Proteome Letters　2017；2：67-73

2017 年奨励賞受賞者論文
総説

プロテオームデータの統合化とバイオインフォマティクス解析

奥　田　修二郎 *

*E-mail: okd@med.niigata-u.ac.jp

新潟大学大学院医歯学総合研究科：951-8510 新潟県新潟市中央区旭町通 1-757

（受付 2017 年 10 月 31 日，改訂 2017 年 11 月 15 日，受理 2017 年 11 月 16 日）

プロテオームデータを公開・共有・再利用するために，データ・リポジトリの必要性が高まっていることから，jPOST リポジトリの開発を進めてきた．日本を初めとするアジア・オセアニア地域からのデータ登録における最大の問題点であるデータ転送速度を劇的に改善する世界初の技術の開発と共に，あらゆる点においてユーザーの視点に立ったリポジトリデータベースの開発を実施してきた．また，このような多くのプロテオームデータによりリン酸化のプロファイリングも進んできており，現在では約 20 万サイトものヒトリン酸化サイトがデータベース化されている．生物が進化する過程でこれらのリン酸化サイトやそのリン酸化モチーフの配列を変化させてきたが，特定の生物種から出現するリン酸化モチーフを網羅的に調べる手法の開発にも従事してきた．ここではこれらのプロテオーム研究についてバイオインフォマティクス研究者という立場で紹介する．

1　序　論

質量分析技術の発展により大量のプロテオミクスデータを取得できる時代になっている．質量分析計から出力されるのはスペクトルの情報であり，その生データは，その後様々な計算機による処理を受けて初めて人が理解できる形にまで変換される．この計算機処理に関わるソフトウェア・ツールの開発はバイオインフォマティクスと呼ばれる分野の研究者を含む情報科学の専門家が開発していることが多い．プロテオミクス領域も，いわゆるオミクス分野であるため，最終的に得られるペプチドやタンパク質の情報が非常に多くなる．そのため，情報科学に精通した研究者がデータ分析・解析を実施している例は非常に多い．このように個々の研究にバイオインフォマティクスの研究者が関与する例もあるが，大量のプロテオミクスデータを収集し，同一プロトコルで再解析するメタ解析のような研究では，バイオインフォマティクス研究者が最初から研究に従事するという例も最近では増えつつある．いずれにせよ，計算機でデータを処理する場合，大きな問題としてデータのフォーマットや用語の問題が挙げられる．一般的にプログラムでのデータ処理はプログラムに書かれたことが書かれたとおりに実行されるだけであり，人間のように柔軟にデータを解釈することはしない．したがって，複数の研究者のデータを合わせて解析するといった場合には，規格に沿った同じフォーマットのデータであることや，利用している専門用語が統一されていることが望ましい．しかしながら，世界中の研究者に対してそれを強制するのは極めて困難である．実際の現場では，その都度，必要なデータのみでフォーマットの調整や用語の変換などを実施していることが多い．プロテオミクスに限らず，こういったデータフォーマットの標準化やオントロジーを利用した用語の統一化が必要であることが認識されており，世界標準を決めるためのコンソーシアムがガイドラインの策定等を実施している．プロテオミクス分野では，Proteome Standard Initiative が HUPO の傘下で構成されており，HUPO-PSI という会議を年に数度開催している[1]．この HUPO-PSI では，プロテオミクスに関わるあらゆる標準規格を定めるため，質量分析やプロテオミクスインフォマティクス等の 5 つのワーキンググループが結成され議論が続けられており，様々なデータフォーマットがこの活動から提案されている．

データフォーマットの標準化と共にデータシェアにとって重要なことが，そのデータを恒久的に保存し続けるためのリポジトリデータベースの存在である．質量分析に基づくプロテオミクスデータは PRIDE[2] あるいは PeptideAtlas[3] というリポジトリに長年の間データ登録されてきた．2011 年に ProteomeXchange（PX）コンソーシアム[4] が正式発足し，プロテオミクスデータの保管・共有についてガイドラインが作成され，パートナーのリポジトリデータベースへの登録データを一定の基準で統合している．PX のデータポータルサイトである Proteome

Table 1 PX partner proteomics data repositories

Database name	URL	Organization
PRIDE	http://www.ebi.ac.uk/pride/	EMBL-EBI, European Bioinformatics Institute, Cambridge, UK
PASSEL	http://www.peptideatlas.org/passel/	Institute for Systems Biology, Seattle, WA, USA
MassIVE	http://massive.ucsd.edu/	University of California San Diego, CA, USA
jPOST repository	http://jpost.org/	Several institutions, Japan

Central[5] では，実験条件と技術条件のメタ情報を記述する PX XML という統一フォーマットを用いて情報共有が図られており，パートナーリポジトリに登録される全てのデータセットの情報が集約されている．現在，PX コンソーシアムに正式に承認されたリポジトリデータベースは，PRIDE，PASSEL[6]，MassIVE およびこの後で解説する jPOST[7] の 4 つが存在し，それぞれ PX リポジトリとしての共通の仕様でデータの受け入れを実施している事に加え，独自の機能の開発を実施している（Table 1）．

2　プロテオーム統合データベースプロジェクト

これまで日本においてプロテオームデータの登録をする場合，質量分析データであれば，英国 PRIDE あるいは米国 MassIVE，SRM データであれば，米国 PASSEL を利用するしか無かった．その際，データのアップロードは FTP 接続を利用することが多く，日本から遙か離れた英米国までの距離でのアップロードではデータ転送の速度が非常に遅く，遅延（レイテンシ）が頻発するという問題があった．また，最近になって PRIDE は Aspera という高速データ転送技術を利用してアップロード出来るように機能拡充を実施したが，この Aspera を利用するには，インターネット上の特殊なポート番号の利用が必要で，セキュリティの厳しい大学等の機関では利用できないこともある．これらの問題は，アジアにリポジトリの拠点が無いことが主な原因であるため，日本にもリポジトリデータベースを構築することは，日本を始めとしたアジア諸国におけるプロテオミクス興隆のためにも必須となっていた．そこで，京都大学薬学研究科石濱泰教授をリーダーとした jPOST（Japan ProteOme STandard）プロジェクトが立ち上がり，日本初のプロテオームリポジトリの開発をするに至った．jPOST プロジェクトは，2015 年度より科学技術振興機構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実施されている．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目指しスタートした本プロジェクトでは，最初にデータの受け皿となるリポジトリを開発すべく活動を開始した．

jPOST リポジトリは，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 7 月からは PX コンソーシアムの正式メンバーとしてデータベースが運用されている（Fig. 1）．jPOST リポジトリには, いくつかの特筆すべき特徴がある. その一つが, 最初に問題になっていたデータ転送の速度問題を解決する全く新しい技術の開発である．プロテオミクス分野以外でも DNA 配列を登録するリポジトリ等においても，通信速度が問題視されてきた．レイテンシの発生により FTP を使った通信は距離と共に通信速度が極端に落ちることが知られているが，無料で使うことが可能なソフトウェアがあることから，FTP は非常によく利用されている．また，IBM 社が提供する Aspera は距離に依存しにくい高速通信を実現しているが，非常に高額な上に，限定的な通信ポートを利用することが実用上問題になることがある．それらの問題点を解決するため，通常のインターネット通信である HTTP を利用した高速データ転送技術を開発した．簡単にアルゴリズムを説明すると，アップロードしたいファイルを，“chunk”と呼ぶ小さなサイズのデータに分割し，それを並列で転送する．レイテンシ問題は HTTP でも発生するが, 並列化することで, 無駄を省くことが可能である. この技術により，jPOST リポジトリへのデータアップロードは 1 GB のデータが平均 3 分で転送が終了する．この速度は世界中からのデータ転送での平均値であり，国内だけに限るとその約 2 倍の速度が得られる（Fig. 2）．この速度は PRIDE 等欧米のリポジトリサイトへの FTP 転送に比べ, 数倍から 10 倍程度高速である．また，jPOST リポジトリへのデータ登録はすべてウェブブラウザ内で完結するように設計されており，登録作業を始めようと思ったところから，最終的に登録完了までを一貫して実施できる環境としている．このようなユーザビリティの向上と，高速なデータ転送速度の実現によりユーザー数は飛躍的に増えている（Fig. 3）．また，アジア・オセアニアでのリポジトリ拠点という当初の目論見以上に世界中のユーザーが，プロテオミクスデータのリポジトリとして jPOST リポジトリを選択している．まだ，リポジトリとして公開してから 1 年半しか経過していないため, 本家である PRIDE のユーザー数・登録数と比べると見劣りするが，近い将来，同等かそれ以上に利用されるリポジトリとして成長する要素を充分に含む非常に優れたデータベースシステムの構築に成功したと言える．また，jPOST プロジェクトではリポジトリに登録された生データやメタ情報を統一したプロトコルで再解

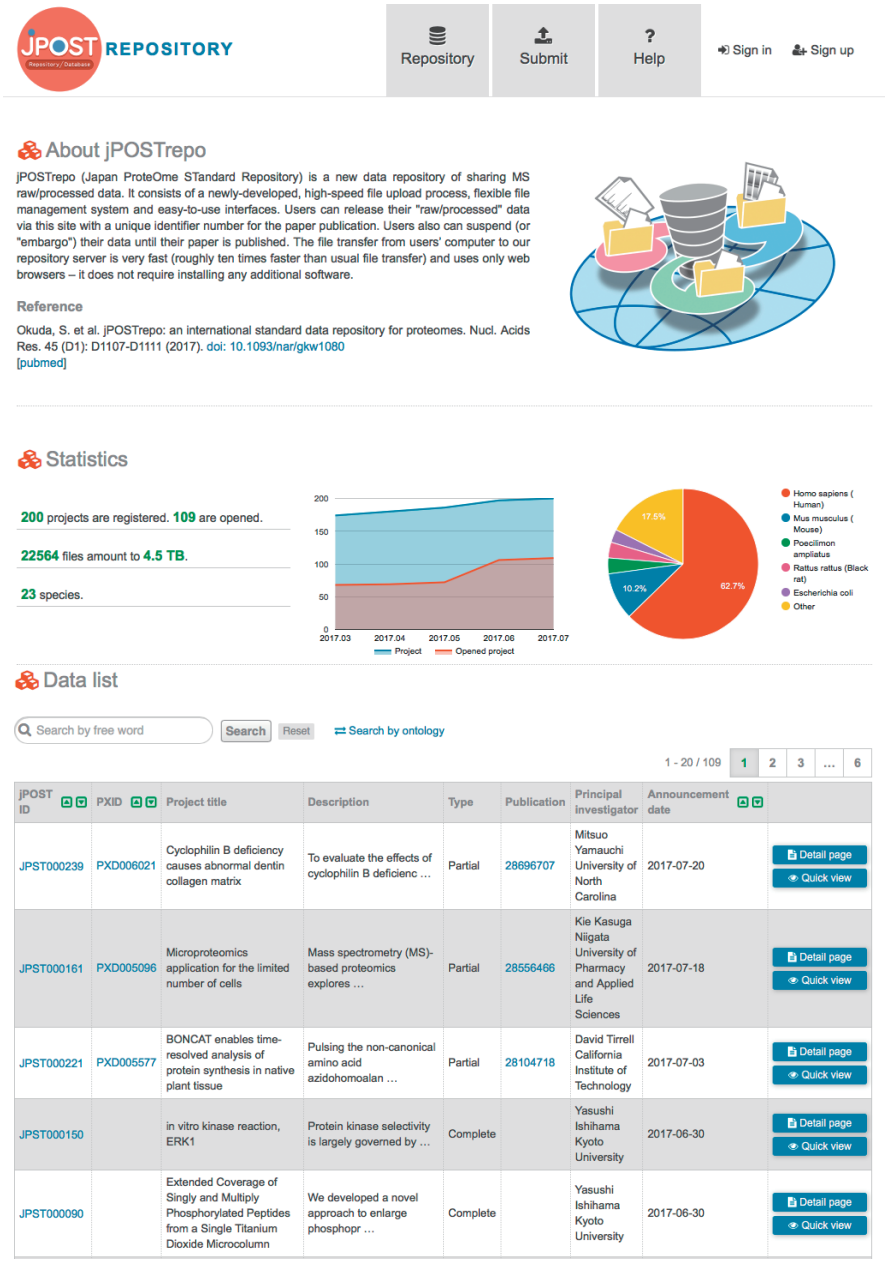

Fig. 1 jPOST repository web site.

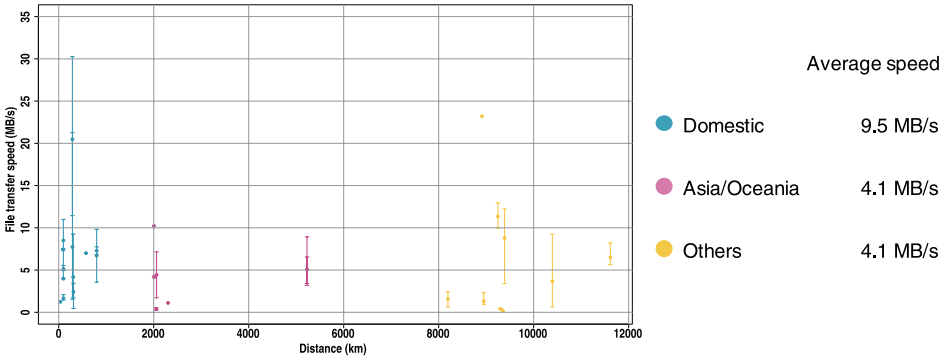

Fig. 2 File transfer speed to jPOST repository.

析する大規模なメタ解析を実施している．これにより同じ基準で得られる大量の結果をあらゆる側面から比較できるデータベースを構築する予定である．

3　リン酸化サイトの進化

これまで述べてきたようにリポジトリデータベースの整備等を通して大量のプロテオミクスデータが公開されている．その中には翻訳後修飾に注目したものもあり，中でもリン酸化の研究は非常に多い．現在，ヒトリン酸化サイトについては 20 万サイト以上が同定され，それらの情報がデータベース化されている．タンパク質のリン酸化は，キナーゼにより触媒される翻訳後修飾であり，転写調節や細胞分化などあらゆる生理活性に関与している．このリン酸化を触媒するキナーゼはヒトにおいて 500 種以上同定されており，これらのキナーゼとリン酸を付与されるリン酸化サイトの関連について多くの研究が進められてきた．多くのリン酸化プロテオミクス研究と，それにより同定された大量のリン酸化サイトの情報が蓄積しているが，それらリン酸化サイトの多くは生理的意義が低い可能性も示唆されている[8]．したがって，膨大なリン酸化サイトの中から意味のあるものを抽出する手法の開発が望まれている．

大量のリン酸化サイトを何かしらの指標でフィルターし，生理的に意味のあるものを絞り込むことができれば，リン酸化プロテオミクスの研究の発展に寄与できるはずである．このフィルターとしてアミノ酸配列の他の生物種での保存性を考慮する，つまり，リン酸化サイトの進化を考える，という考え方がしばしば唱えられてきた．同じ機能のタンパク質配列は種を超えても非常に似ているが，タンパク質中のある一つのアミノ酸であってもそれに生理的に重要な意味がある場合は，長い進化の過程でも変わらない（保存される）可能性が高くなる．つまり，リン酸化サイトでもこのような分子進化の概念を応用することが可能である．いくつかのシステムバイオロジー研究がこういった進化的保存性に着目し，非常によく保存されるリン酸化サイトの生理的重要性を見出してきた[9),10)]．著者らのグループでもリン酸化サイトの生理的重要性を見出してきた．著者らのグループもリン酸化サイトの進化的保存性に着目しているが，これまでの研究とは若干異なるアプローチで生理的に重要なリン酸化サイトの同定方法を開発した．それはリン酸化モチーフとしての進化的保存性を比較するというものである[11),12)]．一般的にキナーゼがリン酸化する場所は，ある特徴的な認識配列が存在する．この認識配列をリン酸化モチーフと呼ぶが，このモチーフ配列を網羅的に同定する．約 20 万サイトあるヒトの既知リン酸化サイトとその周辺領域の配列を抽出し，それらをクラスタリングすることでおよそ 200 種類のリン酸化モチーフの同定に成功した．これらのモチーフはヒトに保存されているものであるが，他の複数の生物種での保存性を評価することとした．16 のリファレンス生物種を設定し，オーソログタンパク質の配列をマルチプルアラインメントした時に，リン酸化モチーフのリン酸化サイトがヒトを起点にしてどの生物種にまで保存されているかを測定する．この時，進化系統の関係はユニバーサルなものを採用し，一定の進化の順序が存在するものと仮定している．各生物種まで保存しているモチーフ数の平均を計算することで，リン酸化モチーフの進化パターンをグラフとして観察することが可能になる．このグラフには 2 つの特徴があることを見出した．一つは，全体的にゆっくりと進化してきたリン酸化モチーフの場合で，このようなリン酸化モチーフはリニアな相関を示す．それに対して，ある生物種が進化上特異的に獲得したリン酸化モチーフの場合，その生物種の前後で保存性が極端に変化する傾向を示す．これをシグモイド型の進化パ

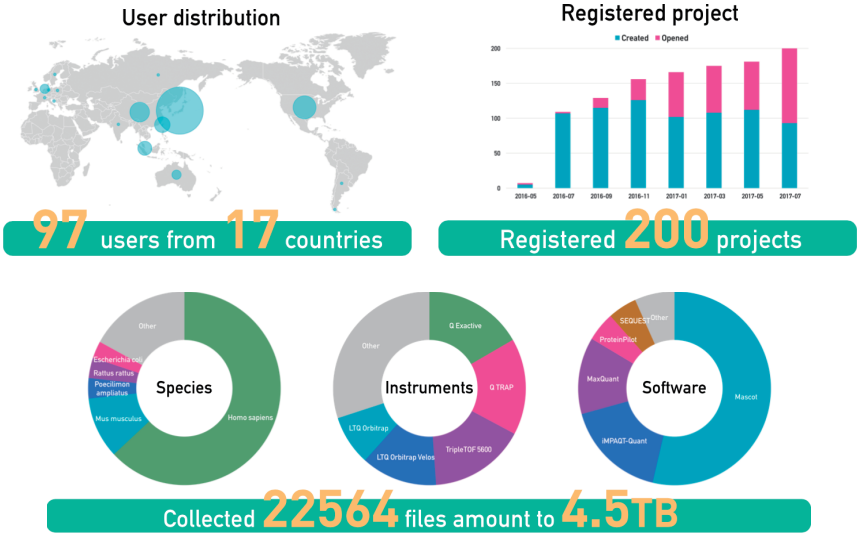

Fig. 3　Statistics of jPOST repository.

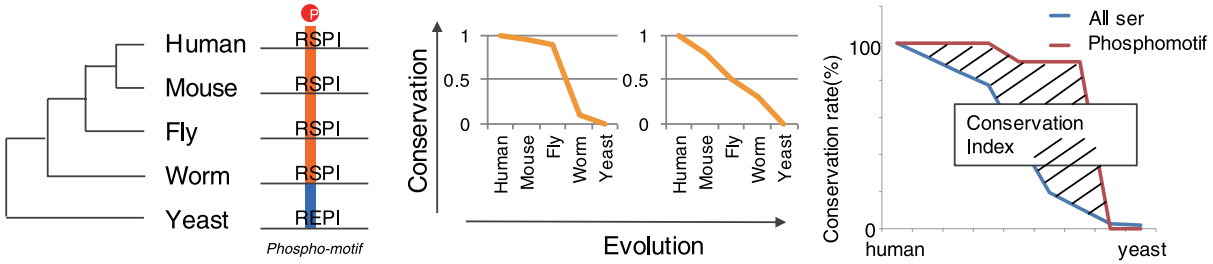

Fig. 4 Method to identifying evolutionary specific phosphomotifs.

ターンと呼んでいる．これらリニア型とシグモイド型のリン酸化モチーフの進化パターンを評価するために，保存指標（Conservation index）として，ヒトにおけるすべてのリン酸化サイトの保存度の進化パターンをリファレンスとして，それぞれのリン酸化モチーフの進化パターンとの差を計算することにした（Fig. 4）．これにより，バックグラウンドとして通常起こりうる進化パターンの影響を除いた保存度について評価できる．このようにして得られたリン酸化モチーフとして，線虫，ショウジョウバエ，ゼブラフィッシュにおける特徴的な進化パターンの発見に至った．これらの進化パターンを示すリン酸化モチーフは，線虫の場合，キナーゼの調節サイトに存在し，そのシグナル制御に関与している可能性が示唆された．また，ショウジョウバエでは，遺伝子発現の調節に関与するジンクフィンガーモチーフ内に特徴的に存在し，細胞内局在にも寄与するモチーフであることを実験的に示した．ゼブラフィッシュの進化パターンは，分子間相互作用ネットワーク解析と合わせ，選択的スプライシングに関するリン酸化シグナルの機能拡大に貢献する可能性が示唆された．このようにリン酸化モチーフの進化パターンの解析は，新規な生理的意味に繋がる情報を得ることができる可能性がある．先に紹介したjPOSTプロジェクトで開発するデータベースにおいても，このような翻訳後修飾サイトにおける進化的特徴を簡便に比較できるシステムの開発を予定している．

4　結　論

バイオインフォマティクスという分野は，新規知識の抽出や技術の開発を実施するため，多様なデータベースを駆使する分野である．これは知識や情報がきれいにまとまっていることが，プログラム的に情報処理するのに必須だからである．ヒトのような臨機応変な対応が可能な時代は，人工知能がヒトの知能に匹敵するまで待たなければならない．そう意味では，今回紹介したリン酸化サイトの進化の解析が実現できた最大の理由は，大量のリン酸化サイトのデータをまとめたデータベースが存在していたからとも言える．数十万というリン酸化サイトの情報を様々な論文やリポジトリのデータから抽出し，リファレンスとなるタンパク質配列にマッピングする，というような作業がリン酸化サイトデータベース構築に必要となる．このような地道な作業の連続からデータベースは成り立っているが，データベース構築という事業はさまざまな研究機関・研究室で実施されており，世界中にありとあらゆるデータベースが開発されている．しかしながら，継続運用のための費用が取れなくなるという理由から途中で開発・運用ができなくなるケースが起こり得る．プロテオミクス分野の場合，Tranche というリポジトリデータベースが存在していたが，維持されなくなった上に，保存されていたデータは一部しか救済されること無く，消えてしまったということがある．データベースの恒常的維持はこのようなデータ損失そのものを防ぐことと，そのデータによる検証・解析の機会損失を防ぐ，という二重の意味で非常に重要なテーマである．著者らのグループが開発しているjPOSTも同様であり，大勢の研究者らによって提供される貴重なデータを永続的に維持できるよう堅牢なシステムを開発するとともにその維持・運用に係る費用等についても持続可能な方法論を常に模索する必要を感じている．

謝　辞

リン酸化サイトの研究は，JSPS 科学研究費補助金の支援の下実施された．共同研究者である金沢医科大学・吉崎尚良講師に感謝する．また，jPOST リポジトリは，JST バイオサイエンスデータベースセンター（NBDC）の「統合化推進プログラム」の支援により開発されたものである．jPOST プロジェクトメンバー全員に心より感謝したい．

著者らに開示すべき利益相反状態は無い．

文　献

1）Deutsch EW, Albar JP, Binz PA, *et al.* Development of data representation standards by the human proteome organization proteomics standards initiative. J Am Med Inform Assoc. 2015;22:495–506.
2）Vizcaino JA, Csordas A, del-Toro N, *et al.* 2016 update of the PRIDE database and its related tools. Nucleic Acids Research. 2016;44:447–456.
3）Deutsch EW, Lam H, Aebersold R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. EMBO Reports. 2008;9:429–434.

4) Deutsch EW, Csordas A, Sun Z, *et al.* The ProteomeXchange Consortium in 2017: supporting the cultural change in proteomics public data deposition. Nucleic Acids Res. 2017;45:D1100–D1106.
5) Vizcaino JA, Deutsch EW, Wang R, *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat Biotechnol. 2014;32:223–226.
6) Farrah T, Deutsch EW, Kreisberg R, *et al.* PASSEL: the PeptideAtlas SRMexperiment library. Proteomics. 2012;12:1170–1175.
7) Okuda S, Watanabe Y, Moriya Y, *et al.* jPOSTrepo: an international standard data repository for proteomes. Nucleic Acids Res. 2017;45:D1107–D1111.
8) Landry CR, Levy ED, Michnick SW. Weak functional constraints on phosphoproteomes. Trends Genet. 2009;25(5):193–197.
9) Pincus D, Letunic I, Bork P, *et al.* Evolution of the phosphotyrosine signaling machinery in premetazoan lineages. Proc Natl Acad Scie U. S. A. 2008;105(28):9680–9684.
10) Tan CS, Pasculescu A, Lim W, *et al.* Positive selection of tyrosine loss in metazoan evolution. Science. 2009;325(5948):1686–1688.
11) Yoshizaki H, Okuda S. Elucidation of the evolutionary expansion of phosphorylation signaling networks using comparative phosphomotif analysis. BMC Genomics. 2014;15:546.
12) Yoshizaki H, Okuda S. Large scale analysis of evolutionary history of phosphorylation motifs in the human genome. GigaScience. 2015;4:21.

Integration of Proteome Data and Bioinformatics Analysis

Shujiro Okuda*

*E-mail: okd@med.niigata-u.ac.jp

Niigata University Graduate School of Medical and Dental Sciences, 1-757 Asahimachi-dori, Chuo-ku, Niigata 951-8510, Japan

(Received: October 2, 2017; Revised: November 15, 2017; Accepted: November 16, 2017)

In order to facilitate the sharing and reuse of promising datasets, it is important to construct appropriate, high-quality public data repositories. The jPOST repository has successfully implemented several unique features such as a high-speed file uploading system. This repository has been launched as a public repository containing various proteomic datasets and is available for researchers worldwide. In addition, our repository has joined the ProteomeXchange consortium. This repository thus contributes to a global alliance to share and store all datasets from a wide variety of proteomics experiments. In addition, comparative genomic analysis of phosphorylation has been performed. A computational method to investigate evolutionary patterns in acquisition of phosphomotifs and relationships between motif structures has been developed and we have reported the kinase substrates associated with phosphoproteins and the evolutionary conservations of kinase groups, revealing physiological roles of unreported phosphosites.

Keywords: database; evolution; jPOST; phosphorylation.