プロテオームインフォマティクス

——プロテオームデータベースの構築とその利用——

戸田年総

◎そもそもプロテオームとは，特定の生物個体が，特定の時期に，特定の組織細胞で発現する全タンパク質の機能的な集団である．したがって，プロテオームを解析するということは，その組織や細胞の生理的な活動状態や病理的な偏倚状態をタンパク質レベルでとらえようとすることにほかならない．しかし，プロテオーム解析で得られる情報は膨大であり，ゲノム科学や細胞生理学・臨床医学などによって得られた情報と結びつけて，診断や医療に役立てるためには，きちんとした形でデータベース化しておくことが不可欠である．これに必要なコンピュータシステムやソフトウェア，情報処理技術などを開発し応用することが，プロテオームインフォマティクス(プロテオーム情報科学)の本質である．プロテオームインフォマティクスは，さらに高いレベルのバイオインフォマティクスのフロントエンドとして利用され，新薬の開発や予防医学の発展に寄与するものと考えられる．

Key word：プロテオーム，データベース，インフォマティクス，タンパク質

プロテオーム研究が実質的にスタートしたのは1995年であるが，その源流は1975年に発表されたO'Farrellの二次元電気泳動[1]に遡る．その後，コンピュータによる画像解析技術[2-4]が加わり，固定化pH勾配等電点電気泳動[5]が生まれ，さらにインゲル消化や質量分析，ペプチドマスフィンガープリント法による新しいタンパク質同定法[6]などが開発されて，現在のプロテオーム研究に至っている．そして最後に，プロテオーム研究で得られた情報をデータベース化し，管理・活用することを目的として誕生したのがプロテオームインフォマティクス[7,8]である．しかし，プロテオームインフォマティクスは産声を上げたばかりであり，関連するゲノムインフォマティクスやトランスクリプトームインフォマティクスとの接続や，上位のバイオインフォマティクス[8]との連携など，課題も多く残されている．

Proteome informatics
Tosifusa TODA：
東京都老人総合研究所プロテオーム共同研究グループ

プロテオーム研究において発生する情報

プロテオーム研究の各ステップで，図1のような情報が発生する．従来，サンプルや方法に関することは研究者が自己のノートに記録し，電気泳動のパターンであれば写真に撮るか，あるいは画像解析用のコンピュータにそのまま残しておくことが多かった．また，質量分析の結果については質量分析計を操作するコンピュータ内にそのままセーブしておくことが多かったが，一般に分析用のコンピュータはメモリー容量が少ないので，定期的に古いデータを削除せざるをえない．しかもこのように非体系的な情報管理のもとでは情報が散逸し，後日あらためて特定のサンプルに関する情報を取り出し，確認をする必要が生じたときに大変苦労をする．さらに，担当者が職場を異動した場合などには元の情報にたどれなくなる事態さえ招きかねない．これではせっかく貴重な病態サンプルを用いて実施されたプロテオーム解析の結果がむだになってしまう．

このような事態を避けるために，それぞれの段階の情報を簡単な操作ですべて電子化し，万一のトラブルに備えて定期的にメモリーバックアップ

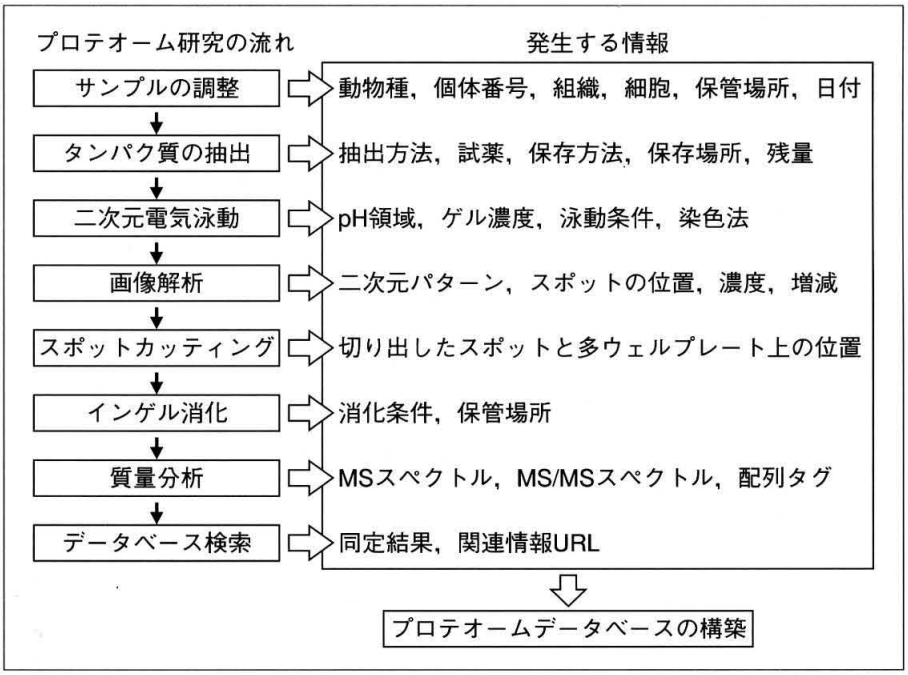

図 1　プロテオーム研究の各ステップで発生する情報
プロテオームインフォマティクスの核となるプロテオームデータベースは，このような情報を集積し管理活用する目的でつくられる．

を行うことのできるコンピュータシステムをつくり上げておくことが望まれている．このような情報の管理は，大学などの独立した研究者が行う研究の場合には個人の責任で行われるが，医療センターや製薬企業などにおいて組織的に行われている検査や研究の場合，情報を組織内で共有する必要がある．これを実現する方法のひとつは LAN の活用である．プロテオーム情報を管理するサーバーシステムを中央におき，検査や研究業務を行うスタッフは手元のパソコンからリモートアクセスし，情報の書き込みや修正，閲覧，ダウンロードなどを随時行えるようにしておくことである．ただしこの場合，個人情報や未発表のデータを含むので，セキュリティーを確保することが重要な課題となる．

データの形式と使用するソフトウェアの問題

データベースの基本は単位となる情報を一定の書式で記述し，ファイルとして蓄えておくことである．二次元電気泳動に基づくプロテオーム解析の場合，電気泳動パターンそのものは画像データであり，文書ファイル化は困難であるが，個々のタンパク質スポットに関する情報(等電点や分子量，相対的発現レベル，ペプチドマスフィンガープリント情報，シークエンスタグ情報など)は，文書ファイルとして保存することが可能である．プロテオミクスにおいてもっとも重要な“電気泳動パターン上のスポットと文書ファイルをリンクする”方法論は後で述べることにして，まず個々のタンパク質スポットに関する情報の電子化の話をしたい．

もっとも単純な文書ファイルの形式は“テキスト形式”(DOS/V の場合，拡張子が.TXT)である．これなら特別なソフトウェアは必要ない．個人で情報管理をするのであれば，好きな項目を設定し，使い勝手のよい書式を決めて自分のパソコンに保存するだけで十分である．グループで情報を共有するためには，すべてのメンバーが満足できるデータ項目と書式を話し合いによって設定する必要がある．情報の共有は LAN 上でたがいに相手のパソコン内のファイルを読めるように設定することによっても実現できるが，セキュリティを考えると，中央に共用のファイルサーバーをおくべきである．

ここでさらに問題となるのが，蓄積されたファイルのなかから必要な情報を選択的に取り出すための仕組みをどうやって実現するかということで

ある．ひとつの方法は，ファイル名やファイルの内容の文字列を検索し，一致するものを表示する簡単な検索ソフトを利用することである．これでもある程度のことはできるが，検索の速度や効率が低く，情報量が多くなるにつれて実用的ではなくなる．あらかじめ“キーワード”を設定しておくことによって検索効率を上げることができるが，これをファイル形式化したものが，最近話題となっているXML[9]である(「サイドメモ」参照)．XMLはもともとインターネット上で情報を共有し交換するために開発されたファイル形式で，キーワードを機能化した“タグ”とよばれる文字列(半角のカギカッコで挟まれた文字列)を情報の先頭におくことで，それに続くものがどのような情報であるかわかるようにしたものである．情報を取り出したり表示したりするときには，この“タグ”をみて判断する．

XMLでは自由にタグを定義できる反面，素人に使いづらいという理由で，基本となるタグをあらかじめ定義したものが，現在ホームページ記述ファイルとして広く使われているHTMLである．文字列や単純な画像を管理するだけならHTMLで十分であるが，プロテオーム情報のように異なる意味をもつさまざまな数値情報を多く含むデータファイルの場合，HTMLの既存のタグだけでは不十分である．プロテオームデータベースの記述言語としてXMLが注目されているのはこのためである．

ただし，XMLにも問題がないわけではない．そもそもXMLでは製作者が勝手にタグを定義できるので，Internet ExplorerやNetscape Navigatorなどの一般のブラウザでファイルを開くと，タグつきの文書がそのまま画面に表示されるだけで，タグを付けた意味がない．XMLのタグを利用して検索を行い見やすい形に並べ換えて表示するには，“スタイルシート”という“タグの利用法を記したファイル”が別途必要となる．スタイルシートは本来製作者自身が用意するものであるが，情報の受取り手が変更を加え，好きな形でデータを取り出すこともできる．まだXML形式のデータベースはあまり普及はしていないが，ゲノムデータベースの一部や，タンパク質一次構造データベースの一部ではXML化の動きがはじまっており，今後プロテオミクスにおいても，JHUPO(Japanese Human Proteome Organization)が中心になってXML化を進める動きがある．

高いレベルのセキュリティ管理が要求される医療センターや公立の研究機関，企業の研究所などがLAN上にデータベースを構築し，プロテオーム研究情報を共有する際にはそれなりのしっかりしたソフトウェアが必要であるが，そのためのシステムがBio-RadやProteome systemsなどから続々と市販されるようになっている．たとえば，Bio-RadのWorksBaseは，UNIXやWindows 2000上で動くオラクルという汎用のデータベースソフトウェアをベースに開発されたものである．オラクルは非常によくできたデータベース管理システムであるが，素人には使いづらいという難点があった．また，実際にファイルを管理する場合には情報の項目や形式など多くの部分を利用者が独自に設定する必要があった．これを，プロテオームを中心とするバイオインフォマティクスに最適化し，必要な設定をすべてすませたものがWorksBaseである．図1に示すように，プロテオーム解析ではサンプルの調製から質量分析，データベース検索までの段階においてさまざまな情報が生ま

サイドメモ

XML

XML(eXtensible Markup Language)は，インターネットに関するさまざまな規格を決定する非営利団体W3Cが制定した“情報を記述するためのメタ言語”のひとつであり，同じようなものにはXHTMLやHTMLがある．HTMLがどちらかというと“情報の表現スタイル”を重視した言語であるのに対し，XMLは“情報の中身”を重視した言語である．XHTMLは両方の特徴を合わせもっている．Internet ExplorerやNetscape Navigatorなどのブラウザを用いると，HTMLはそれ自身のファイルのなかに記載された“表現スタイル”に基づいて情報が表示されるのに対し，XMLでは情報の中身とその階層構造が記載されているだけなので，表現スタイルを記した“スタイルシート”が別途必要となる．一見不便なようであるが，データベースの記述にはXMLのほうが適している．

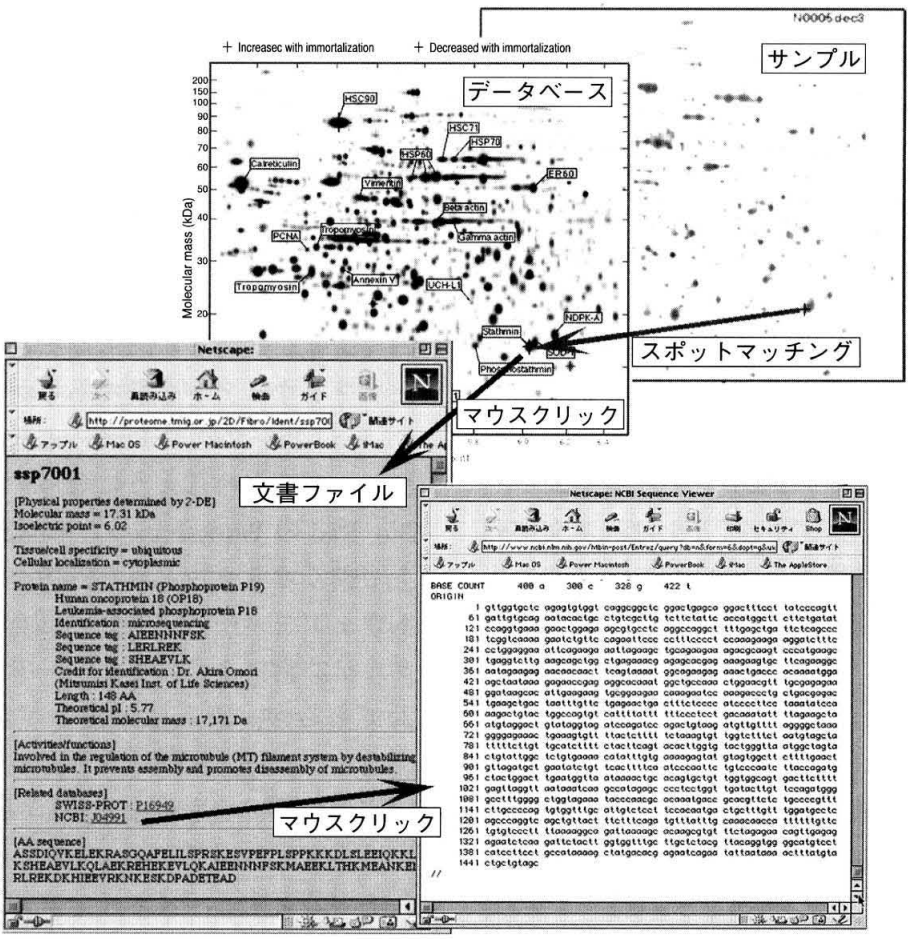

図 2　バイオインフォマティクスにおけるプロテオームインフォマティクスを起点とした情報検索の流れ

れるが, WorksBase ではこれらの情報に加え, 分析機器をコントロールするパラメータや, 機器から吐きだされる生データもそのまま蓄えておけるので, 後日あらためて解析をしなおす必要が生じた場合にも対応ができる.

二次元電気泳動パターンのマッチングによる情報検索

プロテオミクスにおけるもっとも重要な情報管理のひとつは, 電気泳動パターン上のスポットと, それに対応する文書ファイルをリンクさせておくことである, すでに多くのウェブサイトでは二次元電気泳動パターン上のスポットをクリックすることによって情報が取りだせる方式(クリッカブルマップ方式)のプロテオームデータベース[10-12]が製作され公開されているが, 今後さらにこのようなサイトが増えるものと予想される. また, インハウスのプロテオームデータベースを構築する際にも二次元電気泳動パターンは PDQuest などの画像解析ソフトで処理され, 分析結果はアノーテーションの形でスポットにリンクされる.

したがって, データベース構築後に実施されたプロテオーム解析において, 二次元電気泳動パターン上のスポットに関する情報がスポットをマッチングさせるだけで簡単に取り出せたらたいへん便利である(図 2). これを実現するために今後, 市販の画像解析ソフトには, インターネット上で公開されているデータベースの二次元電気泳動パターンとスキャナで読み取った自己の電気泳動パターン上のスポットをマッチさせる機能が組み込まれることをぜひとも期待したい.

バイオインフォマティクスにおけるプロテオームインフォマティクスの位置づけ

現在, 多くの二次元電気泳動データベース(プロテオームデータベース)ではタンパク質情報(プロテオーム情報)側からゲノムデータベースや, 立体構造データベース, 文献データベースなどに向

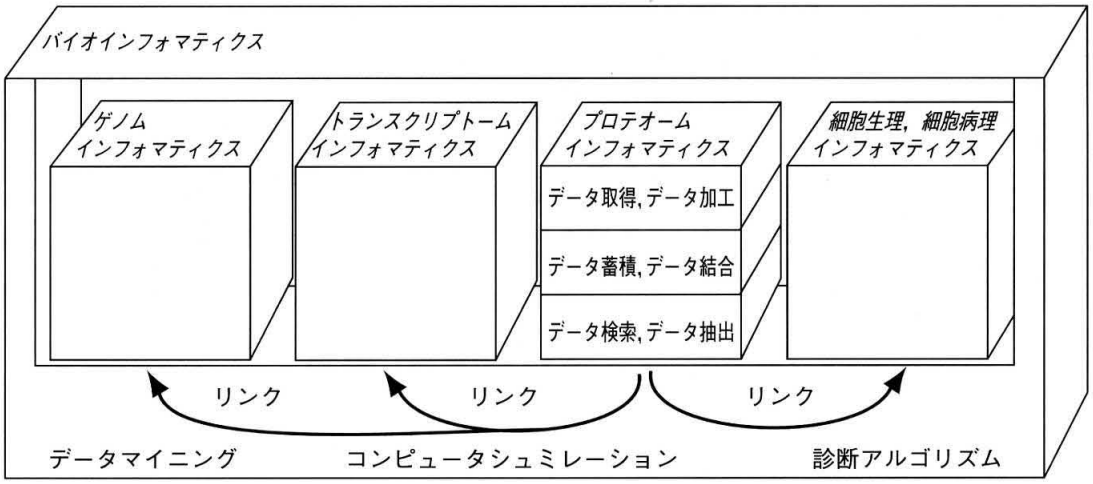

図 3　画像データを起点とするプロテオームインフォマティクス
二次元電気泳動のスポットをデータベースに登録されたスポットにマッチさせることによって，情報が取り出せるしくみ．

かったリンクは張られているが，その逆はほとんどみられない．今後もこのような状況が続くなかで，すべての情報がバイオインフォマティクスで統合された場合，情報検索の方向としてプロテオームデータベースを起点とすることのメリットがますます大きくなる(図 3)．これはプロテオームインフォマティクスにおける検索機能の充実が，バイオインフォマティクス全体の利用効率を左右することを意味する．すなわち，プロテオームインフォマティクスがバイオインフォマティクスのキャスティングボードを握っているといっても過言ではない．

情報を公開する際の問題

生命科学や医科学全般の発展を考えると，プロテオーム情報もゲノム情報と同様に，可能なかぎり多くの研究者間で共有されることが望まれる．このためには自由にアクセスして情報をダウンロードできる一方で，ゲノムデータベースのように，情報の登録も受け付けてくれるパブリックなプロテオームデータベースが現れることが望まれるが，実はこれはかならずしも簡単な話ではない．プロテオームデータベースにはゲノムデータベースにはなかった問題点が残されている．

ゲノムデータベースは単純に配列データベースであったために，すべての生物種や由来組織を区別せずに，ひとつのデータベースにまとめることができた．しかし，プロテオームは，生物種，組織，細胞に固有のものであり，生物種を統合したものをつくることは事実上不可能である．したがって，プロテオームデータベースの構築は，動物種，組織，細胞ごとに別々に行う必要がある．今後ゲノムデータベースのような，情報登録の窓口を設ける場合には，どのサイトがどの動物種，どの組織の情報を担当するのかを決める必要が生じるものと思われる．また，登録を受け付ける際のデータのフォーマットを統一する必要もある．これらの問題は今後，JHUPO の主要な協議事項のひとつになるであることが予想される．

文献

1) O'Farrell, P. H.：*J. Biol. Chem.*, **250**：4007-4021, 1975.
2) Lipkin, L. E. and Lemkin, P. F.：*Clin. Chem.*, **26**：1403-1412, 1980.
3) Anderson, N. G. et al.：*Clin. Chem.*, **27**：1807-1820, 1981.
4) Garrels, J. I.：*J. Biol. Chem.*, **264**：5269-5282, 1989.
5) Bjellqvist, B. et al.：*J. Biochem. Biophys. Methods*, **6**：317-339, 1982.
6) Henzel, W. J. et al.：*Proc. Natl. Acad. Sci. USA*, **90**：5011-5015, 1993.
7) Hoogland, C. et al.：*Electrophoresis*, **18**：2755-2758, 1997.
8) Vihinen, M.：*Biomol. Eng.*, **18**：241-248, 2001.
9) (株)オフィスエム：最新 XML ハンドブック．秀和システム，2001.
10) Appel, R. D. et al.：*Electrophoresis*, **14**：1232-1238, 1998.
11) Toda. T. et al.：*Electrophoresis*, **19**：344-348, 1998.
12) Appel, R. D. et al.：*Electrophoresis*, **17**：540-546, 1996.