ASMS2011 MP445

Peptide search algorithm by selecting and rescoring reliable peaks for MS^n(n > 1) spectra

Kentaro Morimoto[1], Masaki Murase[1], Tsuyoshi Tabata[2], Shigeki Kajihara[1], Yoshiya Oda[2], Koichi Tanaka[1]

[1] Koichi Tanaka Laboratory of Advanced Science and Technology, Shimadzu Corporation, Kyoto, Japan *[2] Biomarkers and Personalized Medicine Core Function Unit, Eisai Product Creation Systems, Tsukuba, Japan*

SHIMADZU

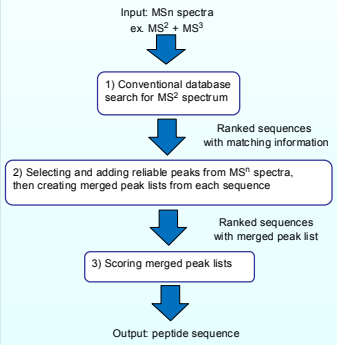
Overview:
A database search algorithm using MS^n (n > 1) spectra obtained by MALDI-IT/MS.

1. Introduction:

Ion trap mass spectrometers have been widely used to analyze the detailed structure of proteins because MS^n (n>1) spectra can be obtained by repeating the isolation and dissociation of precursor ions. MS^n (n>1) spectra, especially where n > 2, include important information and have high potential for analyzing the post translational modification of proteins. However, conventional software doesn't have sufficient ability to utilize this information in MS^n (n>2) because it does not have any functions to compare and combine the results of MS^2 with MS^n (n>2). In this study, we propose a novel algorithm for database search which provides more reliable results by using not only MS^2 but also MS^n (n>2).

2. Method

2.1 Method Overview:



1) Ranked peptide sequences and their matching information are obtained by database search for MS^2 spectra.
2) For all ranked sequences, additional y/b peaks from MS^n(n > 2) spectra are selected and added to each MS^2 spectrum. As a result, every sequence has an individual peak list.
3) The scores are calculated for every peak list. The score histogram is used to compute the expectation value for the top ranked sequence.

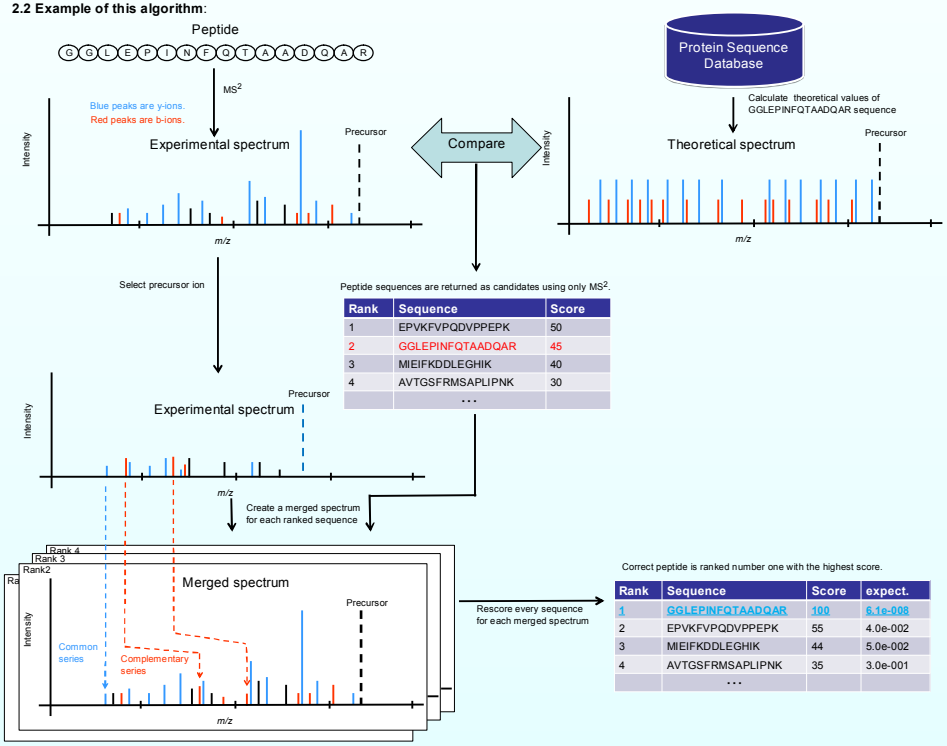
Fig1. Flowchart of this algorithm. In 1), the matching information includes peaks in each spectrum corresponding to theoretical fragments of y/b ion series.

Reference:

[1] Bioinformatics, 20, 1466-1467 (2004), Tandem: matching proteins with tandem mass spectra

Acknowledgement:
This research is granted by the Japan Society for Promotion of Science (JSPS) through its "Funding Program for World-Leading Innovative R&D on Science and Technology (FIRST Program)," initiated by the Council for Science and Technology Policy(CSTP).

2.2 Example of this algorithm:



Rank	Sequence	Score
1	EPVKFVPQDVPPEPK	50
2	GGLEPINFQTAADQAR	45
3	MIEIFKDDLEGHIK	40
4	AVTGSFRMSAPLIPNK	30
	…	

Correct peptide is ranked number one with the highest score.

Rank	Sequence	Score	expect.
1	GGLEPINFQTAADQAR	100	6.1e-008
2	EPVKFVPQDVPPEPK	55	4.0e-002
3	MIEIFKDDLEGHIK	44	5.0e-002
4	AVTGSFRMSAPLIPNK	35	3.0e-001
	…		

Fig2. Example of this algorithm for the peptide sequence GGLEPINFQTAADQAR. Conventional database search is shown above. After comparison of MS^2 experimental spectra with theoretical spectra, merged spectra for each sequence are created by adding common ion series and complementary ion series from MS^3 to MS^2. Common ion series have the same masses in both MS^2 and MS^3. Complementary ion series have shifted masses from MS^2 by the difference between precursor masses in MS^2 and MS^3. When the precursor is a b-ion, y-ions in MS^3 are shifted further by the mass of H_2O.

3. Result:

We applied this algorithm to MS^2 and MS^3 spectra of Trypsin-digested proteins.

Dataset summary.
1) Instrument: AXIMA QIT (Shimadzu/KRATOS, UK), Peak detection software: Mascot distiller (Matrix Science), Sample: Trypsin-digested Bovine Serum Albumin, Lysozyme C, Ovalbumin and Glyceraldehyde-3-phosphate dehydrogenase.
2) The ions with the highest intensity or the second highest intensity in MS^2 were selected as precursors for MS3.

Implementation summary.
1) We implemented this algorithm using X!Tandem[1].
2) We used X!Tandem native scoring as the scoring function.

Table 1. MS^2 column shows correct peptide sequences and expect. (y/b) column shows expectation values and y/b ion count returned by X!Tandem using only MS^2 . MS^3 column shows precursor ions and merged expect. value(y/b) column shows the result after applying this algorithm.

Sample	MS^2	expect. (y/b)	MS^3	merged expect. (y/b)
BSA	LVNELTEFAK	3.7e-003(6/5)	y5(LVNEL**TEFAK**)	1.2e-003(7/5)
	YLYEIAR	4.0e+000(4/4)	y3(YLYE**IAR**)	8.3e-002(5/4)
	RHPYFYAPELLYYANK	5.6e-005(5/9)	b9(**RHPYFYAPE**LLYYANK)	6.6e-006(6/10)
	DAIPENLPPLTADFAEDKDVCK	4.5e-009(12/7)	y21(D**AIPENLPPLTADFAEDKDVCK**)	1.6e-010(12/10)
	LGEYGFQNALIVR	4.3e-004(8/4)	y5(LGEYGFQ**NALIVR**)	2.7e-006(11/5)
	KVPQVSTPTLVEVSR	9.8e-007(8/7)	y8(KVPQVST**PTLVEVSR**)	2.7e-008(11/8)
			b12(**KVPQVSTPTLVE**VSR)	4.0e-006(8/7) (*)
	RPCFSALTPDETYVPK	2.1e-003(3/7)	b10(**RPCFSALTPD**ETYVPK)	1.4e-003(5/7)
			b11(**RPCFSALTPDE**TYVPK)	1.2e-002(3/7) (*)
			y14(RPC**FSALTPDETYVPK**)	7.9e-007(10/7)
LysC	IVSDGNGMNAWVAWR	5.5e-001(5/1)	y11(IVSD**GNGMNAWVAWR**)	6.6e-002(7/1)
	NTDGSTDYGILQINSR	3.6e-001(6/0)	y13(NTD**GSTDYGILQINSR**)	2.3e-003(9/0)
			y9(NTDGSTDY**GILQINSR**)	2.0e-006(11/5)
	NLCNIPCSALLSSDITASVNCAK	2.5e-006(12/3)	y9(NLCNIPCSALLSSDI**TASVNCAK**)	5.9e-009(15/3)
	FESNFNTQATNR	2.6e-004(10/3)	y10(FE**SNFNTQATNR**)	6.1e-008(11/8)
OVAL	AFKDEDTQAMPFR	7.8e-004(9/2)	y7(AFKDEDT**QAMPFR**)	2.5e-004(9/4)
	ISQAVHAAHAEINEAGR	3.8e-010(9/11)	y6(ISQAVHAAHAE**INEAGR**)	5.9e-012(13/11)
	GGLEPINFQTAADQAR	2.6e-004(10/3)	y12(GGLE**PINFQTAADQAR**)	6.1e-008(11/8)
	LTEWTSSNVMEER	3.8e-005(9/4)	y10(LTE**WTSSNVMEER**)	5.8e-005(10/4) (*)
G3P	LISWYDNEFGYSNR	2.4e-001(7/1)	y8(LISWYD**NEFGYSNR**)	6.7e-002(7/3)

In most results, additional peaks of false peptides were few and the scores of false peptides remained low. When testing this algorithm, the differences between the scores of true peptides and false peptides were large.
Although, in a few results (indicated by *), expectation values became worse, because the additional peaks were very few. So only overlapping peaks were detected in both MS^2 and MS^3 spectra in these results; overlapping peaks cannot improve scores and expectation values.

4. Discussion:

There are two outstanding problems.
- How to determine the intensities of additional peaks.
- How to incorporate overlapping peaks in MS^2 and MS^3 spectra in the score calculation.

It is difficult to determine the intensities of additional peaks. The score may not be improved, if the intensity of an additional peak is weak. However, false peptides may have the highest scores, if a noise peak is selected as a precursor and intensities of additional peaks are very high.
When testing this algorithm, many overlapping peaks in MS^2 and MS^3 were detected, though they are not reflected in the score. I think they should contribute to the confidence level of peptide identification.

5. Conclusion:

Applying the proposed database search algorithm using MS^n spectra
- Peptide sequences could be identified with higher confidence
- MS^n spectra are effectively incorporated in the score calculation