Discovery of Unanticipated Protein Modifications Using MS/MS Database Search

ThPA 015

Wilfred H. Tang, Ignat V. Shilov, Sean L. Seymour, Sean P. Keating, Alex Loboda, Alpesh A. Patel, Daniel A. Schaeffer, Applied Biosystems, 850 Lincoln Centre Dr., Foster City, CA 94404

ABSTRACT

We present an MS/MS database search algorithm with the following novel features:
• Novel protein database structure containing extensive pre-indexing
• Zone modification searching, which enables the discovery of protein modifications of known (i.e., user-specified) and unknown delta mass

The ability to search for a large variety of known and unknown modifications allows a significantly greater percentage of MS/MS scans to be identified. In addition, we show examples where the ability to search for unknown modifications allows the scientist to discover:
• Unexpected modifications which have biological meaning
• Amino acid mutations/substitutions
• Semi-tryptic peptides in a sample which has nominally been digested using trypsin
• Salt-adducted peptides in a sample which has nominally been de-salted
• Other unintended consequences of sample handling procedures

INTRODUCTION

Tandem mass spectrometry, also known as MS/MS, has increasingly become the method of choice for identifying proteins in complex mixtures. High-throughput analysis of MS/MS spectra requires automated software, and the most widely-used type of software for automatically interpreting MS/MS spectra is database search software. Database search engines, however, typically have a limited capacity for dealing with modified peptides. We present here a novel database search algorithm which enables the rapid discovery of modified peptides, including peptides with unanticipated modifications.

ALGORITHM

Novel features of the Interrogator™ algorithm (implemented in Pro ID, Pro ICAT, and Pro QUANT software):
• Protein database pre-indexed in two directions: peptide precursor mass, and b & y fragment ion masses
• Zone modification searching (described in detail in manuscript to be submitted for publication)

These two novel features in conjunction enable very efficient database search with the ability to discover a large variety of possible modifications, including unanticipated modifications.

Peptide confidences are computed using an empirical model based on a data set annotated by domain experts. Additional annotated data have been used to verify that the confidences computed by the model accurately reflect the observed probability of a correct result.

Ordinary modification-tolerant database searching	Zone modification searching
Modifications must be explicitly specified by the user in advance	Can find unanticipated modification (no need to specify particular modifications in advance); the modification can be of arbitrary delta mass
Expensive in computer time → can only look for a limited number of modification species	Efficient → can look for a very large variety of different modification species
Can find multiple modifications on a peptide	Designed to find one modification on a peptide (but can find multiple modifications if all the modified residues are close together in the sequence)
Slightly more sensitive	Slightly less sensitive

Table 1. Comparison of ordinary modification-tolerant searching vs. zone modification searching. All database search algorithms (including the Interrogator algorithm) implement ordinary modification-tolerant searching, but zone modification searching is unique to the Interrogator algorithm.

EXAMPLE #1 – CLINICAL SAMPLES

Mutations in transthyretin (TTR) lead to amyloidosis (deposition of abnormal, aggregated protein in tissues and organs). The main clinical symptoms are neuropathy, cardiomyopathy, and vitreous opacities, usually leading to death within 7-15 years following the onset of symptoms. Since the only effective treatment known to date is liver transplantation, accurate diagnosis is critical.

Methods:
• LC/MS/MS experiment performed on a sample purified by using immunoprecipitation to isolate TTR from patient serum[1]
• Data acquired on API QSTAR® Pulsar Hybrid LC/MS/MS System
• Perform zone modification search looking for delta masses in the range of ±400 Daltons

Patient	Delta mass	Modification
1	–73	Tryptophan (W) → Leucine (L) mutation
2	+32	Valine (V) → Methionine (M) mutation
2	+48	Valine (V) → Methionine (M) mutation, followed by oxidation of methionine
2	+80	Sulfonation of cysteine (C)

Table 2. Clinically-relevant modifications. Both of the mutations listed here have been identified previously in the literature as being amyloidogenic.[2] In addition, in vitro experiments suggest that sulfonation of cysteine may play a role in amyloidosis.[3]

EXAMPLE #2 – BETA-GALACTOSIDASE

If a sample comes from an organism whose genome/proteome has not been sequenced, protein identification by database search becomes a much more difficult problem. We simulate this problem by deleting the protein in our sample from the database we search against.

Methods:
• LC/MS/MS experiment performed on a trypsin-digested sample of beta-galactosidase (Escherichia coli)
• Data acquired on API QSTAR® Pulsar Hybrid LC/MS/MS System
• Delete beta-galactosidase (Escherichia coli) from the database, then perform zone modification search using the amino acid substitution matrix

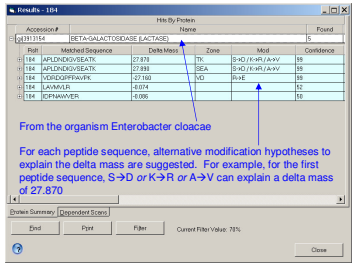

Figure 1. Identification of the homologous protein from the organism Enterobacter cloacae. Performing a zone modification search provides tolerance to evolutionary amino acid substitutions and also increases the confidence of the protein identification. Here, the top 3 most confident peptide identifications have non-zero delta masses.

Note that we can go back to the real protein sequence from beta-galactosidase (E. coli) and verify that the predicted amino acid substitutions are valid. The actual tryptic peptides are:

• APLDNDIGVSEATR (which means that, for the top two peptide identifications, one of the hypothesized substitutions (namely, K→R) is indeed correct)

• VDEDQPFPAVPK (which means that, for the third peptide identification, one of the hypothesized substitutions (namely, R→E) is indeed correct).

EXAMPLE #3 – SIMPLE PROTEIN MIXTURE

In many experiments, a large percentage of MS/MS spectra go unidentified by database search. The zone modification search provides an opportunity to identify significantly more spectra.

Methods:
• LC/MS/MS experiment performed on a simple protein mixture which was denatured, reduced with dithiothreitol, alkylated with iodoacetic acid (standard protocol for carboxymethylation of cysteine), and digested with trypsin
• Data acquired on API QSTAR® Pulsar Hybrid LC/MS/MS System
• Perform zone modification search looking for delta masses in the range of ±400 Daltons. (Note: carboxymethylation of cysteine is built into the pre-indexed database, or SSD file → thus, in the results reported below, "unmodified" peptides can contain carboxymethylated cysteines)
• Data set has undergone extensive manual curation → know with a reasonable degree of certainty which peptide identifications are accurate

Results:
• 473 MS/MS spectra correctly identified (out of 1292)
• 315 spectra correspond to unmodified peptides (delta mass ≈ 0)
• 158 spectra correspond to modified peptides (delta mass ≠ 0)
• Modified and unmodified peptides discovered together in a single search

Delta mass	# of times found	Modification
+58	22	Carboxymethylation of N-terminus
+44	18	Carboxymethylation of methionine
+1	13	Deamidation
–18	13	Loss of water, or pyroglutamic acid conversion of N-terminal glutamic acid
–17	9	Loss of ammonia, or pyroglutamic acid conversion of N-terminal glutamine
–48	8	Decomposed carboxymethylated methionine

Table 3. Common modifications. Algorithm discovers common modifications without prior knowledge.

Example of a doubly-modified peptide.
The peptide MAALKDQLIHNLLK is identified with a delta mass of –89 Daltons. There are actually two modifications:
• Loss of methionine (delta mass of –131 Daltons), followed by
• N-terminal acetylation (delta mass of +42 Daltons)

Both of these modifications are well-known and commonly-occurring processes whose physiological importance is still incompletely understood.

Delta mass	Peptide
–381	FFSASCVPGATIEQK
–380	SELQDAIGSLHSR
–362	CACSNHEPYFGYSGAF
–333	LGYVLTCPSNLGTGLR
–330	GGDDLDPHYVLSSR
–243	SELQDAIGSLHSR
–228	MPCTEDYLSLILNR
–228	VEDIWSFLSK
–227	AVGKVIPELDGK
–214	TLGLYGKDQR
–198	VVSVLPIQHQDWLK
–171	GNPTVEVDLHTAK
–101	TGPNLHGLFGR
–71	ADDGRPFPQVIK

Table 4. Semi-tryptic peptides. Semi-tryptic peptides can be discovered by zone modification searching. The amino acid sequence in black represents the semi-tryptic peptide actually present in the sample. The amino acid sequence in red has mass equal to the magnitude of the delta mass and represents the difference between the semi-tryptic peptide and the fully-tryptic peptide obtained by in silico tryptic digestion of the protein database.

EXAMPLE #4 – HUMAN HEART MITOCHONDRIAL SAMPLE

The ability to find unanticipated modifications allows the scientist to make discoveries that in the past required manual examination of spectra.

Methods:
• LC/MS/MS experiment performed on purified human heart mitochondrial sample[4]
• Data acquired on API QSTAR® Pulsar Hybrid LC/MS/MS System
• Look at aggregate of 12 runs
• Perform zone modification search looking for delta masses in the range of ±400 Daltons

Delta mass	# of times found	Modification
+16	19	Oxidation of methionine
+38	15	Potassium ion adduct
+1	7	Deamidation
+22	5	Sodium ion adduct

Table 5. Discovery of salt adducts. The table shows the most common modifications found for the protein NADH dehydrogenase [gi: 31377539]. Zone modification searching enables the unanticipated discovery of sodiated and potassiated peptides. This is consistent with the conclusions reached by manual examination of the data.[4]

CONCLUSIONS

The novel features implemented in the Interrogator™ algorithm enable the rapid discovery of a large variety of peptide modifications, including unanticipated modifications. We illustrate the power of the algorithm with a series of examples in which unanticipated modifications are discovered by the algorithm and later verified by independent means (such as consideration of biological context or manual examination of spectra).

REFERENCES

1. McComb, M. E.; Lim, A.; Prokaeva, T.; Connors, L. H.; Skinner, M.; Costello, C. E. *Proceedings of the 50th ASMS Conference on Mass Spectrometry and Allied Topics*, Orlando, Florida, June 2-6, 2002.
2. Connors, L. H.; Lim, A.; Prokaeva, T.; Roskens, V. A.; Costello, C. E. *Amyloid* **2003**, *10*, 160-184.
3. Kishikawa, M.; Nakanishi, T.; Miyazaki, A.; Shimizu, A. *Amyloid* **1999**, *6*, 183-186.
4. Gaucher, S. P.; Nimkar, S.; Fahy, E.; Taylor, S. W.; Gibson, B. W.; Ghosh, S. S. *Proceedings of the 51st ASMS Conference on Mass Spectrometry and Allied Topics*, Montreal, Quebec, Canada, June 8-12, 2003.

ACKNOWLEDGEMENTS

We thank the Amyloid Treatment and Research Program and the Mass Spectrometry Resource at Boston University School of Medicine for providing the TTR data. We thank Sara P. Gaucher and coworkers for providing the human heart mitochondrial data.

TRADEMARKS/LICENSING