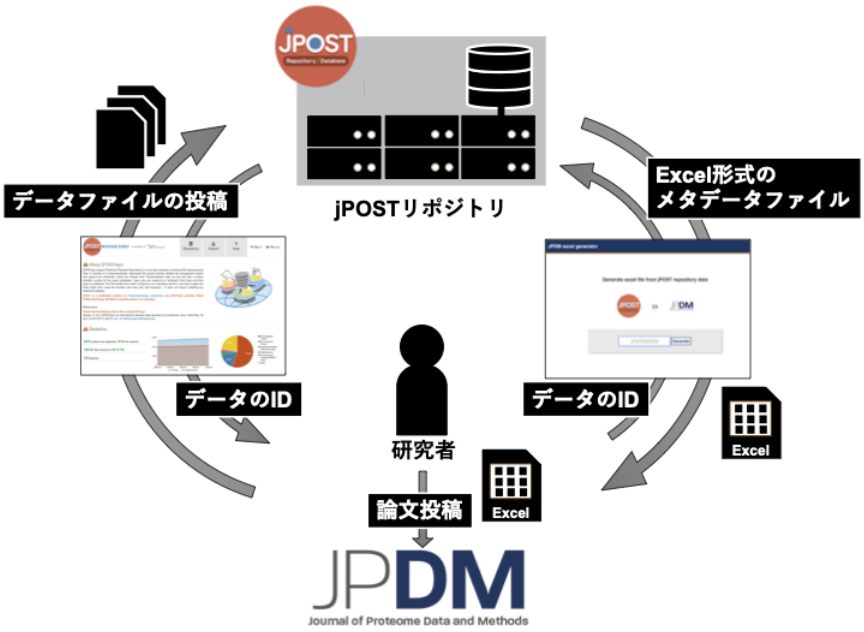


図 1　研究者と jPOST・JPDM の関係

**III．今後の展開**

今後は *JPDM* のデータ論文によるメタデータ収集を jPOST の「第 4 の柱」と位置づけていくことになります。*JPDM* は、jPOST 以外のリポジトリに収録されたデータのメタデータも論文掲載できるように、今後一段と国際化を進めていきます。また *JPDM* を中心としたメタデータ収集システム、及び再解析の半自動化パイプラインが動作を始めることによって、広範な研究対象のデータが再解析され jPOST データベースに収録されていく見通しです。生物種や研究対象を限定しない、基礎研究から疾病研究などへの応用研究まで、幅広い研究フィールドの基盤として、プロテオミクスの研究者に限定しない広い分野の研究者が利用できる世界最大のプロテオーム･データベースの構築を目指していきます。

**IV．研究成果の公表**

本研究成果は、2024 年 11 月 11 日、科学誌「Nucleic Acids Research」のデータベース特集号に掲載されました。

【論文タイトル】jPOST environment accelerates the reuse and reanalysis of public proteome mass spectrometry data

【著者】Shujiro Okuda, Akiyasu C. Yoshizawa, Daiki Kobayashi, Yushi Takahashi, Yu Watanabe, Yuki Moriya, Atsushi Hatano, Tomoyo Takami, Masaki Matsumoto, Norie Araki, Tsuyoshi Tabata, Mio Iwasaki, Naoyuki Sugiyama, Yoshio Kodera, Satoshi Tanaka, Susumu

Goto, Shin Kawano, Yasushi Ishihama
【doi】10.1093/nar/gkae1032

## Ⅴ．謝辞

jPOST の研究・開発は科学技術振興機構(JST)・NBDC 事業推進室による統合化推進プログラム予算によって実施されました。また *JPDM* の発行は、日本学術振興会の科学研究費補助金・研究成果公開促進費（国際情報発信強化(B)）（採択課題番号：21HP2004）の支援を受けて行われています。

## 【用語解説】

（注 1）プロテオーム･プロテオミクス：タンパク質は生命の構造や機能で決定的に重要な役割を持ち、その構造は遺伝子に記録されています。遺伝子に記録された遺伝情報を基にタンパク質が合成されることをタンパク質の発現と呼び、発現しているタンパク質全体の集合のことを**プロテオーム**と呼びます。ただし遺伝子と異なり、タンパク質は特定の条件の下でのみ発現するものもありますから、組織･環境･時間などが異なれば、実際に発現しているタンパク質の種類、プロテオームの中身は変動することもあります。

タンパク質が 1 個のみで生命現象を担っていることは稀で、通常は複数個のタンパク質が関与します。この観点からは、タンパク質の機能は同時に発現している（共発現している）タンパク質を同時に調査することが有効で、例えば疾病の対策などにもプロテオームの研究は重要と考えられます。このようにプロテオームを研究することをプロテオーム解析、或いは**プロテオミクス**と呼んでいます。

（注 2）メタデータ：メタデータとは一般に、何かのデータ（の属性）について説明するデータのことをいいます。例えば携帯電話で写真を撮影すると、その写真を撮影した時刻や撮影地点の緯度経度が自動で記録されますが、これらはその写真のメタデータです。図書館に行くと収蔵されている書籍には図書館分類法に基づく分類ラベルが貼られていますが、この分類情報もその書籍のメタデータと言えます。

プロテオミクスなど生命科学の場合は、そのデータはどの生物種の試料なのか、試料はどのように採取し処理したか、測定ではどのようなパラメータを用いたか、などの多くの情報がメタデータとして必要になります。これらはデータベースに収録して検索用のキーワードとして用いると同時に、データを再利用するために再び解析作業を行う（再解析する）ときにも必要で、この情報が不充分な場合は結果も不充分になる可能性があります。

（注 3）データジャーナル：一般に、データ論文を掲載する科学論文誌をデータジャーナルと呼んでいます。データ論文とは「(公開されている）データを再利用するために必要な情報」すなわちメタデータ（など）を報告するためのもので、データジャーナルとしては、引用数が非常に多いジャーナルとして知られる「Nature」の姉妹ジャーナル「Scientific Data」などが挙げられます。

（注 4）オープンサイエンス：広義には「研究者･非研究者を問わず、誰もが研究成果などの情報にアクセスし、研究活動に寄与できるようにする」考え方･運動を意味します。より狭義な定義としては、例えば日本学術振興会は「オープンサイエンスとは、オープンアクセスと研究データのオープン化を含む概念です」と述べています（https://www.jsps.go.jp/j-policy/open_science/）。具体的には「研究成果は無料で読めるような形で論文発表する」「論文発表した研究のデータは公開する。この結果、他の研究も促進される」といった内容を意味します。例えばプロテオームのデータを公開すると、公開データを多数集積して利用することで、より信頼性の高い解析･同定手法が発明される可能性があります。

（注 5）リポジトリ：一般的には「データの蓄積場所」を意味します。蓄積されるのはデータ（内容）のこともあれば（この場合、一つのファイルに複数のデータが蓄積される）、データファイルのこともあります（この場合、複数のファイルが一カ所に蓄積される）。データベースと異なるのは、データベースは後で（複雑な）検索をすることを念頭に、検索に適した形で全データを予め整理して格納しているのに対し、リポジトリは最小限の記載内容のルールに従うデータがそのまま蓄積されていることです。オミクス解析、特にプロテオミクスの場合はもう少し限定された意味で使われます。多くの学術ジャーナルはオープンデータの原則に従っており、これに基づいて、論文投稿時に（その論文の根拠となる）生データの公開を要求します。これに対応するためのデータ公開場所をリポジトリと呼びます。