Unrestrictive Identification of Multiple Post-translational Modifications from Tandem Mass Spectrometry Using an Error-tolerant Algorithm Based on an Extended Sequence Tag Approach*[S]

Seungjin Na‡§, Jaeho Jeong¶, Heejin Park‖, Kong-Joo Lee¶, and Eunok Paek‡**

Identification of post-translational modifications (PTMs) is important to understanding the biological functions of proteins. MS/MS is a useful tool to identify PTMs. Most existing search tools are restricted to take only a few types of PTMs as input. Here we describe a new algorithm, called MOD^i (pronounced "mod eye"), that rapidly searches for all known types of PTMs at once without limiting a multitude of modified sites in a peptide. MOD^i introduces the notion of a tag chain, a combination structure made from multiple sequence tags, that effectively localizes modified regions within a spectrum and overcomes *de novo* sequencing errors common in tag-based approaches. MOD^i showed its performance competence by identifying various types of PTMs in analysis of PTM-rich proteins such as glyceraldehyde-3-phosphate dehydrogenase and lens protein. We demonstrated that MOD^i innovatively manages the computational complexity of identifying multiple PTMs in a peptide, which may exist in a greater variety than usually expected. In addition, it is suggested that MOD^i has great potential to discover novel modifications. ***Molecular & Cellular Proteomics 7: 2452–2463, 2008.***

Most proteins undergo PTMs[1] at multiple sites. The types and sites of PTMs in a protein vary widely and affect its cellular functions. Identification of all PTMs present in a protein is a key step toward understanding its biological functions and interactions inside a cell (1, 2). MS/MS (3, 4) allows rapid identification of many types of PTMs. However, data analysis and interpretation of MS/MS spectra for identification of PTMs remain a major challenge.

Early approaches to PTM identification using MS/MS involved exhaustive searches of all possible combinations of PTMs for each peptide from a protein database (5, 6). Because the search space grows exponentially as the number of PTMs increases, these early approaches performed a restrictive search that takes into account only a few types of PTMs during data analysis, ignoring all others. Investigators were obliged to guess the PTMs expected to exist in a sample prior to a search, and many potentially important PTMs may have been overlooked.

Various new approaches have been developed to increase the number of PTMs that can be identified during data analyses. VEMS (7) introduced an improved algorithm to reduce the search space, OpenSea (8) implemented a mass-based sequence alignment between database peptides and *de novo* interpretation, and TwinPeaks (9) improved the basic scoring scheme of SEQUEST (5), a popular database search program. But none of these approaches fully addressed the current limitations in the number of PTMs. A few tools were recently introduced for blind PTM search. MS-Alignment (10) predicts PTMs expected in a sample by spectral alignment between a database peptide and a spectrum followed by InsPecT (11) search. ModifiComb (12) introduced a ΔM histogram between unassigned spectra and base peptides found in a database. These blind approaches predict PTMs based on the frequency of mass shifts (indicating potential PTMs) in a sample. Thus, they all have the intrinsic weakness of missing rare PTMs infrequently observed that might provide important clues to understanding the function of a protein. Although many approaches have been developed to take into account many types of PTMs, most assume that there will be a single variable PTM per peptide and ignore multiply modified peptides. On the contrary, our studies with human glyceraldehyde-3-phosphate dehydrogenase (GAPDH) showed that there are many multiply modified peptides in a biological sample.

Here we describe a new algorithm, named MOD^i, that identifies multiple PTMs in a peptide while placing virtually no limit

From the ‡Department of Mechanical and Information Engineering, University of Seoul, Seoul, 130-743, Korea, ¶Center for Cell Signaling and Drug Discovery Research, College of Pharmacy and Division of Life and Pharmaceutical Sciences, Ewha Womans University, Seoul, 120-750, Korea, and ‖Computer Division, College of Information and Communications, Hanyang University, Seoul, 133-791, Korea

Received, March 11, 2008, and in revised form, August 6, 2008
Published, MCP Papers in Press, August 12, 2008, DOI 10.1074/mcp.M800101-MCP200

[1] The abbreviations used are: PTM, post-translational modification; GAPDH, glyceraldehyde-3-phosphate dehydrogenase; ISB, Institute for Systems Biology; Δmass, mass difference.

on the number of PTM sites and types. MOD[i] is essentially a sequence tag approach (13, 14). It constructs a partial sequence of a peptide from an MS/MS spectrum using *de novo* sequencing (15–17). MOD[i] differs from previous approaches in that it simultaneously uses multiple sequence tags derived from a spectrum by introducing a notion of a *tag chain*, a combination structure of multiple sequence tags. A tag chain offers an effective localization of modified regions within a spectrum and thus allows rapid identification of multiple PTMs in a peptide, obviating search space explosion by inspecting PTMs only in the modified regions of a peptide. The tag chain algorithm is robust against *de novo* sequencing errors, whereas most tag-based approaches depend critically on good *de novo* interpretations. This approach is scalable and performs well even when more than 400 modification types are considered and the number of potential PTMs in a peptide increases. Compared with established tools, MOD[i] reliably identifies a greater variety of modification types in multiply modified peptides and even detects modifications of low abundance.

MOD[i] has the potential to discover unknown modifications. As MOD[i] can take into account all known modifications, it effectively localizes parts of a spectrum that cannot be interpreted by the existing set of modifications. The localization algorithm of a tag chain method enables MOD[i] to discover novel modifications even when other PTMs exist in a peptide, unlike other blind PTM search methods. An example is shown to demonstrate the potential utility of MOD[i] in discovering novel modifications.

MATERIALS AND METHODS

Experimental MS/MS Data Sets

We analyzed three MS/MS data sets obtained from Q-TOF mass instruments. Details of sample preparation, separation, and MS/MS spectra acquisition have been described in previous work (8, 18, 20). We first tested data sets from high mass accuracy spectrometers to develop a sophisticated algorithm to exactly localize modification types and sites.

GAPDH—2863 MS/MS spectra were acquired analyzing GAPDH from transiently overexpressing cells using immunoprecipitation (18). The peak lists were generated using Micromass ProteinLynx 2.1 software. The effectiveness of MOD[i] to detect multiply modified peptides and those with uncommon modifications was demonstrated in this sample. Mascot (6) search (version 2.1) was done against the Swiss-Prot human database (version 50.5, 14,518 entries) to compare competencies. The search parameters were as follows: 0.5-Da tolerance for peptide and fragment ions; tryptic peptides; up to five missed cleavages; and propionamide (Cys), acetyl (N terminus), dimethyl (Lys), deamidation (Asn and Gln), oxidation (Met), phospho (Ser, Thr, and Tyr), and pyro-Glu (N-terminal Glu and N-terminal Gln) as variable modifications. After the first search, the error-tolerant Mascot search was additionally performed against GAPDH protein (Swiss-Prot Number P04406).

Lens Tissue—12,800 MS/MS spectra were acquired from the lens proteome of a 93-year-old human male with nuclear cataracts (8). The peak lists were generated using Micromass ProteinLynx software. This sample was chosen because crystallin proteins often become substantially modified post-translationally as this tissue ages. PTM identification for this sample had been studied extensively by others (8, 10, 19). We were thus able to compare the results of MOD[i] analysis with the results reported by others. A Mascot search was done against the Swiss-Prot human database. The search parameters were as follows: 0.5-Da tolerance for peptide and fragment ions; tryptic peptides; up to two missed cleavages; and carbamidomethyl (Cys), acetyl (N terminus), dimethyl (Lys), deamidation (Asn and Gln), oxidation (Met), phospho (Ser, Thr, and Tyr), and pyro-Glu (N-terminal Glu and N-terminal Gln) as variable modifications. After the first search, the error-tolerant Mascot search was additionally performed against the lens proteins.

ISB—The ISB standard protein mixture (20) was used to evaluate and optimize our scoring model. The raw data were converted into mzXML format and searched by Mascot. The search was done against the standard protein database appended with reverse sequences of the Internation Protein Index (human, version 3.36, 69,012 entries). The search parameters were as follows: 0.2-Da tolerance for peptide and fragment ions, carbamidomethyl-Cys as fixed modification, and no enzyme. Peptide assignments matched to one of the standard proteins above the homology score threshold were adopted as a training set (gold standard) to develop our scoring model.

MOD[i] Search

MOD[i] uses a protein database identified by unmodified peptides. This is based on the assumption that at least one unmodified peptide is present in sample proteins (10, 21). MOD[i] searches were conducted against GAPDH and lens proteins with search parameters of more than 400 variable modifications from Unimod, 0.5-Da tolerance for peptide and fragment ions, tryptic termini, and up to five missed cleavages. The mass range for peptide modification was specified from −100 to 250 Da. GAPDH data were searched against a database of human GAPDH protein appended with 29 *Escherichia coli* proteins (similar in size and mass to human GAPDH) as a decoy. Lens protein data were searched against a database of 13 crystallin proteins and their reverse sequence proteins. Even with hundreds of PTM types and a multitude of PTM sites as inputs, MOD[i] runs in a reasonable amount of time. The GAPDH search (2863 spectra, 30 proteins) on a regular Pentium IV personal computer (2.4 GHz, 1-GB memory) took about 600 s.

A false discovery rate of 1%, based on the target-decoy approach, was estimated, and a threshold score for peptide identifications was determined (22) to analyze the results. Manual validation was performed for all peptide identifications with an MOD[i] probability score of 0.5 or higher.

Overview of the MOD[i] Algorithm

The MOD[i] work flow is summarized in Fig. 1, including the sequential steps involved: 1) sequence tag generation (local *de novo* sequencing), 2) database search, 3) tag arrangement, 4) tag chain generation, and 5) PTM interpretation.

Sequence Tag Generation—We perform *de novo* sequencing to identify all the sequence tags (partial amino acid sequences that do not contain PTMs). To reduce noise in a spectrum during *de novo* sequencing, we use only high intensity peaks selected globally or locally from a spectrum. First top N peaks are selected according to their intensities over the entire m/z range of a spectrum where N is related to a precursor ion mass (global selection). Second we perform local selection. Additional peaks are selected by sliding a window of 70 Da (window increment, 35 Da) when fewer than two peaks are selected in any window during the global selection. Thus, we retain at least two peaks in every window. Four virtual peaks are added to represent starting (1, 19) and ending (MW − 17, MW + 1) positions in *de novo* sequencing where MW is a precursor ion mass.

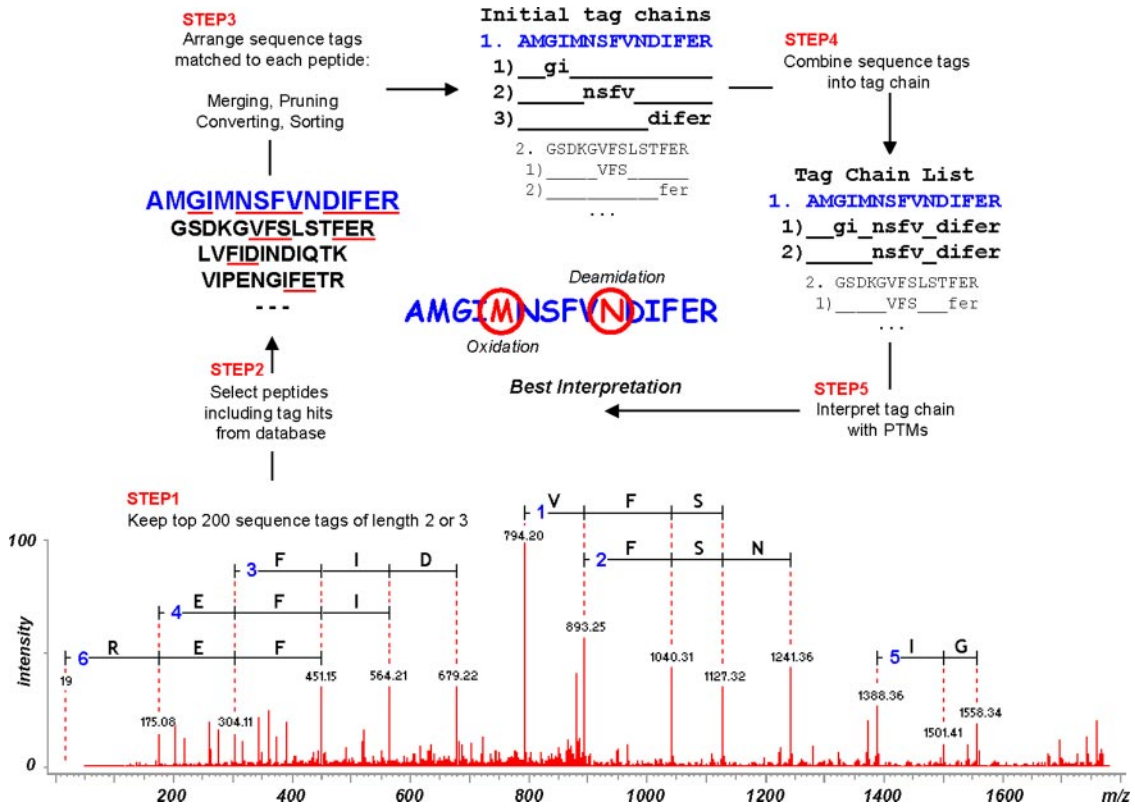

FIG. 1. **Overview of MOD[i] algorithm.** The program work flow is shown. *Step 1*, local *de novo* sequencing is used to generate sequence tags. The top 200 sequence tags of length 2 or 3 are selected by their scores. *Step 2*, the generated tags are matched with peptides in both forward (b ion series) and reverse (y ion series) directions. Subsequently forward and reverse tags are written by capital and small letters, respectively. In this figure, six tags matching the peptide AMGIMNSFVNDIFER were determined as y ion type. Tag hits are *underlined* in *red*. *Step 3*, tags matched to each candidate peptide are arranged by merging, pruning, and sorting by position. For example, tags 3 (dif), 4 (ife), and 6 (fer) are merged to a longer tag (difer) because they overlap in residues and share the peaks. This operation constructs an initial set of tag chains. *Step 4*, more complex tag chains are produced based on predefined rules according to alignment relationship between two tags. *Step 5*, for gaps (*underlined* regions within a tag chain) with non-0 Δmass, mass ambiguities are interpreted using a PTM database.

After peak selection, we construct a spectrum graph (15), a directed acyclic graph, using the selected peaks where a vertex represents a mass of a fragment ion peak and there is an edge when a pair of vertices differs by a certain amino acid in mass. A subpath in the graph is a possible sequence tag. A subpath can start at any vertex. We extract 200 subpaths of length 2 or 3 from the graph by their score. A score of a subpath is the sum of confidence measures of each vertex belonging to the subpath where a confidence measure of a vertex is the sum of normalized intensities of the peak corresponding to the vertex and its supporting peaks ($-H_2O$, $-NH_3$, isotope, and complementary). In this step, Ile and Leu can be substituted for each other. This also applies to Gln and Lys.

Database Search (Candidate Generation)—We search the peptide database using the identified sequence tags in the *de novo* sequencing stage. The peptide database does not include any information on PTMs. All the peptides matched with any tag are obtained as candidates. We look not only for peptides including the identified tags (called "forward" tags) but also for those including the reverse sequences of the identified tags ("reverse" tags). This is because we do not know whether component peaks of the tags are N-terminal or C-terminal peaks. The type (forward or reverse) of a tag can only be determined when a tag is matched to a candidate peptide. When a tag matches a peptide, it requires only the match to the subsequence of the peptide, not the position (mass distance from peptide termini) of a sequence tag within the spectrum. We preprocess the database using a suffix tree data structure (23) for rapid scanning of the peptide database. The suffix tree allows us to match a tag to any peptide in the database in a constant time regardless of the database size.

Tag Arrangement—After the database search, each candidate peptide has a list of sequence tags matched to it. Here we add to the list of each peptide one special tag, a forward N-terminal tag of length 2, if there exists a b2 ion of a peptide in the spectrum. b2 ions are commonly observed in the lower mass range of a spectrum. The N-terminal tag including b2 ion is almost never identified because of the absence of a b1 ion. Near the N-terminal end of a spectrum, ion peaks normally have very low intensity or are not observed at all (24).

We then start merging the sequence tags matched to each candidate peptide. If two tags are of the same type (the type of a tag was determined during the database search step), overlap by their residues, and share the same peaks in the overlapping parts, they are merged to a new single longer tag and are removed from the list. Redundant tags inferred from both b and y ions are removed (only the highest scoring tag survives), whereas redundant tags likely to be inferred from neutral losses are retained. Then the masses of the reverse tags are converted to N-terminal masses according to the prefix residue masses, and the tags are sorted by their positions within the peptide.

Tag Chain Generation—In a candidate peptide under study, a tag chain is an alternating list of sequence tags (subsequences of the peptide) and gaps where a gap is a region of the peptide that does not

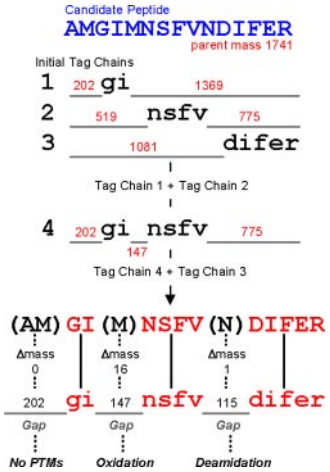

FIG. 2. **Tag chain generation.** Three reverse tags are shown. Both flanking regions, marked by *gray underlines*, represent gaps (there is only one flanking region when a tag starts from the N terminus or ends in the C terminus), and *red numbers* in these gaps represent the size of the flanking regions in terms of mass. Combination of tags 1 and 2 into tag chain 4 introduces a new intermediate gap with a mass size of 147, calculated by 519 − (202 + 170: mass of the subsequence gi in tag 1). From this, Δmass of the gap is calculated. The gap has the subsequence M and mass size 147, thus Δmass (16) is obtained by subtracting its subsequence mass (131) from its mass size (147).

correspond to any tags in the tag chain but may contain modifications. Fig. 2 shows how to combine multiple sequence tags into a tag chain and identify multiple modifications using tag chains. Three tags, gi, nsfv, and difer, are aligned to a candidate peptide, AMGIMNSFVNDIFER, and form a tag chain, __gi_nsfv_difer. Gaps are introduced in between the tags during tag chain generation. Each gap has two main attributes: 1) a subsequence of a candidate peptide that is not covered by the two neighboring tags and 2) a mass difference (Δmass) obtained by subtracting the mass of this subsequence from the difference between the flanking mass regions of the two tags. If Δmass is 0, the gap can be explained by its subsequence without PTMs. If Δmass is not 0, the gap is likely to include PTMs. Fig. 2 depicts three gaps (<AM, △mass: 0>, <M, △mass: 16>, and <N, △mass :1>) in the tag chain __gi_nsfv_difer.

PTM Interpretation—Once tag chains are generated, PTMs are interpreted for all the gaps in a tag chain. For gaps with non-0 Δmass, we retrieve, from a PTM database, a set of PTMs that can best interpret Δmass of the gap. The PTM database contains a list of known post-translational and chemical modifications. These are indexed by mass difference, occurrence residues, and occurrence positions (N terminus, C terminus, or anywhere in the peptide or protein). In the final tag chain of Fig. 2, the first gap <AM, △mass: 0> is estimated to have no PTMs, and the second gap <M, △mass: 16> and the third gap <N, △mass: 1> are estimated to have an oxidized Met and a deamidated Asn, respectively.

If a gap contains multiple residues, several interpretations can be inferred from various combinations of PTMs occurring at multiple sites. For example, a gap <KK, △mass: 28> can have three interpretations: 1) dimethyl at Lys^1, 2) dimethyl at Lys^2, or 3) monomethyl at both Lys^1 and Lys^2. Of these possible interpretations, the one that best explains the candidate peptide is selected. (A detailed description can be found under "Candidate Peptide Scoring"). During PTM identification, to avoid rechecking the same attribute of gaps, we cache gaps once they are analyzed (The same peptides in a sample are likely to reproduce similar spectra (25), resulting in construction of similar tag chains.). Only when an analyzing gap is not found in the cache (mishit) do we calculate PTM interpretations using the PTM database, and a new gap is added to the cache. If a gap cannot be explained by any set of PTMs although the spectrum is of high quality, we can generate a profile of such unexplainable gaps based on their attribute values. This profile can be used to predict novel modifications.

Error-tolerant Tag Chain Generation

A tag chain method is effective in identifying multiple modifications. During the tag chain generation, we align all the identified sequence tags pairwise. Possible alignment relationships between tags are defined as follows: 1) *separated*, 2) *overlapping*, 3) *adjacent*, and 4) *single long tag*. Depending on the alignment relationship between tags, one can infer whether identified tags include errors incurred by *de novo* algorithms. *De novo* sequencing approaches based on greedy algorithms typically yield errors because of missing or randomly matching peaks in a spectrum. We observed that *de novo* sequencing errors are more likely to occur in modified spectra. This is mainly because of peaks related to different types of ions (24) or prompt loss of modification from modified residues (26) during collision-activated dissociation. Pairwise alignments between two tags offer important information in correcting these *de novo* sequencing errors and predicting PTMs. We defined the combination rules for tag chain generation according to the alignment relationship between two tags. In Fig. 2, three tags were combined into a tag chain using the rule associated with a "separated" relationship for example.

Separated Relation—In this relation, when two tags are aligned, there is a gap (a subsequence of a candidate peptide) between the two tags. That is, two tags cannot be linked into a single sequence of residues in a candidate peptide. In the candidate peptide VVKQASEGPLK of Fig. 3*a*, there is a subsequence K between the two tags vv and qaseg. This indicates that the residue K (which becomes the subsequence corresponding to an in-between gap when the two tags are combined) is modified. Unlike this example, if a subsequence corresponding to a gap is longer than two residues, it can indicate that fragment ion peaks are missing. When two tags satisfy separated relations, the combination of two tags is relatively simple. A tag chain is constructed by defining a subsequence between the tags as a new gap. Fig. 3*a* shows the combination of the two tags into a tag chain.

Adjacent Relation—In this relation, when two tags are aligned, there is no gap between the two. In Fig. 3*b*, the two sequence tags can be concatenated to a longer subsequence (INGNPITIFQ) of the candidate peptide. However, such a case must contain an inconsistency because we have already merged all the sequence tags that are linkable by residues during the tag arrangement step. This arises when the two tags do not share a fragment peak. We observed that this was often due to random peak matching at either end of a sequence tag.

Fig. 3*b* shows the identification of a peptide with succinimide at the Asn^6 residue. Initially we start with two sequence tags, ingn and pitifq. At this point, it seems possible to identify all the fragment peaks for the subsequence INGNPITIFQ of the candidate peptide. The rightmost N of the tag ingn was identified by the peak at 986 *m/z*, and the leftmost P of the tag pitifq was identified by the peak at 1003 *m/z*. Both tags cannot be simultaneously matched to the candidate peptide as it includes a mass inconsistency of −17 Da at the same fragmentation site between N and P of the two tags. This has arisen from matching N (from 1100 *m/z*) of the tag ingn to a neutral loss ($-NH_3$) peak of P (at 1003 *m/z*) of the tag pitifq during *de novo* sequencing.

When observing this adjacent relation, we assume that either side of the two adjacent tags has come from a random match. Assuming one or the other is a random match, we construct two tag chains from the two sequence tags as shown in Fig. 3*b*.

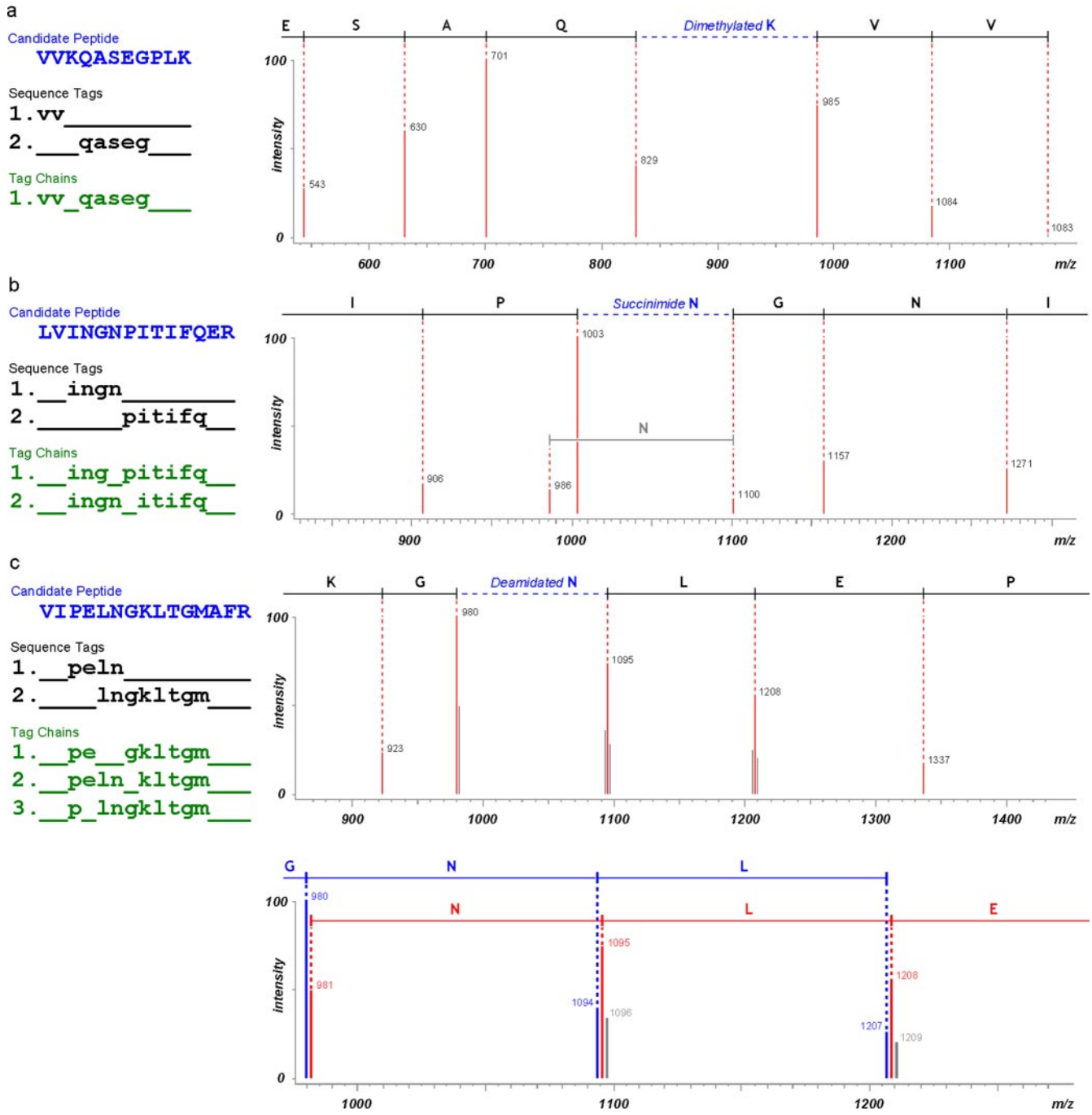

FIG. 3. **Combination rules of tags into a tag chain.** *a*, separated relation. For the candidate peptide VVKQASEGPLK, its tag chain and interpretation of PTMs are shown. After *de novo* sequencing followed by tag merging, we obtained two arranged sequence tags, vv and qaseg (sequence tags 1 and 2). The tags were identified from y fragment ions. When the two tags are combined into a tag chain, a newly defined gap has the subsequence K and Δmass of 28. As a result, the gap is interpreted as dimethylated Lys. *b*, adjacent relation. *De novo* sequencing often accompanies erroneous sequence tags that complicate the subsequent analysis. In this figure, the tag ingn contains a random match at its end. However, this can be corrected by the relation between two neighboring tags. From two sequence tags, we construct two tag chains, __ing_pitifq__ and __ingn_itifq__. The second gap of tag chain __ing_pitifq__ has the subsequence N and Δmass of −17. As a result, the gap is correctly interpreted as succinimide on Asn[6]. *c*, overlapping relation. From tag chain 1 __pe__gkltgm___, we identified the candidate peptide VIPELNGKLTGMAFR deamidated on its Asn[6]. The second gap of the tag chain has the subsequence LN and Δmass of 1. The spectrum at the *bottom* shows the initial peak matching that resulted in an overlap over LN.

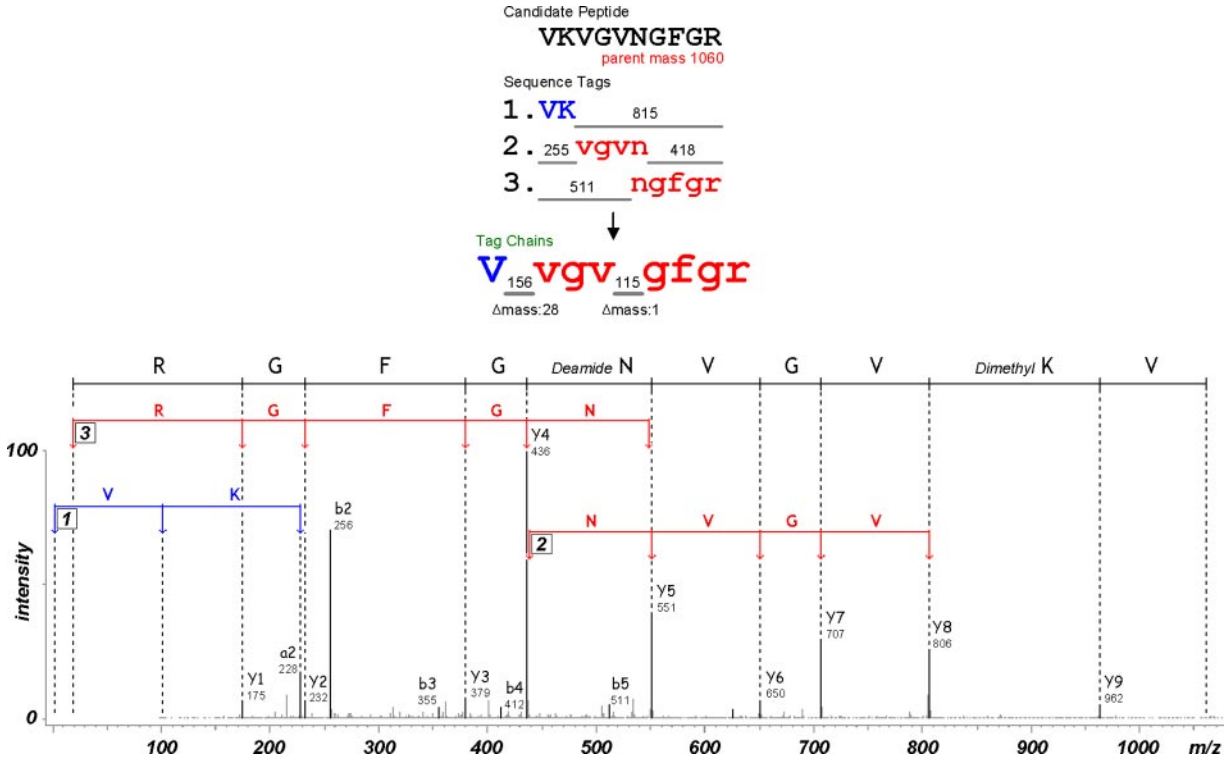

FIG. 4. **Tag chain generation using combination rules.** Sequential steps are shown by which a tag chain is generated to identify the peptide VKVGVNGFGR dimethylated at residue Lys[2] and deamidated at Asn[6]. Initially of the three tags (b ion and y ion tags are shown in *blue* and *red*, respectively), tag 1 has an erroneous match to a2 ion of the correct identification. Tags 2 and 3 have erroneous matches (around 500 *m/z*) by peak location shown in Fig. 3*c*. Our algorithm for gap generation sequentially corrected these errors by the "adjacent" rule (tags 1 and 2) and "overlapping" rule (tags 2 and 3) and constructed the final tag chain that led to the correct identification. For modified regions in the spectrum, our algorithm generated the two gaps, <K, △mass: 28> and <N, △mass: 1>.

Overlapping Relation—An overlapping relation is very much like an adjacent relation. In this relation, although two tags overlap by their residues, they do not share the fragment peaks in the overlapped part. In Fig. 3*c*, the two sequence tags overlap by residues LN. In this relation, we assume three possible situations for the two aligned tags. 1) The overlapping part of both tags is incorrect. 2) One tag is incorrect. 3) The other tag is incorrect. For case 1, we define the overlap of the two tags as a new gap as exemplified in tag chain 1 of Fig. 3*c*. In cases 2 and 3, the inconsistency of the overlap region is attributed to either one of the two tags. As a result, tag chains 2 and 3 of Fig. 3*c* are constructed as in the case of the adjacent relation. One might argue that we could simply have constructed only a single tag chain __p____kltgm___; this includes all possible interpretations derivable from the three tag chains above. However, in view of the complexity of the possible combinations for gap interpretation, constructing three tag chains while retaining shorter multiple gaps is more advantageous than a single tag chain with a longer gap. We cache gaps once they are analyzed to avoid rechecking the same attribute (subsequence and Δmass) of gaps. The shorter the gap length, the higher the hit ratio of gaps in the cache.

Single Long Tag—As the length of a sequence tag gets longer, its probability of being correct decreases (17). To avoid a potential error in a long sequence tag, we may add one shorter tag if there is only one single long tag and no other tags exist (that is, in the tag arrangement step, all the tags matched to the peptide were merged into one longest tag). If either gap flanking this single long tag has a non-0 Δmass, we do the following. First we scrutinize the terminal peak of the long tag that borders the gap of non-0 Δmass. If the intensity of the peak is not superior compared with the other peaks around it, the peak is removed, and thus the length of the tag is reduced by 1. This process is iterated until we are left with credible peaks flanking any non-0 gaps while requiring that the length of a new shorter tag be at least 3.

Application of Combined Rules—Fig. 4 shows the application of our algorithm to a multiply modified peptide, VKVGVNGFGR, dimethylated at Lys[2] and deamidated at Asn[6]. We initially obtained three sequence tags (VK, vgvn, and ngfgr) from the *de novo* sequencing step, but all three included random matches that had arisen from modifications. Any of these errors can cause one to miss the exact peptide identification, but our algorithm detected the errors, corrected them, and successfully identified the correct modifications and their sites. The final tag chain consisting of b ion and y ion tags was constructed, and it exactly localized the modified regions in the spectrum. Although most tag-based approaches depend crucially on good *de novo* interpretations, we used the information from *de novo* sequencing errors and developed a robust algorithm against them. This is possible because MOD[i] takes advantage of both *de novo* sequencing and the database search approach.

Candidate Peptide Scoring

Each interpretation of a tag chain is scored and ranked. Clearly scoring is a key issue for any computational method for MS/MS interpretation. Numerous works on this subject exist (25, 27). However, proteomics analyses of samples with many types of PTMs result in a large number of false positives because of the combinatorial increase in the number of possible matches (28). To overcome this

problem, we developed a robust scoring function that penalizes the number of PTM occurrences and different PTM types assigned to a peptide (Occam's razor). That is, if one interpretation involves a single PTM occurrence whereas another uses two PTM occurrences, we prefer the former interpretation. If the number of PTM occurrences is the same, we prefer interpretation involving the same PTM over those involving different PTMs. In addition, the presence of known immonium ions for each PTM is checked. For example, peptides containing oxidized Met can be supported by a loss of methane sulfenic acid (64 Da), and peptides containing phosphorylated Ser and Thr can be supported by a loss of phosphoric acid (98 Da) (29).

A match quality of peptide identification can best be evaluated taking into account various properties of fragmentation patterns in MS/MS spectra (29, 30). Our scoring model consists of four scoring components: 1) ion score, 2) standard deviation of mass errors of matched fragment ions, 3) score of explained intensities, and 4) score of explained highest peaks. Finally the four component scores are combined by a logistic regression model. This represents a weighted linear sum of components as a probability of a correct match using a logistic function. The scoring model was established and validated using the ISB standard protein mixture data set (20).

Ion Score—This score is calculated from the matched peaks between the experimental spectrum and the theoretical spectrum. Our theoretical spectrum is made only of b/y ions and a single a2 ion (−28 Da) because a2 ion is often observed as the characteristic a2/b2 ion pair in the lower mass range (4). A binomial distribution is assumed to regulate the probability of matching k peaks among N total peaks of a theoretical spectrum. The cumulative binomial probability P is calculated using the total number of fragment ions within a theoretical spectrum (N), the number of ions matched to the MS/MS spectrum (n), and the probability of matching ion (p) as in Equation 1,

$$P(X \geq n) = \sum_{k=n}^{N} \binom{N}{k} p^k (1-p)^{N-k} \qquad \text{(Eq. 1)}$$

where P represents the probability of randomly matching at least the given number of fragment ions (n) to the MS/MS spectrum.

First an MS/MS spectrum is separated into windows of 100 m/z units. Within each window, only top i peaks are retained according to their intensity where $1 \leq i \leq 10$. For each i, P is calculated using the uniform probability of matching ion peak ($p = i/100$), and ion score is calculated as $-10\log(P)$. This process is repeated for each i (This model was suggested by Beausoleil *et al.* (31)). The final ion score is an average of all 10 scores. If the given peptide is modified, the final score is penalized by the number of PTM occurrences and the number of PTM types.

Standard Deviation of Mass Errors of Matched Fragment Ions—In MS instruments with high mass accuracy, mass measurement errors can be systematically predicted (32). For example, for a Q-TOF machine, mass errors between observed and theoretical fragment peaks often grow as their m/z values increase (33). Weighted linear regression based on the intensity of the fragment ion is conducted for mass errors of matched fragment ions (ion matches were forced to be one of the top 10 peaks by intensity within a 100-Da window), and observed fragment masses are recalibrated to compensate for this error. The intensity of an ion is normalized by the highest intensity in radius 50 Da around itself. After the first regression, the standard deviation of new mass errors between recalibrated and theoretical peaks is obtained (Here we assumed that mass errors follow a normal distribution with a mean of 0.). Matched peaks with mass errors that are 3 times larger than the standard deviation are removed as outliers and are not used in calculating the ion score. Then with only the remaining peaks, the final standard deviation is obtained again.

Score of Explained Intensities—This is the fraction of the total ion current of annotated peaks in a spectrum. In this step, supplementary ion peaks are annotated. Only for matched b/y ions are their supporting peaks ($-H_2O$, $-NH_3$, and isotope) retrieved. Immonium ions for amino acids and PTMs in a candidate peptide (7) are also a part of the annotated peaks.

Score of Explained Highest Peaks—This is the fraction of high intensity peaks that are annotated. An MS/MS spectrum is separated into windows of 100 m/z units. Within each window, only the top i peaks are retained according to their intensity. For each i, this score is calculated as the ratio between the number of the annotated peaks and the number of the total retained peaks. This process is repeated for each i for $1 \leq i \leq 10$. The final score is a sum of all 10 scores.

Logistic Regression Model—Fig. 5 shows performance characteristics of each component score in distinguishing a correct peptide from a false one. To calculate a match score for a candidate peptide, four component scores are combined by the logistic regression model, the result of which represents a probability of a true match by the use of a logistic function over a weighted linear sum of components. The MS/MS data set from ISB protein mixture was used to train the logistic regression model. The training set consisted of 5520 correct and 5520 incorrect (top scoring) peptide matches by Mascot search. Fig. 5*e* shows the performance of the final score.

False Positives by PTM Combinations?—MOD^i considers so many types of PTMs during peptide identification that its results may be accompanied by many false positives because of the combinations of many PTMs. We evaluated MOD^i against 1157 unmodified peptides from the lens sample confidently identified by Mascot search. 1150 MOD^i peptide identifications were consistent with Mascot results. Of them, 1094 were identified confidently above our score threshold (0.5), and 56 were not. Conversely of seven peptides inconsistent with Mascot, three were identified as having PTMs (but below the score threshold), and four were missed. These results show that the tag chain approach and our scoring model are robust against false positives by PTM combinations despite the fact that MOD^i takes so many types of PTMs. Modified peptides identified by MOD^i are listed and are compared with Mascot in the supplemental Table 2.

Software Implementation

MOD^i is implemented in Java programming language and is available on line. A graphical tool to annotate MS/MS spectra is available (34). It is recommended to use MOD^i for data sets from high mass accuracy instruments.

RESULTS

MOD^i Application to PTM-rich Data Sets—We analyzed richly post-translationally modified data sets, GAPDH and lens proteins, to validate our tag chain approach for identifying multiple modification sites. MOD^i search was conducted against the two data sets, allowing more than 400 variable modifications. For comparison with an established tool, a Mascot search was also conducted, allowing nine common modifications, followed by an error-tolerant search.

GAPDH, an enzyme that plays a pivotal role in glycolytic energy metabolism, is extremely sensitive to modification of the cysteine residue (Cys^{152}) located in its active site. Recent studies indicate that GAPDH, which is distributed throughout the whole cell including nucleus, cytosol, and membrane, plays roles, in addition to glycolysis, in membrane fusion, microtubule bundling, and phosphotransferase activity and is

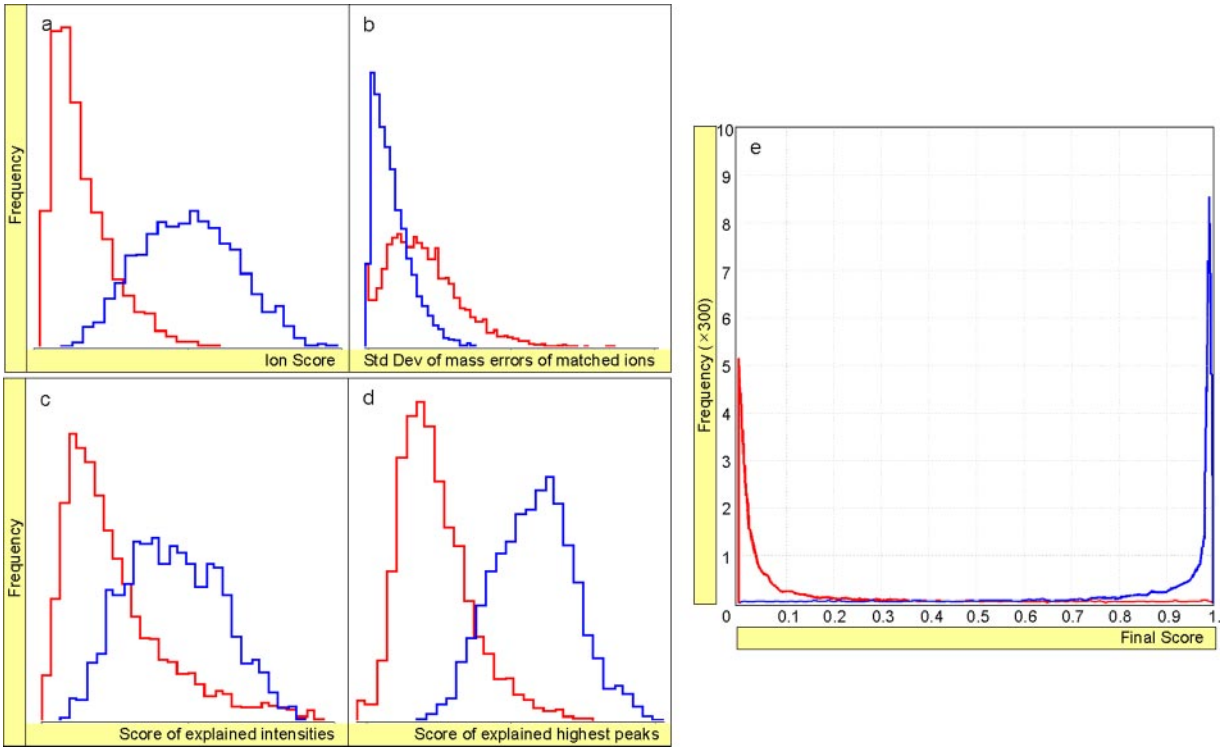

FIG. 5. **Score performance.** Component scores (*a–d*) and the final score (*e*) distinguish correct peptides (*blue*) from false ones (*red*). The combination of four component scores is converted into estimates of probability of a correct peptide match. *Std Dev*, standard deviation.

involved in various nuclear processes. These multiple cellular functions of GAPDH can be attributed to the existence of various structures induced *in vivo* by multiple post-translational modifications recently identified by mass spectrometry using selectively excluded mass screening analysis (18).

MOD^i search results for GAPDH are summarized in Fig. 6*a*. Seventy unique peptides were identified. They consisted of 11 unmodified peptides, 34 with one modification, 19 with two modifications, and six with three modifications, an array of 23 disparate types of modifications on 39 sites. The unique peptides are listed in supplemental Table 1, and their annotated MS/MS spectra are shown in supplemental Data 1. The identification of peptides containing common modifications were validated by comparison with Mascot results (identifications with scores greater than the homology score). New types of modifications were verified by manual inspection. It should be noted that the peptide ^{146}IISNASCTTNCLAPLAK^{162} containing the active site C*XXX*C displayed 11 kinds of modifications, including phosphorylation (Ser^{148}), disulfide linkage (Cys^{152}-Cys^{156}), amino acid substitution (Cys^{152} → Ser), and cysteic acid (Cys^{152}). Another peptide ^{225}VPTANVSVVPLTCR^{248} contained six kinds of PTMs.

UniProt database suggests five phosphorylation sites (Tyr^{42}, Ser^{83}, Thr^{211}, Tyr^{314}, and Tyr^{320}) for GAPDH, but three of these sites (Thr^{211}, Tyr^{314}, and Tyr^{320}) are obtained by prediction based on a similarity search. Peptide containing Tyr^{42} was not detected in our experiment. Phosphorylation at Ser^{83} was not entirely clear in the MS/MS spectra from our oxidized GAPDH. This is not a limitation posed by MOD^i performance. When searched using Mascot, none of the suggested phosphorylation sites were identified either.

GAPDH analysis showed that PTMs in a peptide exist in a greater variety than expected and that our tag chain algorithm can be successfully applied for proteins with multiple PTMs. It effectively localized modified regions from the spectrum and constructed small multiple gaps for multiply modified peptides during the tag chain generation (the gap generation algorithm is presented in detail under "Materials and Methods"). It allowed us to rapidly identify multiple PTMs in a peptide, obviating the large search space arising from enumeration of all possible modifications. Fig. 6*b* shows the possible identification of a peptide with four modifications. The tag chain has four gaps. Each gap has a length of at most 3.

Lens proteins are known to undergo multiple PTMs depending on age. Many modifications have been characterized (19). MS/MS data from human lens proteins on Q-TOF mass instruments (8) were analyzed using MOD^i. The results are summarized in Fig. 7. For modified peptides, MOD^i identified 321 unique peptides, whereas a Mascot error-tolerant search identified 191 unique peptides (Results were verified by manual inspection. MOD^i missed only five peptides among Mascot identifications.). The modified peptides identified in crystallin proteins are listed in supplemental Table 2, and their annotated MS/MS spectra are shown in supplemental Data 2. MOD^i results contain many more modifications of a greater

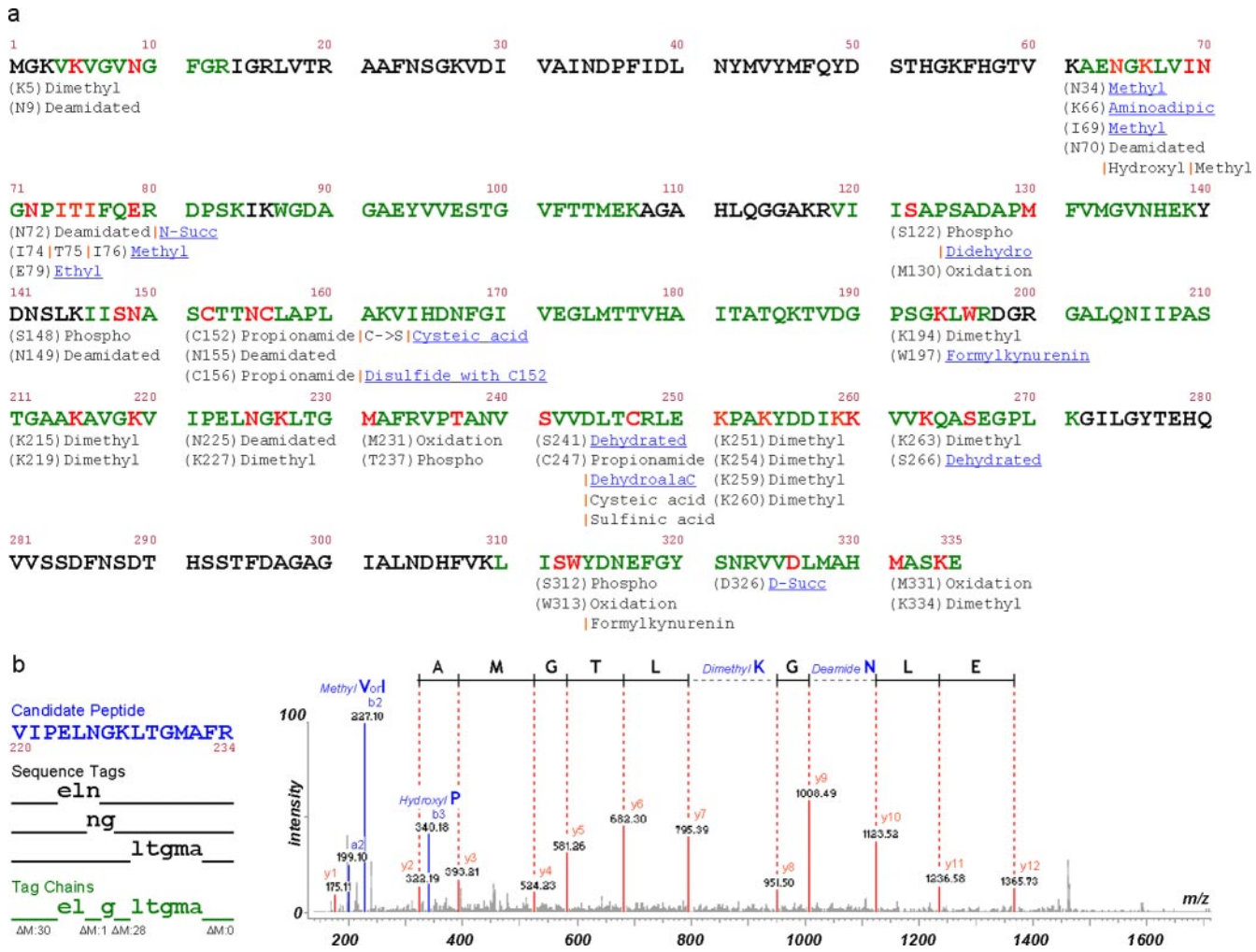

FIG. 6. **Identifications of PTMs in GAPDH protein.** *a*, identified sequences are shown in *green*, and modified sites are shown in *red*. We identified 23 types of modifications on 39 sites. Modifications identified only by MOD[i] and not by Mascot search are *underlined blue* (16 types of modifications on 11 sites). *b*, a tag chain to identify multiple modifications is shown. The peptide has four modified sites. The tag chain localized four gaps of length at most 3 from the spectrum.

variety than reported previously. MOD[i] is sensitive to low abundance modifications in this sample, such as acetylation (Lys, non-N-terminal) and oxidation (His) not reported previously by other PTM tools (8, 10, 19). The presence of such modifications is confidently supported by their MS/MS spectra in Fig. 7. Notably the peptide ^{91}VKVLGDVIEVHGK^{103} in αB-crystallin (water-soluble) is multiply modified as carbamylated at its N terminus and acetylated at Lys^{92} (35). MOD[i] successfully identified peptides modified at consecutive sites, whereas other tools for blind PTM search rejected these peptides. This demonstrates that MOD[i] successfully eliminated the limit on the number of both modification types and sites, but no other tools have simultaneously relaxed these two limitations.

Analyses with many types of modifications suffer from a large number of false positives because of the combinatorial increase in the number of possible matches. However, in comparison with a Mascot search against GAPDH and lens samples, MOD[i] showed its performance competence in localizing modified sites. It also demonstrates that error-tolerant generation of a tag chain by comparing the positions of multiple tags is very sensitive to modified regions of the spectrum.

Discovery of Novel Modification—Use of MOD[i] led to the identification of additional PTMs and revealed the multiplicity of modifications in a proteome sample. Simultaneously we want to stress the potential of MOD[i] in discovery of unknown modifications.

Fig. 8 shows a putative identification of a novel modification (+12 Da at the N terminus, Leu, or Val) (37) in peptide ^{67}LVINGNPITIFQER^{80} of GAPDH, confirming the effectiveness of the tag chain approach in discovering novel modifications. Our tag chain algorithm exactly localized the modified regions within the spectrum. It can be seen that almost all of the y fragment ions of the peptide are assigned to intense peaks in the spectrum and that the b2 fragment ion corresponding to the LV + 12

Modifications	Mass Diff	AA
Glu->pyro-Glu	-18	N-terminal E
Dehydrated	-18	S, T
Gln->pyro-Glu	-17	N-terminal Q
Succinimide	-17	N
Deamidated	1	N, Q
Trp->Kynurenin	4	W
Methyl	14	C, H
Oxidation	16	H, M, W
Formyl	28	H, S, T
Formylkynurenin	32	W
Potassium adduct	38	D, E, C terminus
Acetyl	42	N terminus, K
Carbamyl	43	N terminus, K
Carboxy	44	W
Carboxymethyl	58	K
Carboxyethyl	72	K
Phospho	80	S

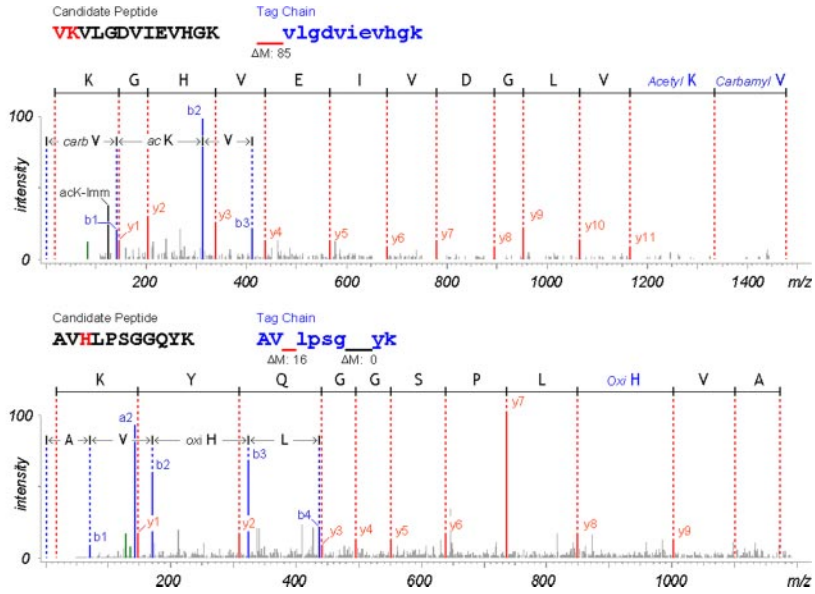

FIG. 7. **Modifications identified in lens proteins from a 93-year-old human male with nuclear cataracts.** This list was confirmed by manual validation, and it includes previously reported lens modifications as well as modifications additionally identified by MOD[i] (see supplemental Table 2). Also an unknown modification (+55 Da) of Arg was discovered (10). Representative MS/MS spectra for acetylated Lys (Lys[92] in αB-crystallin) and oxidized His (His[86] in βS-crystallin) are shown, and their corresponding fragment ions provide confirmation of the modifications. The acetylated peptide is also carbamylated at its N terminus. Modified b1 and b2 fragment ions are observed, and an immonium ion (*acK-Imm*) at 126 *m/z* corresponding to an acetylated Lys is confidently observed (36). These examples demonstrate that each tag chain is sensitive to modified regions. *Diff*, difference; *AA*, amino acids; *carb*, carbamyl.

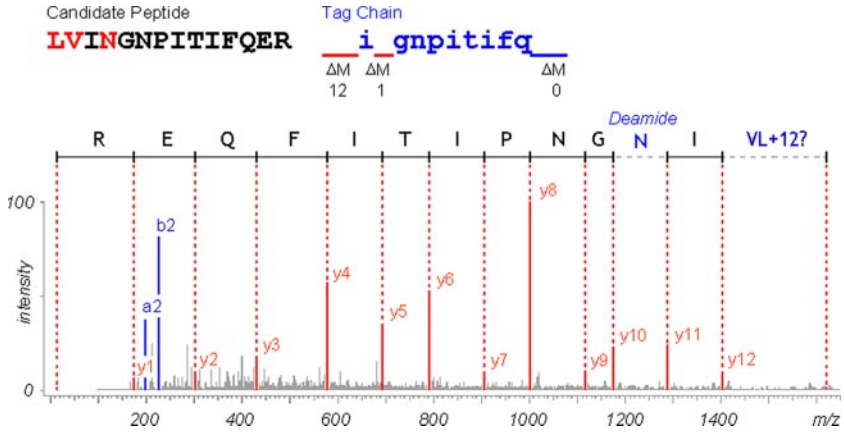

FIG. 8. **Discovery of a novel modification.** An uncharacterized mass of 12 Da at the N terminus, Leu, or Val is shown. This was discovered in the peptide [67]LVINGNPITIFQER[80] of GAPDH. Peptide identifications with +12 Da at their N-terminal LV were observed several times. In this example, most of its y ions were assigned to intense peaks in the spectrum. b2 ion corresponding to the fragment LV + 12 was confidently observed together with a2 ion. Our tag chain exactly localized two regions modified in the spectrum. One of them, <N, △mass: 1>, was interpreted as deamidated Asn. The other <NtermLV, △mass: 12> could not be explained by any combination of known modifications, suggesting a novel modification. Such prediction for various types of PTMs, including unknown ones, was possible by using multiple local gaps.

fragment is confidently observed together with the a2 fragment ion. The appearance of the a2/b2 fragment ion pair strongly supports the existence of a novel modification (4). It is worth emphasizing that the peptide is deamidated at Asn[70]. MOD[i] considers all possible modifications and then more or less pinpoints the potential modification site within a peptide (*i.e.* a single gap in a tag chain). Users can inspect only the unexplained local gaps, not the entire unidentified spectrum. An understanding of the comprehensive maps of the unexplainable gaps within a sample provides a vision for the novel modification discovery. This epitomizes the strength of a tag chain algorithm. Such perspectives cannot be offered by approaches that predict novel modifications by comparing spectra with those from unmodified peptides.

DISCUSSION

Advances in MS/MS allowed rapid generation of peptide MS/MS spectra, but the existing MS/MS search approaches

faced significant computational challenges especially when interpreting modified spectra. We introduced an unrestrictive algorithm to greatly reduce computational complexity in identifying modified spectra and demonstrated the utility of a tag chain for multiple PTM identification. The key idea of a tag chain is the combination of multiple sequence tags from an MS/MS spectrum.

In a candidate peptide under study, a tag chain is an alternating list of sequence tags and gaps where a gap is a region of the peptide that does not correspond to any tags in the tag chain but may contain modifications. Localizing PTMs only to gaps is extremely important in terms of computation because possible combinations of different PTMs grow exponentially as the size of potential regions for PTMs increases. Reduction of a search space, often by several orders of magnitude, is possible by combining as many tags as possible into a tag chain, thus minimizing the sizes of gaps in a tag chain.

The simultaneous use of multiple sequence tags can provide additional advantages. We showed that a tag chain is robust against *de novo* sequencing errors. Local sequence tags are useful in retrieving peptides from a protein database, but the position (mass distance from peptide termini) of a sequence tag has to be determined carefully. During collision-activated dissociation, a peptide is expected to be fragmented at amide bonds, but many supplementary fragment ions can be produced by unexpected dissociation pathways. As a result, polymorphous (having the same partial sequence but shifted by mass) sequence tags can be obtained from continuous internal ions or neutral loss ions, and an erroneous sequence tag can be formed by different types of ion peaks; this may lead to peptide identification pitfalls. The simultaneous use of multiple sequence tags enables each tag to be localized exactly by comparing tag positions.

Although we have shown only the results from applying our approach to MS/MS spectra obtained from Q-TOF instruments, we expect that our approach is applicable to other instrument types. We have tested data sets from various instruments and have confirmed that MOD^{i} could be successfully applied to the data (high accuracy precursor masses and low accuracy MS/MS spectra) from Thermo LTQ-FT and LTQ-Orbitrap mass spectrometers. We are investigating the possibility of using MOD^{i} against low resolution ion trap data with precursor ion mass corrections.

In summary, MOD^{i} enables a rapid search for all known types of PTMs, introducing a novel notion of a tag chain. This enables the management of the computational complexity of multiple PTM identification. The localization algorithm of a tag chain can serve as an effective platform to identify multiple PTMs and discover novel modifications in a peptide.

Acknowledgment—We thank Prof. Seung-Ho Kang of Ewha Womans University, Korea, for helpful comments on statistical analysis.

* This work was supported in part by 21C Frontier Functional Proteomics Project from Korean Ministry of Education, Science and Technology (Grants FPR08-A1-020 and FPR05-A2-480) and by the Korea Science and Engineering Foundation through the Center for Cell Signaling and Drug Discovery Research (Grants R15-2006-002 and R15-2006-020) at Ewha Womans University. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

S The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

§ Supported by the Brain Korea 21 Project and a Seoul science fellowship.

** To whom correspondence should be addressed. Tel.: 82-2-2210-2680; Fax: 82-2-2210-5575; E-mail: paek@uos.ac.kr.

REFERENCES

1. Mann, M., and Jensen, O. N. (2003) Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* **21,** 255–261
2. Cantin, G. T., and Yates, J. R. (2004) Strategies for shotgun identification of post-translational modifications by mass spectrometry. *J. Chromatogr. A* **1053,** 7–14
3. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
4. Steen, H., and Mann, M. (2004) The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* **5,** 699–711
5. Eng, J. K., McCormack, A. L., and Yates, J. R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976–989
6. Perkins, D. N., Pappin, D. J. C., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567
7. Matthiesen, R., Trelle, M. B., Hojrup, P., Bunkenborg, J., and Jensen, O. N. (2005) VEMS 3.0: algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins. *J. Proteome Res.* **4,** 2338–2347
8. Searle, B. C., Dasari, S., Wilmarth, P. A., Turner, M., Reddy, A. P., David, L. L., and Nagalla, S. R. (2005) Identification of protein modifications using MS/MS *de novo* sequencing and the OpenSea alignment algorithm. *J. Proteome Res.* **4,** 546–554
9. Havilio, M., and Wool, A. (2007) Large-scale unrestricted identification of post-translation modifications using tandem mass spectrometry. *Anal. Chem.* **79,** 1362–1368
10. Tsur, D., Tanner, S., Zandi, E., Bafna, V., and Pevzner, P. A. (2005) Identification of post-translational modifications by blind search of mass spectra. *Nat. Biotechnol.* **23,** 1562–1567
11. Tanner, S., Shu, H., Frank, A., Wang, L. C., Zandi, E., Mumby, M., Pevzner, P. A., and Bafna, V. (2005) InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* **77,** 4626–4639
12. Savitski, M. M., Nielsen, M. L., and Zubarev, R. A. (2006) ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. *Mol. Cell. Proteomics* **5,** 935–948
13. Mann, M., and Wilm, M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66,** 4390–4399
14. Tabb, D. L., Saraf, A., and Yates, J. R. (2003) GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal. Chem.* **75,** 6415–6421
15. Chen, T., Kao, M., Tepel, M., Rush, J., and Church, G. M. (2001) Dynamic programming approach to *de novo* peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* **8,** 325–337
16. Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., and Lajoie, G. (2003) PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **17,** 2337–2342
17. Frank, A., and Pevzner, P. (2005) PepNovo: *de novo* peptide sequencing via probabilistic network modeling. *Anal. Chem.* **77,** 964–973
18. Seo, J., Jeong, J., Kim, Y. M., Hwang, N., Paek, E., and Lee, K.-J. (2008) A strategy for comprehensive identification of post-translational

modifications in cellular proteins, including low abundant modifications: application to glyceraldehyde-3-phosphate dehydrogenase. *J. Proteome Res.* **7,** 587–602
19. Wilmarth, P. A., Tanner, S., Dasari, S., Nagalla, S. R., Riviere, M. A., Bafna, V., Pevzner, P. A., and David, L. L. (2006) Age-related changes in human crystallins determined from comparative analysis of post-translational modifications in young and aged lens: does deamidation contribute to crystalline insolubility? *J. Proteome Res.* **5,** 2554–2566
20. Klimek, J., Eddes, J. S., Hohmann, L., Jackson, J., Peterson, A., Letarte, S., Gafken, P. R., Katz, J. E., Mallick, P., Lee, H., Schmidt, A., Ossola, R., Eng, J. K., Aebersold, R., and Martin, D. B. (2008) The standard protein mix database: a diverse data set to assist in the production of improved peptide and protein identification software tools. *J. Proteome Res.* **7,** 96–103
21. Craig, R., and Beavis, R. C. (2003) A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun. Mass Spectrom.* **17,** 2310–2316
22. Elias, J. E., and Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4,** 207–214
23. Gusfield, D. (1997) *Algorithms on String, Trees, and Sequences: Computer Science and Computational Biology*, pp. 87–208, Cambridge University Press, New York
24. Mouls, L., Aubagnac, J.-L., Martinez, J., and Enjalbal, C. (2007) Low energy peptide fragmentations in an ESI-Q-Tof type mass spectrometer. *J. Proteome Res.* **6,** 1378–1391
25. Elias, J. E., Gibbons, F. D., King, O. D., Roth, F. P., and Gygi, S. P. (2004) Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat. Biotechnol.* **22,** 214–219
26. Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J., and Hunt, D. F. (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 9528–9533
27. Havilio, M., Haddad, Y., and Smilansky, Z. (2003) Intensity-based statistical scorer for tandem mass spectrometry. *Anal. Chem.* **75,** 435–444
28. Ong, S., Mittler, G., and Mann, M. (2004) Identifying and quantifying in vivo methylation sites by heavy methyl SILAC. *Nat. Methods* **1,** 1–8
29. Tabb, D. L., Friedman, D. B., and Ham, A. L. (2006) Verification of automated peptide identifications from proteomic tandem mass spectra. *Nat. Protoc.* **1,** 2213–2222
30. Sun, S., Meyer-Arendt, K., Eichelberger, B., Brown, R., Yen, C.-Y., Old, W. M., Pierce, K., Cios, K. J., Ahn, N. G., and Resing, K. A. (2007) Improved validation of peptide MS/MS assignments using spectral intensity prediction. *Mol. Cell. Proteomics* **6,** 1–17
31. Beausoleil, S. A., Villen, J., Gerber, S. A., Rush, J., and Gygi, S. P. (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **24,** 1285–1292
32. Zubarev, R., and Mann, M. (2007) On the proper use of mass accuracy in proteomics. *Mol. Cell. Proteomics* **6,** 377–381
33. Taylor, J. A., and Johnson, R. S. (2001) Implementation and uses of automated *de novo* peptide sequencing by tandem mass spectrometry. *Anal. Chem.* **73,** 2594–2604
34. Kim, S., Na, S., Sim, J. W., Park, H., Jeong, J., Kim, H., Seo, Y., Seo, J., Lee, K.-J., and Paek, E. (2006) MOD[i]: a powerful and convenient web server for identifying multiple post-translational peptide modifications from tandem mass spectra. *Nucleic Acids Res.* **34,** W258–W263
35. Lapko, V. N., Smith, D. L., and Smith, J. B. (2001) In vivo carbamylation and acetylation of water-soluble human lens αB-crystallin lysine 92. *Protein Sci.* **10,** 1130–1136
36. Zhang, K., Yau, P. M., Chandrasekhar, B., New, R., Kondrat, R., Imai, B. S., and Bradbury, M. E. (2004) Differentiation between peptides containing acetylated or tri-methylated lysines by mass spectrometry: an application for determining lysine 9 acetylation and methylation of histone H3. *Proteomics* **4,** 1–10
37. Oses-Prieto, J. A., Zhang, X., and Burlingame, A. L. (2007) Formation of ε-formyllysine on silver-stained proteins. *Mol. Cell. Proteomics* **6,** 181–192