

【日本プロテオーム学会通信】は、日本プロテオーム学会会員の皆様に配信しています。



HUPO-PSI(Proteomics Standards Initiative)ミーティング



参加に関するレポート

日本プロテオーム学会員の皆様

学会員の皆様においてはますますご清栄のこととお慶び申し上げます。

2013年4月15-17日に英国リバプール大学で行われたHUPO-PSI(Proteomics Standards Initiative)ミーティングに参加された河野信会員(情報・システム研究機構 ライフサイエンス統合データベースセンター)がその内容をレポートとしてまとめて下さいましたので、学会員の皆様に配信させていただきます。

ご高覧下さいますようお願い申し上げます。

【日本プロテオーム学会通信】に対するご意見をメールにてお寄せ下さい。

ご意見を【日本プロテオーム学会通信】に掲載希望の場合はその旨お知らせ下さい。

【アドレス変更/配信中止】【ご質問・お問合せ】は、日本プロテオーム学会事務局

(cljhupo@secretariat.ne.jp)にお願いいたします。

## HUPO-PSI/ProteomeXchange meeting 2013 レポート

情報・システム研究機構 ライフサイエンス統合データベースセンター (DBCLS)

河野 信

2013年4月15日から17日にかけてHUPO-PSI(Proteomics Standards Initiative)のミーティングが、18日19日とProteomeXchange(以下PX)ならびにPSIMEx(Proteomics Standards Initiative and International Molecular Exchange)ミーティングが英国リバプール大学で開催された。本稿では筆者が参加したHUPO-PSIミーティングのMass spectrometry/Proteome InformaticsグループとPXミーティングの様子について報告する。

前半のPSIミーティングは約60名が参加して行われた。初日の午前中のセッションは全員が参加して行われ、PSIのチェアであるInstitute for Systems BiologyのEric Deutsch教授による開会宣言に続いて、ホストであるLiverpool大学のAndy Jones教授から、各種標準フォーマットがほぼそろいつつあるPSIの現状と、開発された標準フォーマットを広く使ってもらうための活動の必要性やデータ量の増加に対応するためのデータ圧縮技術など、今後PSIが進むべき方向性についての発表がなされた。続いてLiverpool大学のRob Benyon教授から、実験研究者の立場から見たデータの取り扱いについての話があり、次々と出てくる生データ(raw data)の保存に非常に苦労している点について触れ(現在12台の質量分析計を所有し50TBのデータがあるらしい)、本当に生データをアーカイブする必要があるのか?解析結果だけを公表するのではだめなのか?との提言がなされ、大きな議論をよんだ。さらに、EMBL-EBIのReza Salek教授からメタボロミクス分野の標準化の取り組み(COSMOS=COordination Of Standards In MetabOlomicS)とメタボロームデータのレポジトリ(MetaboLights)、およびこれらのPSIとの連携について講演が行われた。

引き続き、Molecular Interaction (MI), Mass Spectrometry (MS), Proteome Informatics (PI)の各グループから昨年度の進捗状況について報告があった後、MIとMS/PIの2グループに分かれ議論がなされた。筆者の参加した後者のグループでは、

各開発項目の細かな進捗とこれから解決すべき課題についての議論があった。ここで洗い出された課題について、2日目以降に詳細な議論が行われた。

主な議題として、1) プロテオミクス実験に関する最小情報報告ガイドライン (MIAPE: Minimum Information About Proteomics Experiment)、2) mzML などのデータフォーマット、3) データやメタデータを記述する際の標準用語集 (CV: Controlled Vocabularies) の3点について、開発途中で明らかになってきた問題点や新規実験技術への対応などが中心に話し合われた。MIAPE について、これまではあらゆる種類のデータに対して同一のガイドラインを適用していたが、生データ (MIAPE-MS)、ペプチド同定データ (MIAPE-MSI)、定量データ (MIAPE-Quant) それぞれにガイドラインを分離・再整理したことが報告された。データフォーマットに関しては、定量データを記述するための mzQuantML が2月にリリースされたことが報告された。これは、LC-MS を利用したラベルフリー定量法、SILAC、iTRAQ などのラベル化定量法、emPAI に代表されるスペクトルカウントによる定量法を記述可能であり、今後は SRM/MRM を利用した定量法へ拡張していくことで合意した。また、SRM/MRM 実験用のデータフォーマットである TraML についてもその拡張について議論がなされた。また、これらの標準フォーマットはファイルサイズが大きくなる傾向にあるため、データの効率的な圧縮方法について議論され、特に圧縮効率だけでなく圧縮した状態でのデータアクセスの簡便性も視野に入れた圧縮方法について議論された。標準用語については、特にイオンモービリティ MS に関係する用語の追加について話し合われた。その他、メタボロミクス分野との連携や SWATH-MS など新技术への対応、架橋など複数のペプチド間に結合がある場合の取り扱い、オルタナティブスプライシングがある場合の同定タンパク質の取り扱い、バリエーションや修飾を表現可能にする Fasta フォーマットの拡張である PEFF (PSI Extended Fasta Format)、翻訳後修飾の修飾位置が確定できない場合の取り扱い、RNA-seq データのプロテオミクスへの応用などについて活発な議論が展開された。

余談であるが、筆者の参加しなかったもう一方のグループである MI グループの中核となっている PSICQUIC (Proteomics Standard Initiative Common QUery InterfaCe: 世界中のタンパク質相互作用データベースの横断検索を可能としたインターフェイス、25 のデータベースからの1億5千万の相互作用にアクセス可能)

は、筆者の所属する DBCLS が 2008 年から毎年日本で主催しているバイオハッカソン（データベースやツールの開発者による会議）にて、Bruno Aranda 氏が中心となって開発が始まったプロジェクトであり、間接的ながら日本がプロテオミクスデータの標準化へ貢献した例である。

週の後半に行われた PX ミーティングでは約 30 人が参加し、各 Working Package (WP)の進捗報告やプロジェクトに関連するサービスについての紹介がなされた。

WP1 (Management) では、PX プロジェクトは今年末で終了の予定であったが、来年の 6 月まで期間が延長されたことが報告された。WP2 (Data standards development) では、定量プロテオミクスデータの標準である mzQuantML をリリースしたことが報告された。

WP3 (Development of LIMS [Laboratory Information Management System]) では、COLIMS (以前は MS-LIMS X と呼ばれていた) 開発の進捗状況が報告された。このシステムは各種プロテオミクスデータを一括管理可能であるが、標準形式でのデータエクスポート機能も備えており、COLIMS を利用して既に数十のデータが PX にサブミットされたそうである。WP4 (Data deposition) では、MS データについては PRIDE (PRoteomics IDentification Database : ペプチドマスデータのレポジトリ) を、SRM/MRM データについては PASSEL (PeptideAtlas SRM Experiment Library : SRM/MRM データのレポジトリ)を、データのサブミッション先の窓口として整備したことが報告された。会議の時点で 204 のサブミッションがあり、最も多かった国はドイツとアメリカの 24 サブミッションで、日本からも 8 プロジェクトのサブミッションがあったそうである。WP5 (Dissemination pipeline) では、PX データの検索システムである ProteomeCentral についての紹介があった。WP6 (Outreach) では、広報活動と利用拡大のために、去年は講習会を 2 回開催した (ケンブリッジとジュネーブ) こと、e-learning システム

(<http://www.ebi.ac.uk/training/online/course/pride-submissions-proteomexchange>) を作ったことが報告された。

続いて、スペインのグループから HPP (Human Proteome Project) のデータを実際に PX に登録した際の問題点について紹介があった。データサイズが大きいため、転送の途中でタイムアウトしてしまったり、アップロードに時間がかかるなどの問題 (FTP で 27.5GB のデータ転送に 14 時間かかったそうである) が指摘された。やはりデータの大きさに由来するデータの転送速度が大きな問題となっているようである。これは Aspera と呼ばれるファイル転送ソフトウェアを導入することで解決された (79GB のデータを 70 分で転送できた) が、このソフトウェアは有償であり、継続的に使い続けることが難しいとの指摘がなされた。現在、次世代シーケンサのデータを受け入れている ENA (European Nucleotide Archive) のグループがオープンソースのデータ高速転送プロトコルを開発しているとのことで、将来的にはこちらを利用することになるようである。

引き続き、プロテオーム関係のジャーナルの編集者から雑誌の編集方針についての説明があり、これまでと同様に研究の根拠となるデータを公共データベースに登録することを強く推奨する方針が示された。Proteomics 誌の場合、2010 年 1 月から 2012 年 12 月までに出版された 306 報のうち、Dataset Brief では 100%、Research Article, Technical Brief の場合 12-14%のデータが公共データベースに登録されているとのことである。Molecular & Cellular Proteomics 誌では、Tranche プロジェクトの終了によって一時中断している生データの登録義務条項を、Tranche の後継となる massIVE (Mass Spectrometry Interactive Virtual Environment, 後述) が立ち上がり次第、復活させたい意向であることを明らかにした。また、MCP ではグライコミクス分野のデータ登録ガイドラインを新たに制定したこと、今後は SRM/MRM データ登録のガイドラインを今年の秋を目処に作成する意向であることも報告された。

PX プロジェクトに関連するグループからの発表では、PRIDE に登録されているデータから SRM/MRM の transition を推薦する MRmaid、サンプルごとに測定した RNA-seq からペプチドデータベースを作成しマスデータを検索する PIT (Proteomics Informed by Transcriptomics)などが紹介された。これらは公共データベースに登録されているデータを基に開発されたサービス群であり、発表者はそれぞれ公共データレポジトリの重要性を訴えた。

UCSD の Nuno Bandeira 教授からは、Tranche に代わるデータレポジトリとして開発されている massive についての説明がなされた。現在開発中のものは Tranche と同等の機能を提供し、また Tranche データベースからのデータのインポートが進められていることが報告された。ただ残念なことに、既に多数の Tranche サーバが機能を停止しており、全部で 25TB あったデータのうち、最終的にレスキューできたデータはわずか 6TB であったそうである。改めて恒常的なデータベース維持の重要性を感じた。

筆者以外で唯一のアジアからの参加となった中国 Taicang Institute of Life Science Information の Weimin Zhu 教授からは、中国国内で進められているプロテオミクスプロジェクト（Human Liver Proteome Project, Chromosome Centric Human Proteome Project [中国は 1, 8, 20 番染色体担当]、Chinese Human Proteome Project など）の受け皿として、iProX データベースを開発していることが紹介され、PSI 標準の採用やデータ交換など PX コミュニティと協力していく考えが示された。

ディスカッションでは、新たなデータレポジトリが PX コミュニティに参加するためのガイドライン（データフォーマットやオープン性、維持できなくなった場合のデータの取り扱い方法など）や、データ登録に必要なメタデータの改訂について議論がかわされた。特に、再解析データの取り扱い方法（どこにどう登録するかや元データとの関係性の記述方法、PX に登録されていないデータを再解析した場合の取り扱い、再解析データと新規データが混合している場合の取り扱いなど）については活発な議論がなされた。最後に、プロジェクトの今後の課題として、定量データをどのようにして蓄積していくかについて継続的な議論を進めていく方針が確認された。

今回、筆者は初めて HUPO-PSI、PX ミーティングに参加したが、実験研究者・データベース運用者にかかわらず、現状および将来的に予測されるデータ量の膨大さに非常に苦心している様子うかがえた。この問題はもちろん日本も例外ではない。プロテオミクスデータの共有に関して日本の存在感が若干薄いようにも感じるが、

本年は日本で **HUPO** が開催されることもあり、この分野で日本の研究コミュニティからどのような貢献ができるのか考えていく必要性を痛感した会議であった。なお、本会議の発表資料の一部は **HUPO-PSI** のウェブサイトで公開されているので、興味を持たれた方は参考にしていただきたい。



HUPO-PSI ミーティングの参加者。2列目のなぜか真ん中にいるのが筆者。



会議の様子