プロテオームデータベース

戸田年総・森澤 拓

Key Words　プロテオーム　データベース　2次元電気泳動　XML　クリッカブルマップ

●はじめに

従来の蛋白質研究では，大量の抽出液を出発材料とし，イオン交換やゲル濾過，ヒドロキシアパタイト吸着など，さまざまなカラムクロマトグラフィーをくり返して，SDS電気泳動で1本のバンドを呈するまで精製するという，大変手間のかかる操作が必要であった．そのため，情報の質は高いものの，費やされる時間と労力に比して得られる情報量は少なかった．これに対し，1995年に誕生したプロテオミクスは，微量の粗抽出液を直接2次元電気泳動にかけて，そこに存在する全蛋白質を数百～数千個のスポットに分離し，画像処理による比較解析を行なったのちに，高感度の質量分析計を用いて蛋白質を同定するという，ハイスループットな分析法であり，1つのサンプルから得られる情報の量が飛躍的に膨らんだ．その結果，山のような情報のなかから，本当に必要な情報を探しだすためのインフォマティクスが必要になった．また，プロテオミクスは組織的に行なわれることが多いので，研究者間で情報を共有するためのネットワークシステムも必要になってきた．これらの問題を解決する1つの方法は，プロテオーム研究によって得られた情報をデータベース化することである．

本稿では，そもそもプロテオームデータベースとはどのようなものか，データを共有し，いつでも必要な情報を取りだすことができるようにしておくためにはどのようなつくり方をするのがよいかといった基本的な事柄について述べるとともに，“研究施設内で利用するためのプロテオームデータベース”と“インターネット上で情報を公開するためのプロテオームデータベース”について，筆者らが現在行なっている試みを紹介する．

I. プロテオーム解析における分離・分析技術

1995年にWilkinsら[1~3]によって考案されたオリジナルのプロテオーム解析は，蛋白質を2次元電気泳動で分離し，画像解析を行なったのちに，個々の蛋白質スポットをゲル内で消化し，これを質量分析することによって同定するというものであった．この方法は，従来の蛋白質研究の概念を一新する画期的なものであったが，再現性の高い2次元電気泳動パターンを得るには，ある程度の熟練が必要であり，自動化がむずかしいことが普及の足かせとなっており，高性能液体クロマトグラフィー（HPLC）や2次元キャピラリー電気泳動の組合せによる方法[4~7]，蛋白質チップを用いる方法[8~10]など，さまざまな代替法も生みだされている．

よく，どの方法がベストであるか聞かれるが，それぞれに長所と欠点があり，研究の目的に最適な方法を選択するというのが正しい使い方である．ちなみに2次元電気泳動の最大の特長は，蛋白質を事前に消化せず，元の構造（1次構造）のまま，高分解能で分離できるところにある．細胞の分化やアポトーシスの過程において特定の蛋白質が翻訳後の修飾やプロセシングを受け，分子量や等電点が変化するような場合にも，別個のスポットとして分離されるので，修飾の度合いや修飾構造を定量的・定性的に分析することができる．また，泳動後に蛋白質をPVDF（polyvinylidene fluoride）膜に転写し免疫染色を行なったり，リン酸化や糖鎖構造に特異的な染色[11,12]を実施することによって，構造特異的な絞込みプロテオーム解析（targeted proteomics）[13,14]を行なえる

Tosifusa Toda, Hiraku Morisawa, 東京都老人総合研究所プロテオーム共同研究グループ　E-mail : ttoda@tmig.or.jp, morisawa@tmig.or.jp　http://proteome.tmig.or.jp/2D/　http://proteome.tmig.or.jp/TMIG-PCC/Proteome Database

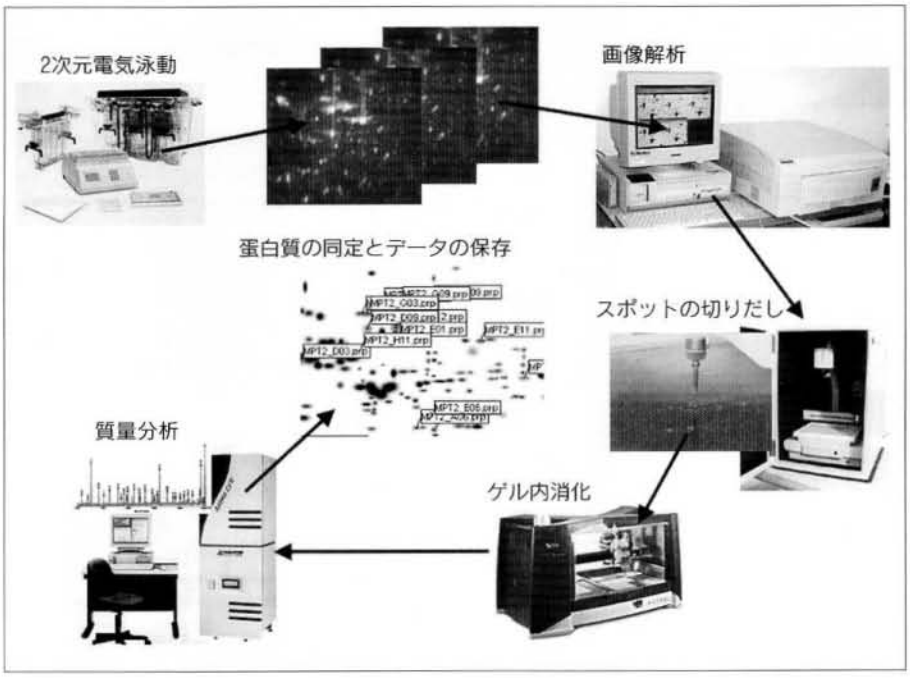

図1　2次元電気泳動に基づいたプロテオーム解析の流れ
比較する数種類のサンプルを2次元電気泳動で分離し画像解析後，スポットを切りだし，ゲル内で消化，質量分析で同定する．

ことも大きなメリットである．筆者らが東京都老人総合研究所で行なっている老化のプロテオーム研究[15~17]では，とくに翻訳後修飾の解析が重要であり，2次元電気泳動に基づいたプロテオーム解析が最も適している．

II. プロテオーム解析によって得られる情報

2次元電気泳動に基づいたプロテオーム解析は，おもに細胞の分化や老化，不死化，各種疾患病態における蛋白質の発現プロファイルの変化を解析し，責任蛋白質を同定することを目的とし，通常**図1**に示す流れにそって行なわれる．比較研究のために調製された1セットの抽出蛋白質を，2次元電気泳動で分離し，CBB（クマシーブリリアンドブルー）などの可視色素やSYPRO Rubyなどの蛍光色素によって染色，スキャナーやCCDカメラなどでパターンを読み取り，画像解析ソフトを用いてスポットのマッチングと定量的ディファレンシャル解析を行なう．現在筆者らの研究所では，Bio-Rad社製の2次元電気泳動画像解析ソフト（PDQuestバージョン7.2）を用いているが，このほかにもImageMaster（Amersham Biosciences）やProFINDER 2D（PerkinElmer），Melanie 3（GeneBio）など，さまざまな解析ソフトが市販されている．

PDQuestでは，最初にバックグラウンドレベルの補正やノイズ，ストリーク（テーリング）の除去などの前処理が行なわれたのちに，蛋白質スポットが検出され，スポットの位置（蛋白質の等電点と分子量に対応する情報）および濃淡の定量値（蛋白質の発現レベルに対応する情報）の数値化が行なわれる．こののち，比較研究のための画像集団（マッチセット）のなかから，標準となるマスターゲルを指定することにより，そのゲル上のスポットの位置に他のゲル上の対応するスポットが自動的にマッチングされる．2次元電気泳動の再現性が高い場合，バージョン7以降のPDQuestでは，ランドマーク（基準点）となるスポットをとくに指定しなくても，大半のスポットは自動的にマッチングされる．位置が多少ずれているためにオートマッチができなかったスポットについては，その近傍で確実に対応しているスポットをランドマークとすることで，正確なマッチングが行なわれる．対応するスポットが見つからないスポットは，アンマッチスポットとして処理されるが，このようなスポットはそのサンプルに特異的に発現する蛋白質である可能性が高く，責任蛋白質を特定するための重要な情報となる．また老化研究や疾患研究では，さまざまな蛋白質において加齢や病態に伴う発現レベルの変動がみられることがある（**図2**）．このような蛋白質の発現レベルの変動が，加齢や病態における細胞機能の変化に関係しているものと考えられている．

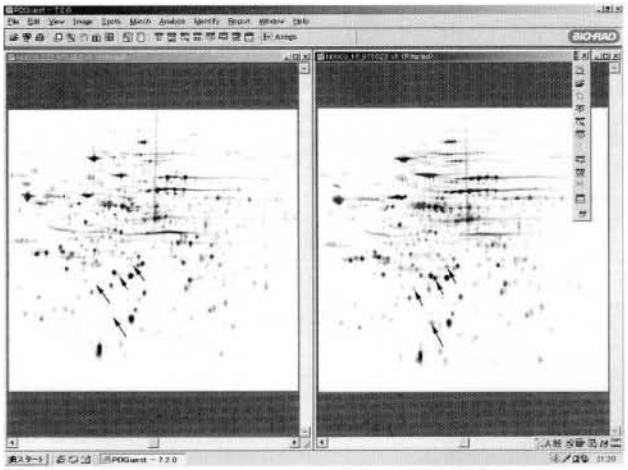

図2　細胞の不死化に伴う蛋白質発現プロファイルの変化
左側が不死化後，右側が不死化前の発現蛋白質の2次元電気泳動パターン．矢印で示した蛋白質スポットが不死化に伴って減少していることがわかる．

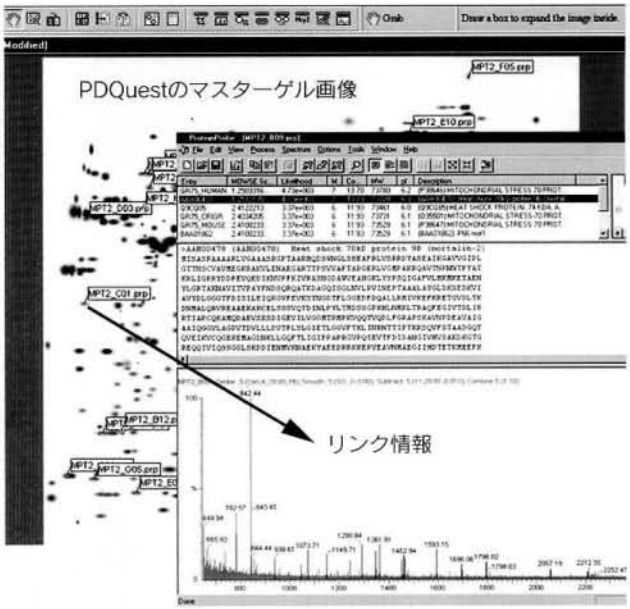

図3　PDQuestのアノテーション機能を利用してマスターゲル画像にリンクされた質量分析による蛋白質同定結果のファイル

次に，マニュアル操作やロボットアームの操作によって，これらのスポットを切りだし，ゲル内でトリプシン消化を行なったのちに，MALDI-TOF型や，MALDI-QIT-TOF型，ESI-Q-TOF型などの質量分析計を用いてペプチド断片の質量を分析する．ここで得られたモノアイソトピックピーク（質量分析によって得られる1価のイオンの質量スペクトルは，^{13}Cの存在比によって1 Daずつずれた一連のピークとなって現われる．このうち一番左のピークはすべての炭素が^{12}Cで構成された成分のピークであり，これをモノアイソトピックピークとよぶ）のリストを手掛かりに，SWISS-PROTなどのデータベースを検索して蛋白質を同定する．データベース検索エンジンとしてよく用いられるMascotは，マトリックスサイエンス社のホームページ（http://www.matrixscience.com/search_form_select.html）にアクセスすると無料で利用できるが，ライセンス料を払って手元のコンピュータにサーバーソフトをインストールし，施設内で検索を行なうこともできる．このほか，インターネット上で利用できる無料の検索エンジンとしては，UCSF Mass Spectrometry Facilityが提供しているProteinProspector（http://prospector.ucsf.edu/）などもある．

質量分析計から出力されるマススペクトルデータや，モノアイソトピックピークリスト，データベース検索の結果なども，PDQuestのマスターゲル画像上のスポットにファイルやURLとしてリンクされる（**図3**）．またリン酸化や酸化，糖化，糖鎖結合などの翻訳後修飾に関する情報も，同様にスポットのアノテーションとして保存される．

PDQuestによって得られた画像解析データと，それにリンクされた質量分析結果などの情報は，基本的にはマッチセット（画像データ群）ごとに独立しているが，PDQuestにはさらに“ハイレベルマッチセット”という機能が備わっており，これを利用してマッチセット間のスポットの位置情報を相互に関連づけ，リンク情報やアノテーション情報を1つにまとめることができる．しかし残念ながら，共同研究やプロジェクト研究のメンバーがネットワークを介してプロテオーム情報を共有するようなことはできない．また，サーバコンピュータに集積された大量のプロテオームデータのなかから，研究者のニーズに合致した情報を選択的に取りだし，創薬の標的探索などに活用できるようなデータベースとしての機能ももっていない．

そこで筆者らは，PDQuestのハイレベルマッチセット機能と，アノテーション機能によってPDQuestに保存された内部情報，および質量分析以後の解析で得られたリンク情報を，研究所内のサーバコンピュータにすべて集め，所内の研究者間で情報を共有したり，データをマイニングしたりする際に利用できる“サーバクライア

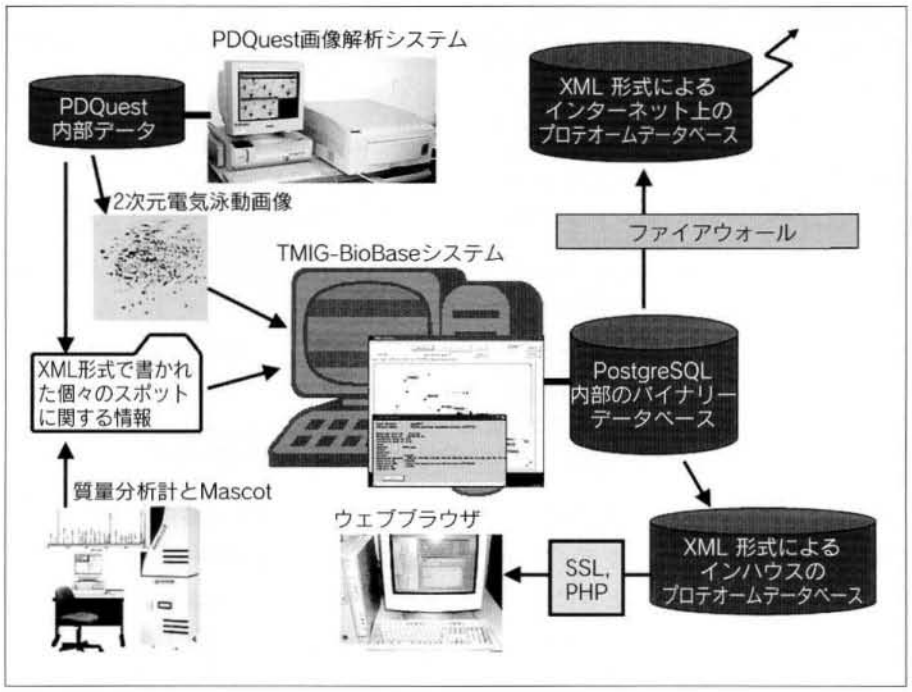

図4
PDQuestで得られたデータや質量分析で得られたデータはTMIG-BioBaseに集められ，所内の研究者間で情報が共有される．一部の情報はインターネットを介して外部に発信される．SSL：secure sockets layer，PHP：正式名称はhypertext preprocessor.

ント型"のプロテオームデータベースの開発を行なっている．またこれと並行し，ファイアウォールの外側におかれた［正確にはDMZ（demilitarized zone）層におかれた］ウェブサーバ上に，情報公開用のプロテオームデータベースを設置し，所外の研究者に向けた情報発信を行なっている（図4）．

III. 研究施設内で利用するためのプロテオームデータベース

データベースを構築する際に最初に考えなければならないことは，データの構造（データモデル）をどのようにするかということであるが，最近はデータ間に相互の関係をもたせられるリレーショナルデータモデルとよばれるものが主流になりつつある．リレーショナルモデルに基づくデータベースをリレーショナルデータベース，リレーショナルデータベースを制御する仕組みをリレーショナルデータベースマネージメントシステム（RDBMS）とよんでおり，現在では，多くのRDBMSがStructured Query Language（SQL）によってつくられている．SQLによるRDBMSでは，データベース検索を行なったり，リレーショナル分析を行なうときには，内部データとしてバイナリーファイルがつくられ，これが利用されるが，周辺の機器などから送られてくるXML形式のデータを取り込んだり，逆に内部データをXML形式に変換してファイル出力することもできる．

筆者らは，研究施設内で利用するためのプロテオームデータベースはSQLでつくるのがベストであると考えており，プロテオーム研究に特化した東京都老人総合研究所独自のRDBMS（TMIG-BioBase System）の開発を始めている．現行バージョンのPDQuestにはXML形式でファイル出力をする機能がないため，現時点ではXMLのスキーマ（本来Schemaとは，XML文書の取りうる構造を記述したものをいう．またスキーマを記述するための言語を，スキーマ言語という．スキーマ言語を用いることによって，XML文書の正しさをある程度まで自動的にチェックしながらXML文書を書き出すことが可能になる）を用いてデータを再入力しなければならないが，PSQuestから直接XML形式で出力できるようになれば，TMIG-BioBase Systemへのデータ転送はさらに容易になる．

IV. インターネット上で情報を公開するためのプロテオームデータベース

インターネット上では，すでに数多くのプロテオームデータベースが公開されており，インターネットイクスプローラーなどのブラウザを用いて，自由にアクセスで

表1　インターネット上で公開されている既存のプロテオームデータベース

TMIG-2DPAGE Database	http://proteome.tmig.or.jp/2D/
SWISS-2DPAGE Database	http://au.expasy.org/ch2d/
SIENA-2DPAGE Database	http://www.bio-mol.unisi.it/2d/2d.html
PMMA-2DPAGE Database	http://www.pmma.pmfhk.cz/
DCHGR 2-D PAGE Database	http://proteomics.cancer.dk/
Proteome 2D-PAGE Database	http://www.mpiib-berlin.mpg.de/2D-PAGE/index-2DPAGE.html
RAT HEART-2DPAGE Database	http://www.mpiib-berlin.mpg.de/2D-PAGE/RAT-HEART/2d/
HEART-2DPAGE Database	http://userpage.chemie.fu-berlin.de/~pleiss/dhzb.html
HSC-2DPAGE Database	http://www.harefield.nthames.nhs.uk/nhli/protein/
Heart High-Performance 2-DE Database	http://www.mdc-berlin.de/~emu/heart/
OGP-WWW 2DPAGE Database	http://proteomewww.glycob.ox.ac.uk/2d/2d.html
Phosphoprotein Database (PPDB)	http://www-lecb.ncifcrf.gov/phosphoDB/
2-DE Map of Cerebrospinal Fluid Proteome	http://www.leelab.org/ASMSCSF/map.htm
Proteome Database of Human T helper Cells	http://www3.btk.utu.fi:8080/Genomics/Proteomics/Database/GelImages/Gels/Geeli1

きるようになっている[18,19]（**表1**）．Proteome Databaseというよび方のほか，2-D PAGE Databaseという名称で公開されているものも，内容的にはプロテオームデータベースと同義である．ここでは筆者らが試験的に公開を行なっているプロテオームデータベースTMIG-2DPAGE（http://www.proteome.jp/2D/）[20]を例にとり，インターネット上で情報を公開するためのプロテオームデータベースのつくり方を簡単に解説する．詳細は，筆者の前著[21]などを参照されたい．

インターネット上でプロテオームデータベースを公開するためには，Apacheなどのhttpd（hyper text transfer protocol daemon）が走っているウェブサーバが必要である．民間プロバイダーのレンタルサーバを利用することもできるが，独自でサーバを立ち上げる場合には，UNIXマシンもしくはLinaxマシンに，Apacheのホームページ（http://httpd.apache.org/）からソースプログラムをダウンロードし，セットアップすることになる．Apacheの立ち上げ方については，文献22などの解説書を参照されたい．

1. クリッカブルマップ用の2次元電気泳動パターン

多くのプロテオームデータベースは，2次元電気泳動パターン上でスポットをクリックすると，そのスポットに関する詳細な情報が表示される仕組み（クリッカブルマップ）になっている．このようなプロテオームデータベースを作成するときに問題となるのが，生画像を用いるか，画像解析ソフトウェアでバックグラウンドやストリークなどを取り除いたマスターゲル画像を用いるかということである．筆者らは，個々のゲルごとに解析結果を記録する際には，生画像を用い，複数のゲルを画像解析でマッチングし，それぞれのゲルで得られた同定結果やディファレンシャル分析の結果をデータベース化するときには，マスターゲル画像を用いるようにしている．

PDQuestからマスターゲル画像をTIFF形式などで取りだし，これをPhotoshopなどのペイント系画像処理ソフトで開いて，ゲルの外周に等電点や分子量などのスケールと数値を書き加え，同定結果などのリンク情報をもたせるスポットには十字やSWISS-PROTのアクセッション番号，スポット番号などの標識をつけ，適当な画像サイズと解像度（通常筆者らは，幅と高さはおおよそ25 cm程度，画像解像度は72程度）にしてGIF形式で保存する．

2. スポットをクリックしたときのジャンプ先を指定するマップファイル

クリッカブルマップを機能させるためには，画像のどの部分をクリックしたときにどのファイルにジャンプするかを指定するための，マップファイル（ファイル名の末尾が.mapとなったテキストファイル）が必要である．

たとえば2次元電気泳動画像上の$(x, y) = (100, 300)$の座標点を中心とする半径5ピクセルの円の内部（スポットの範囲）をクリックしたときに表示されるファイルのURLがhttp://proteome.tmig.or.jp/2D/mousebrain/ssp0023.htmlであった場合，マップファイルには「circle http://proteome.tmig.or.jp/2D/mousebrain/ssp0023.html 100, 300 95, 295」という1行を記入する．複数のジャンプ先を指定するには，このような記入を改行ごとに行なう．最後に拡張子を.mapとし，テキスト形式でファイルを保存する．2次元電気泳動画像上の(x, y)座標を知る最も簡単な方法は，クリッカブルマップ用の2

次元電気泳動画像を実際にブラウザ上に表示をして，その上でマウスを操作するとよい．それにはまずApacheサーバのルートディレクトリ（通常は，/opt/WWW/httpd/htdocs/）の下に，「2D/XML/test」などの新しいディレクトリをつくり，そこにGIF形式のクリッカブルマップ画像と，下記のような内容のテキストファイル（ファイルの拡張子は.htmlとする）をFTP転送する（ここでは例としてファイル名をすべて“test”にしているが，ファイル名は自由）．これにブラウザでアクセスするとクリッカブルマップが表示され，マウスを操作すると（x, y）座標が欄外に表示される．

```
<HTML>
 <HEAD>
  <TITLE>Clickable Image Map for Test</TITLE>
 </HEAD>
 <BODY>
  <A HREF="test.map">
   <IMG ISMAP SRC="test.gif">
  </A>
 </BODY>
</HTML>
```

3. 同定結果などを記述した蛋白質スポットの情報ファイル

スポットをクリックしたときに表示されるデータのファイルで，これがいわばプロテオーム情報の本体である．現在，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の多くは，フラットなテキスト形式か，HTML（Hyper Text Markup Language）形式で情報が記述されているが，これらの形式は，データマイニングソフトを用いて解析を行なえるような論理的な記述構造にはなっていない．これに対し，データベースの記述に適しているといわれるXML（eXtensible Markup Language）形式では，たとえば

```
<proteome_database>
 <spot_protein>
  <spot_id>ssp001</spot_id>
    <isoelectric_point>5.6</isoelectric_point>
    <molecular_mass>13500</molecular_mass>
    <protein_name>stathmin</protein_name>
 </spot_protein>
</ proteome_database>
```

のように，データの意味を示すタグを使ってデータを挟むので，保存された情報の内容が判別しやすく，データマイニングソフトを用いて必要な情報を簡単に取り出すことができる．ただし，XML形式で書かれたデータを直接ブラウザで閲覧すると，タグがついたままの形で表示されてしまうので，通常はスタイルシートとよばれる“表示スタイルを指定するファイル（拡張子が.xslのファイル）”を別に用意し，これを介してデータを取りだす必要がある（**図5**）．誌面の関係でそれぞれのファイル

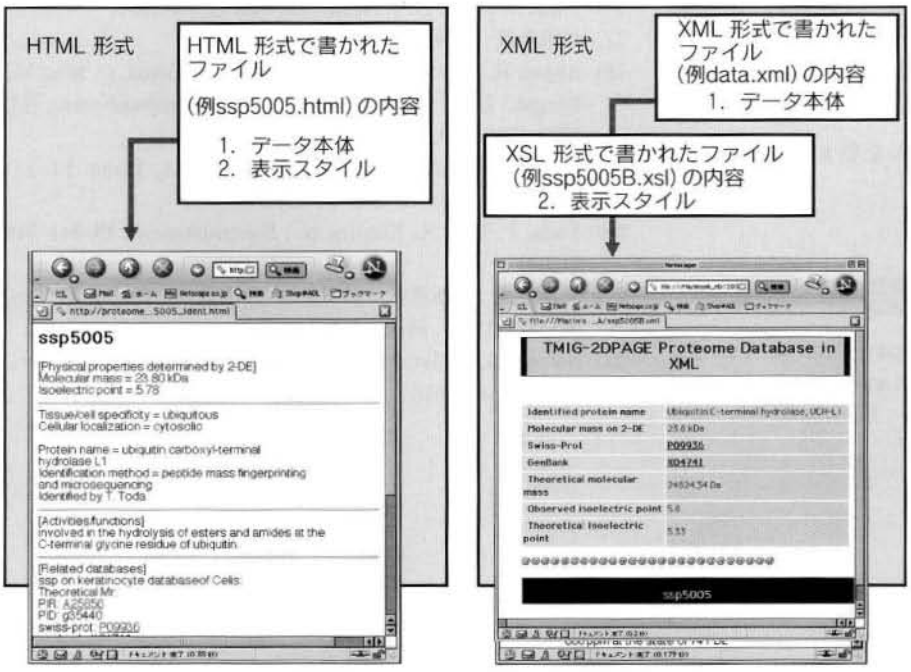

図5　HTML形式とXML形式における表示方法の違い

の具体的な書き方をここで紹介することはできないが，http://www.proteome.jp/2D/XML/test/で，筆者らが作成しているプロテオームデータベースのファイルの一部を公開しているので，興味がある方は，そちらをご覧いただきたい．

●まとめ

プロテオームデータベースは，質・量ともに年々充実してきており，今後さらに重要性が増してくるものと思われる．現在の多くのプロテオームデータベースは単純なテキスト形式か，HTML形式で記述されているが，XML形式のほうが多くの点で有利である．しかし，これにはまだいくつかの問題も残されている．XMLでは自由にタグをつくれる反面，データベースごとに勝手なタグが使われると混乱を生じるので，国際的に統一した基準をつくる必要がある．XML形式によるプロテオームデータベースの標準化については，現在HUPO（Human Proteome Organization；http://www.hupo.org/）のなかで議論が始まっており，近いうちに指針が示されるものと思われる．また，パブリックドメインのプロテオームデータベースからダウンロードしたデータを利用して，創薬研究などの成果や利益が生じた場合に，権利関係をどのように扱うかといったことについても一定のコンセンサスが必要である．いずれにせよ，研究者が個々にウェブサーバを立ち上げてデータベースを構築することは非効率であり，将来的には動物種や組織・細胞種ごとにコアとなる施設がパブリックドメインのプロテオームデータベースを立ち上げ，個々の研究者から投稿されたプロテオーム情報をそれらのコアサーバに登録して，維持管理できるようなシステムを整えることが必要であろう．

文　献

1) Wasinger, V. C., Cordwell, S. J., Cerpa-Poljak, A., Yan, J. X., Gooley, A. A., Wilkins, M. R., Duncan, M. W., Harris, R., Williams, K. L., Humphery-Smith, I.: *Electrophoresis*, **16**, 1090-1094 (1995)
2) Wilkins, M. R., Sanchez, J. C., Gooley, A. A., Appel, R. D., Humphery-Smith, I., Hochstrasser, D. F., Williams, K. L.: *Biotechnol. Genet. Eng. Rev.*, **13**, 19-50 (1996)
3) Wilkins, M. R., Sanchez, J. C., Williams, K. L., Hochstrasser, D. F.: *Electrophoresis*, **17**, 830-838 (1996)
4) Figeys, D., Gygi, S. P., Zhang, Y., Watts, J., Gu, M., Aebersold, R.: *Electrophoresis*, **19**, 1811-1818 (1998)
5) Spahr, C. S., Davis, M. T., McGinley, M. D., Robinson, J. H., Bures, E. J., Beierle, J., Mort, J., Courchesne, P. L., Chen, K., Wahl, R. C., Yu, W., Luethy, R., Patterson, S. D.: *Proteomics*, **1**, 93-107 (2001)
6) Wu, C. C., MacCoss, M. J.: *Curr. Opin. Mol. Ther.*, **4**, 242-250 (2002)；*Electrophoresis*, **17**, 830 (1996)
7) Manabe, T.: *Electrophoresis*, **20**, 3116-3121 (1999)
8) Talapatra, A., Rouse, R., Hardiman, G.: *Pharmacogenomics*, **3**, 527-536 (2002)
9) Lee, Y. S., Mrksich, M.: *Trends Biotechnol.*, **20** (12 Suppl.), S14-18 (2002)
10) von Eggeling, F., Junker, K., Fiedle, W., Wollscheid, V., Durst, M., Claussen, U., Ernst, G.: *Electrophoresis*, **22**, 2898-2902 (2001)
11) Steinberg, T. H., Agnew, B. J., Gee, K. R., Leung, W. Y., Goodman, T., Schulenberg, B., Hendrickson, J., Beechem, J. M., Haugland, R. P., Patton, W. F.: *Proteomics*, **3**, 1128-1144 (2003)
12) Steinberg, T. H., Pretty, On Top K., Berggren. K.N., Kemper, C., Jones, L., Diwu, Z., Haugland, R. P., Patton, W. F.: *Proteomics*, **1**, 841-855 (2001)
13) Higai, K., Shibukawa, K., Muto, S., Matsumoto, K.: *Anal. Sci.*, **19**, 85-92 (2003)
14) Brooks, H. L., Allred, A. J., Beutler, K. T., Coffman, T. M., Knepper, M. A.: *Hypertension*, **39** (2 Pt 2), 470-473 (2002)
15) Toda, T.: *Exp. Gerontol.*, **35**, 803-810 (2000)
16) Toda, T.: *Ann. N. Y. Acad. Sci.*, **928**, 71-78 (2001)
17) 戸田年総：実験医学，**18**，2490-2496 (2000)
18) Appel, R. D., Sanchez, J. C., Bairoch, A., Golaz, O, Miu, M., Vargas, J. R., Hochstrasser, D. F.: *Electrophoresis*, **14**, 1232-1238 (1993)
19) Evans, G., Wheeler, C. H., Corbett, J. M., Dunn, M. J.: *Electrophoresis*, **18**, 471-479 (1997)
20) Toda, T., Kaji, K., Kimura, N.: *Electrophoresis*, **19**, 344-348 (1998)
21) 戸田年総：最新電気泳動実験法（日本電気泳動学会 編），pp. 265-274 (1999)，医歯薬出版 (1999)
22) Helden, G., Keller, M. (IDEC. C訳)：Apache Webサーバー，インプレス (1999)