4　計算プロテオミクス：データベースを利用したタンパク質翻訳後修飾の同定

1　島津製作所　田中最先端研究所
2　エーザイ　エーザイ・プロダクトクリエーション・システムズ

吉沢　明康[1]・田畑　剛[2]・木村　剛之[2]・青島　健[2]・福山　裕子[1]・梶原　茂樹[1]・九山　浩樹[1]・小田　吉哉[2]・田中　耕一[1]

Computational Proteomics: Database-run Identification of Post-Translational Modification

Akiyasu C. YOSHIZAWA[1], Tsuyoshi TABATA[2], Takayuki KIMURA[2], Ken AOSHIMA[2], Yuko FUKUYAMA[1], Shigeki KAJIHARA[1], Hiroki KUYAMA[1], Yoshiya ODA[2], and Koichi TANAKA[1]

1 Koichi Tanaka Laboratory of Advanced Science and Technology, Shimadzu Corporation
2 Eisai Product Creation Systems, Eisai Corporation, Ltd

要　旨

質量分析計のデータを解釈してプロテオームのデータを作り出すComputational Proteomicsの新手法とそのためのデータベース2件，「MSPTM-DB」と「ProteinCarta」について述べる．タンパク質の既知の翻訳後修飾（PTM）情報を基に，PTMがある場合とない場合の配列を事前に作成してデータベースMSPTM-DBに収録し，これを検索対象にすることによって，従来からのPTM検索法の問題点である信頼性の低下を回避し，かつ検索速度を3倍以上に増加させた（MSPTM-DB法）．また，タンパク質のN/C両末端部分の配列合計8残基以上があれば，ヒトの場合ほぼ全てのタンパク質が同定可能であることが計算機実験で明らかになり，同定に用いるためのデータベースProteinCartaを構築した．これらのデータベースはwebからフリーで利用できる．

キーワード：プロテオミクス，データベース，PTM，末端配列，MSPTM-DB，ProteinCarta

プロテオミクスのバイオインフォマティクス

プロテオミクスの最大の特徴・意義は「（閾値以上の）全タンパク質の発現量，及び翻訳後修飾（Post-Translational Modification; PTM）が探知できること」であり，他のomics科学同様，データ処理にバイオインフォマティクスが不可欠である．但し，「プロテオミクスのバイオインフォマティクス」と総称される分野には，位置づけの異なる2種類の研究，“proteome informatics”と“computational proteomics”が共存している．

前者“proteome informatics”は端的に言うと「タンパク質の配列・翻訳後修飾の情報・タンパク質のリストなどのプロテオーム情報を基に，分子生物学的知見を創り出す」研究である．このような研究には，医療応用としてのバイオマーカー

Reprint requests to: Akiyasu C. YOSHIZAWA
Koichi Tanaka Laboratory of Advanced Science and Technology, Shimadzu Corporation
1 Nishinokyo Kuwabara-cho, Nakagyo-ku,
Kyoto 604－8511, Japan

別刷請求先：〒604－8511　京都市中京区西ノ京桑原町1
島津製作所田中最先端研究所　　吉沢明康

探索の他，タンパク質間相互作用のデータと組み合わせたネットワーク解析や，他の omics データ，例えばマイクロアレイの結果と統合したシステム生物学的な解析などがある．このためには各種データの統合が必要になるが，熊本大学腫瘍医学分野の iPeach[1] 開発に著者が参加した経験に基づけば，データ統合での最大の問題は「各データへの重み付け」と「omics データごとに異なる，遺伝子・タンパク質 ID の統一（即ちタンパク質版の“名寄せ”）」に関連する作業であり，後者には UniProt や KEGG などの公共データベースの駆使が必要になる．

もう一方の“computational proteomics”が意味するのは，「測定された生のデータからプロテオーム情報を創り出す」研究である．多くの場合，これは「質量分析計の生データを解釈してアミノ酸配列を推定する」処理を指しており，商用ソフトウェアの Mascot をはじめとする「検索エンジン」の開発・研究が多数行われている．本稿では，この分野の研究のために著者らが行った工夫について述べる．

質量分析法による翻訳後修飾の探知とその問題点

質量スペクトルの測定で「質の高いスペクトルが充分な回数得られた」場合には，アミノ酸配列の推定には *de novo* sequencing 法が利用できる．これは塩基配列決定の Sanger 法と同様の考え方で，試料タンパク質やペプチドをランダムに切断し，アミノ酸 1 個違いの長さの断片を仮想的にラダー状に並べることによって，質量差からアミノ酸の配列を推定する．この推定は高精度だが，現状では信頼性のある結果を得るためには，適切な mass-tag を結合させるなどの化学的手法を援用し，かつ経験者がスペクトルを目視で判断することが必要となる場合がしばしばであり，短い配列の推定しか実用的でない．

このため実際のプロテオーム解析では，タンパク質データベースの検索が用いられる．検索エンジンはデータベースから配列を 1 個読み込むと，ユーザーが実験で用いた酵素でタンパク質を消化したときに生じるペプチドの配列をまずメモリ中に作成し，それと測定スペクトルがどれくらいマッチするか比較してスコアを算出する．本稿では以下，このような検索方法を「通常検索」と呼ぶことにする．

細胞や生体に表現型変化を惹き起こすような生物学的事象としては，タンパク質の発現状態の変化以外に例えば PTM が生じる場合が考えられるが，この場合ペプチドの質量が変化するため，処理が難しくなる．PTM 探知に特化した実験を行わず，通常の測定データを基に計算機的に PTM を探知する場合，その手法は実質的に variable modification 法のみに限定される．即ち検索エンジンは，メモリ中に生成したペプチドそれぞれについて，PTM が生じる可能性のある残基（possible PTM site）全てに対して，PTM が生じた場合と生じていない場合それぞれの配列バリエーションをメモリ中に生成し，これらに対してスペクトルとマッチするか比較する．従って元々 1 個しかないペプチドが，メモリ中では複数個に増幅されて検索が行われるため，結果の信頼性（E-value）には変動が生じ，有意な結果か否かの判定，ひいては「PTM が生じていないため通常検索でも探知できる筈の配列」の探知にまで影響が生じることがある．

専用データベース MSPTM-DB による既知翻訳後修飾の探知

この問題に対応するために，我々は既知 PTM 探知用のデータベース MSPTM-DB を作成した．これは，「公共データベース（UniProt）の情報に基づいて，既知 PTM が報告されている残基（known PTM site）についてのみ，アミノ酸配列のバリエーションを作成し，これをデータベースとして通常検索の対象にする」というものである．

当然，この手法で探知できるのは既知の PTM に限られるが，表現型変化を惹き起こすような変化には，「既知 PTM の実際の修飾の有無」という

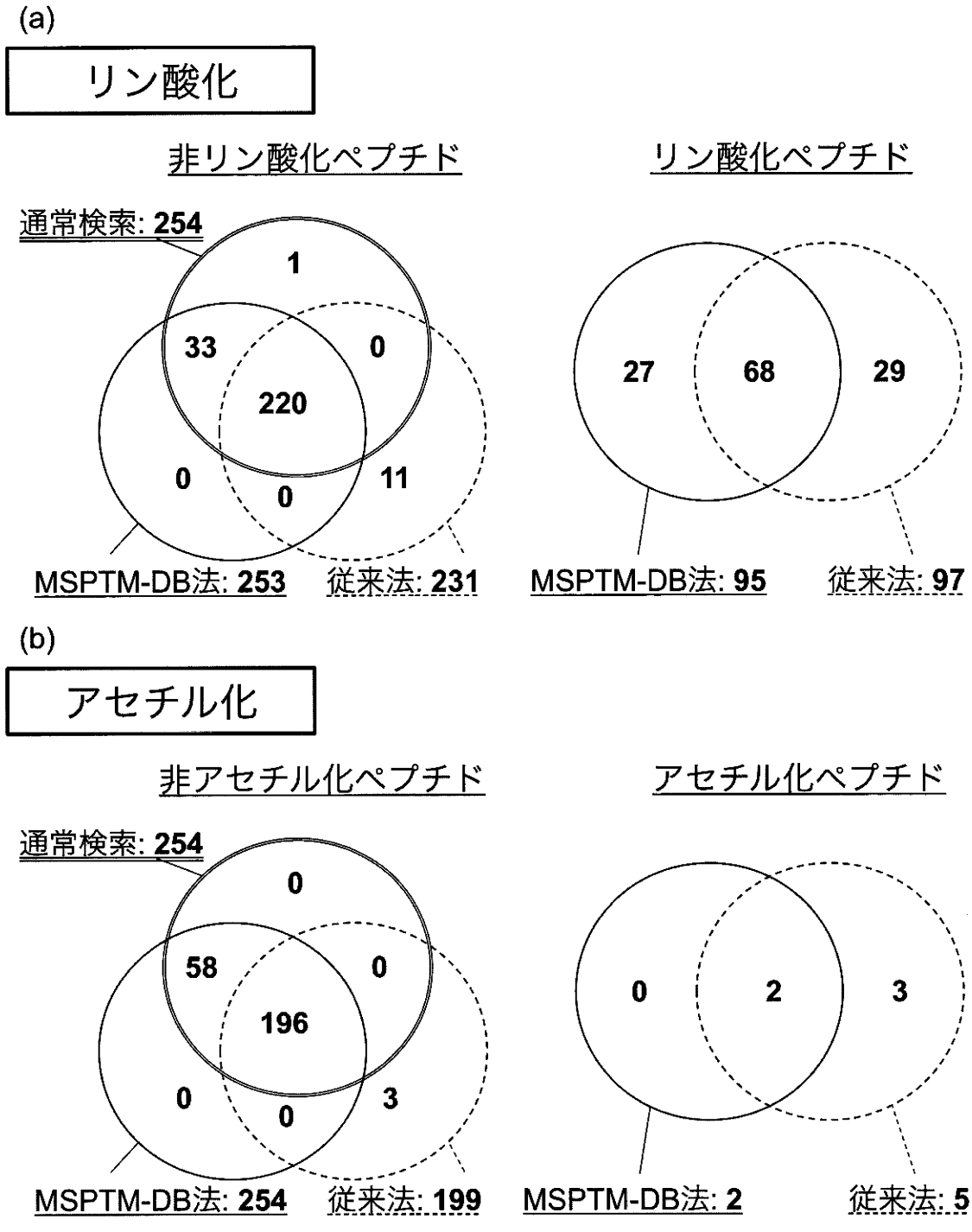

図1　リン酸化（a）とアセチル化（b）探知のために，MSPTM-DB法と従来法で検索した結果．非リン酸化・アセチル化ペプチドの探知結果は通常検索の結果と一致するのが望ましいが，従来法ではかなりの違いが生じている．MSPTM-DB法では，違いは非常に少ない．

場合も考えられる．また known PTM site は possible PTM site よりも当然少ないので，配列のバリエーション数は相対的に少なく，探索空間が減少するため，従来法の「信頼度の低下」という問題を相当程度回避可能であると考えられる．

また既知 PTM 情報は，公共データベース中に既に相当量が蓄積されているが，現時点ではこれを検索に利用する方法はまだ発表されておらず，既存のデータベース検索エンジン（最も代表的な商用エンジン Mascot や，オープンソースのエンジン X!Tandem，freeware のエンジン MaxQuant など）も対応していない．MSPTM-DB は従って，通常検索の対象として構築した．

Swiss-Prot Release 2011_04 に収録された既知リン酸化情報に基づくと，1 個のタンパク質中に存在するリン酸化部位の最大数は 229 個（SRRM2_HUMAN）だった．これに対して，Trypsin で消化して生成したペプチドの場合，最大数は 12 個だった．そこで Swiss-Prot のタンパク質配列からトリプシン消化配列を生成し，既知リン酸化情報に基づいて，これら PTM が報告されている残基についてのみペプチドのバリエーションを作成した．

性能検証として，HeLa 細胞抽出物由来サンプル（還元アルキル化し，Lys-C と Trypsin で消化後，IMAC（immobilized metal affinity chromatography）でリン酸化ペプチドを濃縮）に対して，X!Tandem（ver. 2010.01.01.4）を用いたデータベース検索を行った．Variable modification（従来）法で Swiss-Prot を検索した場合と，Swiss-Prot と MSPTM-DB を通常検索した場合（MSPTM-DB 法）それぞれの検索結果の第 1 位の配列を採用した．**図 1（a）**に示したのは，推定配列の E-value が 10^{-2} 以下だった場合の件数，**図 1（b）**はアセチル化の場合の同様の件数である．

リン酸化の結果で特徴的なのは，非リン酸化ペプチド（PTM がないので通常検索で探知できる）の場合，従来法では 2 割近い配列が通常検索の結果と異なっているということで，E-value の変化によって誤判定が生じていることが推測される．同様の結果はアセチル化の場合でも見られる．また検索速度は，ヒト配列のみを検索対象にした場合従来法の 3 倍以上，Swiss-Prot 全体を対象にした場合は 20 倍以上高速化した．更に，Swiss-Prot の文献情報などを同時に収録することで，既知 PTM の文献情報を検索と同時に表示することも可能になった[2]．

末端配列の探知によるタンパク質同定

MSPTM-DB はデータベース検索法に対する工夫であるが，次に *de novo* sequencing 法に関連した工夫について述べる．著者らは，タンパク質末端の短い配列（末端配列タグ）を用いてタンパク質を同定するための配列データベース ProteinCarta を構築している．収録データは，UniProt Release 2013_3 に登録された，ヒト・マウス・ラット・ショウジョウバエ・線虫・シロイヌナズナ・出芽酵母・大腸菌の 8 生物種の全タンパク質配列から両末端それぞれ 50 残基分で，入力したタグ配列を末端領域内に持つタンパク質の ID や signal peptide 情報を表示する．また入力したタグで一意に同定できなかった場合に「あと何残基のアミノ酸を確定すれば一意に同定できるか」を計算によって推定する．

既に述べたとおり，*de novo* sequencing 法では短い配列タグの同定以外は困難と考えられ，また配列タグは多数の配列中に現れるのが普通であるため，一般に配列タグのみでタンパク質を一意に同定するのは難しい．しかしその配列タグの「タンパク質内での位置情報」が利用できる場合には情報量が増えるため，「末端である」という位置情報を含んでいる末端タグの場合には同定できる可能性がある．

この手法の礎となったのは，1998 年に発表された「タンパク質末端配列タグのみでタンパク質を同定できる可能性が高い」という知見である[3]．末端領域では，signal peptide の除去や官能基の付加などの PTM が生じることが多く，この領域の研究には生物学的な意義がある．また実験的に得られた末端タグとデータベース中の末端配列を比較することによって，新規の PTM（新規の signal

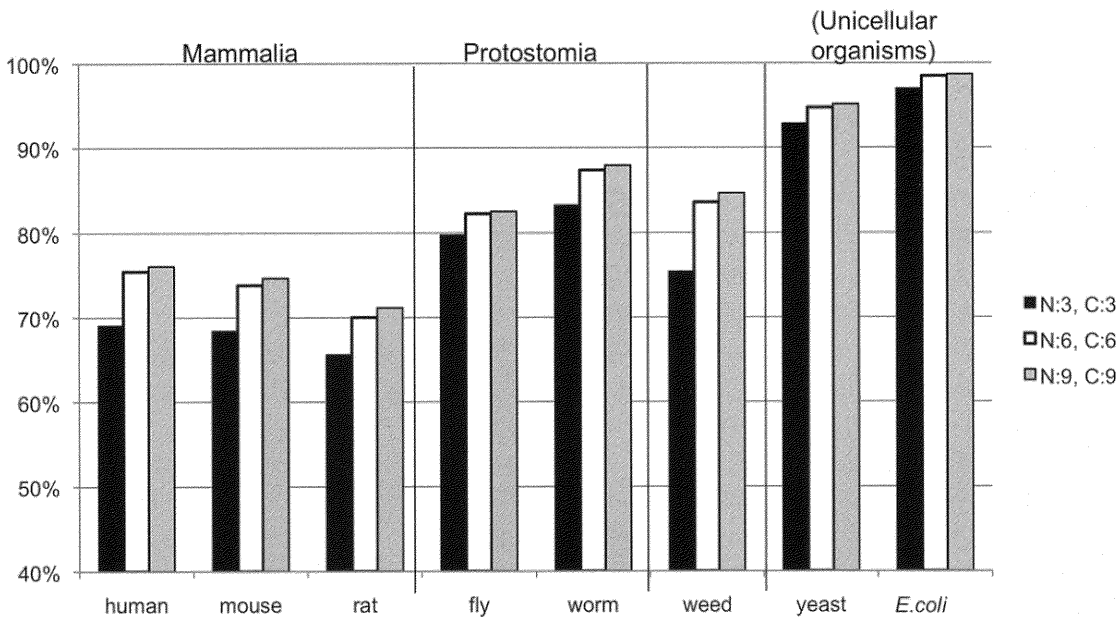

図2　N末端タグとC末端タグの長さを変えた場合（3残基ずつ，6残基ずつ，9残基ずつ）の，一意な末端タグのプロテオーム全体に対する比率．哺乳類（Mammalia），前口動物（Protostomia），単細胞生物（Unicellular organisms）の比率には類似した傾向がある．

peptideの除去）などが発見できる可能性もある．その意味でもタンパク質末端は興味深い領域であるが，98年当時にはN末端配列のEdman分解しか，配列同定手法が用いられていなかった．しかし現在では，両末端とも *de novo* sequencing 法で同定する実験手法が既に確立されている[4)－6)]．

そこで著者らはまず，上記8生物種のプロテオーム配列データから末端領域配列を抽出し，その生物種の全プロテオームに対してどの程度の一意性を持つかを調査した．「N, C両末端3残基ずつ」の配列タグの場合，ヒト全プロテオーム中の69.2％のタンパク質が一意な末端配列タグを持っていた．マウスの場合68.5％，ラットの場合65.7％で結果は類似していたが，ハエでは79.8％，線虫では83.2％と高く，ナズナでは75.4％，酵母では92.8％，大腸菌は97.0％と，分類群（taxon）ごとに差がある傾向が見られた[7)]（**図2**）．

更に，ヒト全タンパク質のアノテーションを調査し，進化的に見て末端タグが同一になってもおかしくない配列，即ちparalogや，同一protein familyのメンバーなど，末端タグのみで区別することが原理的に困難である配列を調査したところ，「全く違う種類のタンパク質が共通の末端タグを持っている例」は皆無ではないが5％以下であることが判明した．この値は末端タグの長さに依存するが，N/C両末端の合計が8残基程度になると，（paralogや同一protein familyのメンバーを除けば）ほぼ全てのタンパク質が同定可能であった．

以上のことから，実験的に決定した末端タグを用いて，専用データベースからタンパク質を同定するこの手法は，通常のデータベース検索では同定が難しい場合などに特に有効であると考えられる．

本研究は，proteome informaticsの基礎研究の結果を，computational proteomicsの手法に直接

応用することによって可能になったもので，学際的分野に於ける基礎研究と応用的技術開発の一体化の一例であると考える．

本稿で紹介したデータベースは，以下の URL からフリーで利用可能である．

http://www.first-ms3d.jp/english/achievement/software

謝　辞

本研究は，日本学術振興会の最先端研究開発支援（FIRST）プログラムによる助成を受けて行われた．また，MSPTM-DB 開発の特に初期段階で，（株）メイズの協力を得た．

参考文献

1）Hirayama M, Kobayashi D, Mizuguchi S, Morikawa T, Nagayama M, Midorikawa U, Wilson MM, Nambu AN, Yoshizawa AC, Kawano S and Araki N: Integrated proteomics identified novel activation of dynein IC2-GR-COX-1 signaling in neurofibromatosis type I (NF1) disease model cells. Mol Cell Proteomics 12: 1377-1394, 2013.

2）Yoshizawa AC, Tabata T, Kimura T, Aoshima K, Oda Y, Kajihara K and Tanaka K: MSPTM-DB: a known PTM database for high-speed and accurate search available on the "Proteo-Analysis" web site. 19th International Mass Spectrometry Conference. Oral Session 34: 1040, 2012.

3）Wilkins MR, Gasteiger E, Tonella L, Ou K, Tyler M, Sanchez JC, Gooley AA, Walsh BJ, Bairoch A, Appel RD, Williams KL and Hochstrasser DF: Protein identification with N and C-terminal sequence tags in proteome projects. J Mol Biol 278: 599-608, 1998.

4）Kuyama H, Shima K, Sonomura K, Yamaguchi M, Ando E, Nishimura O and Tsunasawa S: A simple and highly successful C-terminal sequence analysis of proteins by mass spectrometry. Proteomics 8: 1539-1550, 2008.

5）Kuyama H, Sonomura K, Nishimura O and Tsunasawa S: A method for N-terminal de novo sequence analysis of proteins by matrix-assisted laser desorption/ionization mass spectrometry. Anal Biochem 380: 291-296, 2008.

6）Nakajima C, Kuyama H, Nakazawa T and Nishimura O: A method for N-terminal de novo sequencing bof Na-blocked proteins by mass spectrometry. Analyst 136: 113-119, 2011.

7）Yoshizawa AC, Fukuyama Y, Kajihara S, Kuyama H and Tanaka K: Computational survey of sequence specificity for protein terminal tags covering nine organisms and its application to protein identification. J Proteome Res 14: 756-767, 2015.

5　ヒト腸内細菌叢代謝機構の解明

山田　拓司

東京工業大学大学院生命理工学研究科生命情報専攻