PTMap—A sequence alignment software for unrestricted, accurate, and full-spectrum identification of post-translational modification sites

Yue Chen[a,b,1], Wei Chen[c], Melanie H. Cobb[c], and Yingming Zhao[a,2]

Departments of [a]Biochemistry and [c]Pharmacology, University of Texas Southwestern Medical Center, Dallas, TX 75390; and [b]Department of Chemistry and Biochemistry, University of Texas, Arlington, TX 76019

Contributed by Melanie H. Cobb, November 20, 2008 (sent for review June 12, 2008)

We present sequence alignment software, called PTMap, for the accurate identification of full-spectrum protein post-translational modifications (PTMs) and polymorphisms. The software incorporates several features to improve searching speed and accuracy, including peak selection, adjustment of inaccurate mass shifts, and precise localization of PTM sites. PTMap also automates rules, based mainly on unmatched peaks, for manual verification of identified peptides. To evaluate the quality of sequence alignment, we developed a scoring system that takes into account both matched and unmatched peaks in the mass spectrum. Incorporation of these features dramatically increased both accuracy and sensitivity of the peptide- and PTM-identifications. To our knowledge, PTMap is the first algorithm that emphasizes unmatched peaks to eliminate false positives. The superior performance and reliability of PTMap were demonstrated by confident identification of PTMs on 156 peptides from four proteins and validated by MS/MS of the synthetic peptides. Our results demonstrate that PTMap is a powerful algorithm capable of identification of all possible protein PTMs with high confidence.

acetylation | dehydration | methylation | phosphorylation

Mass spectrometry is the method of choice for mapping sites of protein post-translational modifications (PTMs) that are known to have more than 300 types (1, 2). Efficient sequence alignment algorithms are essential for mapping PTM sites and peptide identification from mass spectrometric data (3). Widely used programs, such as Sequest (4, 5), Mascot (6), and X!Tandem (7), identify PTM sites based on a restricted database search in which tandem mass spectra are aligned with protein sequences bearing one or several specified PTMs at specified amino acid residues. This restricted database search strategy developed in early days has been very useful for identifying peptides bearing a limited number of specified PTMs, but lacks the flexibility to identify unexpected PTMs.

Recently, several algorithms have been developed to extend the capability of shotgun proteomics to allow identification of all possible PTMs and sequence polymorphisms (8–14). These algorithms can carry out unrestricted database searches, identifying any PTM, whether previously known or unknown. To improve the accuracy of peptide identification, different statistical strategies have been proposed in attempts to reduce the number of false positives (10–12, 15–17). These strategies typically involve applying a statistical significance test to score the confidence level of each identification. While useful, the reliability of these strategies has not been critically evaluated, for example, by testing them with manual verification with high stringency, or with MS/MS of synthetic peptides, the gold standard for confirming peptide identification.

Identification of false positives using statistical methods is daunting in unrestricted sequence alignments for several reasons. First, the size of the protein sequence database is exponentially increased. For example, consider a peptide of 12 amino acid residues. If it is assumed that this peptide contains a single PTM at an unknown position, and that this PTM causes a mass shift that could range over the integers between −100 and + 300, a list of 4,800 possible modified peptides is generated from the single peptide sequence. The exponentially increased size of the peptide pool and the high similarity among peptide sequences derived from the full spectrum of possible PTMs make it difficult to remove most of the false positives, let alone all of them. Second, a PTM could happen at several residue [e.g., protein methylation (18)], modified peptides with adjacent PTM sites are likely to have similar theoretical fragmentation patterns, and hence similar statistical scores. Third, *in silico*-generated peptide sequence pools used for sequence alignment are unlikely to include all of the peptide sequences generated by proteolytic digestion. Take, for example, trypsin, a protease often used for shotgun proteomics, generate not only tryptic peptides, but also semitryptic peptides, in which one terminus is generated by chymotryptic activity within the enzyme preparation (19).

The false positives were caused during protein database search, because of misinterpretation of charge states, abnormal enzymatic digestion sites, misinterpretation of protein modifications, wrong assignment of modification sites and modification types, and incorrect use of isotopic peaks (19). Unfortunately, the decoy protein sequence database cannot accurately predicate the false discovery rates as many false positives are sequence-linked to the true hits, such as those cases caused by abnormal enzymatic digestion and wrong assignment of modification sites and types (19).

We have observed that the MS/MS spectra of false positive identifications commonly contain unmatched peaks with significant intensities. Accordingly, we argue that, while identification of candidate peptides should focus on matched peaks, to comprehensively identify false positives, emphasis should be placed on unmatched peaks. We realized that "unexplained peaks" have been used as a penalty parameter in *de novo* peptide sequencing (20). However, it has never been used as a major filter to remove false positive peptide and PTM identifications during database search. Our emphasis on unmatched peaks is based on the rationale that a true peptide identification should explain all major peaks in the spectrum, especially for those MS/MS data that are generated in ion trap types of mass spectrometers that have relatively simple fragmentation patterns (21). We further argue that the modification site should be unambiguously located to confidently identify peptides containing PTMs, and to eliminate false-positive PTM assignments.

Author contributions: Y.C., M.H.C., and Y.Z. designed research; Y.C. and W.C. performed research; Y.C. and Y.Z. analyzed data; and Y.C., M.H.C., and Y.Z. wrote the paper.

The authors declare no conflict of interest.

[1] Present address: Ben May Department for Cancer Research, University of Chicago, Chicago, IL 60637.

[2] To whom correspondence should be sent at the present address: Ben May Department for Cancer Research, University of Chicago, Chicago, IL 60637. E-mail: yingming.zhao@uchicago.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0811739106/DCSupplemental.

Here, we describe PTMap, a sequence alignment algorithm for reliable, full-spectrum identification of mass shifts associated with unspecified PTMs or polymorphisms. To reduce the risk of obtaining false positives when identifying peptides that bear an unknown mass shift at an unknown amino acid residue, the algorithm adopts several strategies: selection of sequence-rich MS peaks, automation of mass-shift adjustment, precise localization of PTM sites, implementation of rules for manual verification, and development of a scoring system based on spectrum quality to remove false-positive sequence alignments. PTMap also reduces searching time by aligning the sequences of only those proteins that are of interest, and by removing isotopic peaks and noise peaks. A unique feature of PTMap is identification of false positives by emphasizing unmatched peaks instead of matched peaks with statistically significant scores. The software is able to use MS/MS data from low-resolution tandem mass spectrometers by implementing automatic mass-shift adjustment. To demonstrate its usefulness, we used PTMap to identify PTM sites in human histone H4, HMG2, mouse SGK1, and BSA. Accuracy of peptide identification by PTMap was confirmed by MS/MS of synthetic peptides for the corresponding peptide identifications with Mascot scores between 15–27 that were usually considered as false positives. Our data demonstrate that PTMap can remove almost all of the false positives while maintaining high sensitivity for identification of peptides and PTM sites.

Results

The Challenge for Statistics-Based Sequence Alignment Algorithms: Identifying False Positives. Statistics-based algorithms have been used to analyze MS/MS data, resulting in putative peptide sequence matches that are statistically significant relative to the peptide pool of interest. Among these statistically significant matches will be both true and false positive hits. Such algorithms have been shown to be powerful for peptide identification and mapping PTM sites when a limited number of PTMs are involved. We argue that such statistical methods cannot efficiently eliminate false positives in an unrestrictive sequence alignment because of the exponentially increased size of the protein sequence database, high sequence similarities among modified peptides that have the same sequence and PTM type but different PTM sites, and the existence of atypical proteolytic peptides (e.g., semitryptic peptides). While statistical methods can efficiently calculate the statistical significance of randomly matched peptides, these methods are less capable of identifying false positives that are derived from nonrandom peptides, such as those arising from proteolytic digestion at peptide bonds other than those corresponding to an enzyme's specificity or peptides from incorrect assignment of a PTM type or PTM sites. Two of such examples are shown in the *SI Appendix*.

Our Strategy for Removing False Positives. The exponential increase in the number of peptide sequences that results from not restricting potential PTMs, and the high degree of similarity among the sequences, will undoubtedly lead to large numbers of false positives that cannot be efficiently eliminated by focusing on matched peaks alone. We reason that a correctly identified peptide should explain all major peaks in an MS/MS spectrum. Accordingly, we developed a new concept that identification of false positives should focus on unmatched peaks, while matched peaks can provide a list of candidate identifications.

Toward this goal, we developed the PTMap sequence alignment algorithm for mapping sites of unspecified PTMs and identifying protein polymorphisms (Fig. 1 and *SI Appendix*). This algorithm incorporates several approaches to improve search speed and reduce false positives, while identifying a large percentage of true positives. First, PTMap searches only protein sequences of interest that have already been identified by general software such as Sequest or Mascot. This targeted sequence analysis reduces the incidence of false matching between MS/MS data and irrelevant protein sequences. Second, PTMap identifies peptide candidates based on matched peaks and then removes false positives from the candidate list based on unmatched peaks. This filtering process is superior to statistical methods, which will typically sacrifice sensitivity of peptide identification for identification accuracy. Third, an identification is only considered true if the PTM site is unambiguously determined. We reason that the precise location of a PTM site is critical in unrestricted sequence alignment because of the large number of possible modified peptides with the same sequence.

To further improve the performance of PTMap, we incorporated three additional strategies (Fig. 1): (*i*) PTMap selects only those peaks with signal intensities that are significant relative to the local noise; (*ii*) PTMap uses only monoisotopic peaks for sequence alignment; and (*iii*) PTMap automates the adjustment of mass shifts. The last feature allows PTMap to analyze data from low-resolution mass spectrometers in addition to the data from high-resolution mass spectrometer.

We use two parameters to evaluate a peptide identification made by PTMap: unmatched peak score ($S_{Unmatched}$) and PTMap score. $S_{Unmatched}$ is determined by the number and intensities of unmatched peaks in the MS/MS spectra, while PTMap score evaluates how well the sequence and the spectrum mutually explain each other. To assess the reliability and effectiveness of these scoring functions for removing incorrect sequence alignment results, we carried out HPLC/MS/MS analysis of tryptic peptides from human HMG2, histone H4, mouse SGK1, and BSA, and analyzed the resulting MS/MS data with both Mascot and PTMap.

Selectivity and False-Discovery Rate of PTMap Analysis. A high $S_{Unmatched}$ score suggests a large number of unmatched peaks with significant intensities. For a singly charged precursor ion, only one $S_{Unmatched}$ score is used. For a precursor ion with two or more charges, two $S_{Unmatched}$ scores are used for each spectrum: one for the high mass range (higher than the precursor ion *m/z*) and one for the low mass range (lower than the precursor ion *m/z*). Both scores must be satisfactory for a positive identification. We routinely use $S_{Unmatched}$ scores of 4.0 and 10.0, for the high and low mass ranges, respectively (notated as 4.0:10.0), to filter out sequence-spectrum alignments of low quality. The $S_{Unmatched}$ score for the low mass range is not as stringent because a greater number of intense noise peaks are usually generated in the low mass range than in the high mass range. To check if the $S_{Unmatched}$ threshold scores will cause the loss of sensitivity of the analysis, we manually analyzed all of the Mascot peptide identifications and plotted the $S_{Unmatched}$ score distributions for correct and incorrect IDs. The results demonstrated that the $S_{Unmatched}$ scores (4.0:10.0) were sufficient for the identification of all of the correct unmodified peptides (*SI Appendix*).

A difficulty arises from spectra having few fragment ions, because these spectra result in low numbers of both matched and unmatched peaks. Such false positives cannot be identified by the $S_{Unmatched}$ score alone. This issue is addressed by the PTMap score.

To evaluate the usefulness of the PTMap score as a second parameter to remove false positives, we searched each MS/MS dataset against the true or randomly scrambled sequence of the corresponding protein. We used 4.0:10.0 as threshold $S_{Unmatched}$ scores, and then generated PTMap scores. PTMap scores for identifications of unmodified peptides from both normal and scrambled protein sequences were calculated (Fig. 2*A*). When a PTMap score cutoff of 0.50 was used, the false discovery rate was ≈1.6% for unmodified peptides (Fig. 2*A*).

Next, we tested whether we could use this cutoff score to evaluate false positives of all peptide identifications, including those containing PTMs. We generated PTMap scores for all peptide identifications (Fig. 2*B*). The number of peptides with scores below the 0.50 cutoff was much higher for modified peptides than for unmodified peptides (Fig. 2*A* and *B*), suggesting a large number of low quality sequence-spectrum pairs derived from unspecified PTMs. Analysis of search results with scrambled protein sequences (Fig.

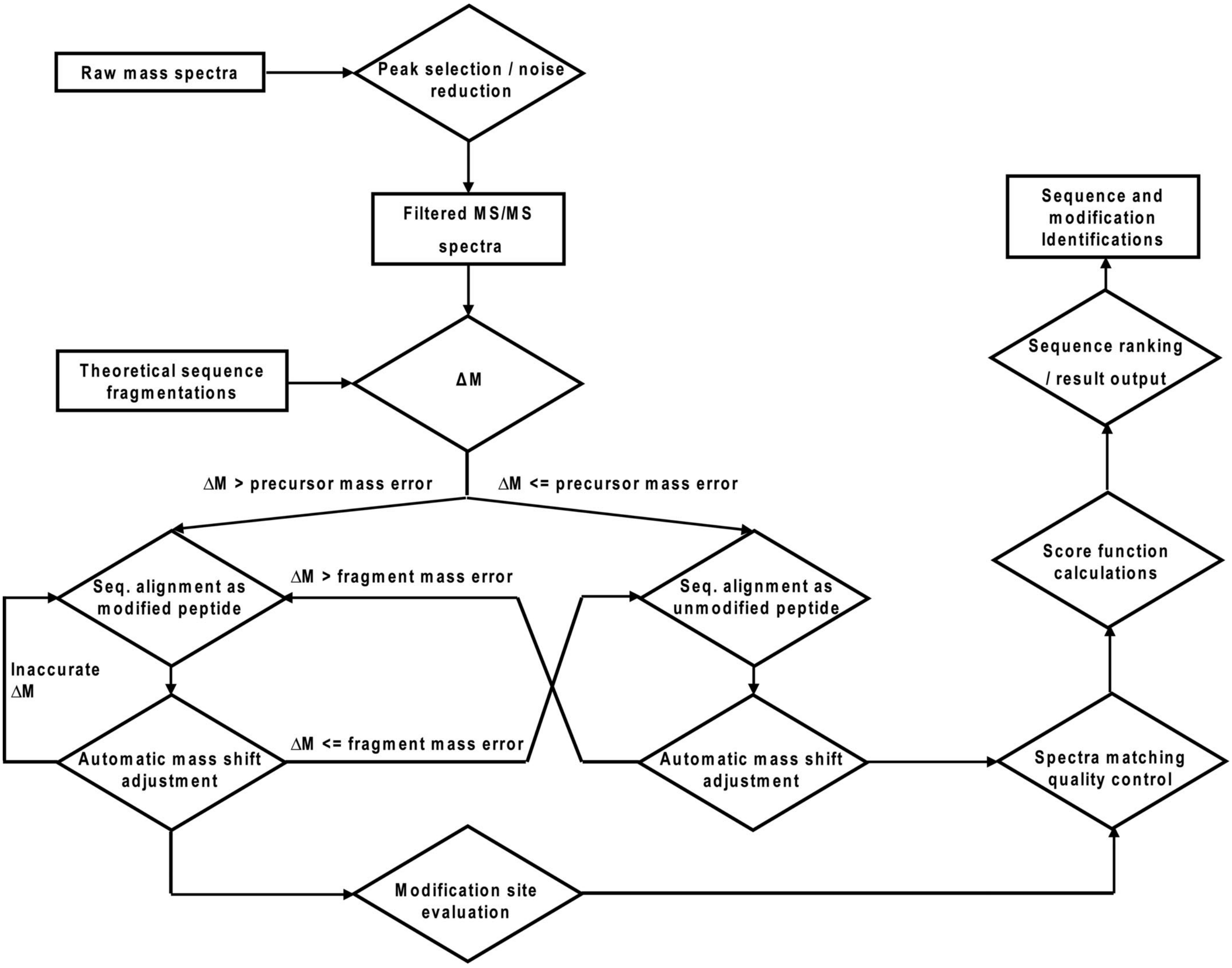
Raw mass spectra
Peak selection / noise reduction
Filtered MS/MS spectra
Theoretical sequence fragmentations
ΔM
ΔM > precursor mass error
ΔM <= precursor mass error
Seq. alignment as modified peptide
ΔM > fragment mass error
Seq. alignment as unmodified peptide
Inaccurate ΔM
Automatic mass shift adjustment
ΔM <= fragment mass error
Automatic mass shift adjustment
Modification site evaluation
Spectra matching quality control
Score function calculations
Sequence ranking / result output
Sequence and modification Identifications

Fig. 1. Flow chart of the PTMap algorithm.

2*B*) shows the false discovery rate for searches including modified peptides was 20.9% using a cutoff score of 0.50. To improve the accuracy of peptide identification, we increased the cutoff score for all peptide identifications to 1.00 for unrestrictive PTM analysis, which caused a loss of sensitivity of ≈18% and an improvement in the false discovery rate to 3.6%.

Sensitivity of PTMap Analysis. To evaluate the sensitivity of the PTMap score for detecting true peptide identifications, we compared the identification results of Mascot and PTMap for unmodified peptides in restrictive analysis (Fig. 2*C*). Because each peptide was usually identified by multiple spectra, only the identifications with the best scores were plotted. When using Mascot, a score of 40 is typically used as a cutoff score for peptide identification. Mascot analysis identified 71 peptides with scores above the threshold score of 40 after manual verification. Over 90% of these peptides had high PTMap scores above 1.0 and all have PTMap score above 0.5 threshold for restrictive analysis. PTMap was able to identify an additional 40 peptides whose PTMap scores were above 1.0 and Mascot scores were below 40, suggesting that the PTMap algorithm was able to boost the sensitivity of the identification by 57%. To evaluate the accuracy with which peptides were identified, we performed MS/MS analysis of corresponding synthetic peptides for 8 of the 40 peptides with Mascot scores from 15 to 27. The results showed that the fragmentation patterns of synthetic peptides completely matched the experimental MS/MS spectra and confirmed the identification of these peptides by PTMap (*SI Appendix*). To identify the difference between PTMap and Mascot scores, we studied the peptide length dependence of the two scores and found that PTMap is able to avoid bias against short peptides while Mascot score is positively correlated with peptide length. This study was described in detail in *SI Appendix*.

Strategies Used by PTMap to Increase Accuracy of Peptide Identification. PTMap implements three main strategies to improve searching speed and accuracy of peptide identification (*SI Appendix*). We systematically evaluated the effectiveness of each strategy. The results demonstrate that the application of these strategies significantly reduces the false discovery rate and boosts the sensitivity of peptide identification.

Peak Selection. PTMap selects only those monoisotopic peaks that are significant when compared with local noise levels. Because noise levels are usually not homogeneous across the whole mass range, the local noise level is used when selecting meaningful peaks. In addition, isotopic peaks do not contain extra sequence information and are therefore removed. To evaluate the effectiveness of peak selection for peptide identification, we compared the distri-

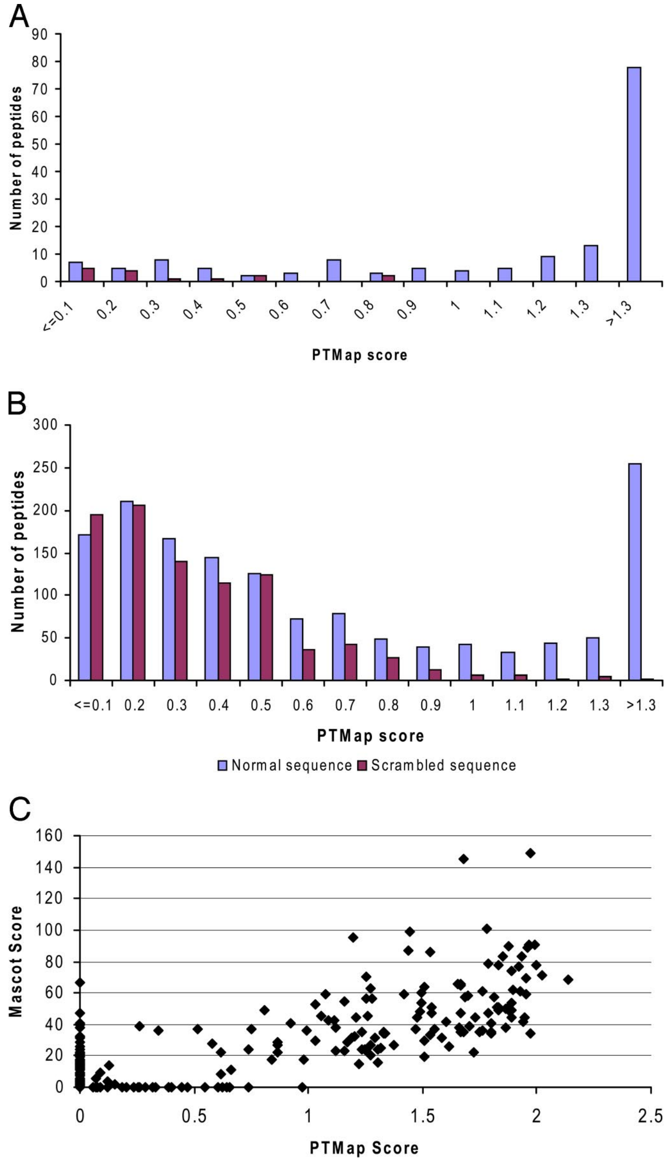

Fig. 2. Evaluation of the PTMap score distribution. (*A*) PTMap score distributions ($S_{\mathrm{Unmatched}}$ = 4.0:10.0 in high and low mass ranges respectively) of unmodified peptides identified from normal and scrambled protein sequences of the four selected proteins—histone H4, SGK1, HMG2 and BSA—using PTMap (155 peptides generated from normal database and 15 peptides generated from scrambled database); (*B*) PTMap score distributions ($S_{\mathrm{Unmatched}}$ = 4.0:10.0) of both unmodified and modified peptides identified from normal and scrambled of the same four protein sequences using PTMap (1,478 peptides generated from normal database and 917 peptides generated from scrambled database); (*C*) Correlation of Mascot scores with PTMap scores ($S_{\mathrm{Unmatched}}$ = 4.0:10.0) for identification of unmodified peptides. The highest Mascot score and PTMap score of each peptide were used. Those peptides identified with Mascot only (PTMap score = 0) were found to be false positives by manual verification methods (21).

bution of PTMap scores of the identified unmodified peptides with or without peak selection.

Because noise peaks are different with or without application of peak selection, we used PTMap score instead of $S_{\mathrm{Unmatched}}$ to evaluate the utility of peak selection. The $S_{\mathrm{Unmatched}}$ threshold was therefore relaxed to 400 in this analysis. PTMap analysis of four protein LC/MS/MS datasets without the application of peak selection identified 147 nonredundant peptides, compared with 126 peptides with the application of peak selection (Fig. 3*A*; PTMap cutoff score = 0.50, for restrictive analysis at 1.6% false discovery rate). The 21 additional peptides identified in the absence of peak selection were manually inspected using the stringent rules described previously (21) and were found to be false positives.

Manual examination of the false positive spectra found that the false positive results were mainly caused by matching to (*i*) peaks in a mass range of 0 to −50 Da relative to the parent ion; (*ii*) noise signals with low intensities; and (*iii*) isotope peaks. Applying peak selection strategies allows PTMap to efficiently eliminate these types of false positives, thereby increasing the accuracy of identification. Additionally, by considerably reducing the total number of peaks, peak selection improves searching speed.

Automatic Mass-Shift Adjustment. To search for undefined PTMs, PTMap scans all mass shifts from −100 to +200 Da in 1-Da increments. In low-resolution mass spectrometers, such as those with the popular ion trap mass analyzer, mass errors are much lower in MS/MS spectra (usually ±0.6 Da) than in the corresponding MS spectrum (usually ±4.0 Da). High-mass errors for precursor ions arise because of low-resolution instruments, space charge in ion trap mass spectrometers, use of isotopic peaks as precursor ions, and identification of parent ions from coincident ions. A large mass error for a precursor ion, if not corrected, could result in identification of a PTM with an incorrect mass shift, or improper identification of the nature of the PTM. To address this concern, PTMap includes a function determining the consensus mass shift based on the MS/MS spectrum. Then the spectrum is realigned, using this consensus mass shift, before spectral alignment quality is evaluated. Automatic mass-shift adjustment allows identification of unmodified peptides with large precursor ion mass errors. We believe that this feature is also useful for analysis of MS/MS spectra from high-resolution instruments because of the possibility that precursor ions may be assigned to isotopic peaks.

To test the effectiveness of this strategy for identification of PTMs, we examined the MS/MS spectra of the identified peptides bearing PTMs (PTMap cutoff score = 1.0, for unrestrictive analysis at 3.6% false discovery rate). The mass differences between the calculated mass shift and the adjusted mass shift were calculated and the distribution of all of the mass differences was plotted (Fig. 3*B*). Among the total number of 1476 identified spectra, 462 spectra or 31.3% were identified with the application of automatic mass shift adjustment. The total number of peptide identifications with PTMs increases by 45.6% after the application of this strategy (Fig. 3*C*). We also examined the mass errors of precursor ions of unmodified peptides (Fig. 3*D*). The results show that PTMap was able to identify unmodified peptides with large precursor ion mass errors.

Exclusive Site Localization. Localization of the site of modification within a modified peptide can be difficult, because peptides that differ only by the site of modification will give highly similar theoretical fragmentation patterns. Incorrect localization of PTM sites will dramatically increase the number of modified peptides identified, a problem that has not been previously addressed. PTMap incorporates a two-step procedure enabling PTM sites to be exclusively located. First, it is assumed that the candidate PTM may reside on any residue of the peptide; theoretical fragmentation patterns of these peptides are aligned with the MS/MS spectrum iteratively. The resulting sequence alignments are compared with each other, and PTMap identifies the peptide isoform with the best PTMap score. If peptides with adjacent modification sites have identical PTMap scores, a strategy illustrated in Fig. 4*A* is used to define the PTM site. For the modification site to be localized to position M and not M + 1 (Fig. 4*A*), PTMap requires that the total intensity of the two PTM-relevant fragment ions, modified b_M and unmodified y_N, be higher than that of unmodified fragment b_M and modified fragment y_N. Second, PTMap requires the PTM site to be identified by consecutive ions in the b or y ion series, or by the

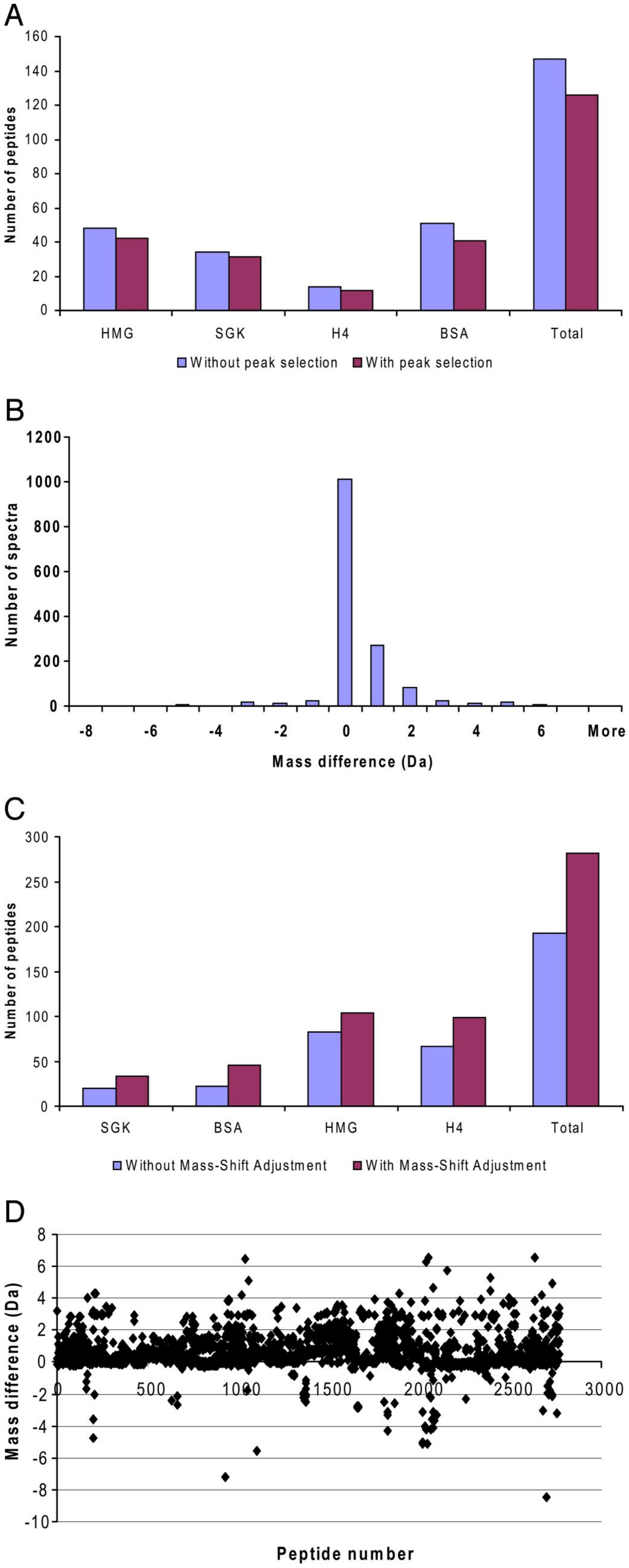

Fig. 3. Evaluation of the two PTMap strategies: peak selection and automatic mass shift adjustment. (*A*) The number of unmodified peptides (PTMap score cutoff = 0.5) identified with or without incorporation of the peak-selection function in PTMap; (*B*) The distribution of the mass changes ($\Delta M_{after} - \Delta M_{before}$) made by PTMap after automatic mass-shift adjustment for all identified spectra bearing PTMs (PTMap score cutoff = 1.0); (*C*) The number of modified peptides identified with or without automatic mass-shift adjustment strategy in PTMap (PTMap score cutoff = 1.0);(*D*) scatter plot showing the distribution of the mass errors of the spectra that identified unmodified peptides in the four proteins (PTMap score cutoff = 0.5)

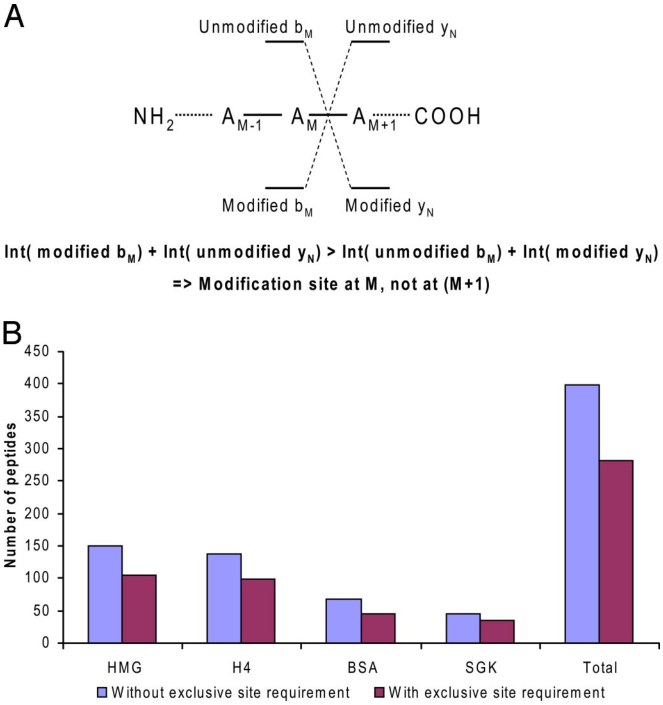

Fig. 4. Evaluation of the PTMap strategy for exclusive site localization. (*A*) A strategy for precise mapping of a PTM site that will distinguish two peptide isoforms that are modified on adjacent sites; (*B*) The number of modified peptides identified by PTMap before and after the implementation of exclusive site localization.

simultaneous appearance of modified b and y ions in which the modified residue is the end residue of each fragment. The site of modification is considered identified if both conditions are met, while hits failing at least one of the filters are considered ambiguous and are removed.

To illustrate the effectiveness of this approach, we compared the peptide modification analysis with or without these criteria (Fig. 4*B*, PTMap cutoff score = 1.0, for unrestrictive analysis at 3.6% false discovery rate). Without exclusive site localization, PTMap identified a total of 399 modified peptides, in large contrast to only 282 peptides when the exclusive site localization requirements were applied. This represents a 29.3% reduction of false positives for modified peptides when ambiguous modification site assignments were removed.

Identification of Modified Peptides in HMG2, Human Histone H4, Mouse SGK1, and BSA. PTMap identified a total of 282 nonredundant modified peptides with a PTMap score higher than 1.0 from the MS/MS datasets of human histone H4, HMG2, mouse SGK1, and BSA. We removed peptides with PTM sites on the N- or C-terminal amino acid for two reasons. First, LTQ mass spectrometer used in our study has low mass cutoff that eliminates modified b_1 or y_1 ions whose *m/z* values are below the cutoff. Second, terminal modifications cannot be mapped through consecutive b or y ion series, raising the possibility that the *m/z* of identified b_1 or y_1 ions coincide with the nominal *m/z* of a combination of terminal amino acids (e.g., *m/z* of one Asn is equal to the total *m/z* of two Gly) and causing false-positive PTM identification. From this dataset, we further removed redundant peptide sequences that identified the same PTM sites. The final results give 156 modified peptides (*SI Appendix*), among which 56 modified peptides were identified by at least three spectra. Seventy-seven of these PTMs (≈50%) can be annotated using the Unimod database (http://www.unimod.org). From the data, we found that the most commonly modified residues were methionine (M), lysine (K), cysteine (C), histidine (H), and glutamic acid (E).

For 18 of the 156 peptides (≈11%), the unmodified peptide counterparts were not observed, which can be explained by three possibilities. First, the modification may have been caused by common chemical reactions (e.g., Cys adduction with acrylamide or iodoacetamide) with high reaction yields. Second, the modification might have completely prevented cleavage by the protease (e.g., lysine acetylation). Third, some identified mass shifts were the result of protein point mutations.

To distinguish among the three possibilities, we synthesized 17 peptides containing a mutation, based on the mass shifts, and then performed MS/MS analysis of the peptides. The MS/MS spectra of 6 protein-derived peptides and their synthetic counterparts were almost same (*SI Appendix*), implying that these protein polymorphisms identified by PTMap are true. The remaining 11 MS/MS spectra did not match those of the synthetic peptides, suggesting that the identified mass shifts on these peptides come from unknown PTMs.

PTMs on histone proteins play critical roles in chromatin structure and gene regulation. In this study, we identified a total of 110 peptides from histone H4, among which 99 peptides were identified as modified peptides through unrestrictive analysis with PTMap (PTMap cutoff score = 1.0). Stringent manual analysis showed that 106 of 110 total peptides and 95 of 99 modified peptides can be manually verified, among which 59 peptides bear modifications that were not reported previously (http://www.unimod.org). After eliminating the PTMs on the N- or C- terminals of the peptides, whose PTM sites cannot be precisely located, we conclusively identified a total of 64 non-redundant PTMs on histone H4 with high confidence (*SI Appendix*). Three peptides identified with potential mutations were further confirmed by fragmentation of synthetic peptides (*SI Appendix*).

Discussion

We describe the development of PTMap software for accurate identification of full-spectrum PTMs and polymorphisms. PTMap incorporates three unique features to improve its function: MS peak selection, automatic mass-shift adjustment, and exclusive site localization. Two logical score systems, $S_{\mathrm{Unmatched}}$ and PTMap score, were developed and were demonstrated to be accurate for evaluating peptide identifications. To remove false positives, the algorithm stresses unmatched peaks and incorporates stringent manual-verification rules. To our knowledge, this strategy has not been previously described for removing false positives.

Minimization of false positives is always accompanied by a sacrifice of sensitivity of peptide identification in sequence alignments based on statistical methods. In contrast, our results demonstrate that PTMap can identify 57% more peptides than statistical methods, while achieving higher accuracy as indicated by the identification of peptides with low Mascot scores. Therefore, PTMap addresses a major problem with statistics-based algorithms, such as Mascot.

Some PTMs lead to unique fragmentation patterns. For example, phosphopeptides tend to generate daughter ions with a neutral loss. PTMap can take such unique fragmentation patterns into consideration to boost the efficiency of identifying peptides with such properties.

While PTMap was used in the analysis of a single, pure protein in our case studies, the algorithm can be easily expanded to the analysis of an MS/MS dataset from complex protein mixtures (e.g., >100 proteins) by initial protein identification and subsequent mapping of PTMs. While more than 200 types of PTM have been identified, the abundance and scope of these PTMs, and interrelationships between them, remain largely unknown. Recent identification of two novel PTMs, lysine propionylation and lysine butyrylation (22), suggests that important undescribed PTMs remain to be discovered. PTMap will provide a powerful technology platform for chemical dissection of cellular PTM networks.

Materials and Methods

Human histone H4 was purified from HeLa cells as previously described (23), except that HDAC inhibitors (2 μm TSA, 50 mM sodium butyrate, and 30 mM nicotinamide) were added during preparation of the core histones to prevent histone deacetylation, depropionylation, and debutyrylation. The Flag-tagged mouse SGK1 fusion protein (fragment 61-431) was purified as previously described (24). Human HMG2 and BSA (BSA) were purchased from ProteinOne and Sigma, respectively. Synthetic peptides were synthesized by GL Biochem and Genemed Synthesis.

Protein digestion, HPLC/MS/MS analysis, and protein sequence database searching were described in *SI Appendix*.

ACKNOWLEDGMENTS. We thank David Fenyo for insightful comments. This work was supported by National Institutes of Health Grant CA126832 (to Y.Z.).

1. Witze ES, Old WM, Resing KA, Ahn NG (2007) Mapping protein post-translational modifications with mass spectrometry. *Nat Methods* 4:798–806.
2. Jensen ON (2004) Modification-specific proteomics: Characterization of post-translational modifications by mass spectrometry. *Curr Opin Chem Biol* 8:33–41.
3. Sadygov RG, Cociorva D, Yates JR 3rd (2004) Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book. *Nat Methods* 1:195–202.
4. Eng JK, McCormack AL, Yates JR 3rd (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.
5. Yates JR 3rd, Eng JK, McCormack AL (1995) Mining genomes: Correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal Chem* 67:3202–3210.
6. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20:3551–3567.
7. Craig R, Beavis RC (2004) TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics* 20:1466–1467.
8. Pevzner PA, Dancik V, Tang CL (2000) Mutation-tolerant protein identification by mass spectrometry. *J Comput Biol* 7:777–787.
9. Bandeira N, Tsur D, Frank A, Pevzner PA (2007) Protein identification by spectral networks analysis. *Proc Natl Acad Sci USA* 104:6140–6145.
10. Hansen BT, Davey SW, Ham AJ, Liebler DC (2005) P-Mod: An algorithm and software to map modifications to peptide sequences using tandem MS data. *J Proteome Res* 4:358–368.
11. Havilio M, Wool A (2007) Large-scale unrestricted identification of post-translation modifications using tandem mass spectrometry. *Anal Chem* 79:1362–1368.
12. Savitski MM, Nielsen ML, Zubarev RA (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. Tsur D, Tanner S, Zandi E, Bafna V, Pevzner PA (2005) Identification of post-translational modifications by blind search of mass spectra. *Nat Biotechnol* 23:1562–1567.
14. Hansen BT, Jones JA, Mason DE, Liebler DC (2001) SALSA: A pattern recognition algorithm to detect electrophile-adducted peptides by automated evaluation of CID spectra in LC-MS-MS analyses. *Anal Chem* 73:1676–1683.
15. Eriksson J, Chait BT, Fenyo D (2000) A statistical basis for testing the significance of mass spectrometric protein identification results. *Anal Chem* 72:999–1005.
16. Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol* 24:1285–1292.
17. Tanner S, *et al.* (2008) Accurate annotation of peptide modifications through unrestrictive database search. *J Proteome Res* 7:170–181.
18. Sprung R, *et al.* (2008) Identification and Validation of Eukaryotic Aspartate and Glutamate Methylation in Proteins. *J Proteome Res* 7:1001–1006.
19. Chen Y, Zhang J, Xing G, Zhao Y (2008) Common types of false-positives identified in shotgun proteomics. Submitted.
20. Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA (1999) De novo peptide sequencing via tandem mass spectrometry. *J Comput Biol* 6:327–342.
21. Chen Y, Kwon SW, Kim SC, Zhao Y (2005) Integrated approach for manual evaluation of peptides identified by searching protein sequence databases with tandem mass spectra. *J Proteome Res* 4:998–1005.
22. Chen Y, *et al.* (2007) Lysine propionylation and butyrylation are novel post-translational modifications in histones. *Mol Cell Proteomics* 6:812–819.
23. Shechter D, Dormann HL, Allis CD, Hake SB (2007) Extraction, purification and analysis of histones. *Nat Protoc* 2:1445–1457.
24. Chen W, *et al.* (2008) Regulation of a third conserved phosphorylation site in serum and glucocorticoid-induced protein kinase 1. *J Biol Chem*, in press.