［特集］ポストゲノム時代に高まるバイオ自然言語処理への期待：バイオ自然言語処理最新事情

2 バイオ研究者からのバイオNLPへの期待

b）大規模プロテオミクスからバイオNLPに望むこと

夏目　徹
natsume@jbirc.aist.go.jp / 産業技術総合研究所

忍耐と労力を惜しまず，経験を積み，知恵と工夫を凝らし，1つ1つのタンパク質に取り組むのが，これまでのタンパク質科学であった．しかし，プロテオミクス技術の飛躍的な進歩に伴い，網羅的・包括的解析がタンパク質レベルであっても夢物語ではなくなった．ものの数時間で数千種類のタンパク質を同定することも珍しくなくなった今，そこで大きな障害となって立ちはだかるのが，タンパク質名の不統一と曖昧性である．仮に近い将来，一義的に統一されたとしても，過去の知識を生命情報工学として活用できない．そこで期待されるNLPの役割は大きい．

プロテオミクス研究・バイオ界の第2次産業革命

プロテオミクス研究☆1にはバイオインフォマティクスが必須である，あるいはバイオインフォマティクスがプロテオミクス研究を牽引するのは間違いのない事実である．教科書によればバイオインフォマティクスとは「生命情報科学」というコトバとほぼ同一だそうだ．特にゲノム科学の進展とともに生み出された「莫大な検索空間（核酸配列，タンパク質名あるいはそれらの定量情報やネットワーク）を狭めるための情報技術と，それを可能とする基礎理論」として位置づけられているのである[1]．またバイオインフォマティクスとは「計算機を実験デバイスとして使い，生物学的な発見をする」学問であると考えている人もいるとも聞く．それはきっと正しいのだろう．しかし，プロテオミクス研究として実際にタンパク質に携わっているウェットな人間が直面する問題には，教科書に書かれたような「高級感」はない．プロテオミクス研究に何らかの形でかかわれば，すぐに数千のスペクトラムが質量分析計からはき出されることを目の当たりにする．1つ1つのスペクトラムはタンパク質の内部アミノ酸配列情報を含んでおり，これらをタンパク質配列データベースに照合することで，他種類のタンパク質があっという間に同定される．そして数百のタンパク質名とその配列情報に呆然とすることこそがリアルな現実だ．それは生命情報工学という学問とはまだ縁遠く，情報生産者に必要な1次情報処理技術そのものが，実に心許ないということが，データを大量に生産してみて初めて気づくという有様だ．

1次情報処理は誰がするのか？

いうまでもなく，情報処理には計算機とソフトウェア（あるいはツール）が必要である．そのソフトウェアの

☆1 プロテオミクス：ゲノムプロジェクトの進展とともに，細胞内で機能している全種類のタンパク質を網羅的かつ統一的に解析しようとする技術・方法論を指す．狭義の意味では，質量分析により，従来のタンパク質科学では不可能であった，多種類のタンパク質を高感度に同定しようとすることを指す．

ソースは以下のようなものであろう．インスツルメント（ここでは主に質量分析計）の付属のソフト，ツール（フリーウェアソフト，あるいはWeb上で提供されるサービス），そして汎用性が高くパッケージソフトとして購入するもの（たとえば検索エンジン等）．これらを駆使し，たとえば質量分析により大量のタンパク質を同定するという1次情報処理について考えよう．測定が終了するとともに1次データの処理がパイプラインとして流れなければ，ハイスループットな大規模解析など意味がない．一般的にアミノ酸の配列情報を持ったMSMSスペクトラを処理し「ピークリストファイル」と呼ばれるテキストに変換する．このファイルを用いサーチエンジンにより，タンパク質配列データベース検索を行う．これをバッチで行えば，タンパク質の同定結果を持ったファイル群が帰ってくる．数千のMSMSスペクトラを一度に取得することも珍しくない昨今であれば，このファイル群はやはり数千の数になるわけである．これらの結果を閲覧処理するため，次にオラクルなどに代表されるリレーショナルデータベースにデータを格納する．格納された後，ビュアを通してフィルタリング・エディティングを行い，最終的に研究者の目に触れることになる．これが最低限の1次情報処理である．ここでの大きな問題は，それぞれの場面で使われるソフト・ツールがマシンの付属ソフトだったり，市販パッケージングソフトであったりして，決してステップごとにソフト間の統合性など期待できないことだ．すなわち各ステップごとに人間の手によるマニュアルの作業が常につきまとう．たとえば測定の終了後にRawデータの処理を自動化し，さらに処理後にサーチエンジンのデーモンがファイルを取得し定められたデータベースに設定された検索条件を元に，自動的に走らせるといったことが意外に難しいのだ．また検索結果をリレーショナルデータベースにシームレスに流し込むことは，市販・付属ソフトを使いこなしたとしても到底不可能だ．さらにリレーショナルデータベース上のデータを効率よく閲覧したり，編集作業を行うプラットフォームなど，たぶんいまだにパッケージング化されていない．逆に言えば，我々の情報処理能力はこれらのパッケージやツールの目的とその処理範囲に限定されており，その範疇を少しでも外れると，それは外注，あるいはインフォマティクスの専門家との共同作業となる．しかし，外注ソフトの制作は金銭面的にはもちろん，思ったほど楽ではない．すなわちウェットのサイドの意図することをプログラマやインフォマティシャンに的確に伝えることが意外に大変であるからだ．幸い潤沢な研究資金があったとしても，外注するには，きちんとした仕様書を作らなければならない．その仕様書を馴れない人間（ずぶぬれのウェット研究者）が行えば，受注サイドとのやりとりは試行錯誤の繰り返しであり，完成させるのに半年など珍しくない．さらにソフトの完成に，バグフィックスを含め1年以上かかることもあろう．そうこうしているうちに，研究者の目標に変更が生じ，仕様の修正などがあろうものならソフトの完成は月より遠い．はたまた，月日のうちに目的自体が陳腐化しまったくの無用の長物に成り下がることもままある．したがって，ウェット研究者サイドに情報処理専門家の緊密なサポートがインハウスに不可欠ということである．

しかし，運良く情報処理に長けた研究者が身近にいたとしても，彼らがプロテオミクスに必要なインフォマティクスの統合化（情報1次処理のパイプライン化）に興味を持ってくれるかは，はなはだ疑わしい．我々にとっては非常に重要でかつ日々直面する問題が，バイオインフォマティクスの専門家の興味の対象とはなり得ないということだ．それはインフォマティクスの立場から言えば当たり前だ．インフォマティクスの統合などとカッコをつけても，それは所詮ウェット実験者の雑用の延長であり，インフォマティシャンの研究対象足り得ず，もちろん論文などは書けもしないからだ．

あるバイオインフォマティクスの入門書の前書きに「ウェット研究者はバイオインフォマティシャンの興味を引き出すため，自己の研究の面白さをなるべく解りやすく伝える努力が必要である」というくだりがあり，私は腰を抜かした．我々の汚れ仕事や雑用を「クリエイティブ」に伝えるとは，それを巷では一般に「詐欺」と呼ぶからだ．プロテオミクス研究の現場で，処理しなければならない問題は，1つ1つは些細だが，数が多くなると侮れないものばかりなのだ．そして，1次処理の統合化をウェット研究者，すなわち「私たち自身が自らの手で行わなければならない」ということだ．

バイオインフォマティクスの壁再び・Who's Michel problem

運良く，ウェット研究者自身の手で，このような1次情報処理の統合化を果たし，本当の大規模な解析を始められたとして，次に直面する問題はもっともっと深刻だ．そして，これは，もはやウェット研究者によって解決で

きる問題ではない．過去の研究から，さまざまなタンパク質がどのような機能をそれぞれ持ち，疾患等とどのような関係を持つかが電子化されオンラインジャーナルやデータベース上に蓄積されている．プロテオミクス研究で得られた大量のデータを，これらの既存の電子化された知識に照らし合わせ，新たな知識を発見しようというのがポストゲノム研究の大きな目標の1つだ．しかし，それは意外にも入り口のところで頓挫するのだ．それはタンパク質が，それぞれ正式名称（Official name），別称（Synonym），略称（Acronym），通称（Jargon）を冗長に持つからである．この問題は，極端に村意識が強いタンパク質研究の世界では，特に酷い．村（分野）によって独特の言い回しや，タンパク質名を使い，他村とのコンセンサスを得ようとはさらさら考えない．それどころか，独特の表現をすることで，同じ村民同士の結びつきを強固にし，よそ者を排除しようという意識すらあるからだ．

たとえば Transforming growth factor beta activated kinase 1というタンパク質がある．この略称はTAK1である．これは発見の経緯から名付けられ，Transforming growth factorの研究者がこう呼ぶ．しかし，タンパク質の構造という面から見ると，Map kinase kinase kinaseというクラスに属し仲間がすでに6個発見されている．したがってMap kinase kinase kinase 7という名前も付けられてしまった（略称はMAPKKK7，MAP3K，MKKK7など）．こちらの名称はMap kinaseを主に研究してきた研究者に支持される名前だ．このような別称問題は電子辞書中にキチンと別称・略称が記載されていれば特に問題なく計算機の中で処理が可能である．しかし困ったことにTR4 Nuclear hormone receptor というタンパク質もなぜか略称がTAK1である．したがって，電子化されたテキスト中でTAK1が出てきた場合，どちらのタンパク質を指すのかが単純に計算機には判断ができない．これなどはまだ単純な分かりやすい例であるが，実情はさらに複雑であり，高度なNLPを駆使し，文脈解析が功を奏さないと，誤ったデータを引き出すか，あるいは何の役にも立たぬノイズを生み出す．さて，ではなぜこのような深刻な問題をはらんだまま生物学の世界は進んできたのであろうか？　多くの生物学者がこのことを認識しているにもかかわらずだ．

たとえば，こんな猫はいないだろうか．ご近所さんの数軒でエサをもらう半ノラ猫である．山田さんちでは三毛猫なので「ミケ」と呼び，縁側から入ってきたら煮干しをやっていた．鈴木さんちでは，最初に現れたのがまだ子猫のころだったので「チビ」なぞと呼び，キャットフードを常備しエサを与え，やっぱり自分が飼い主のつもりでいた．野村さんちでは，人気漫画のキャラクタに似ているのかそうでないかよく分からないが「マイケル」と呼び，やはりかわいがっていた．そしてこのことは山田家，鈴木家，野村家も互いに知らない．しかし，ある日全員が同じネコに別の名前を付けていたことに気づく．そこで誰かが提案するのだ．「同じネコなので名前を1つに決めて，皆でかわいがりましょう」と．この状況は複数の研究グループが，発見したタンパク質を命名する状況に似ている．発見の経緯や，思い入れを込め，研究者たちは同じタンパク質に，実にさまざまなネーミングをすることになるからだ．その結果1つのタンパク質が，分析法や分野の違いによっていろいろな呼ばれ方をする結果となる．そしてその事情というのをよく知るのは，長年その分子に携わった事情通の長老にしか分からなかったりする．新規遺伝子クローニングや，タンパク質の発見は，命名という名誉で締めくくられる．そして，分野が大きな広がりを見せ（村が都市になると）1つのタンパク質が，分析法や分野の違いによっていろいろな呼ばれ方をすることが不都合になり，国際会議が開かれタンパク質名の統一を図る．各命名者は自分の命名がその後も存続することを願う．なぜなら自身のネーミングこそが研究者の「血と汗と涙の結晶」だからだ．ネコの名前のように，「トラと呼んでいたネコをマイケルとしましょう」というのとはわけが違う．なぜなら，残念にも廃止となれば，当然，自分の全研究業績すべてが否定されたも同然の喪失感を味わうこととなるからだ（もちろん，慣れ親しんだペットの名前を変えろと強制されるのもかなり心理的な抵抗があろうが）．また会議に出席できなかった不運な命名者は，すねる．ポストゲノムシークエンスの時代とは，そんな町村統廃合が，それこそ「ゲノム・スケール」で起きていることを意味する．だから，喪失し，すねたタンパク質科学者が，これから大量に生まれるのではないかと私は思う．そして，こんなことが実はポストゲノム研究の最大の抵抗勢力だったり足枷にならないことを切に願う．もし，仮にそうであるなら（ほとんどそうなりつつあるが），真のプロテオミクス研究は，これら遺伝子ハンティングの時代の偉大な功績者たちがすべて死に絶えるまで始まらないとすら思う．もちろん，こんな乱暴な議論も，他人の死を待ちわびるな

どという非現実的なことはできないのだ．だからタンパク質名が早急に統一され曖昧性が排除され，１つのタンパク質名が一義的に１つのタンパク質を指すようになることは当分ないと思われる．また奇跡的にそれが実現されたとしても，これまでの曖昧なタンパク質名で書かれた膨大な過去の知識は利用不可能である．だから，ある情報に含まれるタンパク質名がどのタンパク質を意味し，その関係を正確に抽出したりする高度なNLPがかくも必要であり生命線となるわけだ．

タンパク質科学者のインフォマティクスは豚が木に登るが如しか

タンパク質科学は本来，自給自足の農作業のようなものであったように思う．機械化不可能な急峻な棚田で猫の額ほどの田畑と泥まみれになりながら格闘し，自分たちのお腹を満たすだけのデータをやっとのことで収穫し，生活していく…．しかし最近，疲れた腰を伸ばし遠く山の下の遙かな平野を見やると，そこには大工場が建設されているのが見える．その中では整然と流れるオートメーション生産ラインから次々とゲノムな情報が大量生産されているというのだ．しかし，それは遙か彼方のゲノム・シティの出来事で，タンパク質科学者の住む山の生活とは無縁の出来事であると，皆信じていた．しかし，恐ろしいことにタンパク山はゲノムの力であっという間に切り開かれ，広大な農地を近代的なトラクタ（質量分析計）が行き来し，大型のコンバイン（高速計算機）が次々とデータを収穫し，収穫した農産物を消費者に向け流通しようかという時代になった．まるで夢のような出来事ではないか．しかし，機械化と大規模化による（データの）大量収穫で農民（タンパク質科学者）はついに幸せになるのであろうか．「大収穫イコール幸せ」でないことは，直ちに分かった．収穫した農産物の検品，箱詰め，鮮度保証と輸送という慣れない作業をいやというほどやらなければならないのだ．自分たちで育て収穫し，採れたてを食卓に並べ，家族で囲みその収穫を喜び味わうのとはわけがちがうのだ．目に見えない消費者への責任がつきまとうからだ．だから当然農作業だけに精通していればいいのではなくなった，ということだ．大規模解析が現実のものとなると，次なる新たなボトルネックは分析後のデータ処理とその利用，すなわちインフォマティクスの成否なのである．

参考文献

1）高木利久編：ゲノム医科学と基礎からのバイオインフォマティックス，実験医学増刊，羊土社，Vol.19, No.11（2001）．

（平成17年1月8日受付）