Data Processing Algorithms for Analysis of High Resolution MSMS Spectra of Peptides with Complex Patterns of Posttranslational Modifications*\[S\]

Shenheng Guan‡ and Alma L. Burlingame

The emergence of efficient fragmentation methods such as electron capture dissociation (ECD) and electron transfer dissociation (ETD) provides the opportunity for detailed structural characterization of heavily covalently modified large peptides and small proteins such as intact histones. Even with effective gas phase ion isolation so that a single molecular precursor ion is selected, the MSMS spectrum of a heavily modified peptide may reveal the presence of a mixture of peptides with the same amino acid sequence and the same total number of posttranslational modification (PTM) moieties (same PTM composition) but with different PTM configurations or site-specific occupancy isoforms. Currently available data analysis methods depend on a desotoping procedure, which becomes less effective when spectra (fragmentation patterns) contain many overlapping isotopic distributions. Peptide database search engines can only identify the most abundant PTM configuration (PTM arrangement on different residues) in such mixtures. To identify all the PTM configurations present in these mixtures and to estimate their relative abundances, we extended our fragment assignment by visual assistance program to search for ions representing all possible configurations, subjected to the total PTM composition constraint. This resulted in the identification of PTM configurations supported by unique fragment ions, and their relative abundances were estimated by use of a non-negative least squares procedure. Molecular & Cellular Proteomics 9:804–810, 2010.

Characterization of large peptides and intact proteins by MSMS has initially been driven by the need for determination of complex patterns of PTMs\[1] on histones. Although core histones are among the most conserved proteins in the cell, they bear an extensive population of methylations, acetylations, phosphorylations, and ubiquitination on N-terminal tails of those proteins as well as particular additional key modifications throughout their sequences. The various combinations of different numbers of PTMs at different residue locations allow the cell to generate many different PTM configurations as well as alter particular sites dynamically in response to cellular cues. It has been proposed that those configurations are associated with epigenetic regulations (1–3). However, detailed characterization of core histones and their PTM configurations is thus far a difficult task. Although direct intact protein analysis for all four types of core histones, H2A, H2B, H3, and H4, has been demonstrated (4–7), it is more tractable presently to analyze in detail the N termini of histones where most PTMs are located. The difficulty for analysis of heavily modified histone N-terminal tail peptides arises from the fact that many, sometimes thousands of possible PTM configurations may exist due to the combinatorial nature of PTMs, occupying different sites (8). Significant effort has been spent to physically separate these individual PTM configurations. Chromatographic methods based on hydrophilic interaction (hydrophilic interaction chromatography (HILIC)) have been developed to separate peptides with different numbers of acetylations to near completion and to separate peptides with different numbers of methylations to a lesser extent (9–12).

A complete isotope distribution representing the parent molecular ion from a single elemental composition can be isolated before an MSMS scan. The isolation is typically achieved in a quadrupole or an ion trap device. For a high molecular weight parent ion, in the case that complete isolation of the parent ion may not feasible due to limited mass selectivity of the quadrupole or ion trap devices, stored waveform inverse FT (SWIFT) isolation can be used in an FTICR mass spectrometer (13). Although such optimal parent ion selection can isolate peptide ions of a single PTM composition (PTM configurations contain certain numbers of modification groups), many different arrangements (positional isomers) can still exist. The problem persists even with a combination of highly selective ion isolation and highly effective chromatographic separation (12). An ideal data process-
ing tool should be able to identify all PTM configurations present in these mixtures and to quantify their relative abundances. Currently, all the protein database search engines lack the capability to deal with such a task. They generally treat the PTM configurations as independent identities and only “find” and report the most abundant configuration.

DiMaggio et al. (14) recently introduced a mixed integer-linear optimization or mixed integer-linear programming (MILP) framework to analyze ETD spectra of heavily modified histone peptides. The method utilizes two stages of MILP procedures to produce both identification and quantification of PTM configurations. The first MILP model is constructed to rank all possible configurations, and the second MILP model is used to iteratively retrieve the relative abundances of the PTM configurations in the spectrum.

We present here a new set of algorithms for determination of the identification and relative abundance of PTM configurations in mixtures. Our method is a natural extension of our fragment assignment by visual assistance (FAVA) program (15), a software tool for analysis of high resolution MSMS spectra of large peptides and small proteins. FAVA was developed to overcome the deficiencies of deisotoping methods by requiring the user to provide overall information on peptide sequences and nature and number of PTM moieties. FAVA synthesizes a theoretical isotope distribution for an ion of interest (such as c or z ion) with the Mercury 2 (16) algorithm and directly matches the isotope distribution to raw data to identify the ion. FAVA can produce 15–50% more isotope distribution identifications than the existing deisotoping methods because of incorporation of prior known information such as mass accuracy into the calculation.

Our present method uses the FAVA core procedure to compute non-redundant ions, representing all possible PTM configurations. A matching matrix is constructed to map all possible PTM configurations to the observed non-redundant ions. All the possible PTM configurations are classified into independent configurations, the associated dependent configurations, and configurations lacking support from the observed fragment ions. Only the independent configurations are subjected to quantification. Using intensities of the identified non-redundant ions, the relative abundance problem is modeled as a standard linear least squares problem. The method is demonstrated for complete characterization of a set of ETD spectra from heavily modified peptides with 31 amino acid residues from histone H3.

**EXPERIMENTAL PROCEDURES**

Core histones were extracted from human 293 cells by use of a standard isolation protocol (17). Core histones were fractionated by a RPLC system with a semi-preparative C₁₈ column, and core histone fractions were collected. Histone H3.1 was digested with Glu-C, the digested peptides were fractionated on a microbore RPLC system, and isolated peptides were collected.

All data were acquired on an LTQ/Orbitrap XL instrument equipped with ETD (Thermo Fisher Scientific, Bremen, Germany). Samples of histone H3.1 peptides with a concentration of about 1 pmol/μl in 1% formic acid and 50% methanol in water were infused into the instrument through a static (medium) nanospray tip (Proxeon, Odense, Denmark). Five microliters of sample loading typically allows for a least a 30-min experiment. ETD spectra were acquired with a resolution of 30,000 and using 10 microscans per spectrum. Ten scans were averaged to give the final spectrum, and each of the final spectra took about 1.7 min to acquire. Double isolation steps were used: the first isolation window had a width of 30 m/z units, and the second had a width of 4 m/z units. The ETD reaction time was 25 ms.

Raw ETD spectra acquired in the orbitrap were converted to monoisotopic peak lists (M + H format) with the Xtract program (Thermo Fisher Scientific, Bremen, Germany). The deisotoped peak lists were searched against human sequences in the Swiss-Prot July 7, 2009 database with Protein Prospector (18, 19). Met loss + acetyl(protein N terminus), methyl on Lys, dimethyl on Lys, and trimethyl on Lys were variable modifications. The peptide was allowed to have up to six variable modifications on different residues. All other programs were written in a Matlab version 6 environment (MathWorks, Natick, MA).

A description of the software follows. As shown in Fig. 1, the complete data processing workflow consists of several processing modules (deisotoping, database searching, FAVA for non-redundant ions, classification of PTM configurations, and least squares modeling for relative abundances), which will be described in more detail below. In the deisotoping procedure, the raw data of ETD spectra are converted to monoisotopic peak lists by use of Xtract (part of the Xcalibur software from Thermo Fisher Scientific, San Jose, CA). For the database search, the deisotoped peak lists are searched with Protein Prospector (18, 19) to produce a peptide sequence showing the major PTM configuration. The database search results (peptide sequence and type and number of covalent modifications) allow for generation of all possible PTM candidate configurations, and FAVA code is used to extract non-redundant ions from the original raw ETD data. A classification procedure has been developed to classify all theoreti-

cally possible PTM candidate configurations, and relative abundances of independent configurations (supported by non-redundant, experimentally observed fragment ions) are calculated with a standard linear least squares procedure.

To clearly describe the software, we chose a histone peptide as an example. RPLC fractionation of a Glu-C digest of histone H3 from human HEK-293 cells produced a fraction containing a set of molecular ions whose molecular masses match a peptide of sequence “LATKAARKSAPATGSKKPHRYPRPGTVLRE” with an additional zero to five “14-Da moiety” groups (Fig. 2).
The sequence is consistent with the segment of human H3.1 from residue 20 to 50. Clearly, residue 19 was converted into a glutamic acid before the Glu-C digest. The molecular ion corresponding to the sequence containing four methyl groups was isolated for ETD. If we assume methylation can only occur on lysine residues and a lysine can be modified with zero to three methyl groups, there are 31 possible configurations (Fig. 3).

In this case, molecular weight information is clearly sufficient to identify the peptides and PTM compositions. In general, that information can be obtained by a database search of “normal” deisotoped peak lists. Indeed, the database search of the deisotoped ETD spectrum identified 0220 as the major configuration. To determine which of the 31 possible PTM configurations are actually present in the sample (identification) and to estimate their relative abundances (quantification), we used the core function of FAVA to compute c and z ions, representing all the 31 configurations. The simplest way to characterize the ETD spectrum is to compute all c and z ions of the 31 candidate PTM configurations with FAVA. However, such a brute force approach will result in many redundant computations of “shared” ions. Shared ions include all c and z ions before the first and after the last lysine residue. Some of the “internal” c and z ions between the first and the last lysine residues are also shared among some of the possible PTM configurations. These internal ions constitute non-redundant ions, which contain information for differentiating the PTM configurations. To avoid repeated computations, only non-redundant ions are searched by FAVA. An example of a non-redundant ion is illustrated in Fig. 4.

The FAVA calculation of the ETD spectrum is summarized in supplemental Fig. S1. To determine which PTM configurations are present in the spectrum, a matrix is constructed to match all the possible PTM configurations to the observed non-redundant ions. All the possible PTM configurations are sorted according to the number of observed non-redundant ions. For simplicity, we use only the observed c ions here (50 of them; see supplemental Fig. S1). The PTM configuration that received the most support from the non-redundant ions is used as a reference configuration. All other configurations are compared with the reference configuration to see whether they receive support from other unique ions. If they share some ions with the reference configuration but do not have unique ions, they belong to the dependent configurations associated with that reference configuration. If one does receive support from other unique ions, it becomes an independent configuration. The process is repeated after setting the existing independent configurations to the collection of the reference configurations.

The classification procedure groups all the possible configurations into three classes. The first class consisting of independent configurations will be further investigated for their relative abundances. The second class is the dependent configurations, associated to one of the independent configurations. Although there are no unique ions to support their existence, there is no evidence to establish their absence. In the third class, the configurations do not receive any ions to support their existence. In our example, four configurations (0220, 0310, 0130, and 1300) are declared as independent configurations (Table I). They are associated with 11, seven, seven, and two dependent configurations, respectively. There is no third class configuration in this example.

To estimate the relative abundances of the independent configurations, the following linear equation is established for each of the non-redundant ions,

\[
\sum_{i} \left( \sum_{j} I_{c}^{z_{i}=z_{0}} \right) X_{i}
\]

in which \( I_{c}^{z_{i}=z_{0}} \) is the intensity of non-redundant c_{i} ions of charge state \( z_{0} \), the first summation symbolizes the contribution of all the

---

2 Configuration notation (abcd) indicates a, b, c, d numbers of methyl groups on four lysine residues, Lys^0, Lys^1, Lys^2, and Lys^3, respectively, for histone H3.1 peptide.
independent configurations to the ion, the second summation adds all the \( c_k \) ions of charge \( z_0 \) regardless of the numbers of methyl groups on them, and \( x_i (j \in \text{independent configurations}) \) are the relative abundances of the independent configurations. To provide a better understanding of the above equation, let us take \( c_8 \) ions of charge state 2 as an example (Fig. 5). Three non-redundant ions, \( c_8(\text{Me})^2 \), \( c_8(2\text{Me})^2 \), and \( c_8(3\text{Me})^2 \), were observed in this spectrum, and their relative intensities are 1840.8, 44,364.1, and 3066.6, respectively. As shown in Table I, there are four independent configurations (0220, 0310, 0130, and 1300), and \( x_1, x_2, x_3, \) and \( x_4 \) are their relative abundances. The summation \( \sum c_k' x_i \) is the total intensity of all \( c_k \) ions of charge state 2 or the sum 49,271.5 (1840.8 + 44,364.1 + 3066.6). For the \( c_8(2\text{Me})^2 \) ion, the above equation becomes 44,364.1 = 49271.5 \( x_1 \), because only configuration 0220 can produce this particular ion. The \( c_8^2 \) ions from all other configurations have different \( m/z \) values and therefore do not make a contribution to the ion selected here.

The equation is valid only if fragmentation is not affected by different PTM configurations. Otherwise, a set of weighing factors to account for the difference has to be included. In principle, one can synthesize individual PTM configurations and determine the weighing factors experimentally. Information about the weighing factors may also be “learned” statistically from analysis of a large number of experiments. At present, we are not in a position to obtain the information, and therefore the weighing factors are considered as unity.

In our example, there are four variables (four independent configurations) and 50 (non-redundant c ion) equations, indicating it may be an overdetermined problem. It can also be recognized as a standard linear least squares problem. Because relative abundances should not be negative, we use a standard non-negative least squares procedure (20) to solve it. We note that the use of a least squares method to determine the relative abundance levels of PTM configurations in histone H4 has been demonstrated (21). It should be clear that the matching matrix for mapping non-redundant ions and possible PTM configurations plays a key role in construction of the above equations and for classification of PTM configurations.

Fig. 4. Example of non-redundant c ion. The \( c_{15} \) ions for the nine configurations listed are identical, and they correspond to the single non-redundant \( c_{15}(2\text{Me}) \) ion. Instead of computing all nine identical \( c_{15} \) ions obtained from the nine configurations, FAVA only searches for the single non-redundant ion.

RESULTS AND DISCUSSION

The quantification analysis of the ETD spectrum from histone H3.1 peptide LATKAARKSAPATGGVKKPHRYRPGTVALRE with four methyl groups produces three independent PTM configurations with non-zero relative abundances (shown in Table I). The quantification result is consistent with the raw spectrum partially shown in Fig. 5.

The relative intensities of the three \( c_8 \) doubly charged ions are roughly the same as the quantification result (5.6% for 0130, 79.5% for 0220, and 14.9% for 0310). The relative abundance ratio of configurations 0310 and 0220 is about 3:16, which also agrees with the intensity ratio of \( c_{13}(3\text{Me}) \) and \( c_{13}(2\text{Me}) \) triply charged ions.

The intensities of the molecular ion spectrum of Fig. 2 allow for estimation of relative abundances of different PTM compositions (different numbers of methyl groups). The PTM composition abundances are listed in Table II. For example, the peptide with no methyl group has a relative abundance of 6.1%. For all other PTM compositions, more than one PTM configuration are present. The relative abundance for an individual PTM configuration can be obtained simply by multiplication of the relative abundance of the PTM configuration in the composition and the relative abundance of the composition. Because the peptide with four methyl groups has an abundance of 29.9% and PTM configuration 0220 has a relative abundance of 79.5% in the composition, configuration 0220 has an overall relative abundance of 23.8%. The estimation method described here is based on an assumption that ionization efficiencies are not affected by PTM compositions and that any potential ion suppression effect is minimal.
Table II provides a complete characterization of PTMs on the histone H3.1 peptide (LATKAARKSAPATGGVKKPHRYRPGTVALRE). The configurations containing trimethyl groups are consistent with the high mass accuracies of the precursor ions. The precursor mass deviations (errors) of the peptides containing three, four, and five methyl groups are 0.87, 0.98, and 4.36 ppm compared with the mass errors of 11.78, 9.88, and 6.46 ppm for the peptide assignments with one acetyl, one acetyl and one methyl, and one acetyl and two methyl groups, respectively. The trimethylation assignments are further supported by the lower mass errors of the fragment ions containing trimethyl groups. For example, there are 14 c ions with trimethylation observed in the peptide containing five methyl groups. The maximal mass error for the c ions with trimethylation is −3.47 ppm versus 24.1 ppm with the acetylation assignment. The absolute average mass error of those c ions with trimethylation assignment is 1.10 ppm, supporting the trimethylation assignment. Clearly, Lys^{23} and Lys^{37} are not methylated at all, whereas Lys^{27} is preferentially methylated. We currently do not know the biological significance of this distinctive PTM distribution.

Thus far, we have demonstrated a complete data processing work flow for identification and quantification of complex PTM configurations of peptides from high resolution ETD spectra. Our work flow does not require use of any sophisticated commercial software package(s). Therefore, this entire work flow can be readily automated. More importantly, our work flow only uses deisotoping to extract sequence information and relies on the much more reliable FAVA method for identification and quantification of PTM configurations. Use of database search strategies for identification of protein and peptide sequences is now ubiquitous in any proteomics study. Although any deisotoping method may fail to identify many ions (many false negatives) and may produce many false positives, our work flow can handle this problem by using deisotoping and FAVA methods in parallel.

![Figure 5: Portion of ETD spectrum of 7+ peptide ion with four methyl groups.](image-url) The quantification results are consistent with the observed intensities of the identified c ions.

**Table I**

<table>
<thead>
<tr>
<th>Independent configuration</th>
<th>No. of Me on Lys</th>
<th>Total c ions</th>
<th>Unique c ions</th>
<th>Correlated configurations</th>
<th>Relative abundance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lys^{23}</td>
<td>Lys^{27}</td>
<td>Lys^{36}</td>
<td>Lys^{37}</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

* Initial reference configuration.
incorrect results (many false positives), it typically works well enough to allow a database search engine to identify the peptide sequence with the major PTM configuration. Meng et al. (22) have introduced a Poisson model to explain why only a few ion matches with a small portion of ion matches are needed for identification of protein sequences. For identification and quantification of PTM configurations, optimal extraction of all information present in the fragmentation pattern is required. Any deisotoping method by definition does not incorporate all information available before analysis, and therefore it will miss many ion identifications as well as introduce incorrect assignments. Those “false negatives” and “false positives” have a detrimental effect on establishment of the correct final PTM configuration results, especially for PTM configurations with low stoichiometry or low relative abundances.

Although we have used lysine methylation in the example, the general logic applies to any type of modifications at any amino acid residue. In a high resolution experiment, molecular weight of precursor ion can be measured with high mass accuracy, and PTM composition can often be determined with high confidence. Our method can distinguish acetylation and trimethylation by use of mass accuracies of both the molecular ions and the identified fragment ions. In a high resolution experiment, molecular weight of precursor ion can be measured with high mass accuracy, and PTM composition can often be determined with high confidence. Our method can distinguish acetylation and trimethylation by use of mass accuracies of both the molecular ions and the identified fragment ions.

In conclusion, we have developed a comprehensive strategy and a set of effective algorithms for extracting maximal and accurate information on identities and relative abundances of PTM configurations imbedded in complicated MSMS spectra. To improve the method presented here, we need to have a better understanding of how different PTM configurations influence the ionization and fragmentation efficiencies. HILIC is a very effective separation method to reduce the complexity of heavily modified histone peptides (9–11). The on-line HILIC based on the pH gradient (12) is especially attractive because it significantly simplifies the experimental procedures and reduces sample loss. The work of DiMaggio et al. (14) has demonstrated that incorporation of LC retention time increases the identification confidence. Implementation of such a strategy to high resolution ETD is still challenging because it takes longer to acquire high quality high resolution ETD spectra. It becomes a trade-off between the sampling space (acquire as many ETD spectra possible) and the spectral quality. With increasing MS instrumentation performance capability, we have no doubt that the on-line HILIC/high resolution ETD will provide unparalleled information content on the histone PTMs. Extension of our data processing algorithms to analyze on-line LCMSMS data is similar to our peptide identification-based label-free quantification method (23) in which peptides identified by a database search of the MSMS spectrum are correlated with the survey scans (MS1). Integration of extracted ion chromatograms allows for the quantification of the peptides. Because our method allows for identification and relative quantification of PTM configurations within an MSMS spectrum, integration of the corresponding extracted ion chromatogram would be able to produce the overall quantification of the PTM configurations for the entire sample.

Acknowledgment—We thank Dr. Frank Li for acquisition of the ETD spectra of histone H3.1 peptides.

* This work was supported, in whole or in part, by National Institutes of Health Grants P41RR001614 and RR019934 from the National Center for Research Resources.

† To whom correspondence should be addressed: Mass Spectrometry Facility and Dept. of Pharmaceutical Chemistry, University of California, 600 16th St., Genentech Hall, Rm. N472, Box 2240, San Francisco, CA 94158-2517. Tel.: 415-476-0245; E-mail: sguan@cgl.ucsf.edu.

REFERENCES
Algorithms for Analyzing Peptides with Complex PTMs