プロテオーム統合データベース jPOST を開発
－アジア・オセアニア唯一の国際標準データリポジトリをスタート－

概要

jPOST（Japan ProteOme STandard Repository/Database, http://jpost.org/）[注1] は、京都大学を中心としたオールジャパン体制で開発が進められているプロテオーム[注2] 統合データベースです。国内外に散在している種々のプロテオームデータを標準化・統合化・一元管理したもので、多彩な生物種・翻訳後修飾[注3]・絶対発現量[注4] などの情報を含み、さまざまな解析が可能であるという特徴を有しています。

今回、jPOST データリポジトリ[注5] システムを新たに開発し、全世界に向けて公開しました。本システムは、アジア・オセアニア地域における初めての国際標準プロテオームデータリポジトリであり、2016 年度国際ヒトプロテオーム機構・プロテオミクス標準化構想会議（HUPO-PSI，2016 年 4 月 18-20 日，ベルギー・ゲント市）において、国際標準のデータリポジトリシステムを提供する ProteomeXchange コンソーシアム[注6] への加盟が宣言されました。今後、アジアを中心に世界中のプロテオームデータを jPOST に収集することが可能となります。

１．背景

ポストゲノム時代に入り、生命科学分野において大型国際研究が盛んに行われています。中でも「発現しているすべてのタンパク質」を意味するプロテオームは「生命活動を直接担う」分子群であることから、創薬分野を中心に大きな注目を集め、さまざまな大型研究が行われてきました。成果として得られたプロテオームデータは、データベースの形で欧米を中心に各地で蓄積され、国際連携が進みつつあります。一方、日本においては国際的に連携できるプロテオームデータベースが存在しないため、国産のデータを海外のデータベースに登録せざるを得ないという懸念されるべき状況にありました。

本 jPOST プロジェクトは、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、生命科学系のデータベース統合化を推進する上で今まで抜け落ちていたプロテオームデータベースを開発しようとするもので、2015 年度より開始しました。() 多彩な生物種（ヒト、動物、植物、酵母、細菌など）、翻訳後修飾（リン酸化など）および絶対発現量情報を付加した、世界初の横断的プロテオーム統合データベースの構築を行っています。

世界のプロテオームデータベース開発の現状としては、まず 2010 年に開始された国際ヒトプロテオーム機構（HUPO）によるヒトプロテオームプロジェクト（HPP）が挙げられます。これはヒトの全タンパク質が人体のどこで、いつ、どれだけ発現するのか、それらが疾患においてどのように変化するのかなどの情報を集めた統合データベース（ヒトプロテオームマップ）を国際連携で構築するというものです。しかしこれは世界各国が染色体ごとにデータ収集・解析を分担するシステムであるため、異なるデータ解

析法やフォーマットが混在し、発足から数年たった現在でも、統一されたデータは発表されていません。また 2014 年 5 月に、この HPP とは無関係に、ヒトプロテオームドラフトマップ[注7]が *Nature* 誌で発表されました。しかし、このドラフトマップは、網羅性を上げるためにデータをひたすら寄せ集めた結果、多くの偽陽性情報（実際には発現しないはずのタンパク質が検出される）を含んでおり、HUPO をはじめとするプロテオーム研究コミュニティから、論文や公式の場で「公開すべきではない」という強い批判が繰り返し行われています[注8]。

一方、研究者コミュニティによるデータの検証や研究不正の回避、あるいはデータの再利用による新規の研究などを目的として、「測定データを再利用可能な状態ですべて公開する」ことが国際的に求められています。プロテオーム分野でも、ProteomeXchange コンソーシアム（PXC）が英国・欧州バイオインフォマティクス研究所ならびに米国・システム生物学研究所を中心に結成され、共通のプロテオームデータリポジトリシステムを構築しています。現在のところ、PXC が提供するリポジトリが最も国際的に認知度が高く、日本の研究者もこれらのリポジトリを通じてデータを公開しています。

しかし、PXC 加盟機関は欧米諸国のみであったため、インターネット接続されているとはいえ、リポジトリサイトとの物理的距離によって通信速度が明らかに影響を受け、データの大きさ（ファイルサイズ）によっては、日本からアップロードするのに 1 週間以上かかる場合もありました。すなわち、欧米諸国以外の研究者は数日かけて欧米のサイトにデータをアップロードし、別の研究者はそのデータを再び数日かけてダウンロードする、という極めて効率の悪い状態が続いていました。

２．研究手法・成果

jPOST プロジェクトでは、前述の 2 問題、すなわち

- 「プロテオームデータベースを構築する際、各研究グループからのデータの持ち寄りでは構築が進まず、一方、多数の研究グループが得たデータを収集・解析した場合には、偽陽性情報などデータの品質の問題で批判が生じている」こと、および
- 「データ公開用のリポジトリが欧米にしか存在しないため、日本人研究者にとって非常に効率が悪い」こと

をともに解決することを目指しています。

統合データベースを構築するためにはプロジェクトや研究機関の枠を超えて、大量のプロテオ―ムデータを収集・統合する必要があります。その際に適切な方法でデータを標準化（再解析）することによって、信頼性の問題を解決できます。したがって、「解析結果」ではなく「測定データ」を収集し、これを独自に解析する必要があります。

これらをふまえて、jPOST の構成は、リポジトリ部、データ解析部、およびデータベース部からなっています（図 1）。

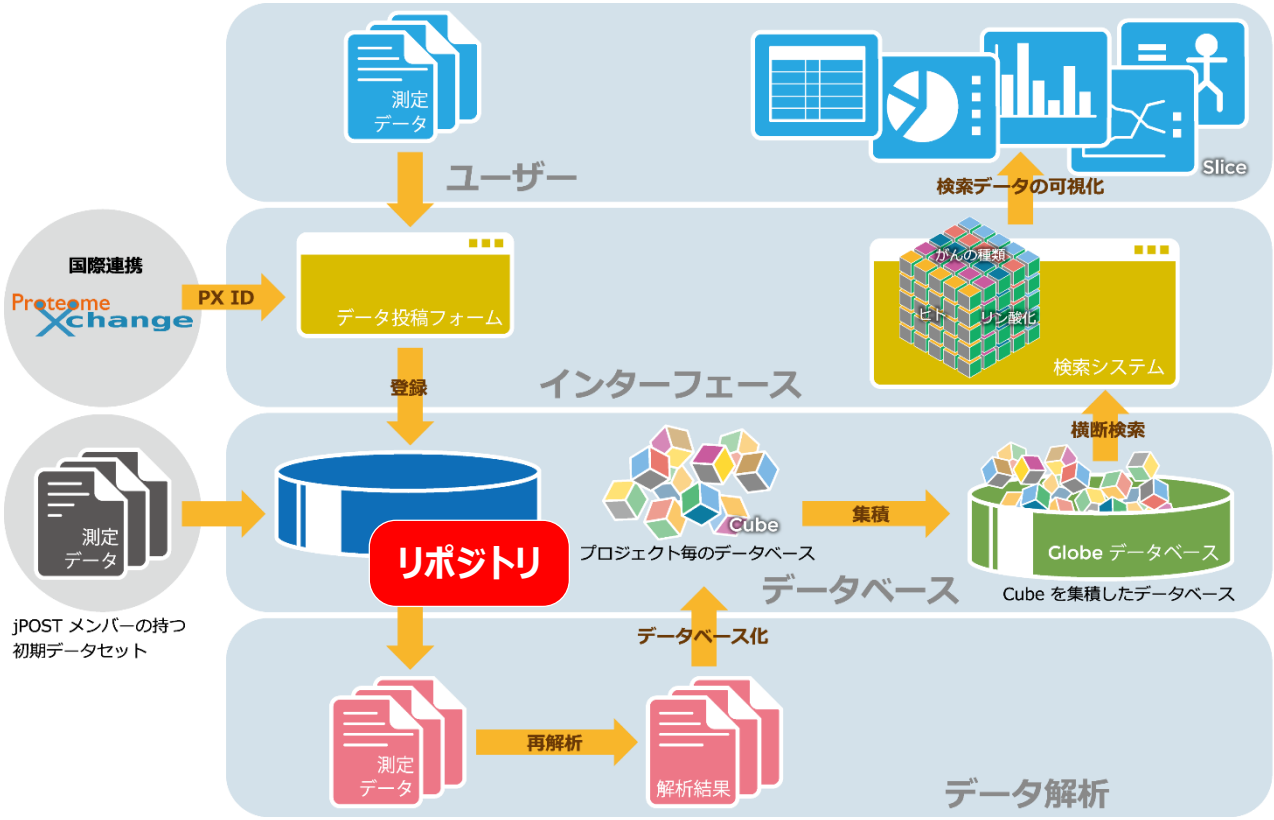

図１　jPOST の構成とデータの流れ

今回運用を開始し、全世界に向けて公開したリポジトリ部には、以下に示すような特徴があります。

- アップロードされるデータについての詳細な属性情報を集めるために、実際の実験プロトコルに即した新規の入力インターフェースを実装しました。既存のリポジトリと比べて、より多くの情報をより少ない手間で入力可能です。
- 2016 年 4 月 18 日～20 日にベルギー・ゲント市にて開催された HUPO-PSI 2016 において、jPOST の PXC への加盟が宣言されました。今後、データのアップロード時にユーザーは PXC が発行する ID（PX ID）を自動で受け取ることができます。
- jPOST リポジトリは、アジア・オセアニア地域で唯一の PXC 加盟プロテオームデータリポジトリであり、これらの地域からのデータのアップロードが極めて高速です。さらに、ファイル送信方式を新たに開発し、既存の PXC 加盟リポジトリと比較して、10 倍以上の高速化を実現しました。

３．波及効果

今回公開したjPOST リポジトリは、「再解析用データの収集」にも大きく寄与します。リポジトリには、一般ユーザーからのアップロード・データに加えて、10万種に及ぶタンパク質リン酸化サイトデータ、ヒトの全タンパク質の絶対定量データなど、jPOST開発チームが有する世界的にも類例のないデータも、初期データセットとして登録されます。これらのデータは、jPOST内部で統一的に再解析され、データベースに収載されます。またデータベースの検索では、生物種や翻訳後修飾、あるいは測定装置など、

データのさまざまな属性情報に基づいて、複雑な絞り込み検索が可能です。さらに、PXC に加盟したことにより、jPOST リポジトリにアップロードされたデータのみならず、他の PXC 加盟リポジトリに登録されたデータも、シームレスに再解析に利用できるようになります（図 2）。

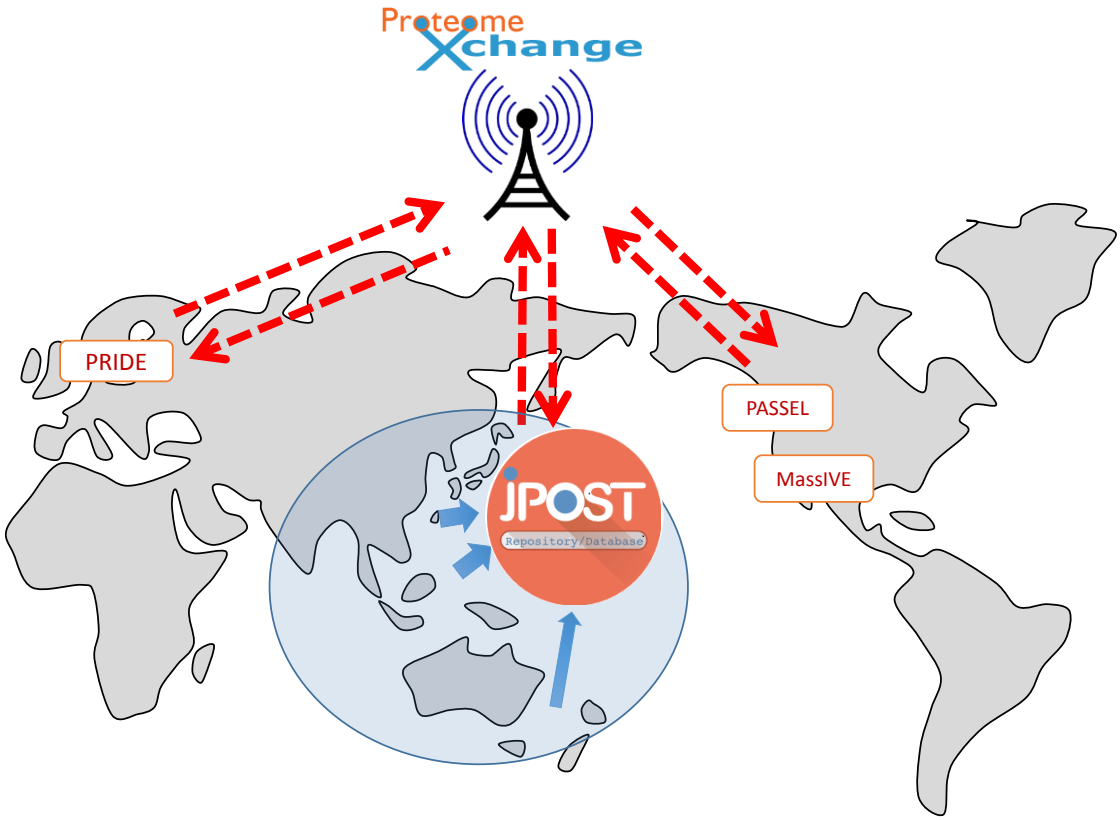

図 2　世界における jPOST の位置づけ

jPOST は、開発当初から生命科学系データベースの将来的な統合化を念頭に置いて設計されています。データアップロードと同時にユーザーに入力してもらう属性情報を用いて、ゲノムデータベースや遺伝子発現データベースなど複数のデータベースを横断した検索が可能になる予定です。したがって、プロテオミクスが関係する、生物学・医学・薬学・農学など生命科学系の、企業・アカデミックを問わずすべての研究者や技術者に対して、本プロジェクトのユニークなプロテオームリポジトリは開かれており、世界的なサイエンスの進歩に貢献できる、信頼性の高いものになることが期待されます。

４．今後の予定

今後は、データの再解析プロトコルを確定し、データベースとして必要な基本システムを開発する予定です。特に、国際的なデータベースであるUniProtKB、neXtProt, HPP、Human Protein Atlas、PeptideAtlas、H-InvDB、DDBJ、PDBや、現在までに日本で構築されたユニークなデータベースである糖鎖科学統合データベース（JCGGDB）、メタボロームデータベース（MassBank）、パスウエイデータベース（KEGG PATHWAY）などのデータベースと連携し、タンパク質単独ではなく、生命科学全般の「データの統合化」の中核となることを目指します。

＜用語解説＞

1. jPOST：：Japan ProteOme STandard Repository/Database の略。2015 年度から科学技術振興機構の支援のもと、6 つの研究機関を中心にオールジャパン体制で開発を行っています。プロテオーム測定生データを収集するリポジトリの提供、データの再解析、再解析データのデータベース化を行っています。
2. プロテオーム：生体中に存在する各種タンパク質のすべて。タンパク質（protein）に「総体」を意味する接尾語（-ome）を組み合わせて、プロテオーム（proteome）と呼ばれるようになりました。
3. 翻訳後修飾：タンパク質が、mRNA 情報に基づいてポリペプチドとして合成（翻訳）された後に受ける様々な修飾のこと。アミノ酸残基への修飾として報告されているものが 300 種以上あり、代表的なものにリン酸化修飾があります。タンパク質は修飾後に本来の機能を発揮するものが多いとされています。
4. 絶対発現量：それぞれのタンパク質について、細胞や組織の中でどのくらい発現しているかを絶対量として表したものです。
5. リポジトリ：データの一元的な貯蔵庫のこと。レポジトリと表記される場合もあります。
6. ProteomeXchange コンソーシアム、PXC：2012 年に英国の欧州バイオインフォマティクス研究所（EBI）および米国のシステム生物学研究所（ISB）の研究者が中心となり、プロテオームデータのリポジトリシステム連合体を組織しました。現在、PXC 加盟リポジトリは、jPOST を含めて EBI が運営する PRIDE、ISB が運営する PASSEL ならびにカリフォルニア州立大学サンディエゴ校（UCSD）が運営する MassIVE の４つが存在します。
7. ヒトプロテオームドラフトマップ：（1） Kim MS, et al., A draft map of the human proteome, Nature, 2014 May 29;509（7502）: 575-81. （2） Wilhelm M, et al., Mass-spectrometry-based draft of the human proteome, Nature, 2014 May 29;509（7502）: 582-7.
   URL：（1） http：//www.nature.com/nature/journal/v509/n7502/full/nature13302.html
   （2） http：//www.nature.com/nature/journal/v509/n7502/full/nature13319.html
8. Ezkurdia I, et al., The potential clinical impact of the release of two drafts of the human proteome, Expert Rev Proteomics. 2015;12（6）: 579-93.