

§2-1-4 データベース・解析ソフトウェアの増強

本プロジェクトで実現する次世代質量分析システムでは、性能の向上に伴って複雑な未知化合物が多種類測定されると期待されるので、より高度な解析を行うための使いやすいデータベース・解析ソフトウェアの開発が必要である。島津製作所のソフトウェアグループでは、未知(*De novo*)物質を同定するため、既知物質・部品からなるデータベースの構築と、未知物質までも対応可能なデータ表示・解析ソフトウェアの開発に取り組んだ。表示・解析ソフトウェアの開発では、2006年からCRESTプロジェクトでエーザイが開発していた質量分析用表示解析ソフトMass++をプラットフォームとして採用し、本プロジェクトで必要な機能をプラグイン形式で開発した。Mass++の概観を図1に示す。

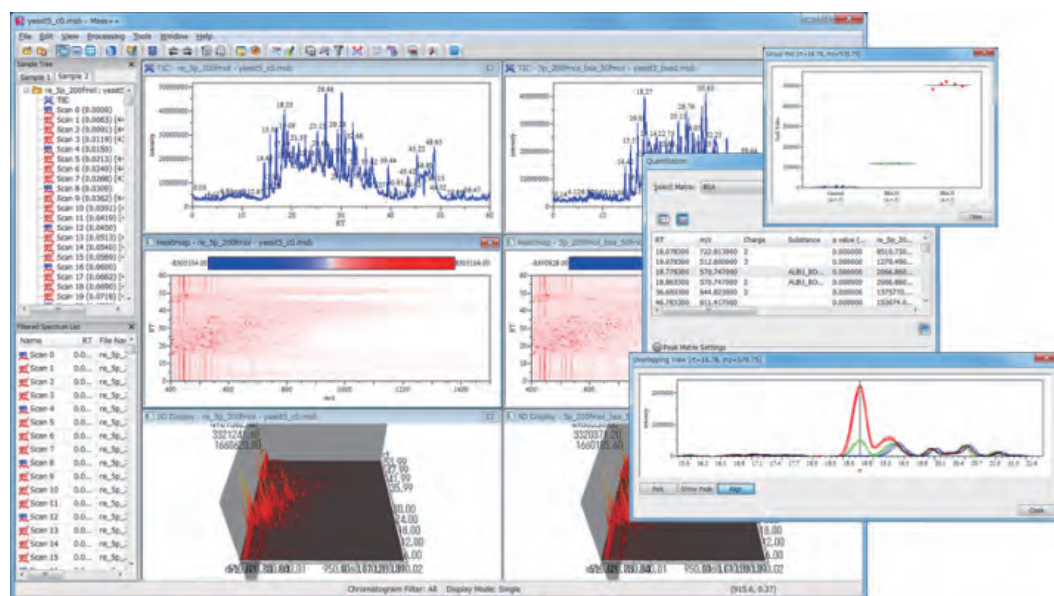


図1 Mass++の概観

Mass++には、スペクトル表示やヒートマップ表示(2Dおよび3D)のほか、ピーク検出、ピーク定量や差異解析(ピークマトリックスの作成、箱ひげ図表示、スペクトルの比較表示)などの機能がある。

本報告書では、以下の開発成果について報告する。

1 Mass++プラグインの拡充

独自の構造解析法、および次世代質量分析システムに必要な機能をMass++のプラグインとして開発した。

1-1 MSⁿを利用したデータベース検索法(MSⁿ Search)

1-2 MSⁿを利用した*de novo* sequencing法(SIMSE)

次世代質量分析システムの特徴の1つであるMSⁿ分析は構造解析に非常に有用な手法であり、本特徴を活用するため、MSⁿを利用した独自の解析ソフトを開発した。

1-3 生理活性ペプチド同定法(Sequence Tag Library Search)

血中や尿中のバイオマーカーを探索するため、従来法では同定が困難な生理活性ペプチドをターゲットとする新規の解析ソフトを開発した。

1-4 次世代質量分析システムへの対応

マスマスペクトル解析やデータ表示、ピーク検出・定量など、次世代質量分析システムにおいて必要な機能の開発・改良を行い、従来法との性能比較を行った。

2 Mass++プラットフォームの改良と普及

Mass++を開発プラットフォームとして効率的に利用するための整備を行い、一般へも公開した。

2-1 開発プラットフォームとしての整備

多様なユーザーニーズに対応するため、ユーザー自身が容易に機能を追加できるようなソフトウェア構成、ライセンス、プログラミング支援機能を整備した。

2-2 Mass++の一般公開とその特徴

マルチベンダー対応のフリーウェアとして一般公開し、ユーザーからの要望や質問を参考に改良を行った。また、他のフリーソフトと比較したMass++の特徴をまとめた。

3 Mass++を利用したアプリケーションソフトの開発

Mass++のプラグインを利用し、測定と連動した解析を行う自動分析ソフトや、膨大な測定データの解析を並列処理するジョブ管理ソフトを開発した。

3-1 試料中の物質の網羅的な自動分析法 (LDIA for Pep)

3-2 目的物質の詳細解析を行う自動分析法 (LDIA for GP)

バイオマーカー探索のためにはハイスループットの分析システムが必須である。Mass++の解析機能を利用し、解析結果から次の測定条件を立案し測定を実行する処理を繰り返す自動分析ソフトを開発した。

3-3 ビッグデータ対応 (JobRequest & ProteoAnalysis)

膨大な測定データを一括してルーチンで解析するため、Mass++と組み合わせたジョブ管理ソフトJobRequestと、クラウドコンピュータ環境でJobRequestを運用するソフトProteoAnalysisを開発した。

4 データベースの開発と拡充

バイオマーカー探索用新規データベースシステムの構築、高度な構造解析を目的とする既存データベースの活用、MassBankデータベースとの連携、などを行った。

4-1 尿中ペプチドデータベースシステム

バイオマーカー探索のため、独自の尿中ペプチドデータベースシステムを開発した。

4-2 翻訳後修飾データベース (MSPTM-DB)

4-3 末端配列データベースシステム (ProteinCarta)

公的データベースを有効利用してタンパク質の翻訳後修飾データベースおよび末端解析データベースシステムを構築した。

4-4 マススペクトルデータベースMassBankとの連携

日本質量分析学会の公式データベースMassBankを利用するソフトウェアを開発し、データベースの拡充に貢献した。

上記の開発成果(ソフトウェア)は、原則、一般利用可能な形で公開した結果、公開開始から2年でMass++のユニークユーザーが1,000人を超え、質量分析用表示解析ソフトウェアとして高い知名度を得ている。

1 Mass++プラグインの拡充

次世代質量分析システムの特徴の1つである MS^n 測定は、複雑な未知化合物の構造解析を行う上で多くの情報を得られる方法である。しかし、本プロジェクトで開発されるような高分解能の MS^n ($n \geq 3$)測定を行う装置は普及していないため、 MS^n データを有効に利用する解析ソフトウェアの開発は進んでいない。そこで、本研究では、 MS^n データを解析するための新規手法を考案し、 MS^n データ用のデータベース検索ソフトと*de novo sequencing*ソフトを開発した。

さらに、本プロジェクトにおいて、血中や尿中の生理活性ペプチドをバイオマーカーとする場合、従来の構造解析法ではこのような生理活性ペプチドを同定することが困難であった。そこで、我々は、生理活性ペプチドを同定するため、独自の新規ソフトウェアを開発した。

また、次世代質量分析システムの開発をデータ解析を進める上で必要となった新規機能をMass++のプラグインとして実装し、また既存プラグインの改良を行った。

1-1 MS^n を利用したデータベース検索法 (MS^n Search)

【研究目的】

MS^n ($n \geq 2$)データを利用したデータベース検索法として、階層的検索法[1]や検索結果マージ法[2]が提案されている。しかしながら、階層的検索法では、正しいペプチドを同定できるのは、 MS^2 検索で正しいペプチドが候補となる場合に限られる。検索結果マージ法では MS^2 と MS^3 の両方で正しいペプチドがハイスコアでヒットしなければ、ペプチド同定の信頼度が向上しないという問題がある。また、これらのソフトウェアは一般に利用できる形態で公開されていない。

本研究では、 MS^n データを利用する独自のデータベース検索法 (MS^n Search法)を開発し一般に公開した。

【研究成果】

方法

MS^2 と MS^3 データを利用する場合の MS^n Search法の概念図を図1に示す。解析手順は以下の通りである。

- ① 実測 MS^2 から作成した MS^2 ピークリストを使ってデータベース検索により、複数の候補ペプチドを選出する。
- ② 各候補ペプチドについて、 MS^3 の理論フラグメントピークを求め、実測 MS^2 には存在せず実測 MS^3 に存在するピークを抽出する。
- ③ 抽出した MS^3 ピークを、①で使用した MS^2 のピークリストに追加して修正ピークリストを作成する。
- ④ 修正ピークリストを使用して、全ての候補ペプチドのスコアを再計算した後に、それらを比較してスコア順に候補ペプチドを提示する。

結果

4種類のタンパク質 (Bovine Serum Albumin, Lysozyme C, Obalbumin, Glyceraldehyde -3-phosphate dehydrogenase) のtrypsin消化物をMALDI-QIT-TOFMS (島津製作所) で測定し、従来法 (MS^2 を利用したX!Tandem) と、 MS^2 と MS^3 を利用した MS^n Search法の比較を行った。検索結果の信頼度 (期待値) を表1に示す。 MS^n Search法を使用すると20例中17例で検索結果の信頼度が向上した。

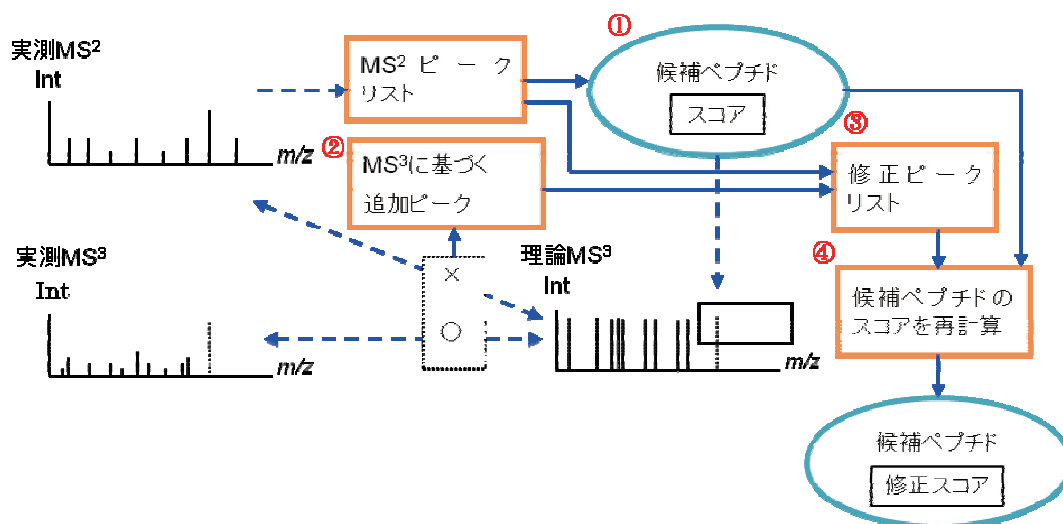


図1 MSⁿ Search法の概念図

MS²によるデータベース検索でヒットした候補ペプチドの配列とMS³のピークを照合し、ペプチド由来のフラグメントと考えられるMS³のピークをMS²ピークリストに追加して、再度、データベース検索を行うことで検索結果の信頼度を向上させる。

表1 従来法とMSⁿ Search法による検索結果の信頼度(期待値)

Sample	Peptide sequence	従来法による検索結果の期待値	MS ² Search 法による検索結果の期待値
BSA	LVNELTEFAK	3.7e-003	1.2e-003
	YLYEIAR	4.0e+000	8.3e-002
	RHPYFYAPELLYYANK	5.6e-005	6.6e-006
	DAIPENLPPLTADFAEDKDVCCK	4.5e-009	1.6e-010
	LGEYGFQNALIVR	4.3e-004	2.7e-006
	KVPQVSTPTLVEVSR	9.8e-007	2.7e-008
	RPCFSALTPDETYVPK	2.1e-003	4.0e-006
LysC	IVSDGNGMNAWVAWR	5.5e-001	1.4e-003
	NTDGSTDYGILQINSR	3.6e-001	1.2e-002
	NLCNIPCSALLSSDITASVNCCK	2.5e-006	7.9e-007
	FESNFNTQATNR	2.6e-004	2.3e-003
OVAL	AFKDEDTQAMPFR	7.8e-004	2.0e-006
	ISQAVHAAHAEINEAGR	3.8e-010	5.9e-009
	GGLEPINFQTAADQAR	2.6e-004	6.1e-008
	LTEWTSSNVMEER	3.8e-005	5.8e-005
G3P	LISWYDNEFGYSNR	2.4e-001	6.7e-002

16個の試料のMS²を測定し、さらに各試料について1~3個のMS³を測定した。従来法とMSⁿ Search法によるペプチド検索の期待値を示す。期待値が小さいほど検索結果の信頼度が高いことを意味しており、従来法よりMSⁿ Search法の信頼度が向上した結果を赤字で示した。MSⁿ Search法を使用すると、20例中17例で検索の信頼度が向上した。

【結論】

MS²とMS³データを利用するデータベース検索ソフトMSⁿ Searchは、MS²データのみを利用する通常の方法と比較して、検索結果の信頼度が向上する。従来、一般に利用可能なMSⁿデータ用データベース検索ソフトはなかったが、本ソフトはMass++プラグインとして一般に公開されているので、今後、複雑な構造解析を行うためMSⁿ測定を行う多くのユーザーが利用すると期待される。

【成果一覧】

学会発表

- (1) Kentaro Morimoto, Masaki Murase, Tsuyoshi Tabata, Shigeki Kajihara, Yoshiya Oda, Koichi Tanaka. "Peptide search algorithm by selecting and rescoring reliable peaks for MSⁿ(n>1) spectra", 59th ASMS Conference on Mass Spectrometry and Allied Topics, 2011
- (2) 森本健太郎, 村瀬雅樹, 田畑 剛, 梶原茂樹, 青島 健, 小田吉哉, 田中耕一, "MSⁿを利用する新規データベース検索法の評価", 第59回質量分析総合討論会, 2011年
- (3) 森本健太郎, 村瀬雅樹, 田畑剛, 梶原茂樹, 青島健, 小田吉哉, 田中耕一, "MSⁿを活用したデータベース検索によるm/zが近接するペプチドイオン解析", 第10回日本プロテオーム学会2012年大会, 2012年

特許出願

- (1) 森本健太郎, 村瀬雅樹, 質量分析データ解析方法及び解析装置、特願2011-124792
- (2) Kentaro Morimoto, Method and System for Analyzing Mass Spectrometry Data, US 13/656505

【参考文献】

- [1] Ulintz PJ et al. Investigating MS²/MS³ matching statistics: a model for coupling consecutive stage mass spectrometry data for increased peptide identification confidence. Mol Cell Proteomics, 2007, Vol.7 No.1, pages 71-87
- [2] Hua Xu et al. A Hierarchical MS²/MS³ Database Search Algorithm for Automated Analysis of Phosphopeptide Tandem Mass Spectra. Proteomics. 2009, Vol. 9 No.7, pages 1763–1770

1-2 MSⁿを利用したde novo sequencing法(SIMSE)

【研究目的】

一般に、データベースに登録されていない化合物の構造を推定するため、de novo sequencing法が利用される。しかしながら、de novo sequencing法で、信頼度が高い結果を得るためには、多くのフラグメントイオンが測定されている必要がある。このとき、MS²だけでなくMSⁿ(n≧3)も利用できれば、解析に利用できるフラグメントイオンの数が増え、信頼度の高い結果が得られると期待できる。Bandeiraらは、MS²とMS³を結合して、ペプチドのde novo sequencingを行う方法を提案したが[1]、一般に公開され利用可能なソフトウェアは存在しない。

我々は、本プロジェクト以前に開発したde novo sequencingソフトウェアSIMSE(Shimadzu Sequence Explorer)[2] をMass++プラグインとして発展させるとともに、ペプチドだけでなく様々な化合物に対する配列推定評価を行い、開発したソフトを一般に公開した。

【研究成果】

方法

de novo sequencing法は、マスペクトルから検出したピークリストの中からピーク間の質量差に一致する化合物を見つける手法である。本プロジェクトで開発したSIMSEソフトの実行結果(図1)を参照し、操作・処理手順を説明する。

- ① ユーザーは、起点となるピーク(図1ダイアログボックスのStart Peak: m/z 1,476.71)、化合物の種類(アミノ酸、糖鎖、核酸の中から選択。図1ダイアログボックスのCompounds: Sugar)、化合物を探索する方向(高質量方向、低質量方向の中から選択。図1ダイアログボックスのSequence Direction: To High Mass)、質量許容誤差(図1ダイアログボックスのTolerance: 0.3 Da)を指定する。
- ② プログラムは、起点ピークから探索方向に向かって化合物質量と一致するピークを探索し、さらにそのピークを起点として化合物質量と一致するピークを探索する処理を繰り返す。ピークの探索には、動的計画法を利用して、漏れなく候補化合物の配列を見つけ、独自のスコアリング方法により順位づけを行う。探索処理が完了すると、ダイアログ上に候補化合物とスコアが表示される。図1では、Rank 1位で、スコア855のCompoundとしてHexNAcの環開裂(x02+, X02-)、HexNAc、Hex、Hex、Hex、HexNAc、Hex、HexNAcが表示されている。
- ③ さらに、スペクトル上に候補化合物と帰属ピークが表示される。図1では、Rank 1位からRank5位の候補Compoundが帰属ピーク上に表示されている。

結果

ヒト・タンパク質Transferrinから得られた糖ペプチド(CGLVPVLAENYnK+4HexNAc 5Hex: 小文字nは糖鎖結合部位を示す)のMS¹のマスペクトル解析結果を以下に示す。

N結合型糖ペプチドのMS²のマスペクトルでは、HexNAcの環開裂(x02)に伴うTriplet peaks(質量差120 Daと83 Da)が現れる。 m/z 3,099.40をプリカーサとして測定したMS²のマスペクトルにおいて、Triplet Peaksの最も低質量側にあるピーク m/z 1,476.71を起点とし高質量方向に糖を探索した結果、ランク1位で正しい糖鎖組成(4 HexNAc, 5 Hex)を推定できた(図1 a))。次に、 m/z 1,476.71を起点とし低質量方向にアミノ酸を探索したが、正しいアミノ酸配列を得ることはできなかった(図1 b))。

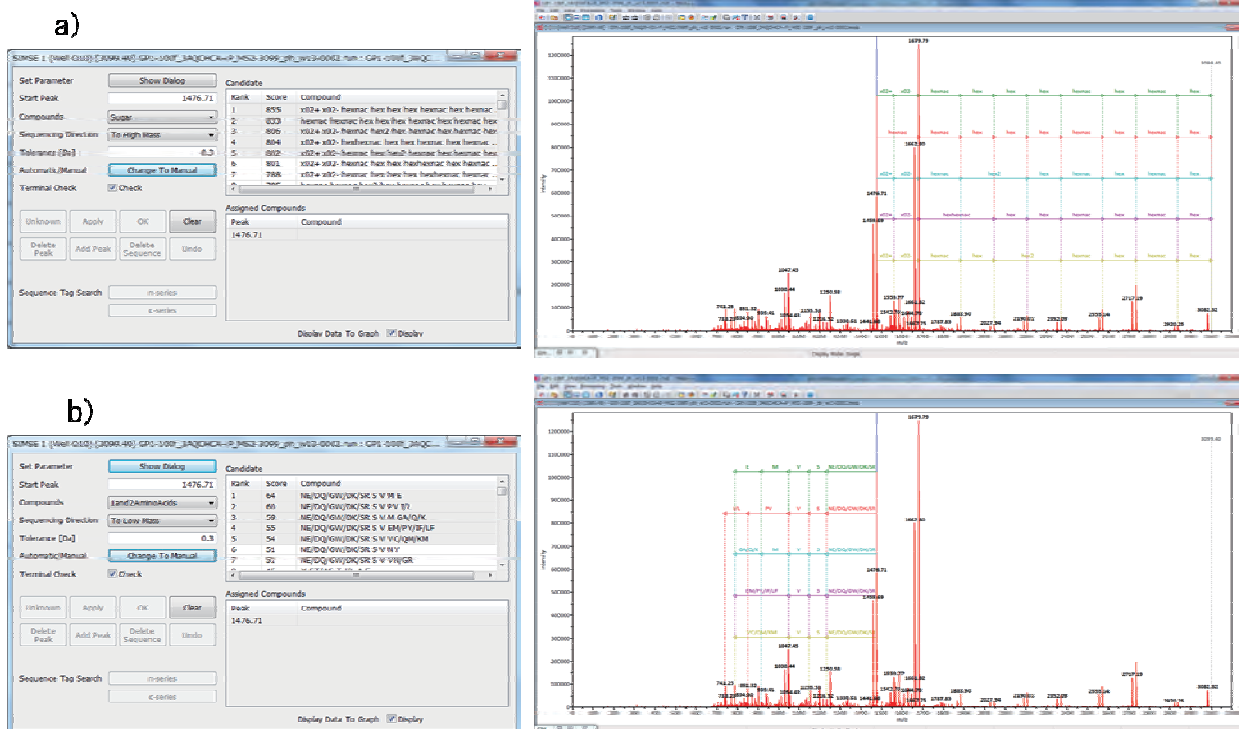


図1 MS²の*de novo* sequencing結果

de novo sequencing法により、HexNAcの環開裂(x02)に伴うTriplet peaksを起点とし、高質量側で糖鎖、低質量側でペプチド配列を推定した。a) 高質量領域では、糖鎖由来のピークを帰属し、正しい糖鎖組成(4 HexNAc, 5 Hex)を推定できた。b) 低質量領域では、ペプチド配列を正しく推定できなかった。

そこで、 m/z 1,476.71をプリカーサとして MS^3 を測定し、さらに MS^3 上の大きなピーク(m/z 1,047.30)をプリカーサとして MS^4 を測定し、これらのスペクトルを結合して解析を行った。その結果、ランク1位でアミノ酸の全配列を正しく推定できた(図2)。

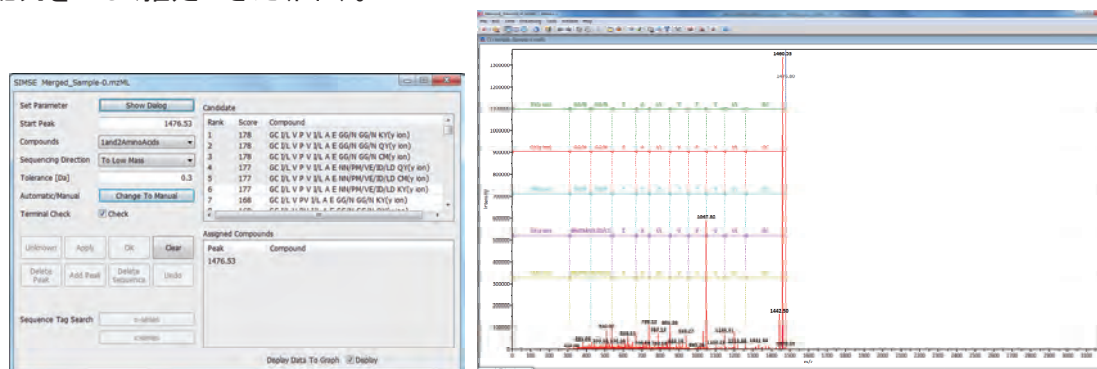


図2 MS^3 & MS^4 の*de novo* sequencing結果

さらに MS^3 と MS^4 を測定し、それらを結合して*de novo* sequencingを行った結果、Rank 1位の候補配列としてGC-I/L-V-P-VI/L-A-E-GG/N-GG/N-KYが得られた。配列表示において、- が無い2文字のアミノ酸の順番は未決定であり(ex.GCはG-C-かC-G-)、/で区切られたアミノ酸はいずれかのアミノ酸(ex.I/LはI-かL-)を意味する。よって、Rank 1位の候補配列は正しい配列CGLVPVLAENYNKと一致している。

【結論】

MS^n データ用*de novo* sequencingソフトSIMSEを開発し、Mass++上で利用可能なプラグインとして公開した。従来、一般に利用できる MS^n データ用配列推定ソフトは無かったが、今後、 MS^n 測定と本ソフトを利用することで、データベースに登録されていないペプチドや糖鎖、核酸などの多様な化合物の解析が可能になる。

【成果一覧】

学会発表

- (1) Kentaro Morimoto, Takashi Nishikaze, Satoshi Tanaka, Masaki Murase, Shin-ichi Utsunomiya, Shigeki Kajihara, Tsuyoshi Tabata, Ken Aoshima, Yoshiya Oda, Koichi Tanaka. "New plug-ins for freely available Mass++ software to identify biomolecules", 61th ASMS Conference on Mass Spectrometry and Allied Topics, 2013
- (2) 森本健太郎、西風隆司、田中聡、村瀬雅樹、船越なつ美、福山裕子、児嶋浩一、宇都宮眞一、梶原茂樹、高橋健太郎、福田充、青島健、小田吉哉、田中耕一、「Mass++: 質量分析データによる構造解析機能 — 糖鎖データベース検索を中心に」、第61回質量分析総合討論会、2013年

特許出願

- (1) 森本健太郎、質量分析データ解析方法及び解析装置、特願2011-231435

【参考文献】

- [1] Nuno Bandeira et al. Multi-spectra peptide sequencing and its applications to multistage mass spectrometry. ISMB, 2008, Vol. 24, page416-423
- [2] Masaki Murase et al. Integrated MALDI MS approach for structural characterization of intact N-linked glycopeptides using de novo sequencing and protein database search. 58th ASMS abstract, 2010, TP148

1-3 生理活性ペプチド同定法 (Sequence Tag Library Search)

【研究目的】

本プロジェクトにおけるバイオマーカー探索のターゲットの1つである生理活性ペプチド(尿中ペプチドなど)は、体内で任意のアミノ酸部位で切断されているうえ、未知の翻訳後修飾 (PTM: Post Translational Modification) を受けている場合が多いため、従来のデータベース検索ソフトで同定することが困難である。生理活性ペプチドを同定するため、いくつかの方法が提案されているが[1][2][3]、データベースに登録している配列と完全に一致するものしか同定できない、信頼性の高い計算方法が確立されていない等の問題がある。

本研究では、生理活性ペプチドを同定するため、同系列のプロダクトイオン群で生成されるSequence Tagを推定し、PTMの可能性を考慮してSequence Tag Libraryを検索しペプチド配列を同定する新規手法を開発した。

【研究成果】

方法

Sequence Tag Library Search法のブロック図を図1に示す。解析手順は以下の通りである。

① Sequence Tag Libraryの構築

従来法で同定された生理活性ペプチドのスペクトルの特徴を、3残基長のSequence TagとTag構成主要ピークで構成されるTag情報として抽出し、同定されたペプチドおよびタンパク質の情報とともにデータベース (Sequence Tag Library) に登録する。

② Sequence Tagの生成

未同定スペクトル中の主要ピークをSequence Tag Library上のTag構成主要ピークと照合し、Sequence Tagを生成する。

③ Sequence Tag Libraryの検索

②で求めたSequence Tagを含むペプチドを、①で構築したSequence Tag Libraryに対して検索し、タンパク質配列情報とプリカーサ質量を参考に、PTMを考慮してペプチド配列を同定する。信頼度が高いペプチドが同定された場合、①のSequence Tag Libraryに追加登録する。

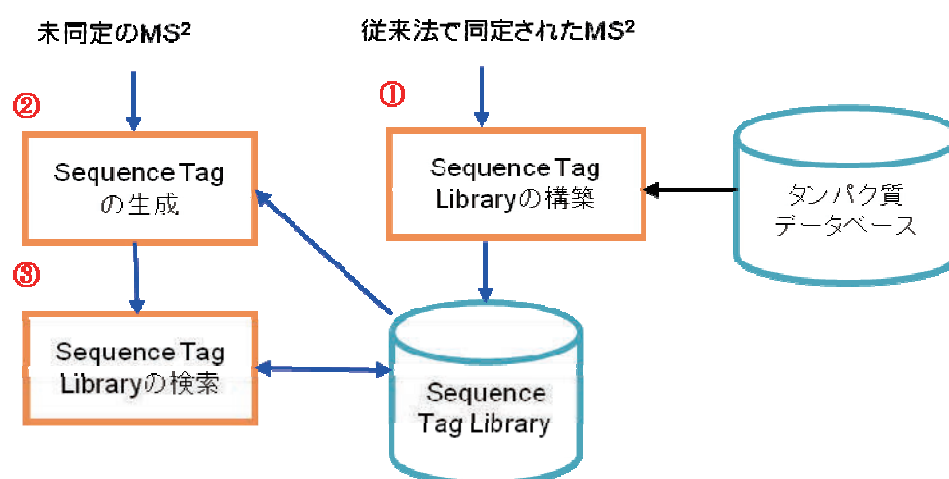


図1 Sequence Tag Library Searchソフトのブロック図

あらかじめ、同定されたペプチドのTag情報を登録したSequence Tag Libraryを構築する。解析対象のMS²からSequence Tagを求め、Sequence Tag Libraryに対して検索を行う。

結果

Sequence Tag Library Searchソフトの生理活性ペプチド検索性能を評価するため、表1に示す3種類の生理活性ペプチドを準備し、MS²およびMS³のマスペクトルを測定した。

表1 性能評価用生理活性ペプチド

No.	Sample	Peptide sequence	Precursor Mass
1	obestatin (Human)	VGIKLSGVQYQQHSQAL	1856 Da
2	obestatin (Human)	FNAPFDVGIKLSGVQYQQHSQAL	2547 Da
3	obestatin (Rat)	FNAPFDVGIKLSGAQYQQHGRAL	2517 Da

性能評価に使用した3つの生理活性ペプチド(購入品)を示す。No.1はNo.2が切断されたものであり、No.3はNo.2と比べ一部のアミノ酸配列(V→A、SQ→GR(表中の下線部))が変異している。

表1のNo.1のペプチドは、従来法(MASCOTのMS/MS Ion Search)で同定できたので、BSA(Bovine Serum Albumin)のトリプシン消化物ペプチドなどとともに、Sequence Tag Libraryに登録した。

No.2のペプチドのMS²のマスペクトルを解析した結果を図2に示す。Sequence Tag LibraryのTag情報をもとに候補ペプチドを抽出し、Sequence Tag Libraryの配列情報をもとに候補ペプチドの配列を拡張し、正しいペプチドが同定された。同定したペプチドは、Sequence Tag Libraryに追加登録した。

No.3のペプチドのMS²を解析した結果、最もScoreが高いSequence Tag を起点として、アミノ酸変異を考慮した配列の拡張を行い、正しいアミノ酸配列を同定できた(データ省略)。

Index	Tag	Score	Peptide	Ion Mass
1	QQH	54	VGIKLSGVQYQQHSQAL	1855.17
2	HSQ	52	VGIKLSGVQYQQHSQAL	1855.17
3	QHS	52	VGIKLSGVQYQQHSQAL	1855.17
4	SQA	52	VGIKLSGVQYQQHSQAL	1855.17
5	QAL	40	VGIKLSGVQYQQHSQAL	1855.17

Index	Peptide	Score	Mod Site	Mod Mass
1	FNAPFDVGIKLSGVQYQQHSQAL	263		
2	PGSGKSLLLSKWQLQKHSPNT	69		
3	VDNVEPDEEFSVGRIYQKHAKSI	66		
4	WKGAWAKLKGKAGKQHDHDKR	54		
5	DGDRQIGKTAVIDAINGKHTGIK	45		

Index	Protein Id	Description
142894	sp Q9UBU3	GHRL_HUMAN Appetite-regulating hormone OS=Homo sapiens
142898	sp A0MLS4	GHRL_PAPHA Appetite-regulating hormone OS=Papio hamadasi

Protein
MPSPGTVCSL LLLGMLWLDL AMAGSSFSLP EHQRVQQRKE SKKPPAKLQP RALAGWL RPE DGGQAEAGD ELEVRFNAPF DVGIKLSGVQ YQQHSQAL GK FLQDILWEEA KEAPADK

図2 Sequence Tag Library Search による解析結果

図1のNo.2を測定したMS²をSequence Tag Library Search法で解析した結果を示す。Sequence TagとしてQQH、HSQ、QHSなどが生成され、Sequence Tag Libraryに登録されたペプチドVGIKLSGVQYQQHSQAL(図1のNo.1)がヒットした(左上: Tag Matched from Library欄)。最もScoreが高いSequence Tag'QQH'について、Sequence Tag Library中のタンパク質配列情報をもとにブリーカー質量と一致するようにSequenceを拡張した結果、正しいペプチドFNAPFDVGIKLSGVQYQQHSQALが同定された(右上: Matched Peptide欄)。Sequence Tag Library中のタンパク質情報として、AccessionとSequence(同定されたペプチドは赤字)が表示されている(下: Protein Accession欄、Protein欄)。

【結論】

従来は、生理活性ペプチドの構造解析を行うソフトウェアが存在しないため、生体試料内のバイオマーカーを同定することが非常に困難であったが、Sequence Tag Library Searchソフトを用いて、未知のPTMを受けた生理活性ペプチドを同定できる例を示した。本ソフトは新規バイオマーカー開発時に非常に強力なツールになると期待され、今後、実データに対する適用・評価を継続する。

【成果一覧】

学会発表

- (1) Jingwen Yao、森本健太郎、村瀬雅樹、船越なつ美、福山裕子、岩本慎一、梶原茂樹、田中耕一、「生理活性ペプチド同定用シーケンスタグ検索ソフトウェアの開発」、第61回質量分析総合討論会、2013年

特許出願

- (1) ヤオジンウェン、質量分析データ解析装置及び解析方法、特願2013-178769

【参考文献】

- [1] Fälvh, M et al. Validation of Endogenous Peptide Identifications Using a Database of Tandem Mass Spectra. J. of Proteom. Res. 2008, Vol. 7 No. 7, pages 3049-3053
- [2] Menschaert, G. et al. A Hybrid, de Novo Based, Genome-Wide Database Search Approach Applied to the Sea Urchin Neuropeptidome. J. of Proteom Res. 2010, Vol. 9 No. 2, pages 990-996
- [3] Menschaert, G. et al. Spectral Clustering in Peptidomics Studies Allows Homology Searching and Modification Profiling: HomClus, a Versatile Tool. J. of Proteom. Res. 2012, Vol. 11 No.3, pages 2774-2785

1-4 次世代質量分析システムへの対応

【研究目的】

Mass++はメタボローム・プロテオーム解析用ソフトウェアとして開発され、スペクトル表示、ヒートマップ表示、ピーク検出、ピーク定量などの多様なデータ解析プラグインが準備されていたが、次世代質量分析システムの開発を進める中で、さらなる性能向上や新規のプラグイン開発が必要となった。ここでは、本研究で開発した主要なプラグインをリストアップし、従来法との性能比較を行った結果を示す。

【研究成果】

本研究で開発した新規プラグイン、および既存のプラグインに追加した新規機能の中で主要なものを表1に示す。以下、スペクトル解析用プラグインの性能評価結果について説明する。

1) ノイズ低減(Wavelet)

マススペクトル用のノイズ低減法としてWavelet変換を利用したプラグインを開発した。Wavelet変換は時間-周波数解析用の数学変換であり、通常のフィルタリングや周波数解析用のFourier変換を使った手法より、ノイズ低減効果とピーク形状維持の限界性能が原理的に高い[4]。従来法としてよく利用されるGaussianフィルタと比較した結果、より高い雑音除去効果が副作用なしに得られることが確認できた(図1)。さらに、データに応じたパラメータ調整も容易となった。

2) ベースライン除去 (BLT)

マススペクトル用のベースライン除去法として、BLT(Bottom Line Tracing)法を利用したプラグインを開発した。本手法は、分光光度計の吸光度スペクトル補正に使用されていた包絡線検出法[5]を改造し、マススペクトル中の大小のピーク底辺形状を追跡する形でベースライン変動を検出して除去するものである。ベースライン変動の大きなMALDI/ISDデータに適用し、ベースラインを精度よく除去できることを確認した(図2)。本法はパラメータ調整も容易である。

表1 主要な新規開発プラグインおよび追加機能

データ表示	
スペクトル表示形式 (Stack、3D)	単独表示、重ね表示に加え、Stack 表示(縦に並べて表示)と 3D 表示機能を追加した。
複数スペクトルの縦軸／横軸同期表示	複数のスペクトルを表示している時に、縦軸と横軸を同期して表示できる機能を追加した。
表示ウィンドウの自動整列	多くのウィンドウが表示されているときに、ワンクリックでウィンドウを整列する機能を追加した。
MS ⁿ データの表示	MS ⁿ (n≥3) のスペクトル表示機能を追加した。
マススペクトル解析	
ノイズ低減 (Wavelet)	ノイズ低減法として Wavelet 法を追加した。(☞ § 2-1-4 1-4 1))
ベースライン除去 (BLT)	ベースライン除去法として BLT 法を追加した。(☞ § 2-1-4 1-4 2))
ピーク検出 (GION、MWD)	独自のピーク検出法を開発した。(☞ § 2-1-4 1-4 3))
ピーク定量 (AB3D)	定量解析用のピーク検出・定量法を開発した。(☞ § 2-1-4 1-4 4))
LC/MALDI-MS への対応	LC/MALDI-MS 用に RT 補正機能を改良した。(☞ § 2-1-4 1-4 5))
データベース検索	
MS ⁿ 用データベース検索	MS ⁿ を利用する独自の DB 検索法を開発した。(☞ § 2-1-4 1-1)
MS ⁿ 用 <i>de novo</i> sequencing	MS ⁿ を利用する <i>de novo</i> sequencing を開発した。(☞ § 2-1-4 1-2)
生理活性ペプチド検索	生理活性ペプチドを同定するため、独自の DB 検索法を開発した。(☞ § 2-1-4 1-3)
MassBank 連携	MassBank(公的なスペクトル DB)を利用する機能を開発した。(☞ § 2-1-4 4-4)
糖鎖データベース検索	Negative 測定 of MS ² から DB 検索により糖鎖を同定する機能を開発した。
X!Tandem の組込み、MASCOT の呼び出し	フリーの DB 検索ソフト X!Tandem、商用 DB 検索ソフト MASCOT を利用する機能を組み込んだ。
検索結果の統合 (TPP-prophet の組込み)	TPP の prophet ソフト[1][2][3]を使用し、複数の検索結果を統合する機能を開発した。
検索結果のデータベース化	検索結果を DB に登録し、検索・閲覧する機能を追加した。
その他	
統計解析	差異解析用のピークマトリクスに対して、t/U 検定、ANOVA、箱ひげ図表示を行う機能を開発した。
データ入出力	次世代質量分析システム、KRATOS、Agilent のデータを表示解析する機能を追加した。(☞ § 2-1-4 2-2)(☞ § 2-1-4 3-2)
バッチ処理	スペクトル解析やデータベース検索などの複数処理を一括して行う機能を追加した。

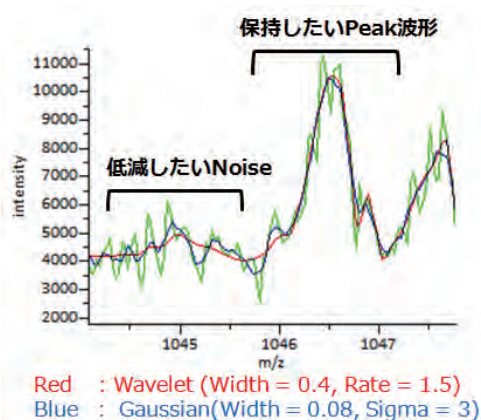


図1 ノイズ低減法の比較結果

生のスペクトル(緑)に対して、Wavelet フィルタによる処理波形(青)は、Gaussianフィルタの処理波形(赤)と比べ、ピーク形状を保持したうえで、ノイズを十分に低減している。

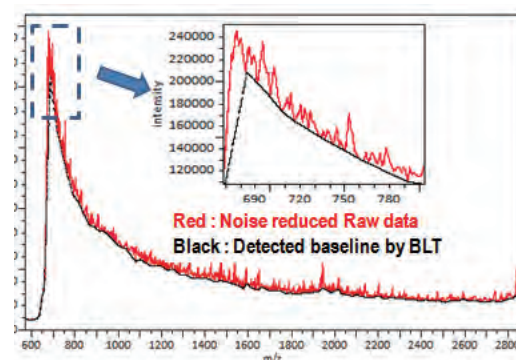


図2 ベースライン除去の結果

BLTが抽出したベースライン(黒)は、生スペクトル(赤)の底面部を成す信号波形に追従しており、特に変動が大きい領域においても正しくベースラインを追跡している。

3) 定性解析用ピーク検出(MWD)

定性解析用のピーク検出アルゴリズムMWD(Multi Window Detection)法を開発した。MWD法には以下の特徴がある。

- 1) 質量範囲ごとに、孤立点検出演算によって適切なピーク検出レベルを定めるため、弱いイオンピークの検出漏れやと強いノイズピークの誤検出を防ぐことができる。
- 2) 解析対象がタンパク質、ペプチドであることを利用したモデルデータを使用することにより、高質量範囲においても同位体ピーククラスタを正確に判別できる。

既知ペプチドのMALDI計測データに同手法を適用し、市販のピーク検出ソフトMascot Distiller(Matrix Science)と比較した結果を表1に示す。13回中12回の実験でMWD法のほうが検索結果の信頼度(スコア)が高かった。

表1 MASCOTによる検索スコアの比較結果

Exp.	# 1	# 2	# 3	# 4	# 5	# 6	# 7	# 8	# 9	# 10	# 11	# 12	# 13
MWD	1072	955	890	701	440	650	823	1121	590	503	325	244	362
Distiller	951	689	473	338	234	378	438	1224	338	378	175	156	82

BSAを酵素消化した試料をLC/MALDI-MSで分析し、MASCOT検索で得られたBSAの同定スコアを示す。試料濃度や測定条件を変えて13回の実験を行い、MWD法および市販のピーク検出ソフト(Distiller)を使用した場合の結果を比較した。MWD法のスコアが高かった結果を赤太字で示す。13回中12回の実験でMWD法を利用したほうが検索結果のスコアが高かった。

4) ピーク検出・ピーク定量(AB3D)

バイオマーカー探索において患者群と健常者群のMSデータの差異を解析するためには、ピーク定量が非常に重要な機能である。独自のピーク検出および定量法AB3D(Algorithms for Biomarker Discovery in Diagnostics and Drug Development)をエーザイにて開発しMass++プラグインとして実装した。AB3D法はLC-MSでのラベルフリーピーク検出および定量用に開発され、以下の手順でピークを検出する。

① ピーク検出

指定したRetention Time(RT)許容範囲と m/z 許容範囲でXIC(extracted ion chromatogram)を作成し、ピーク検出を行う。また、XICの合計値を定量値として用いる。

② ピークフィルタリング

検出したピークから、A) FWHM(Full Width at Half Maximum)、B) RT span、C)凹凸の数、D) データポイント数、E) ベースラインの高さ、F) S/N比を利用して、適切なピークを選択する。

③ 同位体ピークのクラスタリング

検出したピークから、 $(1/z)$ Daずつ離れたピーク群を識別し、 z 値の同位体ピークとしてクラスタリングする。

従来ソフト(XCMS[6]、MZmine2[7])と正答率を比較した結果を図3に示す。図3より、AB3Dは従来ソフトと比べピーク検出の信頼度が高いことがわかる。

さらに、AB3Dの計算は従来ソフトより高速であることを確認した(XCMSの1~3倍、MZmine2の6~9.5倍)。

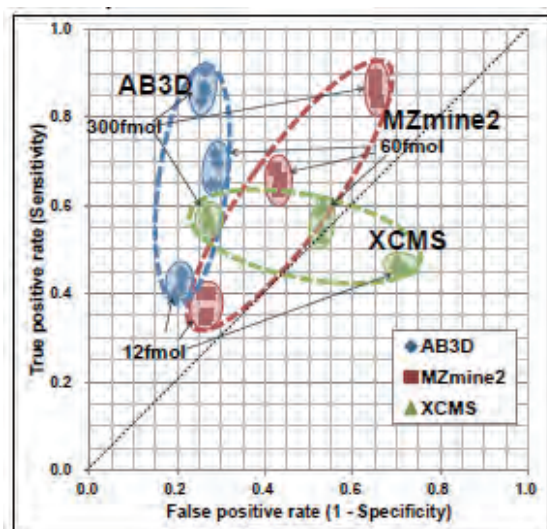


図3 性能比較のためのROCプロット

濃度 (300 fmol、60 fmol、12 fmol)を変えた12個の試料において、各ソフトで検出したピークの正答率を、横軸 False Positive Rate、縦軸 True Positive Rateとしてプロットした (ROC: Receiver operating characteristic)。図の左上にあるほうがピークの検出性能が良いことを意味し、AB3Dは従来法より性能が高いことがわかる。

5) LC/MALDI用RTアライメント

バイオマーカー探索のためサンプル間(患者群と健常者群)の差異を解析する場合、測定データ間でLCの Retention Time (RT)のずれを補正する必要がある。このため、Mass++にはTIC (total ion chromatogram) 波形の類似性を利用した動的計画法によるRT補正(RTアライメント)が実装されている[8]。しかしながら、LC/MALDI-MSデータにおいてはスペクトラム強度の再現性が低いためTIC波形の類似性も低く、RTアライメントが精度良く行われれないという問題があった。

本問題を解決するため、TIC波形の類似性を向上させる下記スケーリング付きTIC計算法を考案した。

$$TIC'(t) = \sum_{i=1}^N Int_i \frac{a(t)}{v(t)}$$

ここで、N: RT = tのスペクトル中のピーク数

.i: ピーク番号

Int_i : i 番目ピークの強度

$a(t)$: RT=tのピーク強度の平均

$v(t)$: RT=tのピーク強度の分散

LC/MALDIのマスペクトルに対して、等量の既知物質(内部標準)による正規化処理を施した後、本手法を適用しRTアライメントを行った結果を図4に示す。従来法(スケーリング無しのTICでRTアライメントを実行)に対して、本手法を適用するとTIC波形の類似性が向上し、RTのずれが良好に補正できたことが分かる。

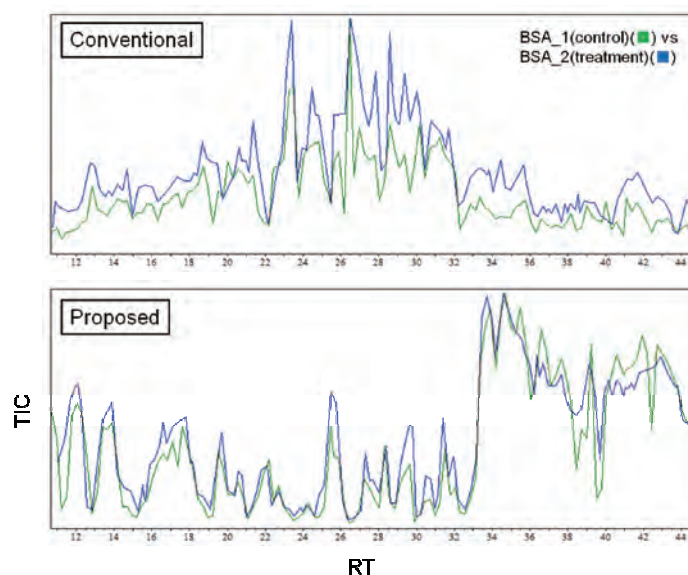


図4 RTアラメントの結果(従来法と提案手法の比較)

RTアラメントでは2つの試料のTICが一致するようにRTを補正する。すなわち、ピークの山や谷が2つの試料間で一致しているほうが好ましい。提案手法はスケーリングにより形状を変えたTICに対してRTの補正を行う。従来法に比べ、提案手法ではRT補正により2つのTICが良く一致している。

【結論】

本研究では多くの新規プラグインを開発したが、それらは既存のソフトウェアと比べ、優れた性能を有する。これらのプラグインは次世代質量分析システムのデータ解析に利用されるとともに、一般に公開しているので、今後、質量分析データの高精度な解析に貢献すると期待される。

【成果一覧】

学会発表

- (1) Jingwen Yao, Satoshi Tanaka, Howell Parry, Shin-ichi Utsunomiya, Shigeki Kajihara, Koichi Tanaka, Tsuyoshi Tabata. "Peak Detection Method in Mass++ – A Protein Analysis Platform in Mass Spectrometry", The 6th AOHUPO Conference, 2012
- (2) Ken Aoshima, Koikegami Shigeru, Fukuda Mitsuru, Takahashi Kentaro, Matsuura Kentaro, Watanabe Hideki, Sato Yoshiaki, Uehara Taisuke, Kimura Takayuki, Nakamura Tatsuji, Parry Howell, Tanaka Satoshi, Utsunomiya Shin-ichi, Kajihara Shigeki, Tanaka Koichi, Oda Yoshiya. "Ab initio peak identification for SRM/MS/MS data", IMSC2012, 2012
- (3) Yuichiro Fujita, Natsumi Funakoshi, Yoshihiro Yamada, Satoshi Tanaka, Shin-ichirou Kawabata, Shinichi Iwamoto, Shin-ichi Utsunomiya, Shigeki Kajihara, Tsuyoshi Tabata, Ken Aoshima, Yoshiya Oda, Koichi Tanaka. "Novel preprocessing method to align retention time of LC-MALDI and new implemented functions in Mass++ for differential analysis", 61st ASMS Conference on Mass Spectrometry and Allied Topics, 2013
- (4) 宇都宮眞一、姚精文、藤田雄一郎、田中聡、山田賢志、梶原茂樹、川畑慎一郎、田畑剛、高橋健太郎、青島健、小田吉哉、田中耕一、「Mass++: 質量分析のピーク情報検出技術の開発」, 第61回質量分析総合討論会, 2013年
- (5) Ken Aoshima, Masayuki Ikawa, Takayuki Kimura, Mitsuru Fukuda, Kentaro Takahashi, Satoshi Tanaka, Yuichiro Fujita, Akiyasu C. Yoshizawa, Shin-ichi Utsunomiya, Shigeki Kajihara, Koichi

Tanaka, Yoshiya Oda. "A simple and fast label-free quantitation algorithm for LC-MS", HUPO2013, 2013

特許出願

- (1) ヤオジンウェン、質量分析におけるピーク検出方法及びそのシステム、特願2012-243103
- (2) 森本健太郎、糖鎖構造解析方法及び装置、特願2013-040523
- (3) ヤオジンウェン、質量分析データ処理方法及び質量分析データ処理装置、特願2013-078841
- (4) 宇都宮真一、グラフ表示処理装置、特願2012-202325
- (5) 田中聡、ピーク検出装置、特願2013-088530
- (6) 宇都宮真一、グラフ表示処理装置、特願2013-093280
- (7) 藤田雄一郎、梶原茂樹、クロマトグラフ質量分析データ処理装置、特願2013-078428
- (8) ヤオジンウェン、質量分析におけるピーク検出方法及びそのシステム、特願2013-092951
- (9) Yao Jingwen, Peak detection in mass spectrometry, US 13/872466
- (10) Yao Jingwen, Accurate Ion Peaks in Mass Spectrometry, US 61/639757
- (11) Shinichi Utsunomiya, Graph Display Processing System, US 14/024914

【参考文献】

- [1] A. Keller et al., Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 2002, Vol. 74, pages 5383-5392
- [2] D. Shteynberg et al., iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Mol Cell Proteomics, 2011, Vol. 10, M111.007690
- [3] A. Nesvizhskii et al., A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 2003, Vol. 75, pages 4646-4658
- [4] S.Utsunomiya, Development of Signal and Image Processing Techniques Using Wavelet Transform, SHIMADZU REVIEW, 2000, Vol.57, pages 133-139
- [5] 宇都宮真一、他、FTIRの水蒸気補正を行う新手法の開発、日本分析化学会年会講演要旨集、2002, Vol.51、page 19
- [6] Pluskal et al., MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometrybased molecular profile data. BMC Bioinformatics 2010, Vol.11, page 395
- [7] C.A. Smith et al., XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. Analytical Chemistry, 2006, Vol. 78 No.3, pages 779-787
- [8] K. Aoshima et. al., AB3D: A suite of Algorithms for Biomarker Discovery in Diagnostics and Drug Development using LC-MS. 56th ASMS, 2008, TP-659

2 Mass++プラットフォームの改良と普及

質量分析用表示解析ソフトウェアMass++は、プロテオーム、メタボローム解析をターゲットとして、2006年にCRESTプロジェクトにて開発が開始された。その特長は以下の3点である。

- 1) 質量分析装置ベンダー複数社、複数機種の分析データの入出力対応
- 2) 生命科学分野に的を絞った解析機能と操作性
- 3) ユーザーによる独自開発機能追加を可能とする拡張性

Mass++はプラグイン構造のソフトウェアなので、既に開発済みのスペクトル表示などの基本機能のうえに、ユーザーが開発した独自の機能をプラグイン形式で容易に追加できる。このため、本研究では、次世代質量分析システムで必要となる表示解析機能を追加するためのプラットフォームとしてMass++を利用した。

ここでは、Mass++に関する以下の項目について報告する。

- Mass++のプラットフォームとしての特徴を活用する仕組みの整備によるプログラム開発効率の向上
- 開発したMass++を一般公開した結果と他ソフトウェアに対する特徴

2-1 開発プラットフォームとしての整備

【研究目的】

創薬・診断分野の質量分析装置ユーザーは、課題ごとに必要となる機能が異なり、ベンダーから供給される装置付属ソフトウェアでは対応できない場合がある。Mass++はこのような課題を解決するために開発された質量分析用表示解析ソフトウェアであり、最大の特徴は機能の着脱が容易なプラグイン構造を採用している点にある。基本機能がプラットフォームとして提供され、ユーザーは研究課題特有の応用機能だけをプラグインとして開発すればよいので、開発時間と費用の大幅削減が期待できる。しかし、(1)プラグイン開発のプログラミング面での煩雑さ、(2)ソフトウェア肥大化による操作簡便性と保守性の低下、(3)ライセンスや保守に対する懸念、がユーザーによるプラグイン開発に際しての問題点だった。これらを解決して開発プラットフォームとしての基盤を確立し、プロジェクト内外での開発を促進するための対策を行った。

【研究成果】

1) プラグイン開発効率の向上

- プログラマー開発支援

Mass++プラグイン開発のための独特の作法を覚え開発環境を整えるのが、Mass++を初めて使うソフトウェアプログラマーにとっては難しく、プラグイン開発の起ち上げの阻害要因になっていた。そこで、主要開発環境であるVisual C++およびC#.Net を使用した開発者のプログラミング作業負担を軽減するため、開発支援用Wizradを作成した。図1に作業フローを示す。これにより、通常のプログラム開発と異なる知識を要する部分はWizardが代替することになり、プログラマーは通常のアプリケーション開発と同等の負担で、プラグイン開発を実施することができるようになった。

- プラグイン開発者層の拡大

プラグインの開発者層をプログラマー以外の生物・化学・医学者に拡大するため、Script言語で簡易的にプログラミングができるScript Console機能を開発した。言語は、Microsoftの.Net Frameworkの豊富な機能を利用可能なIronPythonを採用した。これらの開発の結果、本研究で開発したMass++のプラグインは122個となった。また、一般公開で提供されたMass++を使い、本プロジェクト外のユーザーが応用機能を開発した事例が2013年日本質量分析総合討論会でポスター発表された[1]。

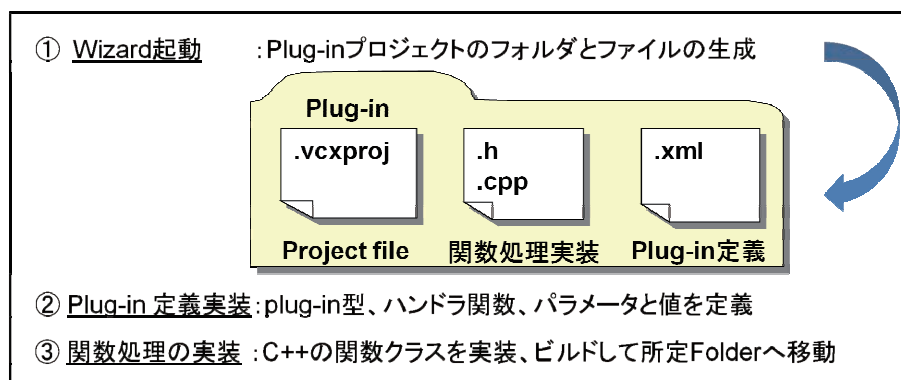


図1 開発支援用Wizardを使ったプラグイン作成手順

Visual C++やC#.Netを使用してMass++プラグインのプログラムを開発する場合、Wizardにより必要なフォルダ・ファイルが自動的に作成される。その後は、通常のプログラム開発と同様に、プラグインの定義や処理を実装できるので、プログラム開発者は特別な知識を必要としない。

2) プラグイン構成の再構築

一方で、Mass++の機能が増えすぎたため、これらをA)プラットフォーム部とB)本研究で開発した特定プラグイン部に分離して再構築した(図2)。データ読み込みや波形表示などの基盤としてのプラットフォームの上に、ピーク検出、DB検索、定量解析などの応用機能プラグインがあり、ユーザーはIdentification、Quantitation、DataProcessing、MassBankなどの7種のプラグインセットを追加・削除できる形態とした。

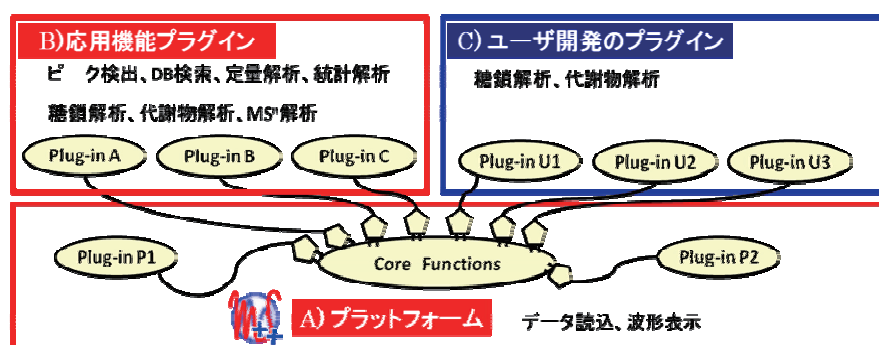


図2 再構築後のMass++プラグインの構成

プラットフォーム部と応用機能プラグインを分離し整理した。ユーザーは解析目的に応じて、7種類のプラグインセットを追加・削除できる。また、ユーザー開発のプラグインをプラットフォーム部に結合することも可能である。

3) ライセンスと開発保守体制

Mass++をフリーウェアとして一般公開するに際して、開発保守体制とソフトウェア使用ライセンスを定めた。以下にその基本ポリシーを示す。

- エーザイと島津製作所田中最先端研究所が共同で開発・保守に当たる。
- ユーザー(装置ベンダーを含む)による機能開発を促進するため、ユーザー開発プラグイン(図2のC)には販売、頒布、貸与、譲渡の権利を認める。
- 一般公開先のユーザーからも要望情報を得て、開発、保守に活用する。
- 継続的に開発・保守体制を維持する予定である。

【結論】

上記の通り、(1)プラグイン開発効率の向上、(2)規模肥大化と保守性低下の是正、(3)ライセンスと保守体制の明確化、を実施することで、Mass++の特徴である機能拡張の容易性が強化された。

【成果一覧】

学会発表

- (1) Howell E Parry, Satoshi Tanaka, Tsuyoshi Tabata, Ken Aoshima, Yoshiya Oda, Shinichi Utsunomiya, Shigeki Kajihara, Koichi Tanaka. "Rapid development of functional extensions for mass spectrometry using freeware Mass++", IMSC2012, S2012
- (2) 宇都宮眞一、田中聡、村瀬雅樹、田畑剛、青島健、小田吉哉、梶原茂樹、田中耕一、「Mass++ の開発:質量分析解析用"Platform Viewer"」, 日本医用マススペクトル学会第37回年会, 2012年
- (3) Shin-ichi Utsunomiya, Satoshi Tanaka, Masaki Murase, Shigeki Kajihara, Tsuyoshi Tabata, Ken Aoshima, Yoshiya Oda, Koichi Tanaka. "A platform for mass spectrometry to construct suitable software to achieve user's own purposes", 61st ASMS Conference on Mass Spectrometry and Allied Topics, 2013

【参考文献】

- [1] 山本敦史、他、高分解能質量分析を用いたジチオカーバメート農薬の反応生成物の同定, 第61回質量分析総合討論会講演要旨集, 2013年, page 156

2-2 Mass++の一般公開とその特徴

【研究目的】

Mass++は、CREST プロジェクト(2005～2010年)でエーザイ・小田吉哉氏が担当するテーマ「定量的メタボロミクスとプロテオミクスの融合」の中で開発が開始された。このプロジェクトの中でもそのコンセプトを引き継いだ。

本研究では、国内・国外にユーザーの数を増やし、MS表示解析用標準ソフトウェアとなることを目指した。

【研究成果】

1) マルチベンダー対応の推進

取扱い可能なデータとして、表1に示す通り、複数ベンダーのフォーマットに対応した。なお、島津製ソフトウェアへの対応は島津製作所が担当し、それ以外のベンダーへの対応はエーザイが担当した。

表1 Mass++ 対応フォーマット

ベンダー	ソフトウェア	拡張子	読み込み	書き込み
Shimadzu	LCMSsolution	.lcd	✓	
Shimadzu	GCMSsolution	.qgd	✓	
KRATOS	Launchpad	.run	✓	
AB Sciex	Analyst, AnalystQS	.wiff	✓	
Thermo	Xcalibur	.raw	✓	
Waters	MassLynx	.raw	✓	
Agilent ^(*)	Mass Hunter	.d	✓	
共通フォーマット	(mzML)	.mzML	✓	✓
	(mzXML)	.mzXML	✓	✓
	(AIA)	.CDF	✓	
独自フォーマット	Mass++	.msb	✓	✓

Mass++から読み込み・書き込みが可能なファイルフォーマットを示す。(*)ただし、Agilentは2014年3月に対応予定。

2) ユーザーからの要望対応

当初、本研究ではMass++の公開準備までを行なう予定であったが、ユーザーの要望・ニーズを収集してソフトウェアの改良に反映させるため、2011年9月に Mass++ ver 2. として公開を開始した。ユーザーの要望はwebサイト上で要望投稿コーナーを設置し、Google GroupsのMass++のユーザーグループを運営した他、学会でのポスター発表を中心に積極的に発表を行ない ユーザーから意見を聞き取った。2012年10月までは共同研究先の京大薬学部、それ以降はコア・サポートユーザー(京都大学薬学部・石濱泰教授ら6機関) から要望を収集し優先的に対応した。

3) 他フリーウェアと比較した特徴

質量分析分野の他フリーウェアとの比較を表2に示す。

本比較結果から、今後、標準ソフトウェアとして認知されるため、下記のMass++の強みをアピールする方針である。

- プラグイン構造による機能拡張、機能制限容易性
装置付属ソフトウェアには実装が困難な要望をユーザー自身が実装したりVendorが特注品として開発したりする場合に、開発期間と費用を圧縮できる効果がある。
- 質量分析装置ベンダーとユーザー連合による網羅的Multi-Vendor対応
計測データが大容量になると汎用書式へのデータ変換に膨大な処理時間がかかるため、Multi-vender format直接読み込みが有利になる。
- 糖鎖解析、PTM解析アプリケーションと代謝物解析DB連携
MALDI対応、定性解析(DB検索、代謝物DB連携、糖鎖解析)など、特徴的なアプリケーション群、優れた性能を示す基本解析は差別化機能として使用の動機づけとなる。

表2 質量分析用途フリーウェア比較

	TPP[1]	ProteoWizard [2]	OpenMS[3]	MaxQuant[4]	Mass++
開発機関	Seattle Proteome Center	Spielberg Family Center for Applied Proteomics	German Universities. in Tübingen, Berlin and Saarbrücken	MaxPlank	島津製作所 & エーザイ
ライセンス	LGPL2.1	Apache v2	LGPL2.1	独自(配布 OK、課金不可)	独自(本体配布不可、拡張部販売可)
ソースコード公開	有	有	有	無	無
開発ツール	無	有 (データアクセスのみ)	有 (ソフトウェア拡張可)	無	有 (ソフトウェア拡張可、スクリプト対応)
対応 OS	Windows UNIX MacOS	Windows UNIX MacOS	Windows UNIX MacOS	Windows	Windows UNIX (機能限定)
対応データ	mzXML mzML	mzXML mzML AB Sciex Thermo Waters Agilent Bruker	mzXML mzML	mzXML Thermo	mzXML mzML AIA Shimadzu KRATOS AB Sciex Thermo Waters Agilent
対応同定エンジン	Mascot X! Tandem SEQUEST	—	Mascot X! Tandem OMSSA Inspect MyriMatch	Andromeda (独自)	Mascot X! Tandem MassBank MS ⁿ (独自) 糖鎖 DB (独自)
その他の機能 ・Non label 定量 ・Label 定量 ・スムージング ・Baseline 補正 ・Background 補正 ・Recalibration ・統計解析	✓ ✓ ✓ ✓ ✓ ✓	✓	✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓
備考	パイプラインのさきがけとなったソフトウェア。サーバにインストールし、ユーザーがブラウザを介して操作する形を取っている。	複数ベンダー対応ソフトとして定評のあるソフトウェア。mzXML/mzML 変換ツールと簡単なビューアを提供している。	多機能なパイプライン ソフトウェア。一括解析のツールがフローチャートを編集する様な形で使いやすい。	定量用ソフトとして定評のあるソフトウェア。ほぼ自動で複数サンプルに対し定量・同定を行なう。	プラグイン構造を特徴とするソフトウェア。カスタマイズが容易。MALDI-MS データ解析、糖鎖解析、MassBank 連携といった機能が特徴。

【結論】

Mass++を一般公開し、学会発表、web上でのアナウンス、質問対応等を行ってきた結果、表3の反応を得ており、対応フォーマットと機能の豊富さで世界トップクラスのフリーウェアとなっている。

表3 Mass++一般公開に対する反応

項目	人数
サイト閲覧者数	2,905 人 (日本語サイト: 1,809 人, 英語サイト: 1,096 人)
ユニークユーザー数	1,092 人 (日本語サイト: 397 人, 英語サイト: 695 人)

サイト閲覧者数は、ダウンロードサイト閲覧者のIPアドレスの個数をカウントした重複のないユーザー数を示す。ユニークユーザー数は、実際にソフトウェアをダウンロードしたユーザーの数であり、登録されたE-mailアドレスをカウントした重複のないユーザー数を示す。上記は2011年9月12日(公開開始日)から2013年11月12日までの集計結果である。

【成果一覧】

学会発表

- (1) Howell Parry, 田中 聡, 田畑 剛, 青島 健, 小田吉哉, 二瓶義人, 西岡孝明, 宇都宮真一, 梶原茂樹, 田中耕一, “Mass++ 2.0.0: 質量分析解析フリーウェアの新バージョン”, 第59回質量分析総合討論会, 2011年
- (2) Yuichiro Fujita, Natsumi Funakoshi, Yoshihiro Yamada, Satoshi Tanaka, Shin-ichirou Kawabata, Shinichi Iwamoto, Shin-ichi Utsunomiya, Shigeki Kajihara, Tsuyoshi Tabata, Ken Aoshima, Yoshiya Oda, Koichi Tanaka. “Novel preprocessing method to align retention time of LC-MALDI and new implemented functions in Mass++ for differential analysis”, 61st ASMS Conference on Mass Spectrometry and Allied Topics, 2013
- (3) 藤田雄一郎、船越なつ美、田中聡、山田賢志、岩本慎一、宇都宮真一、梶原茂樹、青島健、小田吉哉、田中耕一、「Mass++: 差異解析のための統計・多変量解析機能」, 第61回質量分析総合討論会, 2013年
- (4) 田中聡、藤田雄一郎、吉沢明康、福田充、宇都宮真一、梶原茂樹、青島健、小田吉哉、田中耕一、「Mass++: 質量分析データの一括解析」, 第61回質量分析総合討論会, 2013年
- (5) Satoshi Tanaka, Yuichiro Fujita, Akiyasu C. Yoshizawa, Shin-ichi Utsunomiya, Shigeki Kajihara, Mitsuru Fukuda, Masayuki Ikawa, Kentaro Takahashi, Tsuyoshi Tabata, Ken Aoshima, Yoshiya Oda, Koichi Tanaka. “Identification and label-free quantitation of mass spectrometric data via freely available plug-in software, Mass++”, HUPO2013, 2013
- (6) 田中聡, “質量分析を取り巻くソフトウェアの現状 –Made in Japanを世界へ–”, 第7回 明日の質量分析を創る若手討論会, 2013年

【参考文献】

- [1] Andrew Keller et al. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol. Syst. Biol., 2005, Vol. 1, page 0017
- [2] Darren Kessner et al. Open source software for rapid proteomics tools development. Bioinformatics, 2008, Vol. 24, pages 2534-2536
- [3] Marc Sturm et al. OpenMS – An open-source software framework for mass spectrometry. BMC Bioinformatics, 2008, Vol. 9, page 163
- [4] Jürgen Cox et al. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification: Nature Biotechnology, 2008, Vol. 26, pages 1367-1372

3 Mass++を利用したアプリケーションソフトの開発

近年、高性能化したLCおよびMS装置を利用し、各試料に含まれる物質を網羅的に測定し、健常者・患者の試料間での差異解析を行い、バイオマーカー候補を探索することが可能になった。このバイオマーカー候補物質の同定を行う場合、MALDIのMSⁿ (n ≥ 2) 測定は有効なツールであるが、マニュアルで測定するにはオペレータの負担が大きいとともに、オペレータの習熟度が同定できるかどうかに影響する。これらの問題を解決するため、MSⁿ⁻¹による構造解析結果からプリカーサ選択やMS測定の条件を決定しながら、自動的にMSⁿ測定を行うintelligentな自動分析システムが必要である。

ここでは、Mass++の解析機能を利用して開発した下記の2タイプの自動分析ソフトウェア (LDIA: Logical Data-Influenced Acquisition) について報告する。

- 試料中の物質の網羅的な自動分析 (LDIA for Pep)
試料に含まれるペプチドおよびその翻訳後修飾を網羅的に自動分析する。
- 目的物質の詳細解析を行う自動分析 (LDIA for GP)
糖ペプチドの糖鎖及びペプチド配列を自動分析する。

また、装置がハイスループット化し大量のデータ(ビッグデータ)を解析する必要が生じ、さらに解析方法が高度化していくことへの対応として、Mass++の解析機能をクラウドコンピュータで利用するソフトウェア (Job Request & Proteo Analysis)を開発した。

3-1 試料中の物質の網羅的な自動分析法 (LDIA for Pep)

【研究目的】

生体試料に含まれるペプチドを網羅的に分析することは、プロテオミクスにおいて重要な役割を果たす。MALDIは、イオン化法の相違によりESI-MSでは測定できないペプチドを測定できる。また、MALDIではESIと異なり、試料を繰り返し測定できる。従来のMALDI-MS用自動分析ソフトは、これらの利点を十分に生かすことができていない。

我々は、MALDIの特徴を活かし、ペプチド及びその翻訳後修飾を網羅的に自動分析するソフトウェア LDIA for Pepを開発した。

【研究成果】

方法

従来の自動分析ソフトウェアは、ユーザーが指定した閾値以上のMS¹ピークをMS²用プリカーサとして選択する。これに対し、LDIA for Pepソフトは、同定ペプチド数を最大化するため、線形計画法を用いて最適なプリカーサ選択とスペクトル積算回数を決定する。このため、Mass++の解析機能を利用するとともに、プリカーサ選択プラグインを追加した。

以下、図1のフローチャートに従い、分析手順を説明する。

① 同定確率モデル作成

あるピークをプリカーサとしてMS²を測定した時に、そのスペクトルからペプチドが同定できる確率は、プリカーサピークのS/N比から推定できる。このため、モデル作成用試料に対するMS¹、MS²測定・同定結果から、同定確率推定用のモデルを作成する。作成した同定確率モデルの例を図2に示す。

② MS¹測定

解析対象の試料をLCにて分離し試料プレート上のウェルにRT (Retention Time) 順に滴下した後、全ウェルのMS¹を測定する。各ウェルのMS¹のマスペクトルをRT順に並べて、スペクトルの強度を濃淡で表示すると、図3 a)のようなヒートマップが得られる。

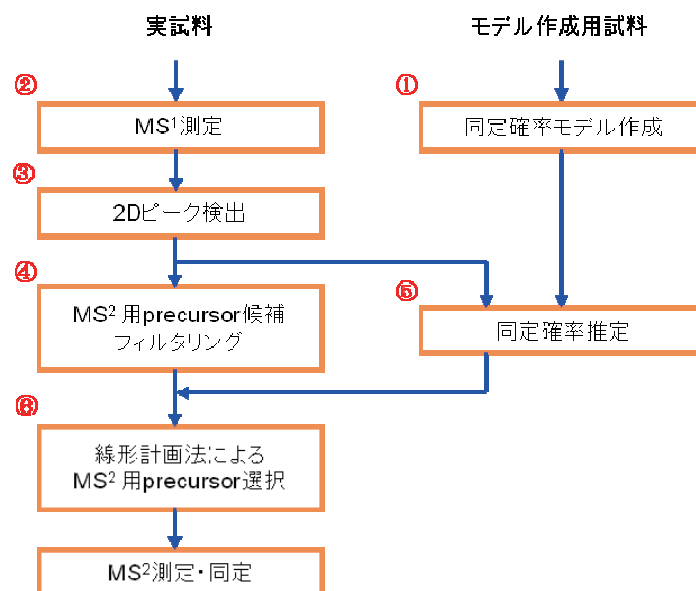


図1 自動分析ソフトLDIA for Pepのフローチャート

プリカーサピークのSN比から同定確率を求めるモデルを事前に作成しておき、実試料のMS¹を測定した時に、同定確率が最大となるMS²用プリカーサを線形計画法で求めてMS²の測定を行う。

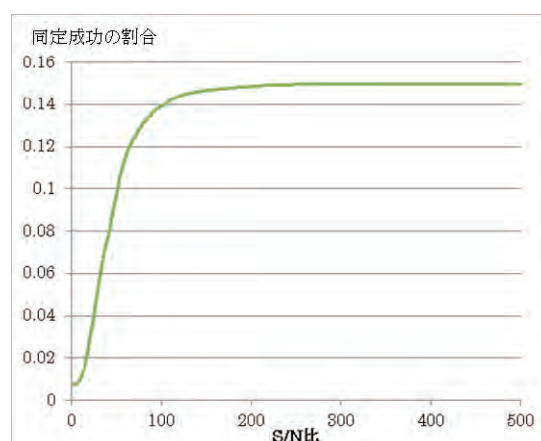


図2 プリカーサピークのS/N比と同定に成功する割合の関係

横軸はプリカーサピークのMS¹におけるSN比、縦軸はMS²によるデータベース検索で同定に成功する割合を示す。この関係から同定確率モデルを作り、MS¹のピークのS/N比をもとに同定できる確率が高いものをプリカーサとして選択する。

③ 2D ピーク検出

元の試料に含まれていたペプチドを、図3 b)のように、ヒートマップ上の2次元(2D)ピークとして検出し、その位置(ウェル・m/z)をMS²用プリカーサ候補とする。

④ MS²用プリカーサ候補フィルタリング

③で選択したMS²用プリカーサ候補から、ペプチド同定に不適切なもの(夾雑物由来のピーク、特定の翻訳後修飾に相当する質量差を持つペアピークの一方など)を除外する。

⑤ 同定確率推定

①で作成した同定確率モデルを使用して、各プリカーサ候補に対して、そのS/N比から同定確率を推定する。

⑥ 線形計画法によるMS²用プリカーサ選択

MS²用プリカーサ候補の中から、⑤で設定した同定確率の和が最大になるように、線形計画法を利用

して実際に測定するプリカーサを、下記の制約を考慮して選択する。

- 各ウェルに対する総積算回数は指定値以下：同一ウェルで多数回のMS²測定を行うと試料の枯渇が生じるため
- 全ウェルを通じた総積算回数は、指定値以下：図2に示す通り、積算回数を増やせば(S/N比を上げれば)同定確率が増えるが、測定時間も増えるため

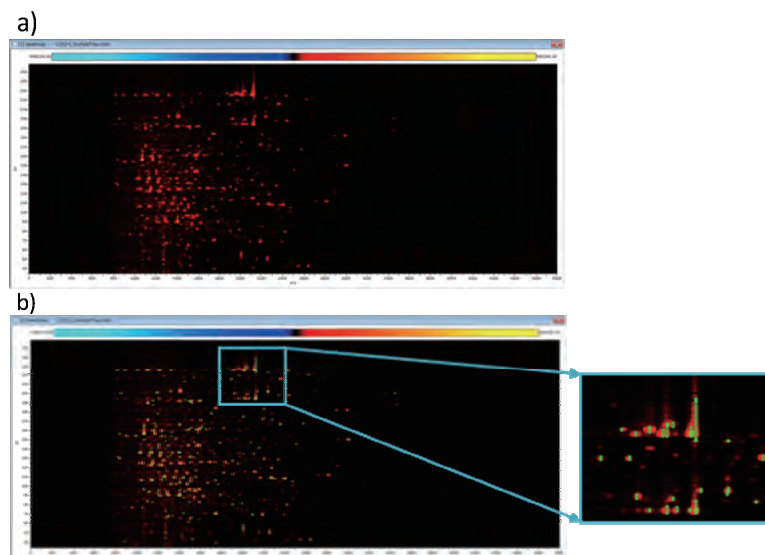


図3 ヒートマップとその上で検出された2Dピーク

a) 測定データのヒートマップを示す。横軸 m/z 、縦軸RTでカラーは信号強度を表わす。b) 2Dピーク検出結果を示す。 m/z とRTの連続性を考慮し2次的に検出したピークをヒートマップ上に緑色で表示した。

結果

エーザイにて、HeLa細胞の抽出物をLys-CとTrypsinで酵素消化しIMAC (immobilized metal affinity chromatography)によりリン酸化ペプチドを濃縮した。本試料をESIおよびLC/MALDIで網羅的に分析し、同定したペプチドの数を図4に示す。従来ソフトに比べ、開発したソフトでは約1/3の測定で約1.6倍のペプチドを同定できた。また、MALDI(LDIA for Pep)分析では、ESIによる分析(エーザイで実施)で同定されなかったペプチドが同定できた。

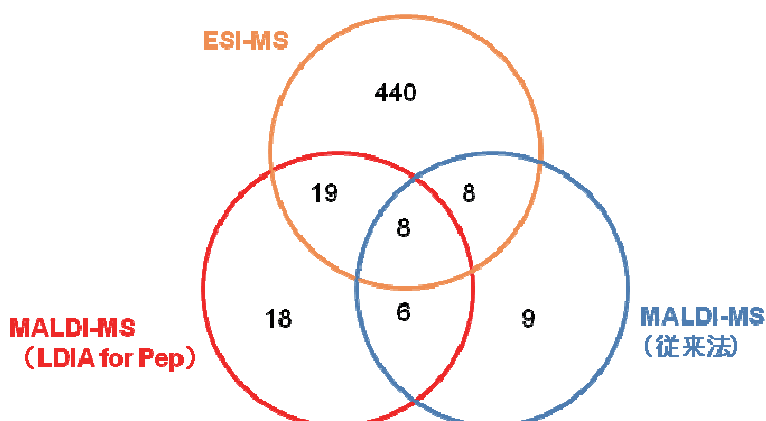


図4 MALDI-MS(LDIA、従来法)、ESI-MSによるペプチド同定数のベン図

従来のMALDI-MS用自動分析ソフトを利用すると、1,593回のMS²を測定し31(=8+8+6+9)個のペプチドを同定した。これに対し、本プロジェクトで開発したMALDI-MS(LDIA for Pep)ソフトを利用すると、591回のMS²を測定し51(=19+8+6+18)個のペプチドを同定した。また、MALDI-MS(LDIA for Pep)分析では、ESI-MSによる分析で同定されなかったペプチドが24(=18+6)個同定できた。

【結論】

同じ試料を繰り返し測定できるというMALDI法の特徴を生かし、自動分析ソフトLDIA for Pepでは、 MS^{n-1} を全て解析した後に MS^n 測定の最適条件を決定する。従って、本ソフトを利用し、ESI-MSで同定できないペプチドを少ない測定回数で見つけることが可能になった。

【成果一覧】

学会発表

- (1) Yoshihiro Yamada, Yuko Fukuyama, Natsumi Funakoshi, Hidenori Takahashi, Masaki Murase, Shigeki Kajihara, Koichi Tanaka. "Data dependent acquisition with peptide identification probability estimation during comprehensive analysis", 60th ASMS Conference on Mass Spectrometry and Allied Topics, 2012
- (2) 山田賢志、船越なつ美、村瀬雅樹、梶原茂樹、木村剛之、青島健、小田吉哉、田中耕一、「MALDI-MSによるリン酸化ペプチドの網羅的解析」, 第61回質量分析総合討論会, 2013年

特許出願

- (1) 山田賢志、質量分析装置、特願2011-227620
- (2) 山田賢志、質量分析データ処理方法、質量分析データ処理装置、及び質量分析装置、特願2011-244600
- (3) 山田賢志、物質同定方法及び該方法に用いられる質量分析システム、特願2012-064806、
- (4) 山田賢志、質量分析データ処理方法及び該方法を用いた質量分析装置、特願2013-178768
- (5) 山田賢志、梶原茂樹、物質同定方法及び該方法を用いる質量分析装置、特願2013-178770
- (6) Yoshihiro Yamada, Method and System for Processing Mass Spectrometry Data and Mass Spectrometer, US 13/670396
- (7) Yoshihiro Yamada, Substance Identification Method and Mass Spectrometry System used for the Same Method, US13/848324

3-2 目的物質の詳細解析を行う自動分析法(LDIA for GP)

【研究目的】

糖タンパク質の糖鎖構造・定量情報はバイオマーカー探索に有用であるが、糖鎖修飾によるペプチドのイオン化効率の低下や糖鎖修飾の不均一性のために、前処理として化学修飾や糖鎖分解を必要とする[1][2]。次世代質量分析装置の特徴の1つであるMALDI- MS^n を使用する場合、前処理なしで、標的とする糖ペプチドを高感度で分析可能であるが、 MS^n 用プリカーサ選択や MS^n データを用いた糖ペプチド構造解析に手間と習熟を要していた。

そこで、簡便・迅速な糖ペプチド構造解析の実現を目的として、*N*結合型糖ペプチドの全自動構造解析システムLDIA for GPソフトを開発した。

【研究成果】

方法

LDIA for GPソフトは、次世代質量分析装置付属ソフトDIT-FP Consoleの自動測定用プラグインとして実装した(図1)。本ソフトは、 MS^n 用 プリカーサ自動選択機能、糖ペプチド構造自動解析(糖鎖組成推定、ペプチド・糖鎖結合部位同定)機能を備えている。これらの機能は、Mass++に実装している独自開発の*de novo* sequencingソフトSIMSEとフリーソフトX!Tandem[3]を利用して実現した。

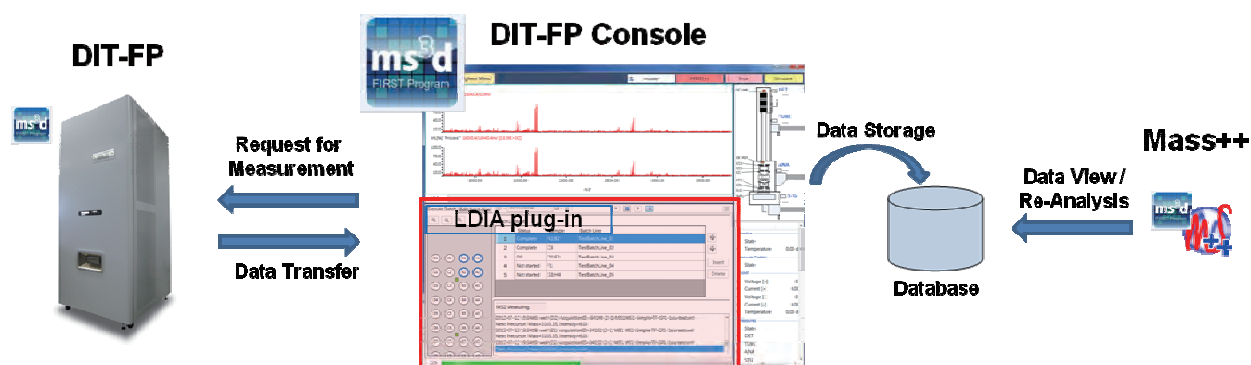


図1 LDIA for GP のシステム構成

LDIA for GPソフトは次世代質量分析装置付属の測定ソフト(DIT-FP Console)に組み込まれ、Mass++を利用してDIT-FP機の測定データを解析し、次の測定条件を決めて、DIT-FP機をコントロールして測定を実施する。

LDIA for GPソフトを用いた糖ペプチド分析手順を以下に記す。

① MS² 用プリカーサ選択

MS¹測定を実施し、MS¹のマススペクトル中に検出される糖のニュートラルロス解析し、糖ペプチド由来のイオンピークをMS²用プリカーサとして選択する。

② MS²による糖鎖組成推定とペプチド同定

①で選択されたプリカーサを対象にMS²測定を行い、MS²のマススペクトル中のピークを糖鎖の開裂に由来する領域(高質量側)とペプチドの開裂に由来する領域(低質量側)に分割し、糖鎖組成推定とペプチド・糖鎖結合部位の同定を試みる。

③ MS³によるペプチド同定

②でペプチド配列が十分な信頼度で同定できなかった場合は、さらにMS³測定を行い、ペプチドと糖鎖結合部位を同定する。

結果

ヒト・タンパク質Transferrin由来の2分岐N結合型糖ペプチド 500 fmol について、LDIA for GPソフトを用いて自動分析した結果を以下に示す。

まず、MS¹のマススペクトルから糖ペプチドのMS²用プリカーサを選択した(図2)。次に、MS²のマススペクトルから正しい糖の組成を推定できたがペプチドが十分な確度で同定されなかった(図3)ので、さらにMS³測定を実施し、MS³のマススペクトルからペプチド配列ならびに糖鎖結合部位を正しく同定できた(図4)。

【結論】

自動分析ソフトLDIA for GPは、手間と習熟を要していた MALDI-MSⁿによるN結合型糖ペプチドの高感度測定・構造解析を全自動で実行できる。本ソフトを利用することで、従来のように試料の前処理(ラベル修飾や糖鎖分解など)を行うことなく、糖ペプチドの構造解析をハイスループットで行うことが可能になった。

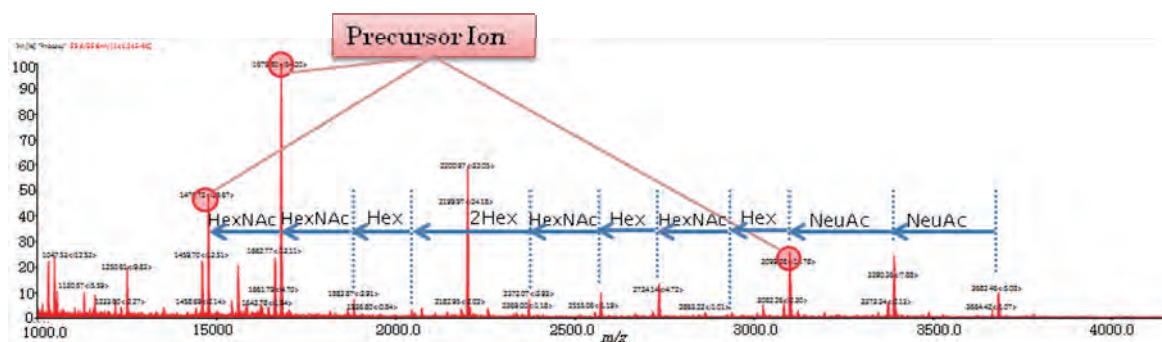


図2 MS² プリカーサの選択結果

糖ペプチドのMS¹スペクトルにおいて、高質量領域から*de novo* sequencingにより糖のニュートラルロス解析し、帰属された糖ペプチド関連イオンをMS2用プリカーサとした。

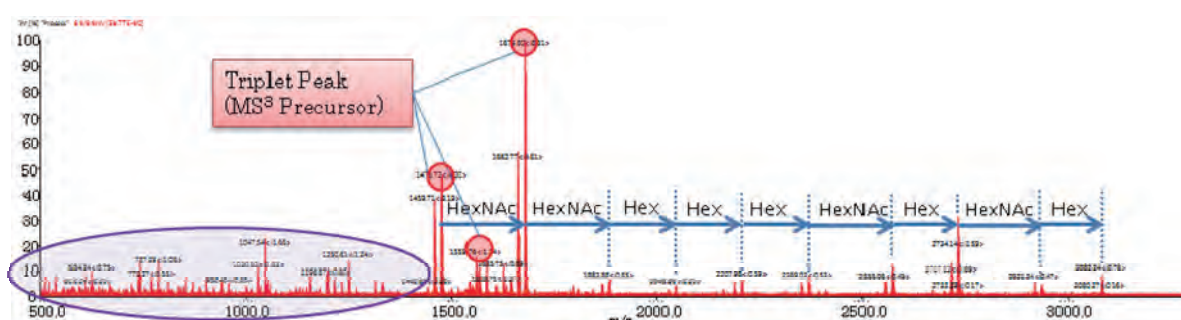
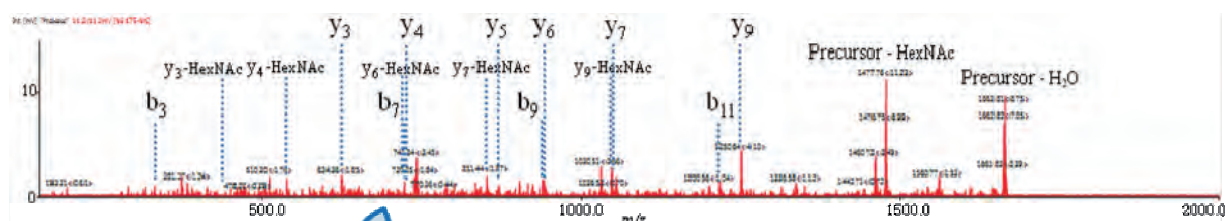


図3 糖鎖組成推定およびMS³プリカーサ選択結果

m/z 3099.28をプリカーサとしたMS2スペクトルにおいて、HexNAcの環開裂から生じるTriplet Peaksにより高質量側と低質量側に分割し、高質量領域において*de novo* sequencingで正しい糖の組成を推定できた(4HexNAc, 5Hex)。低質量領域からペプチドの同定を試みたが、十分な確度で同定されなかつ



Result of DB Search
by X!Tandem

precursor mass	amino acid residues	e-value	hyper score
1679.8	CGLVPVLAENYnK	7.1E-04	31.6

図4 ペプチド配列・糖鎖結合部位の同定結果

m/z 1,679.80をプリカーサとしたMS³スペクトルをX!Tandemで検索したところ、CGLVPVLAENYnKが同定された(e-value=7.1E-04, hyper score=31.6)。小文字nは糖鎖結合部位、小文字cはcysteineのcarbamidomethylationを示す。

【成果一覧】

学会発表

- (1) Masaki Murase, Hidenori Takahashi, Yoshihiro Yamada, Sadanori Sekiya, Shigeki Kajihara, Shinichi Iwamoto, Koichi Tanaka. "Data-dependent acquisition system for N-linked glycopeptides using MALDI-DIT-TOF MS", IMSC2012, 2012

特許出願

- (1) 村瀬雅樹、関谷禎規、梶原茂樹、谷口謙一、九山浩樹、ペプチド構造解析方法及び装置、特願2012-266980
- (2) 村瀬雅樹、糖ペプチド解析装置、特願2013-244499

【参考文献】

- [1] R. I. Viner et al. Quantification of post-translationally modified peptides of bovine α -crystallin using tandem mass tags and electron transfer dissociation. *J. Proteomics*, 2009, Vol. 72 No. 5, page 874
- [2] K. Sparbier et al. Analysis of Glycoproteins in Human Serum by Means of Glycospecific Magnetic Bead Separation and LC-MALDI-TOF/TOF Analysis with Automated Glycopeptide Detection. *J. Biomol. Tech.*, 2007, Vol. 18, pages 252-258
- [3] R. Craig et al. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 2004, Vol. 20, page 1466

3-3 ビッグデータ対応 (JobRequest & ProteoAnalysis)

【研究目的】

質量分析装置のハイスループット化により、装置から得られるデータ量は増加しており、近い将来、既にDNAの新型シーケンサーが陥っているのと同様、「データの保存場所が充分確保できない」という事態に陥る危険が高くなっている。また同時に、質量分析・プロテオミクス解析のフローは複雑化の一途を辿っているため、Mass++によるデータ解析が分析フロー全体の中でボトルネックになる可能性がある。これらの問題を解決する可能性の一つとして考えられているのは、クラウド・コンピューティングの利用である。

本研究は、一括して大量のデータにルーチン解析を行うため、Mass++と組み合わせたジョブ管理ソフトJobRequestと、クラウド・コンピューティング環境でJobRequestを運用する解析ソフトProteoAnalysisを開発した。

【研究成果】

JobRequestは、複数台のPC上で動作し、全体の処理の流れと計算実行ハードウェアを管理する。JobRequestは、ユーザーが指定したデータ、解析プログラムの実行順序、解析処理のパラメータに従って、個々の計算処理をCPU負荷の低いPCに割り振りながら、所定の処理を並列実行する。

さらに、JobRequestを利用するウェブアプリケーションProteoAnalysisを開発し、クラウド上に環境を構築して一般公開した。ProteoAnalysisでは、複数の生データをアップロードすることによって、(1)ピークリストの作成、(2)タンパク質の同定、(3)スコアに拠る結果のフィルタリング、(4)Gene SymbolやGene Ontologyなどのアノテーション付加、を並列で実行し、最後に結果を単一のファイルに統合して表示する。

ProteoAnalysisのシステム構成を図1に、概観を図2に示す。ピークリストの作成のためには、Mass++をクラウド上で実行する。また検索エンジンにはX!Tandem(あるいはMASCOT)を用い、更に既知PTM高速探知のために、MSPTM-DBとPTM-annotation-DB(4.2章参照)を利用する。

本課題における主要ソフトウェアであるJobRequestは、エーザイによってCRESTプロジェクト(2005～2010年)で開発が進められていた。このため本研究でも、JobRequestとProteoAnalysisの実装はエーザイが担当した。島津製作所はMSPTM-DBとPTM-annotation-DBの開発を行い、更にMass++その他のソフトウェアと連携させた。

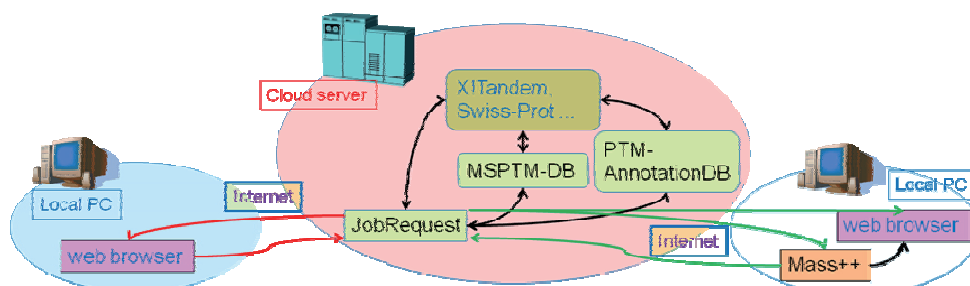


図1 ProteoAnalysisの構成

インターネットを介してリクエストした解析は、クラウドコンピュータ上のJobRequest、Mass++、X!Tandem、MSPTM-DBなどを利用して実行され、ウェブブラウザで解析結果が表示される。また、Local PCでMass++による処理を実行した後、データベース検索のみをクラウドコンピュータで行うことも可能である。

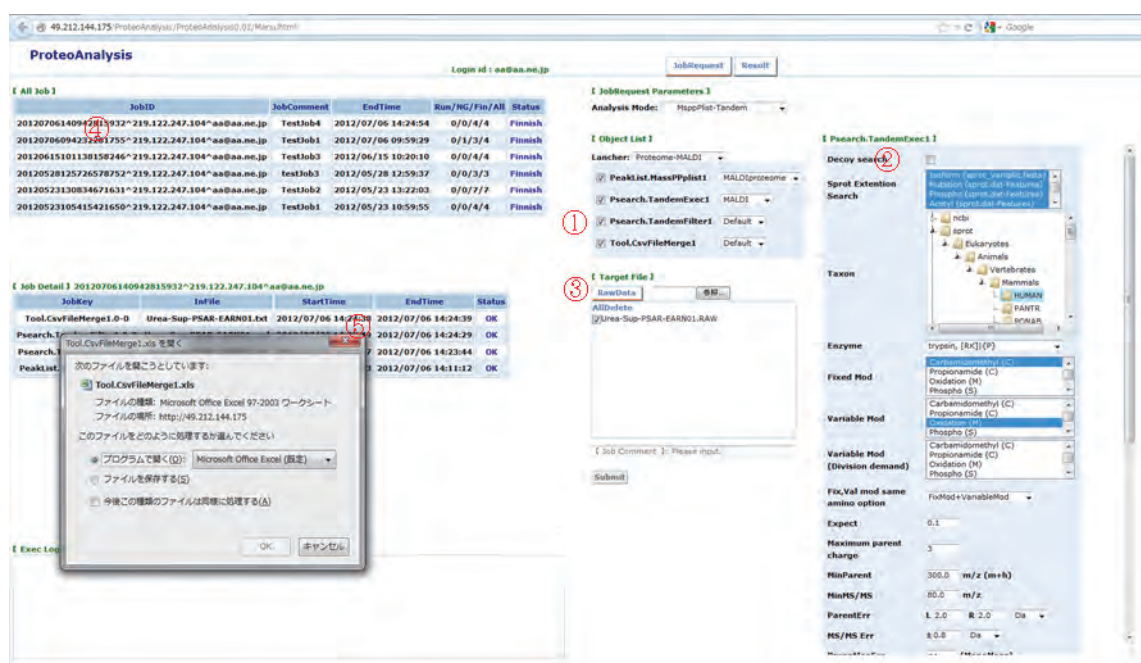


図2 ProteoAnalysisの概観

画面右で、バッチ管理ソフトJobRequestに対するピークリスト作成(①)、データベース検索(②)、検索結果の整理方法(③)のパラメータを設定する。画面左には、クラウドコンピュータ上でのバッチ処理の状況(④)が表示され、処理が完了したものをリスト上(⑤)でダブルクリックすると、解析の結果をExcel等で表示する。

【結論】

JobRequestとMass++を組み合わせ、いわゆるビッグデータに対応する解析システムをクラウド上に構築した。今後、数千検体によるバイオマーカーの検証やコホート研究において実験データ量が増大し、パソコンの処理能力を超えるようになったとき、クラウド上の処理システムにデータをアップロードして、より強力な処理能力をオンデマンドで用いて計算を行う、というスタイルが増加することが予想される。本システムは、そのための最初の試行であり、今後、より実用的な形に改良していくための出発点が構築できた。

【成果一覧】

学会発表

(1) Tsuyoshi Tabata, Akiyasu C Yoshizawa, Takayuki Kimura, Tatsuji Nakamura, Ken Aoshima, Yoshiya Oda, Shigeki Kajihara, Koichi Tanaka. "JobRequest – an easy-to-use software platform

for proteomic analysis – and ProteoAnalysis, its application for protein identification”, IMSC2012, 2012

- (2) 吉沢明康、田畑剛、木村剛之、青島健、小田吉哉、梶原茂樹、田中耕一、「既知翻訳後修飾情報データベースMSPTM-DBと web実行環境 ProteoAnalysis」, 第35回日本分子生物学会年会, 2012年
- (3) 田畑剛、吉沢明康、山本昇、青島健、小田吉哉、梶原茂樹、田中耕一、「ユーザーによる処理の追加可能なproteome解析プラットフォームJobRequestと、それを利用したタンパク質同定ツール ProteoAnalysis」, 第61回質量分析総合討論会, 2013年

4 データベースの開発と拡充

本研究では、ソフトウェアの開発とともに、新規データベースシステムの構築と既存データベースの拡充を行った。ここでは、以下の成果について報告する。

- バイオマーカー探索用尿中ペプチドデータベースシステムの開発。
- タンパク質の高度な構造解析用翻訳後修飾データベースおよび末端配列データベースの構築。
- 日本質量分析学会の公式データベースMassBankと連携するソフトウェアの開発。

4-1 尿中ペプチドデータベースシステム

【研究目的】

疾患バイオマーカーとなる尿中ペプチドを探索するため、臨床試料からCE-MSで網羅的に同定された尿中ペプチドのデータベースが構築されているが[1]、次世代質量分析システムで利用できるMALDI用のデータベースは存在しない。

本研究では、尿試料のマススペクトルと同定ペプチドを登録したデータベースを構築し、スペクトルから得られた配列情報やSequence Tagをデータベースと照合し、尿中ペプチドを同定する新規のデータベース検索システムを開発した。

【研究成果】

方法

開発した尿中ペプチドデータベースシステムのブロック図を図1に示す。

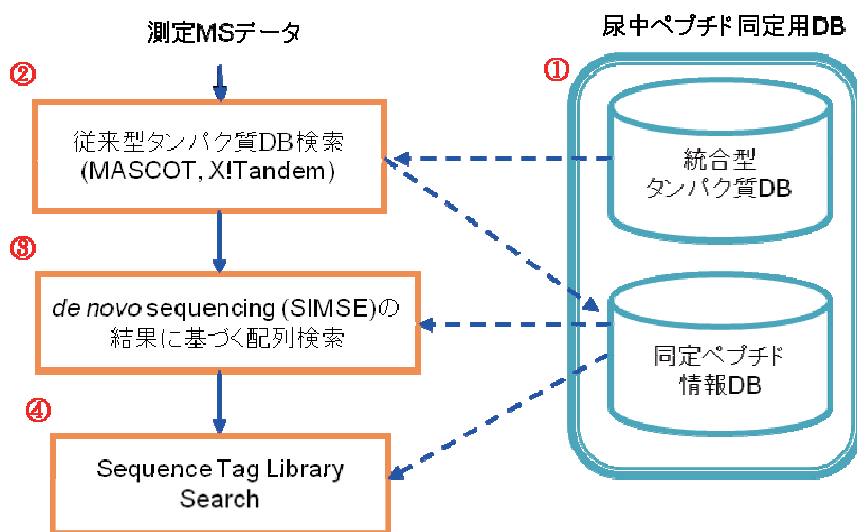


図1 尿中ペプチドデータベースシステムのブロック図

本システムでは、公的なデータベースを重複なく統合した「統合型タンパク質データベース」と同定したペプチドのSequence Tagなどの情報を登録した「同定ペプチド情報データベース」に対して、データベース検索ソフト (MASCOT、X!Tandem)、*de novo* sequencingソフト (SIMSE)、Sequence Tag Library Searchなどの解析ツールを利用して、尿中のペプチドを同定する。

解析の手順を以下に説明する。

- ① 尿中ペプチド同定用データベース (DB) として、同定ペプチド情報DBと統合型タンパク質DBを構築する。同定ペプチド情報 DB には、従来法により高い信頼度で同定されたペプチド情報 (アミノ酸配列、シーケンス・タグ、スペクトル) を登録する。統合型タンパク質 DB は、UniProt と NCBI nr, および

Ensembl DB を重複なく統合したデータベースで、同定情報の Accession や注釈情報の参照にも利用する。

- ② 従来法(MASCOTやX!Tandem)を用いて統合型タンパク質DBの検索を行い、同定された場合はペプチド情報を同定情報DBへ登録する。
- ③ ②で同定できない場合、*de novo* sequencing(§ 2-1-4 2-2参照) により得られた配列を、同定ペプチド情報DBのタンパク質アミノ酸配列と照合する。
- ④ ③で同定できない場合、同定ペプチド情報DBを用いたSequence Tag Library Search(§ 2-1-4 2-3参照)を行う。

結果

同定ペプチド情報 DB の有用性を確かめるために、一例としてX!Tandemで同定した alpha-1-microglobulin(AMBP)由来のペプチド4 個について、MS²のマスペクトルから抽出したTag 構成主要ピークをDBに登録し、未同定スペクトルに対してSequence Tag Library Search を行ったところ、未同定スペクトルから新たに 2 個の AMBP 由来のペプチドを同定できた(図2)。このように、同定ペプチド情報DB活用の有用性としてペプチド同定数向上を示唆する結果が、実試料への適用例からも得られつつある。現在、尿試料を粗精製して得たペプチド混合物試料からMASCOTやX!Tandemにより約30個のペプチドを同定しており、これらを登録した同定ペプチド情報DBを使い、他の未同定スペクトルについても検証を進めている。

Spectrum Name	Precursor Ion Mass	Title	Date	Peptide sequence	Accession	Search Engine	Score	E-value
1 (Well N2)	1,303.70	urine (SeqTag Lib + ProteinDB)	16 Jan 2014 19:17:20	YGRAPQLRETL	JP11P00022426.1 SWISS-PROT:P02760 TRE...	Search Local DB	94.00	-1
1 (Well N2)	2,218.80	urine (SeqTag Lib + ProteinDB)	16 Jan 2014 19:17:20	TKKFSRHHGPTITAKLYGRAPQLRETL	JP11P00022426.1 SWISS-PROT:P02760 TRE...	Search Local DB	203.00	-1
1 (Well N2)	1,815.90	urine (Mascot)	16 Jan 2014 18:36:59	AKLYGRAPQLRETL	30KG	MASCOT	31.95	3.52E+001
1 (Well N2)	1,822.90	urine (Mascot)	16 Jan 2014 18:36:59	LTKKFSRHHGPTIT	30KG	MASCOT	47.01	1.08E+000
1 (Well N1)	1,739.90	urine (Mascot)	16 Jan 2014 18:36:59	FRVVAQGVGIPEDGIF	30KG	MASCOT	103.71	2.00E+006
1 (Well N1)	1,386.70	urine (Mascot)	16 Jan 2014 18:36:59	FRVVAQGVGIPED	30KG	MASCOT	67.52	1.00E+002

図 2 尿中DBシステムでのペプチド同定結果

MASCOTで4個のペプチドを同定して同定ペプチド情報DBに追加登録した。MASCOTで同定できなかったMSスペクトルに対して、Sequence Tag Library Searchを行ったところ、同定ペプチド情報DBに登録されたペプチドAKLYGRAPQLRETLのTag情報を参考に、新たに2個のMSスペクトルからペプチドYGRAPQLRETLとTKKFSRHHGPTITAKLYGRAPQLRETL(下線部はSequence Tag)を同定することができた。

【結論】

開発した尿中ペプチドデータベースシステムは、尿中ペプチドに限らず生理活性ペプチド全般を対象としたペプチドーム解析における同定システムとしても利用可能であり、診断・創薬への応用が期待される。

【参考文献】

- [1] Joshua J. et al. CE-MS analysis of the human urinary proteome for biomarker discovery and disease diagnostics. Proteomics Clin Appl. 2008, Vol. 10 No. 2, page 964

4-2 翻訳後修飾データベース(MSPTM-DB)

【研究目的】

タンパク質の翻訳後修飾(PTM)の検索は、プロテオミクスの主要な目的の一つであるが、PTM解析を念頭に置いて設計された実験を行わずに、通常の測定データを基に計算機的に検索する場合、メモリ中に仮想の配列を大量に生成するため結果の信頼性に影響が生じるなど、扱いが難しい。一方、公共データベース中には実験的に同定された相当量のPTM情報が蓄積されているが、これを利用する手法は発表されておらず、既存のデータベース検索エンジンもこれに対応していない。

我々は、従来法の問題を回避するため、(1) 既知PTM情報を利用し、(2) 既存の検索エンジンで利用できる、(3) PTM探知のための新しいデータベースMSPTM-DBを開発した。

【研究成果】

方法

一般に、ペプチドのデータベース検索エンジンは、variable modificationとして指定されたPTMについて、可能性のあるアミノ酸残基全てに対して、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で生成し、これらに対して測定スペクトルとマッチするかどうかを比較している(図1 a))。そこで、これと同様の配列バリエーションを、「既知のPTM」に対してのみ事前に生成しておき、これをデータベースMSPTM-DBとして構築し検索対象とする(図1 b))。

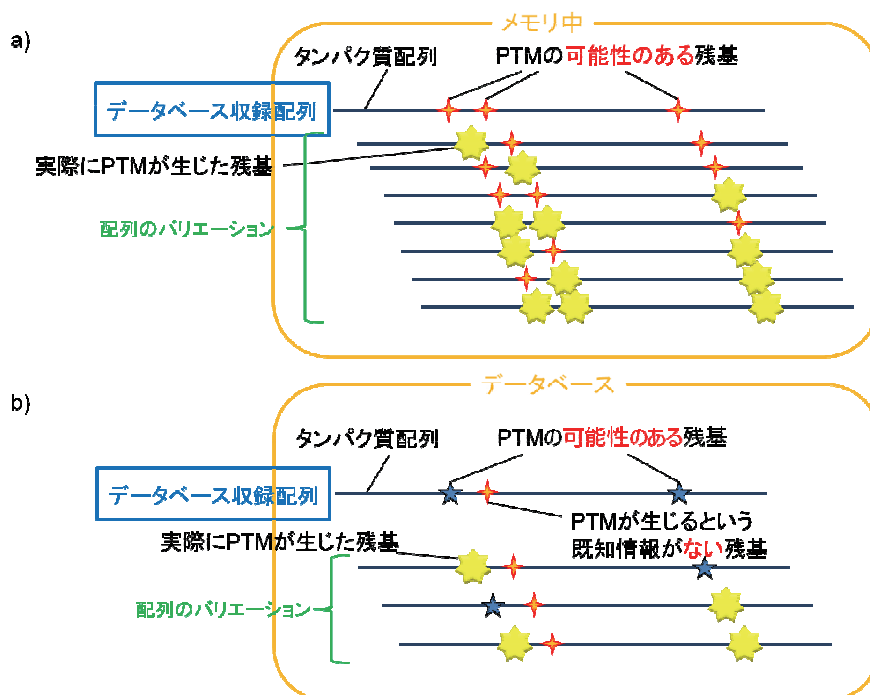


図1 従来法(左)とMSPTM-DB法(右)の比較

a) 従来法ではPTMの可能性のある残基全てに対して、PTMが生じた場合と生じていない場合の配列バリエーションをメモリ中で生成する。b)MSPTM-DBはPTMが生じることが既知の残基のみで配列バリエーションを生成してデータベースを構築する。

以下、リン酸化を解析する場合を例に、手順を説明する。

- ① 公共データベースに対して、ペプチドのアミノ酸S, T, Yがリン酸化を受けた場合のシンボルをそれぞれB, O, Zで置き換え、MSPTM-DBを構築する。
- ② 既存の検索エンジン(MASCOTやX!Tandemなど)を利用し、modification無しの条件で、MSPTM-DBに対して解析対象のMS/MSに適合するペプチドを検索する。

- ③ 検索でヒットしたB, O, Zを含むペプチドはリン酸化ペプチドであり、そのリン酸化について報告した文献情報を別途構築したAnnotation-DBから抽出して表示する。

結果

検証実験として、HeLa細胞抽出物からリン酸化ペプチドを濃縮した試料をLC-MSで測定し、公共データベースSwiss-Prot から構築したMSPTM-DBとSwiss-Protに対して、X!Tandemによるデータベース検索を行った。これに対し、従来法でvariable modificationとしてリン酸化を指定して検索を行った結果、および検証用の通常検索(従来法でmodification無し)を行った結果の比較を図2に示す。これらの比較において、非リン酸化ペプチドの同定は検証用の検索を行った場合が最も信頼度が高いと考え、従来法では33個が同定されず11個が余分に同定されている(図2 a))。これは、図1 a)のように、可能性がある配列を全て生成したため探索空間が広くなり、その結果、検索結果の信頼度の指標が変化して、非リン酸化ペプチドの同定結果も変わってしまったことを示している。一方、MSPTM-DBでは、非リン酸化ペプチドの同定数が検証用の検索結果と同程度であったことからわかるように、信頼度が高い同定結果が得られたと言える。従って、リン酸化ペプチドの同定結果(図2 b))においても、従来法よりMSPTM-DBのほうが信頼度が高い検索結果を得られていると推測できる。

さらに、検索に要した時間は、従来法では約1,400秒、MSPTM-DB法では約420秒であり、MSPTM-DBでは非常に高速な検索が可能であった。

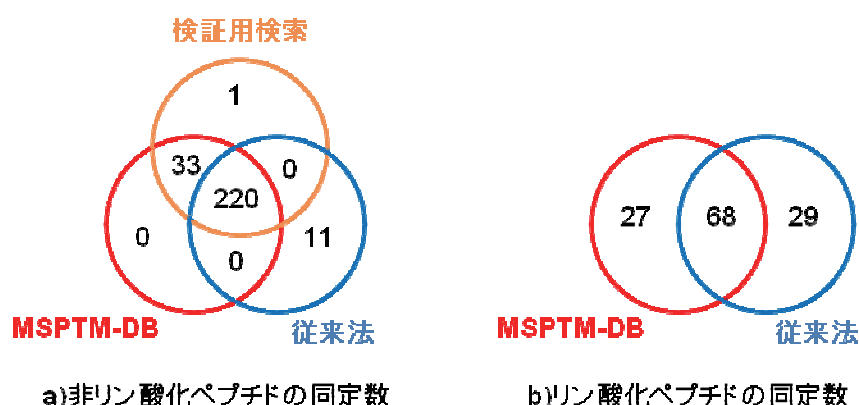


図2 従来法とMSPTM-DBによる同定数の比較

まず、variable modification無しの条件で検索を行った結果、254個の非リン酸化ペプチドが同定された(a)の検証用検索)。次に、variable modificationとしてリン酸化を指定して従来法で検索を行った結果、231(=220+11)個の非リン酸化ペプチドと97個のリン酸化ペプチドが同定された。a)の検証用検索と比べると、非リン酸化ペプチドでは33個が同定されず11個が余分に同定されている。最後に、MSPTM-DBで検索した結果、253(=220+33)個の非リン酸化ペプチドと95(=68+27)個のリン酸化ペプチドが同定された。a)の検証用検索と比べると、従来法より差異が少ないので、MSPTM-DBのほうが信頼度が高い検索結果であると考えられる。

【結論】

既知のPTMを高速高精度に探知できるMSPTM-DBは、基礎的研究に留まらず、特に「既知のタンパク質の既知のPTMが病状に応じて変化する」という観点からのバイオマーカー探索において、強力なツールになると考えられる。

【成果一覧】

論文発表

- (1) 吉沢明康、田畑剛、木村剛之、青島健、福山裕子、梶原茂樹、九山浩樹、小田吉哉、田中耕一、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、新潟医学会、印刷中

学会発表

- (1) 吉沢明康, 田畑剛, 木村剛之, 青島健, 小田吉哉, 梶原茂樹, 田中耕一、“質量分析データベース検索への翻訳後修飾情報の利用”, 第34回日本分子生物学会, 2011年
- (2) Akiyasu C. Yoshizawa, Tsuyoshi Tabata, Takayuki Kimura, Ken Aoshima, Yoshiya Oda, Shigeki Kajihara¹, Koichi Tanaka. “MSPTM-DB: a known PTM database for high-speed and accurate search available on the “ProteoAnalysis” web site”, IMSC2012, Sep. 15 – 21, 2012
- (3) 吉沢明康、“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”、第694回新潟医学会、2013年

特許出願

- (1) 吉沢明康、田畑剛、梶原茂樹、質量分析を用いた修飾タンパク質同定方法及び同定装置、特願2011-185711

4-3 末端配列データベースシステム(ProteinCarta)

【研究目的】

タンパク質の翻訳後修飾はタンパク質の機能に重要な影響を与えるが、特に末端領域では、シグナルペプチドの除去のような、構成アミノ酸が削除される顕著な変化が生じることがある。また、タンパク質の末端領域の数残基程度のペプチド配列だけでタンパク質を同定する、という先行研究も行われている[1]。そこでこの手法を支援し、かつ末端領域配列の精密調査に資するため、タンパク質末端配列データベースシステムProteinCartaを開発した。

【研究成果】

方法

公共データベースUniProtに登録されたヒト・タンパク質全配列からProteinCartaを構築した。ProteinCartaは、実験的に同定された各末端部分の配列を入力すると、それに合致する末端を持った配列を検索する。更に、配列を一意に絞ることができなかった場合、それぞれの末端をあと何残基同定すれば、一意にタンパク質を同定できるかを予測する機能を有する。

図1に実行例を示す。図1では、N末端としてA、C末端としてARのアミノ酸をもつタンパク質を検索した結果である。本結果は以下のことを表わしている。

- ① 図中のマス目の数字は、N末端(縦方向)とC末端(横方向)の配列数に対して絞り込めるタンパク質の最大数を表す。
- ② 水色のマス目は、現在入力した末端配列をもつタンパク質の数を表す。本例ではA.....ARという配列のタンパク質が14個あることを示している。
- ③ 赤色のマス目では、どんなアミノ酸が追加されてもタンパク質が一意に同定できないことを表す。本例では、N末端のアミノ酸が1つ追加されただけでは、追加されたアミノ酸の種類によらず、タンパク質を一意に絞ることができない。
- ④ 黄色のマス目では、追加されるアミノ酸の種類によってはタンパク質が一意に同定できることを表す。本例では、C末端にアミノ酸AあるいはKが追加されればタンパク質を一意に絞ることができるが、Sが

追加されれば2個のタンパク質が存在し、A, K, S以外のアミノ酸が追加された場合は3つのタンパク質が存在することを示しているため、ここでは最大数である3が示されている。

- ⑤ 緑色のマス目では、タンパク質が一意に絞れることを表わす。本例ではC末端に3つ以上のアミノ酸を追加すれば、アミノ酸の種類によらずタンパク質が一意に同定できることを示している。

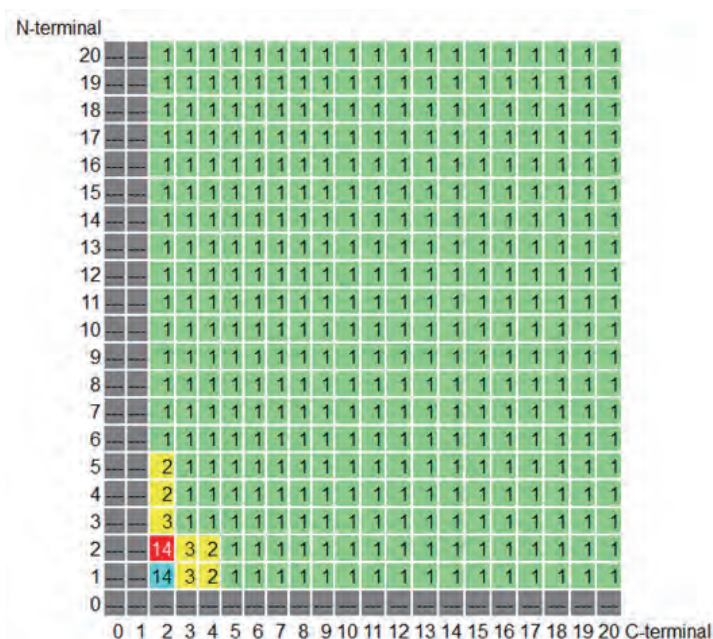


図1 ProteinCartaの実行例

例えば、タンパク質のN末端はA、C末端はARのアミノ酸であるとわかったとする。ProteinCartaでタンパク質のデータベースを検索すると、このようなアミノ酸をもつタンパク質は14個存在した(水色枠)。さらに末端のアミノ酸情報が追加された場合を予想すると、N末端の2個目のアミノ酸がわかったとしてもタンパク質の数を絞り込むことはできないが(赤色枠)、3個目のアミノ酸で3個、4個目および5個目のアミノ酸で2個のタンパク質に絞ることができ(黄色枠)、6個目のアミノ酸がわかると一意にタンパク質を同定できる(緑枠)。同様に、C末端の5個目のアミノ酸までがわかると一意にタンパク質を同定できる。あるいは、N末端の3個目までのアミノ酸とC末端の3個目までのアミノ酸がわかった場合にも一意にタンパク質を同定できる。

結果

ProteinCartaは、web上で一般に公開し、実際に末端配列の実験的研究でも利用されている(<http://www.first-ms3d.jp/english/achievement/software/proteometools>)。ここでは、ProteinCarta開発のために行ったfeasibility studyの結果を報告する。KEGG GENES(2011年6月版、20,834配列)を対象として、N-/C-末端のアミノ酸数と一意に同定されるタンパク質の割合を調べた結果を図2に示す。図2の折れ線種類はN末端のアミノ酸の数を、横軸はC末端のアミノ酸の数を示し、縦軸は一意に同定されるタンパク質の割合を示している。図2より、約87%のタンパク質がN-/C-両末端3残基で一意に同定できることがわかる。ただし、N-/C-両末端20残基のTagを用いても一意に同定できないタンパク質が約5%あるが、これらは配列が同一になってもおかしくない配列(isoformやparalog)が含まれているためと考えられる。従って、タンパク質の末端領域の数残基程度のペプチド配列だけでタンパク質を同定するという先行研究[1]に基づく手法は、ProteinCartaを利用することで有効な手法となりえることがわかった。

【結論】

ProteinCartaの利用によって、末端解析においてタンパク質を一意に同定するための追加実験の必要性に関する情報を得ることができる。よって、末端解析によるタンパク質同定効率の改善が期待されると

もに、タンパク質末端配列の種別間の特徴解析など、多方面への応用・発展が考えられる。

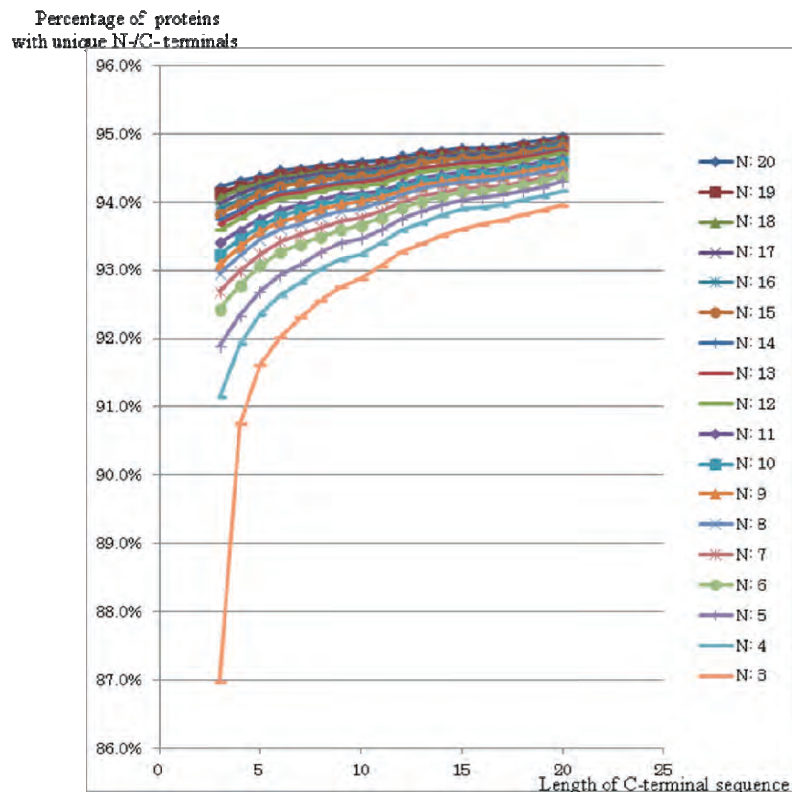


図2 単一のN/C末端配列をもつタンパク質の割合

横軸はC末端のアミノ酸の数を、縦軸は一意に同定されるタンパク質の割合を示し、折れ線種類はN末端のアミノ酸の数を示している。本結果から、約87%のタンパク質がN-/C-両末端3残基で一意に同定できることがわかる。また、N-/C-両末端20残基のTagを用いても一意に同定できないタンパク質が約5%あることがわかる。

【成果一覧】

学会発表

- (1) 吉沢明康、福山裕子、梶原茂樹、九山浩樹、田中耕一、「ヒトタンパク質末端配列データベースの構築と末端配列の一意性の推定」、第61回質量分析総合討論会、2013年
- (2) Akiyasu C. Yoshizawa, Yuko Fukuyama, Shigeki Kajihara, Hiroki Kuyama, Koichi Tanaka. "Cross-organism comparison of the uniqueness in protein terminal region sequences", HUPO2013, 2013
- (3) 吉沢明康、福山裕子、梶原茂樹、九山浩樹、田中耕一、「Cross-organism comparison of protein terminal sequences and development of terminal sequence database」、第36回日本分子生物学会年会、2013年

特許出願

- (1) 吉沢明康、九山浩樹、タンパク質の同定方法及び同定装置、特願2013-162008

【参考文献】

- [1] Wilkins MR et al. Protein identification with N and C-terminal sequence tags in proteome projects. J Mol Biol. 1998, Vol. 8, pages 599-608

4-4 マススペクトルデータベースMassBank との連携

【研究目的】

MassBank[1] は、奈良先端科学技術大学院大学の西岡孝明教授らが開発しているマススペクトル用データベースで、日本質量分析学会の公式データベースである(<http://www.massbank.jp>)。しかしながら、MassBankの課題として、MassBankレコード作成時に、サンプルデータ読み込み、ピーク抽出、検索データの加工・編集など、多くのソフトと手順を必要とし、付加情報(試料情報、実験情報、構造式)の手入力にも多大な労力を要していた。

そこで、これらの煩雑な操作をなくしてMassBankを効率良く利用するため、Mass++上にMassBank連携プラグインを開発した。

【研究成果】

方法

多くのMassBankレコードを簡単に作成するため、MassBankレコード作成機能を開発した。概念図を図1に示す。本プラグインの機能の特徴は、①Wizardによる多様な入力支援、②複数試料の同時処理、③試料固有の測定情報の自動取得、④外部DB(CAS等)から化合物情報の自動取得、⑤MassBankフォーマット(登録形式)への自動変換である。

さらに、MassBankの検索機能を強力にサポートするため、MassBank検索機能を開発した。概念図を図2に示す。本プラグインの機能の特徴は、①スペクトル情報をピークデータに変換することなく、単一スペクトル検索や複数スペクトルの一括検索、②検索結果をローカルDBに蓄積し高速処理、③キーワード検索によるSearch Resultsの絞り込み、④同定結果や詳細情報の比較、などの点である。

島津製作所ではMassBankを導入してデータを登録しMassBankユーザーからヒアリングを行なうなどして仕様を決定し、開発はエーザイが行なった。

結果

本研究ではMALDIで測定した代謝物を同定するため、in-houseのMassBankを構築し、使用するマトリックスが同定に与える影響を調べた。まず、表1 a)に示す4つの代謝物についてDHBマトリックスを用いて測定したMS/MSをin-houseのMassBankに登録した。次に、表1 b)に示す3つの代謝物についてCHCAマトリックスを用いて測定し、in-house MassBankに対して各MS/MSをSpectrum Search機能で検索した。Spectrum Searchの検索スコアを表2に示す。CHCAを用いて測定したLPC [524.3 Da]やTG [879.6 Da]は、DHBを用いて測定した同じ代謝物にハイスコアでヒットした。これに対して、PC [790.6 Da]は同じ代謝物がMassBankに登録されていないのでスコアが低かった。以上より、MassBankは、異なるマトリックスを用いて測定したMALDIのスペクトルにも利用可能であることを確認した。

開発したMass++連携プラグインは一般公開し、その利用方法についてMassBankと共同で講習会(第133回日本質量分析学会関東談話会)を実施した。MassBank には40,064個のスペクトルが登録され(2013年10月)、ユニーク訪問者は2012年には104,761人に達している。MassBankのホームページではMass++が紹介されており、多くのMassBankユーザーがMass++を利用している。

【結論】

MassBank連携プラグインにより、少ない手順で効率よくデータ登録ができるようになり、Mass++においてMassBankの利用が容易になった。さらに、本プラグインを一般公開したことで、MassBankの普及・拡充が促進されると同時に、Mass++のユーザーも増加した。

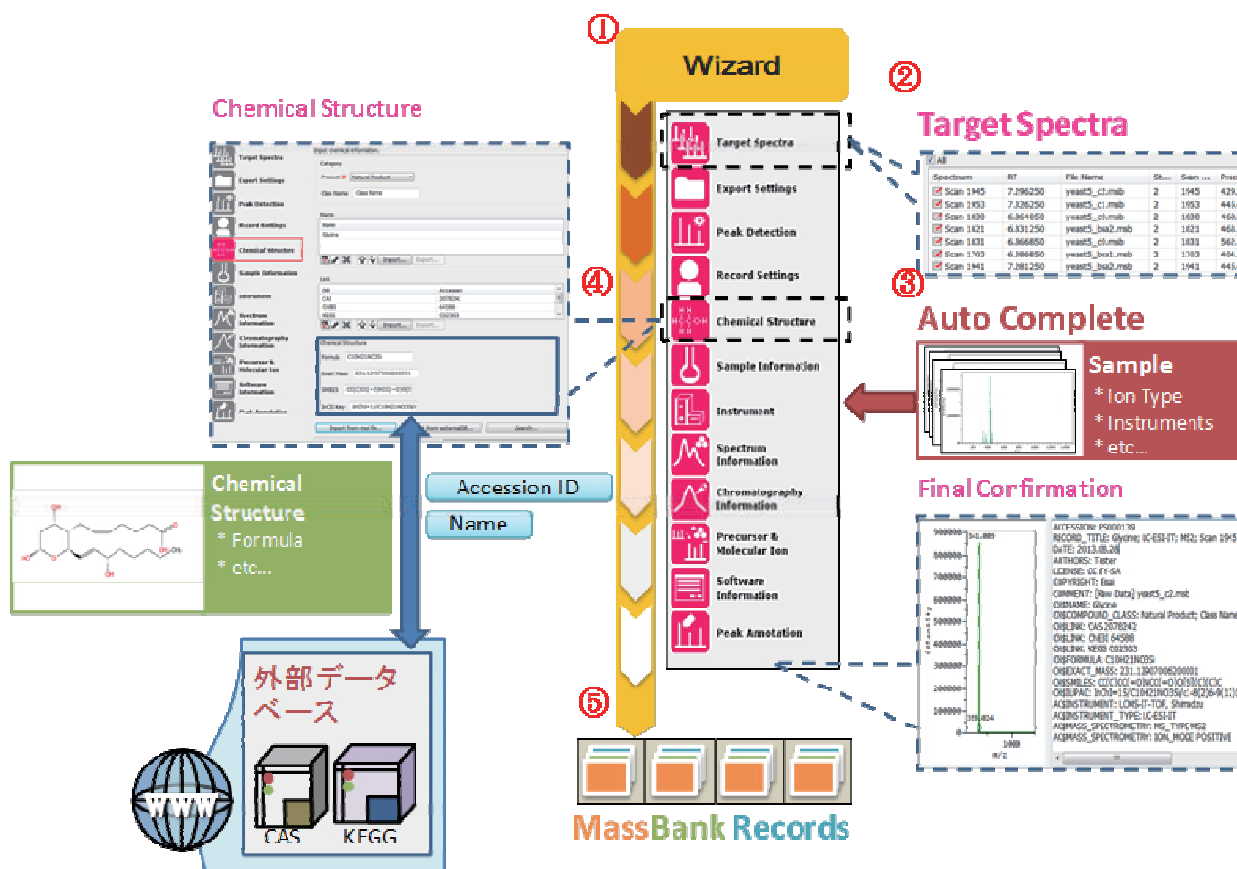


図1 MassBankレコード作成プラグインの概念図

Wizardに従い、対象スペクトル、ピーク検出パラメータ、試料情報、測定情報などを入力しMassBankレコードを作成する。手入力の労力を軽減するため、データファイルから測定情報を、外部DBから化合物情報を自動的に取得する。

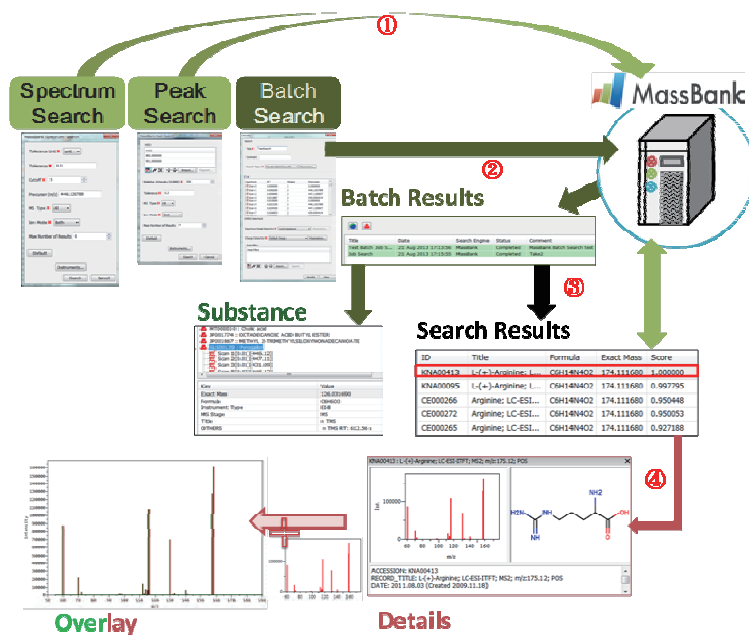


図2 MassBank検索プラグインの概念図

Mass++で表示したスペクトルあるいはピークリストをMassBankで検索し化合物を同定する。検索結果はローカルDBに登録され、化合物情報の確認やスペクトルの比較などが行える。

表1 分析した代謝物

a) DHB で測定し MassBank に登録した代謝物

Substance (Lipid)	Precursor [Da]
1-Stearoyl-sn-glycero-3-phosphocholine; LPC(18:0)	524.3
1-Palmitoyl-2-oleoyl-3-linoleoyl-rac-glycerol; TG(16:0/18:1/18:2)	879.6
1,2-Dioleoyl-sn-glycerol; DG(18:1)	643.4
1-Stearoyl-sn-glycero-3-phosphocholine; LPC(18:0)	546.2

b) CHCA で測定し MassBank で検索した代謝物

Substance (Lipid)	Precursor [Da]
1-Stearoyl-sn-glycero-3-phosphocholine; LPC(18:0)	524.3
1-Palmitoyl-2-oleoyl-3-linoleoyl-rac-glycerol; TG(16:0/18:1/18:2)	879.6
1,2-Distearoyl-sn-glycero-3-phosphocholine; PC(18:0/18:0)	790.6

a) DHBでMS²を測定しMassBankに登録した4個の代謝物を示す。b) CHCAでMS²を測定しMassBankに対して検索を行った3個の代謝物を示す。LPC(18:0)とTG(16:0/18:1/18:2)は、DHBとCHCAの両方で測定した。その他の3つの代謝物は、DHBあるいはCHCAのいずれかで測定した。

表2 MassBankに対するSpectrum Searchのスコア

	LPC (DHB) [524.3 Da]	TG (DHB) [879.6 Da]	DG (DHB) [643.4 Da]	LPC (DHB) [546.2 Da]
LPC (CHCA) [524.3 Da]	0.43	(no hit)	(no hit)	0.04
TG (CHCA) [879.6 Da]	(no hit)	0.56	(no hit)	0.01
PC (CHCA) [790.6 Da]	0.15	0.06	(no hit)	(no hit)

CHCAで測定した3個の代謝物(表1 b))をMassBankで検索し、MassBank中のDHBで測定した4個の代謝物(表1 a))に対して得られたスコアを示す。CHCAで測定したLPC(18:0)やTG(16:0/18:1/18:2)は、DHBで測定した同じ物質が高いスコアでヒットした。

【成果一覧】

学会発表

- (1) 青島健、福田充、井川幹之、高橋健太郎、木村剛之、梶原茂樹、宇都宮眞一、田中聡、吉沢明康、田中耕一、池田奨、二瓶義人、西岡孝明、小田吉哉、「質量分析用データ解析ソフトウェアMass++とマススペクトルデータベースMassBank連携プラグインの開発」、第61回質量分析総合討論会、2013年
- (2) 田中聡、吉沢明康、草野麻衣子、藤田雄一郎、福田充、宇都宮眞一、梶原茂樹、田畑剛、高橋健太郎、青島健、小田吉哉、二瓶義人、西岡孝明、田中耕一、「Mass++, MassBank を用いた脂質同定システムの構築」、日本医用マススペクトル学会第38回年会、2013年

【参考文献】

- [1] Hisayuki Horai, et.al. MassBank: a public repository for sharing mass spectral data for life sciences. J. Mass. Spectrom. 2010, Vol. 45, pages 703-714