

ぶ ん せ き

別 刷

No. 12

2005

プロテオミクス解析の分析技術

プロテオームデータベース

戸田 年総・森 澤 拓

社 団 日 本 分 析 化 学 会
法 人

東京都品川区西五反田 1 丁目 26 番 2 号 五反田サンハイツ 304 号



プロテオームデータベース

戸田 年 総, 森 澤 拓

1 はじめに

プロテオミクスは新しい形のタンパク質研究手法の一つであり、これによって従来のタンパク質研究がすべてプロテオミクスに置き換えられるというものではない。しかし、特定の組織や細胞で発現されるタンパク質を網羅的に分離同定したり、分化や癌化^{がん}などの細胞変化に伴って変動するタンパク質を探索し同定するといったような研究では、これまでのように大量のサンプルから調製のクロマトグラフィーを繰り返し、タンパク質を精製する必要が全くなくなったために、圧倒的に微量のサンプルで、極めて短期間に結果が出せるようになった。

そもそもプロテオーム¹⁾とは、特定の細胞が特定の条件に置かれた時に発現する全タンパク質の総称で、プロテオミクスはこれを網羅的に解析する研究手法のことである。したがってプロテオミクスは、高分解能のタンパク質分離法と高感度のタンパク質同定法の組み合わせによって成立していると言って良い。最近ではさらに、構造プロテオミクスや機能プロテオミクスといった領域も生まれ、網羅的に構造解析を行うということも加わったが、いまだにその中心となっている発現プロテオミクスや疾患プロテオミクスでは、二次元電気泳動によるタンパク質の分離とディフュゼンシャルな画像解析、ゲル内消化、質量分析による同定が実施されている。筆者らも、老化のバイオマーカーを探索するための手法としてプロテオミクスを利用しているが、プロテオミクスでは一つのサンプルから得られる情報量が極めて多く、大量に発生する情報を一旦デジタルデータとして保存し、その中から必要な情報を取り出して解析するためのプロテオーム情報管理システムが必要になってきた。これに加え、他の研究者が行ったゲノム研究やプロテオーム研究の成果もインターネット上で容易にアクセスできるよう

になっており、それらの膨大な情報を有効に活用し複雑な生命現象を解明しようとするバイオインフォマティクスがますます重要になってきている。

インターネット上で公開されている生命科学関連のデータベースは、内容を相互に補完しながら発展しており、タンパク質関連のデータベースもゲノムデータベースや相互作用データベース²⁾と連携をとり合いながら、さらに充実されていく方向に進んでいる。実際筆者らは、質量分析によって得られたペプチドの質量情報（ペプチドマスフィンガープリント）をもとにタンパク質を同定する際に、Matrix Science 社の Mascot や、UCSF の ProteinProspector などのサーチエンジンを利用して、検索対象としては Swiss-Prot のオリジナルのデータベースのほか、DNA の塩基配列をアミノ酸配列に置き換えた TrEMBL や NCBIr などの翻訳データベースも利用しており、さらに EST (Expresses Sequence Tag) の情報をもとに作られた dbEST などのデータベースも加わって、利用できる情報の範囲はますます広がっていく傾向を示している。

そもそもタンパク質の同定は、二次元電気泳動のパターン上で興味深い変化を示したものに対して実施されることが多く、その際には蛍光色素などで染色された二次元電気泳動ゲル（2D ゲル）をスキャナーや CCD カメラで読み取り、スポットの検出、マッチング、定量化などが行われる。したがって、二次元電気泳動に基づいたプロテオーム解析では、2D ゲル画像とこれをコンピュータ解析して得られた数値データがベースに存在する。さらにゲルのスポットを切り出し、ゲル内消化を行って質量分析を行った場合には、消化を行ったマルチウエルプレート上での配置と MALDI のターゲットプレート上での配置、質量スペクトル、MS-Fit や Mascot 検索によって同定されたタンパク質名などの情報が加わる（図 1）。さらに、使用したサンプルに関する情報やタンパク質の抽出・分離条件等に関する情報、関連

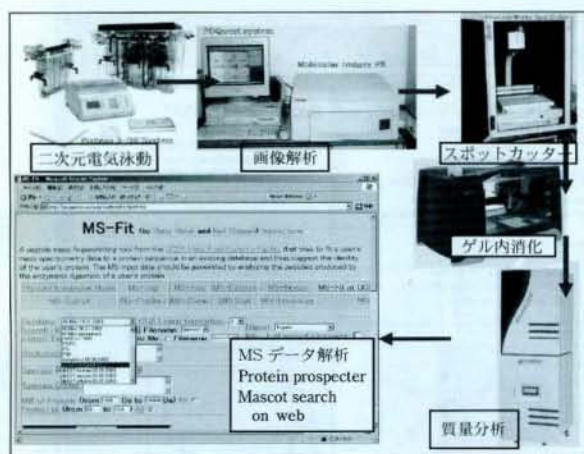


図1 二次元電気泳動に基づいたプロテオーム解析の流れ

する研究の文献情報などもプロテオーム解析に関連する重要な情報である。このようにプロテオーム研究においては、様々な形の情報が大量に発生するので、これらの情報を電子化し、ネット上で管理、共有および検索を行うことのできるシステムが必要となってくる。

すでに筆者ら自身が、インターネットサーバー上に『プロテオームデータベース』を作成し、成果の一部を公開しているが、それに加えて所内のLANを通じて最新のプロテオーム研究情報を共有し、共同研究を円滑に行っていくための研究情報管理システム (laboratory information management system, LIMS) の開発も行っている。本稿では、現在筆者らが開発中のシステムについて紹介する。

2 公開プロテオームデータベース

最近では、「ショットガン法」と呼ばれるような、二次元 LC/ESI-TOFMS によるプロテオーム解析や、SELDI (surface-enhanced laser desorption/ionization) 法などのプロテインチップを用いる方法なども行われているが、現在公開されているプロテオームデータベースの多くは、2Dゲル画像上のスポットにタンパク質の情報をリンクさせた2DPAGEデータベース (two-dimensional polyacrylamide gel electrophoresis database) の形態をとっている。その最大の理由は、2DPAGEデータベースが、タンパク質の発現レベルと翻訳後修飾に関連する情報などの多様な情報を効果的に参照することに最も適しているからである。最もよく利用されている2DPAGEデータベースの一つは、Swiss Institute of BioinformaticsのExPASyサーバー上で公開されているSWISS-2DPAGE³⁾である。ExPASyサーバーではSWISS-2DPAGEのほか、SWISS-PROT (アミノ酸配列データベース) や、PROSITE (タンパク質ファミリーやドメイン構造に関するデータベース) なども公開されており、SWISS-2DPAGEで

表1 インターネット上で公開されているプロテオームデータベース

SWISS-2DPAGE	Swiss Institute of Bioinformatics
SIENA-2DPAGE	Molecular Biology, University of Siena, Italy
PMMA-2DPAGE	Purkyne Military Medical Academy, Czech
RAT HEART-2DPAGE	German Heart Institute Berlin
HEART-2DPAGE	German Heart Institute Berlin
HSC-2DPAGE	Heart Science Centre, Harefield Hospital
PPDB	PhosphoProtein DataBase Inc.
The UAB proteomics Database	UAB Comprehensive Cancer Center, University of Alabama
Database of Human T Helper Cells	Turku Centre for Biotechnology, Finland
TMIG-2DPAGE	Tokyo Metropolitan Institute of Gerontology

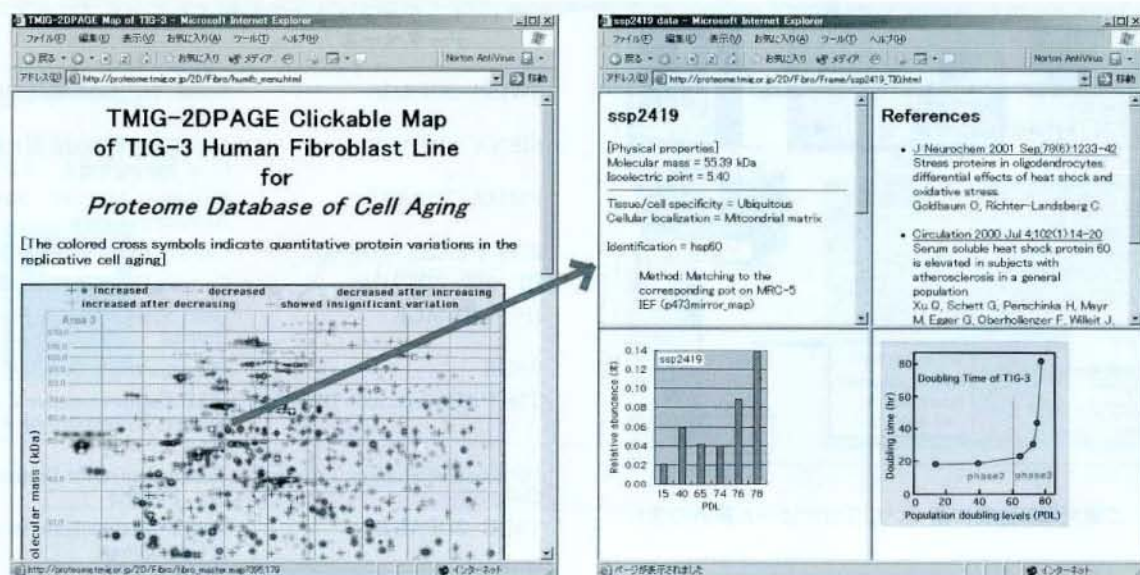
はこれら各種データベースにリンクする形がとられている。また、データベース作成用のパッケージであるMake2D-DBII⁴⁾などソフトウェアも提供されている。

このほかにも、表1のように世界各地のサーバー上で、様々な2DPAGEデータベースが公開されている。各々のサイトでは特定の生物種、臓器 (組織)、細胞株について、限られた範囲内でデータベースの構築が行われているが、インターネット上で公開されている他のデータベースと相互にリンクを掛け合うことによって、情報の連結が実現されている。ここでは、筆者らが運用を行っているTMIG-2DPAGEを例にとって、2DPAGEデータベースの構築方法と運用システムの概要を説明する。

3 2DPAGE データベースの構築方法

公開用のプロテオームデータベースを構築するためには、民間プロバイダーのレンタルサーバーを利用するか、専用ウェブサーバーを立ち上げる必要がある。レンタルサーバーを利用する場合、データベースの構築と公開に必要な機能が備わっていることを確認する必要がある。一方、専用のウェブサーバーを立ち上げる場合には、ファイアーウォールなどセキュリティの確保に留意しなければならない。セキュリティ向上のためには、データベースの構築と運用に必要なプロセスは極力停止すべきである。筆者らは研究所内の独自サーバーとレンタルサーバーを併用している。

ウェブサーバーのソフトウェアとしては、Apacheが最も広く利用されており、筆者らもこれを利用している。Apacheを利用してプロテオームデータベースを構築する場合、Apacheのクリックابلマップ機能が利用できるため、比較的簡単に2Dゲル画像上のスポットから個々のタンパク質の情報にリンクを張ることができ、と



クリックブルマップ機能 (Apache の imap 機能を利用して各スポットの詳細情報が表示される)

図2 HTML形式で作られたTMIG-2DPAGE

りあえずプロテオーム解析結果をインターネット上で公開する場合などには大変便利である。実際にクリックブルマップ機能を利用するためには、Apacheの動作設定ファイルであるhttpd.confファイルを若干書き換え、拡張子がmapのファイルを扱うことできるようにした上で、htmlファイル内に下記のような一文を書き入れることにより、mapファイルに書かれたスポット情報にURLと二次元電気泳動画像上の実際のスポットが連結され、2Dゲル画像上の特定のスポットをクリックしたときに、そのスポット情報が表示されるようになる(図2)。

```
<A HREF="二次元電気泳動.map"><IMG ISM
AP SRC="二次元電気泳動画像.gif"></A>
```

mapファイルの各行には、下記のようにcircle+スポット情報を表示するファイル名+スポット位置の円の中心ピクセル位置(x座標,y座標)+円周上のピクセル位置(x座標,y座標)を列記する。

```
circle スポット情報.html 167,455 164,452
```

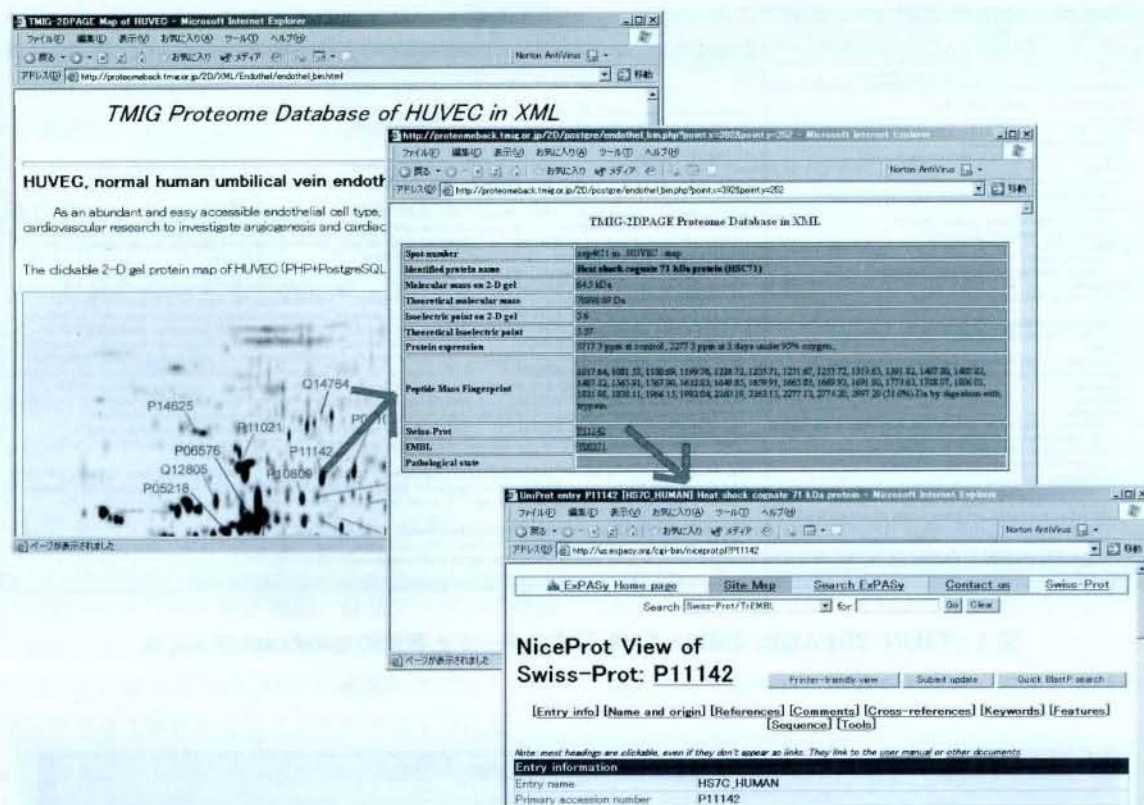
このように、スポットの位置情報を一つのテキストファイル(mapファイル)に書き込むだけで、位置情報の数だけファイルリンクが作れ、二次元電気泳動画像から各種タンパク質関連情報ページへのアクセスポイントを容易に作成することができるので、初期の頃は筆者らも専らこのやり方で情報公開を行っていた⁵⁾⁶⁾。しかしこのやり方では、スポットの数だけスポット情報ファイル(スポット情報.html)を用意しておく必要があり、複数の電気泳動画像上の多数のスポットを関連付けて管理したり、キーワード検索機能を付加することが難し

い。そこで現在では、後述のRDBMSを組み合わせた方法に切り替えている⁷⁾⁸⁾。

クリックブルマップ用の画像としては、GIF形式やJPEG形式などの2Dゲル画像が利用される。PDQuestなどの二次元電気泳動画像解析ソフトでは、スキャナーで取り込んだ生画像をRaw 2D Imageと呼び、これに対しバックグラウンドレベル補正やノイズ、ストリークの除去処理が行われた画像をFiltered Imageと呼んでいる。スポットの検出や、濃淡の数値化、マッチング処理、ディファレンシャル分析などは、通常このFiltered Imageに対して行われる。筆者らはPDQuestのエクスポート機能を用いて、Filtered Imageを、TIFF形式などの画像ファイルとして出力し、これをPhotoshopなどの画像編集ソフトで開いて等電点や分子量などの目盛や数値などを書き加え、画像のサイズ調整、圧縮変換などを行った後にJPEGフォーマットのファイルとして保存し、クリックブルマップ用の画像として用いている。

4 リレーショナルデータベース管理システム(RDBMS)の導入

Apacheのクリックブルマップ機能を用いてウェブサーバー上に構築されたデータベースに、さらにキーワード検索機能を追加する場合、一般にはPerlなどのプログラム言語で作られた検索ソフトが用いられているが、複数の2Dゲル画像上の多数のスポットを連結させて管理することは難しい。これに対しMostaguir⁴⁾は、structured query language (SQL)の利用を提案している。そこで筆者らも、オープンソースのOSであるLinux上に、カリフォルニア大学バークレー校で開発されたSQLの一種であるPostgreSQLを導入し、リレー



ここでは、Apache-PHP-PostgreSQL の組み合わせによってクリックابلマップ機能が実現されている。

図3 XML形式で作られたTMIG-2DPAGE

ジョナルなデータベース管理システム (RDBMS) の構築を試みた。実際にウェブサーバー上で RDBMS の機能を実現しようとする、Apache と PostgreSQL を仲介するためのインターフェースが必要となる。これに対し筆者らは、サーバーサイド埋め込み型スクリプト言語である Hypertext Preprocessor (PHP) を用い、Apache-PHP-PostgreSQL の 3 者の連携により、リレーショナルなデータベースの開発環境を実現している (図3)。PHP には画像をハンドリングするための GD ライブラリを装備することができるため、ウェブサーバー上で圧縮フォーマットの画像ファイルを用いて、作図等の画像編集が行える。筆者らのプロテオームデータベースでは、クリックابلマップ機能を実現するために、map ファイルによる方法と GD ライブラリによる方法を共存させている。さらに、PHP のプログラミングにより XML フォーマットへの対応も行えるようにしており、どの方式がベストであるか比較検討を行っている。

5 XML フォーマットへの対応

Human Proteome Organization (HUPO) は、2004 年 2 月プロテオーム研究のための標準情報交換フォーマットとして XML フォーマットを発表した。eXtended Markup Language という名の通り、テキストに対する注釈という形態を基本にする XML フォーマットは、

データのツリー構造を表現することに適しており、様々なデータ構造に柔軟に対応できる。そのため、情報交換フォーマットとして優れ、インターネット上のデジタルデータフォーマットとして確固たる地位を占めている。HUPO が提案した XML フォーマットに対しては、情報交換の目的だけでなくプロテオームデータベースの標準化を先導する役割が期待されているが、それには、データのマイニングや解析に耐えられる汎用のプロテオームデータベースのデータフォーマットとの互換性が前提となり、今後プロテオームデータベースのデータ出力の XML 化およびフォーマット変換アプリケーションの開発が鍵を握っている。筆者らは、以前から簡易的な方法で制作されたプロテオームデータベースを公開しているが、さらにデータの再解析に有利な独自の XML フォーマットでのデータの出力を可能にする新しいウェブインターフェースの開発を行っている。XML フォーマットは、インターネット上での情報交換技術と親和性が高く、XSL ファイルによって表示形態を容易に変更することができる。図4は、筆者らが運用している RDBMS 方式のプロテオームデータベースから出力される XML フォーマットのデータのサンプルと、その表示方法を指定する XSL ファイルの一例である。XML フォーマットの特徴は、データの意味を示すタグが使われていることである。直感的にデータ構造を理解しやす

データを含んでおらず、ディファレンシャルな解析結果をどのように扱うかという課題が残されている。これに対し、筆者らが開発中のプロテオームデータベースでは、2Dゲル画像間ディファレンシャルな解析データも取り扱えるような構造を目指している。さらに筆者らは、各研究者がデータベースからダウンロードしたXMLフォーマットのデータを、自分たちの二次元電気泳動パターン解析に活用するためのビューア（TMIG-2DPAGE Viewer）を開発した¹²⁾。図5は、プロトタイプのビューアの画面である。プロテオームデータベースの画面からXMLフォーマットのデータをWindowsパソコンにダウンロードし、さらに対応する二次元電気泳動画像をダウンロードする。TMIG-2DPAGE ViewerでXMLフォーマットデータと画像を一旦Windowsパソコン上に取り込んだ後は、オフラインでクリックマッピング機能やキーワード検索機能、簡易マッチング機能などを利用して、詳細な情報を取り出すことができる。またこのとき、自分たちの2Dゲル画像のスポットの位置とマッチさせて解析を行うこともできるようになっている。このようにXMLフォーマットは、長所と短所を理解したうえでプロテオームの情報の交換に用いることが重要である。

6 laboratory information management system (LIMS)

近年、特に創薬標的の探索を目的とするプロテオーム研究では、実験データの保守と管理が要求され、よりセキュリティの高い研究情報管理システムLIMSを利用することの重要性が指摘されている。

一般的にLIMSには、実験データの保持機能に加え、実験操作の検証機能、データのバックアップ機能、データの抽出と交換機能、データマイニング機能、公共データベースとの情報連結機能などが求められる。そこで筆者らは図6に示すように、一方ではSWISS-PROTなどの公開データベースとのリンクを充実させ

ながら、他方では高いレベルでセキュリティを保って施設内の実験データを管理することのできる専用のLIMSを開発し、機能の検証を行っている。Windowsパソコンを用いて、筆者らが開発したLIMSを利用する際の操作画面の一例を図7に示す。ここでは、2Dゲル画像のスポットとゲル内消化用プレートのウェル番号、および質量分析用サンプルプレート上でのウェル番号をマッチングさせた状態でデータを管理できるように設計されている。LIMSについては現在開発中であり、詳細は完成後に報告したいと考えている。

7 まとめ

生命科学研究の中でも、とりわけ創薬研究の分野では開発競争が激しく、その中で確実に成果を上げて行くためには、情報の収集能力と管理、解析能力が必要である。今後さらにプロテオーム研究が盛んになり、大量の解析データが蓄積されるようになると、プロテオーム研究によって得られた情報を効率良く収集し管理すること

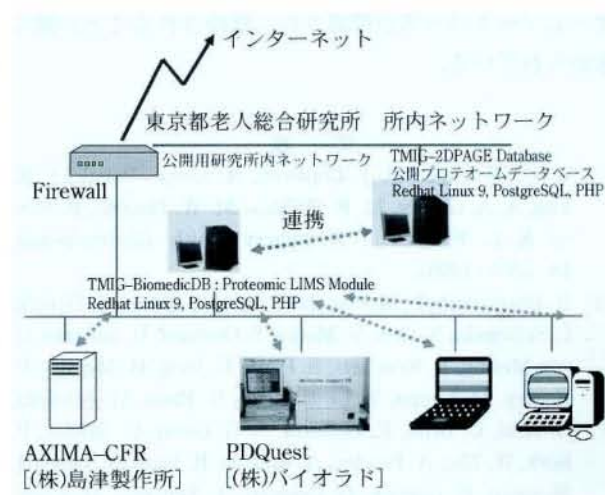


図6 TMIG-2DPAGE DatabaseとProteomics LIMS Module

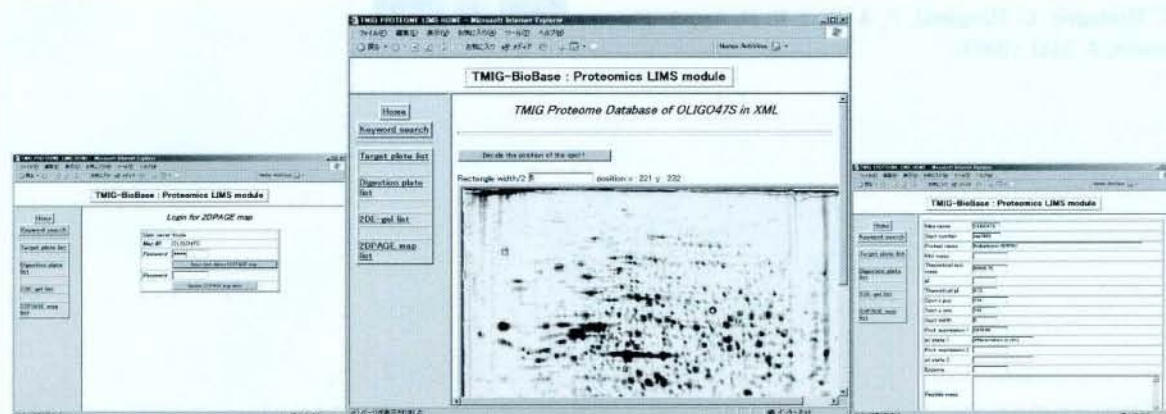


図7 Proteomics LIMS Moduleのインターフェース画面

のできるコンピュータシステムを整備することが、重要な課題の一つになってくる。

その一方で生命科学の基礎研究を推進するためには、当該分野の研究者間でできる限り多くの情報を交換し共有することも重要であり、そのためには、公的な機関が中心となってゲノムデータベースのようにパブリックドメインな公開プロテオームデータベースを立ち上げ、そこに個々の研究者が持っている独自の情報を登録できるようなシステムを構築する必要がある。またそれには、まずプロテオームデータを交換するためのXMLフォーマットの標準化が必要である。近年バイオインフォマティクスの重要性が叫ばれており、生命科学情報のデータベース化と公開システムの開発においても激しい開発競争が始まっている。今後さらにデータベースの内容を充実させ、安定的にサーバーを運用していくためには、その財源を確保する必要上、アクセスの有料化や一部の研究協力者に限定したアクセス制限など、様々な制約が発生することも予想されている。我が国における生命科学研究のレベルをアップし、国際競争力を高めていくためにも、国家的な経済支援の下で、公共性の高いプロテオームデータベースが構築され、維持されることが強く求められている。

文 献

- 1) V. C. Wasinger, S. J. Cordwell, A. Cerpa-Poljak, J. X. Yan, A. A. Gooley, M. R. Wilkins, M. W. Duncan, R. Harris, K. L. Williams, I. Humphery-Smith: *Electrophoresis*, **16**, 1090 (1995).
- 2) H. Hermjakob, L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. von Mering, B. Roechert, S. Poux, E. Jung, H. Mersch, P. Kersey, M. Lappe, Y. Li, R. Zeng, D. Rana, M. Nikolski, H. Husi, C. Brun, K. Shanker, S. G. Grant, C. Sander, P. Bork, W. Zhu, A. Pandey, A. Brazma, B. Jacq, M. Vidal, D. Sherman, P. Legrain, G. Cesareni, I. Xenarios, D. Eisenberg, B. Steipe, C. Hogue, R. Apweiler: *Nature Biotechnology*, **22**, 177 (2004).
- 3) C. Hoogland, J. C. Sanchez, D. Walther, V. Baujard, O. Baujard, L. Tonella, D. F. Hochstrasser, R. D. Appel: *Electrophoresis*, **20**, 3568 (1999).
- 4) K. Mostaguir, C. Hoogland, P. A. Binz, R. D. Appel: *Proteomics*, **3**, 1441 (2003).

- 5) T. Toda, K. Kaji, N. Kimura: *Electrophoresis*, **19**, 344 (1998).
- 6) 戸田年総: “最新電気泳動実験法”, 日本電気泳動学会編, p. 265 (1999), (医歯薬出版).
- 7) 戸田年総, 森澤 拓: “蛋白質核酸酵素増刊号バイオ高性能機器・新技術利用マニュアル”, 小原 収, 谷口寿章, 市川哲生, 猪飼 篤編集, p. 1886 (2004), (共立出版).
- 8) 戸田年総: 生化学, **76**, 1328 (2004).
- 9) C. F. Taylor, N. W. Paton, K. L. Garwood, P. D. Kirby, D. A. Stead, Z. Yin, E. W. Deutsch, L. Selway, J. Walker, I. Riba-Garcia, S. Mohammed, M. J. Deery, J. A. Howard, T. Dunkley, R. Aebersold, D. B. Kell, K. S. Lilley, P. Roepstorff, J. R. Yates, 3rd, A. Brass, A. J. Brown, P. Cash, S. J. Gaskell, S. J. Hubbard, S. G. Oliver: *Nature Biotechnology*, **21**, 247 (2003).
- 10) K. Garwood, T. McLaughlin, C. Garwood, S. Joens, N. Morrison, C. F. Taylor, K. Carroll, C. Evans, A. D. Whetton, S. Hart, D. Stead, Z. Yin, A. J. Brown, A. Hesketh, K. Chater, L. Hansson, M. Mewissen, P. Ghazal, J. Howard, K. S. Lilley, S. J. Gaskell, A. Brass, S. J. Hubbard, S. G. Oliver, N. W. Paton: *BMC Genomics*, **5**, 68 (2004).
- 11) Y. Yoshida, K. Miyazaki, J. Kamiie, M. Sato, S. Okuizumi, A. Kenmochi, K. Kamijo, T. Nabetani, A. Tsugita, B. Xu, Y. Zhang, E. Yaoita, T. Osawa, T. Yamamoto: *Proteomics*, **5**, 1083 (2005).
- 12) H. Morisawa, H. Hisatomi, M. Hirota, T. Toda: *J. Electrophoresis*, **49**, 35 (2005).



戸田年総 (Tosifusa TODA)

東京都老人総合研究所老化ゲノムバイオマーカー研究チーム (〒173-0015 東京都板橋区栄町35-2)。大阪大学大学院修了。理学博士。＜現在の研究テーマ＞老化と老年病に関わるタンパク質のプロテオーム研究。



森澤 拓 (Hiraku MORISAWA)

東京都老人総合研究所老化ゲノムバイオマーカー研究チーム (〒173-0015 東京都板橋区栄町35-2)。筑波大学大学院環境科学研究科修士課程修了。＜現在の研究テーマ＞プロテオーム研究のためのバイオインフォマティクス。