9-6-2013

ISPTM: an Iterative Search Algorithm for Systematic Identification of Post-translational Modifications from Complex Proteome Mixtures

Xin Huang
University of Nebraska Medical Center

Lin Huang
University of Nebraska Medical Center

Hong Peng
University of Nebraska Medical Center

Ashu Guru
University of Nebraska - Lincoln

Weihua Zue
University of Nebraska Medical Center

See next page for additional authors

Published in final edited form as:
J Proteome Res. 2013 September 6; 12(9): 3831–3842. doi:10.1021/pr4003883.

ISPTM: an Iterative Search Algorithm for Systematic Identification of Post-translational Modifications from Complex Proteome Mixtures

Xin Huang[†,‡,‖], **Lin Huang**[†,‖], **Hong Peng**[†,‡,‖], **Ashu Guru**[¶], **Weihua Xue**[†], **Sang Yong Hong**[†], **Miao Liu**[†], **Seema Sharma**[#], **Kai Fu**[†], **Adam Caprez**[¶], **David Swanson**[¶], **Zhixin Zhang**[†], and **Shi-Jian Ding**[†,§,*,⊥]

[†]Department of Pathology and Microbiology, University of Nebraska Medical Center, Omaha, Nebraska 68198

[‡]Department of Environmental, Agricultural & Occupational Health, University of Nebraska Medical Center, Omaha, Nebraska 68198

[§]Mass Spectrometry and Proteomics Core Facility, University of Nebraska Medical Center, Omaha, Nebraska 68198

[¶]Holland Computing Center, University of Nebraska–Lincoln, Lincoln, Nebraska 68588

[#]Thermo Fisher Scientific, San Jose, California 95134

Abstract

Identifying protein post-translational modifications (PTMs) from tandem mass spectrometry data of complex proteome mixtures is a highly challenging task. Here we present a new strategy, named iterative search for identifying PTMs (ISPTM), for tackling this challenge. The ISPTM approach consists of a basic search with no variable modification, followed by iterative searches of many PTMs using a small number of them (usually two) in each search. The performance of the ISPTM approach was evaluated on mixtures of 70 synthetic peptides with known modifications, on an 18-protein standard mixture with unknown modifications and on real, complex biological samples of mouse nuclear matrix proteins with unknown modifications. ISPTM revealed that many chemical PTMs were introduced by urea and iodoacetamide during sample preparation and many biological PTMs, including dimethylation of arginine and lysine, were significantly activated by Adriamycin treatment in NM associated proteins. ISPTM increased the MS/MS spectral identification rate substantially, displayed significantly better sensitivity for systematic PTM identification than the conventional all-in-one search approach and offered PTM identification results that were complementary to InsPecT and MODa, both of which are established PTM identification algorithms. In summary, ISPTM is a new and powerful tool for unbiased identification of many different PTMs with high confidence from complex proteome mixtures.

[*]Corresponding author: Dr. Shi-Jian Ding, Diabetes and Obesity Research Center, Sanford Burnham Medical Research Institute at Lake Nona, Orlando, Florida 32827. Phone: +1-407-745-2149. Fax: +1-407-745-2032. sjding@sanfordburnham.org.
[⊥]Current affiliation: Diabetes and Obesity Research Center, Sanford Burnham Medical Research Institute, Orlando, Florida 32827
[‖]These authors contributed equally to this work

Conflict of interest: the authors have declared no conflict of interest.

Supporting Information Available: This material is available free of charge via the Internet at http://pubs.acs.org.

Keywords

ISPTM; Post-translational modifications; Database search; OMSSA; DNA Damage; Nuclear Matrix

INTRODUCTION

Post-translational modifications (PTMs) of proteins play an extensive and pivotal role in eukaryotic signal transduction, gene regulation, and metabolic control in cells.[1, 2] PTMs determine protein conformation, activity, and localization, as well as stability.[1] Abnormal PTMs are often a cause or consequence of many pathological and disease conditions.[3] Although they are important, system-wide identification of PTMs remains a highly challenging task for many reasons. First, PTMs display enormous diversity and complexity.[4] There are more than 300 PTMs that are known to occur physiologically. Vertebrate proteins often undergo multiple PTMs at the same time. It was estimated that for human proteins there are 8~12 modified versions for each unmodified tryptic peptide.[5] Second, PTMs generate complex fragmentation patterns in tandem mass spectrometry. This complexity poses a significant challenge for subsequent data analysis. Third, PTMs are usually present at low stoichiometry and low-abundance. Fourth, global proteomic studies are often limited to a specific PTM due to the prerequisite of effective enrichment strategies that employ specific PTMs.[2] An unbiased approach for system-wide identification of many different PTMs in complex proteome mixtures is highly desirable.

Currently, liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) is the central method for identifying proteins with PTMs.[1, 6] A particular advantage of this technique is that the MS/MS spectra contain information both on the intact full length peptide and on the masses of fragment ions from which amino acid sequences with specific sites of PTM can be derived. Typically, a modified peptide is identified through a process of peptide-spectrum-match (PSM) using programs such as SEQUEST[7], Mascot[8], or OMSSA[9] to compare the observed spectral data to a protein database. Identification by these algorithms is based on a restricted database search in which MS/MS spectra are aligned with protein sequences bearing a few specified PTMs attached to specific amino acids. These approaches are not very effective at identifying large numbers of PTMs from complex proteome mixtures because the database search space expands exponentially as the number of PTMs increases. This increases the search time and false positive rate. For these reasons, it is generally advisable to include a limited number of variable modifications during database searches using conventional database search engines such as SEQUEST and Mascot.

To overcome the drawbacks of conventional database search methods, a number of strategies for unrestricted PTM identification have been developed, such as the *de novo* sequencing approach, sequence-tag approach, and spectral matching approach.[10] Each approach has its own strengths and weaknesses. For example, the blind search methods, such as InsPecT[11], can identify all possible PTMs at once, especially the unknown and unexpected PTMs. However, it is sensitive to the size of protein database, and a double pass strategy is recommended to increase the specificity when the database is over ten million amino acids.[12] The double pass strategy identifies proteins in the sample using unmodified peptides (or minimally modified peptides) in a first pass, and then reduces the database to include only those identified proteins and search it for a wide selection of modified peptides in a second pass. Recently, Na *et al* developed a novel blind search tool named MODa, which can perform fast and unrestrictive searches for large scale databases of the human proteome.[13] Using a dynamic programming method, MODa solved the limitation of the

number of unrestrictive PTMs that can be allowed in each peptide. However, MODa was not designed to address the accurate localization of modifications to specific sites in the identified peptide sequences.[13] Other approaches, such as ModifiComb and DeltAMT, can identify both known and unknown PTMs from complex mixtures in a quick fashion.[10,14] However, because these spectrum match algorithms are based on the similarity of mass shifts and retention times between the unmodified form and its modified counterpart, they are insensitive to the quality of MS/MS spectra. Thus they may not accurately localize the modification site for a PTM that may occur on different amino acid residues in the same peptide.

Here we report a novel strategy, named ISPTM (iterative search for peptide identification with PTMs), for the systematic identification of PTMs with site-specific confidence from complex MS data. The iterative search strategy concept has been applied to some conventional search engines (such as X!Tandem, Mascot and SEQUEST) by early developers, but with the double pass strategy.[14, 15] However, our ISPTM approach differs from these iterative search approaches by refining the MS/MS spectra instead of refining the database. ISPTM enables the identification of PTMs from complex peptide mixtures without prior identification of the proteins in the sample. The performance of ISPTM was evaluated using three datasets with different levels of protein complexity. Our results indicated that the ISPTM approach substantially increased the MS/MS spectral identification rate, demonstrated significantly better sensitivity for global PTM identification than the conventional all-in-one search approach, and provided PTM identification results complementary to those from InsPecT and MODa. Using ISPTM, we found that many biologically meaningful PTMs, as well as some chemical modifications, occurred on the nuclear matrix (NM) proteins of mouse pro-B 2A cells after ADR-induced DNA damage.

MATERIALS AND METHODS

Preparation and LC-MS/MS analysis of the synthetic peptide samples

Two synthetic modified peptide mixtures were received from the Proteomics Standards Research Group at the Association of Biomolecular Resource Facilities. One (sample#1) contained a lyophilized mixture of 70 synthetic modified peptides, and the other (sample#2) contained the same mixture combined with a tryptic digest of six proteins from which the synthetic peptides were derived. More details of these samples are available at www.abrf.org/sprg, survey project 2011. Peptide samples were analyzed using an Easy nanoLC, equipped with a 75 Vm × 10 cm, Magic C18 AQ LC column, coupled to a Q-Exactive mass spectrometer (Thermo Scientific, San Jose, CA) as previously described.[16]

18-protein Standard mixture data

Datasets for an 18-protein standard mixture were downloaded from the Institute for Systems Biology (ISB) website.[17] These datasets were analyzed by InsPecT and MODa. Briefly, the mgf files were searched against the ISB database of 18 standard proteins plus 92 contaminant proteins and 1709 *Haemophilus influenzae* RD proteins as background (obtained from ISB website). Carbamidomethylation of cysteine was used as the fixed modification. Only one modification was allowed in each peptide. In InsPecT, all the identified spectra were collected by applying a filter of *p-value* less than 0.05. An FDR cutoff of 0.01 was applied for the filtered spectra based on the F-score. In MODa, a probability score > 0.95 was applied, and an FDR cutoff of 0.01 was used to filter the spectra by the PSM score. In ISPTM, the OMSSA outputs were collected and filtered by *p-value* < 0.05, and an FDR cutoff was applied based on the OMSSA E-value.

Preparation and LC-MS/MS analysis of the NM samples

Abelson virus-transformed mouse pro-B cell line 2A was maintained in RPMI media supplemented with 10% fetal bovine serum and 12.5 VM β-mercaptoethanol (Invitrogen, Carlsbad, CA). Pro-B 2A cells were treated with either 1 μM Adriamycin (Sigma-Aldrich, St Louis, MO, dissolved in DMSO) or DMSO alone for 4 hours. All buffers for NM sample preparation contained 1% protein phosphatase inhibitor cocktail 1, 1% protein phosphatase inhibitor cocktail 2, 1% protease inhibitor cocktail, and 1.2 nM phenyl methane sulfonyl fluoride (all from Sigma). Cell lysates were re-suspended in CSK buffer (10 mM PIPES pH 6.8, 100 mM NaCl, 300 mM sucrose, 3 mM $MgCl_2$, 1 mM EGTA, 0.5% Triton X-100, 1 mM dithiothreitol (DTT), all from Sigma) and incubated on ice for 5 min. After centrifugation at 9000 rpm for 1 min, the pellet was re-suspended in low salt extraction buffer (42.5 mM Tris HCl pH 8.3, 8.5 mM NaCl, 2.6 mM $MgCl_2$, 1% Tween 40, 0.5% deoxycholic acid) and incubated on ice for 5 minutes, followed by centrifugation at 9,000 rpm for 1 min. Again, the pellet was re-suspended in digestion buffer (10 mM PIPES pH 6.8, 50 mM NaCl, 300 mM sucrose, 3 mM $MgCl_2$, 1mM EGTA, DNase (500 u/mL, Roche, Indianapolis, IN), RNase (500 u/mL, Ambion, Austin, TX), 0.5% Triton X-100, and 1 mM DTT) and incubated at room temperature for 1 hour with gentle shaking. The concentration of ammonium sulfate was adjusted to 0.25 M and the preparation was incubated for a 10 min extraction period, at room temperature. By centrifugation at 13,000 rpm for 1 min, the pellet was washed with 2 M NaCl buffer (2 M NaCl, 10 mM PIPES, pH 6.8, 10 mM EDTA). After a final centrifugation at 13,000 rpm for 1 min, the pellet was re-suspended in lysis buffer (8 M urea, 50 mM ammonium bicarbonate) and incubated at 37°C overnight to dissolve the nuclear matrix-associated proteins. All centrifugations were performed at 4°C.

Methods for sample preparation, trypsin digestion, strong cation exchange (SCX) chromatography, and LC-MS/MS analysis of the NM protein digests were described previously.[18] Briefly, two NM protein samples were reduced with DTT and carboxyamidomethylated with iodoacetamide (IAA) at room temperature. Tryptic digestion was performed and the resulting peptides were desalted by solid phase extraction. SCX separation was performed and twenty fractions were obtained from each NM sample. These fractions were analyzed by LC-MS/MS on a nanoLC coupled with LTQ-Orbitrap-XL mass spectrometer (Thermo Scientific).

Database and computing resources

For the synthetic peptide data, we used a protein database containing the six proteins that the modified peptides belong to, plus four contaminate proteins, 1,990 background proteins from both the human and bovine proteomes, and the reverse sequences for all of these 2,000 proteins (obtained from the sPRG 2011 survey project). For the ISB data, the same database was used. For the NM data, MS/MS spectra were searched against i) a concatenated database containing 55,303 proteins from the international protein index (IPI) mouse database (version 3.52), ii) the commonly observed 262 contaminants (forward database), and iii) the reversed sequences of all 55,565 proteins from i and ii (reverse database). The OMSSA engine (v2.1.9, Linux version) was used for the database searches. The initial mass deviation tolerance of precursor ion was set to 0.02 Da and fragment ion tolerance was set to 0.5 Da for the NM data. The initial mass deviation tolerance was 0.02 Da and the fragment ion tolerance was 0.05 Da for the synthetic peptide mixture data. A maximum of 2 missed cleavages were allowed in peptide identification. We also employed a multi-blind search with the MODa software to analyze the NM data, using the same settings of ISPTM. Both ISPTM and MODa searches were performed using the computing resources available at the University of Nebraska Holland Computing Center (HCC).

ISPTM analysis of the data

The ISPTM approach consists of four steps. Step 1: the MS/MS raw data were pre-processed by DeconMSn and DtaRefinery as previously described.[18] MS/MS spectral data were then stored in mgf format ready for the OMSSA search. Step 2: a basic search was performed with carboxyamidomethylation of cysteine as a fixed modification (no variable modifications in this step). The identified unmodified peptides were filtered by OMSSA E-value cutoff of < 0.01. Step 3: the identified MS/MS spectra from Step 2 were removed from the initial spectrum pool, and the remaining spectra were used for iterative searches. In this step, each cycle tested a small number of variable modifications (1, 2 or 3) until all combinations of the modifications were tested. For the synthetic peptides data, ISPTM searches were made in five variations: 1) testing 13 known modifications taken two-at-time (IS-13, 13×12/2 = 78 runs), 2) testing these 13 authentic modifications plus 13 false modifications (Supplemental Table 1) taken two-at-time (IS-26, 26×25/2 = 325 runs), 3) testing all (207) modifications in the OMSSA database, one-at-time (IS-Single, 207 runs), 4) testing all OMSSA modifications two-at-time (IS-Double, 207×206/2 = 21,321 runs), and 5) testing all OMSSA modifications three-at-time (IS-Triple, 207×206×205/(3×2) = 1,456,935 runs). For the ISB data, iterative searches were performed by the IS-single strategy, testing 207 modifications one-at-time. For the NM data, we removed 46 modifications (chemical modifications using stable isotope labels, Supplemental Table 2) that cannot occur in our biological samples, and iteratively searched the rest of the modifications using the IS-double strategy. The modified peptides that were identified were filtered by OMSSA E-value cutoff of < 0.1. In the case of multiple peptide sequences identified from the same MS/MS spectrum, the peptide sequence with the smallest E-value was retained. Step 4: all identification results were combined and exported with a fixed FDR, followed by calculation of site confidence score for each modification site.

Site confidence score for identification of peptides with PTMs

To provide an empirical measure of confidence that a PTM site was correctly localized a probability-based significance was calculated using the site-determining product ions. Briefly, a probability distribution ***P(X)*** is based on the hypothesis that random sampling of fragment ions in an MS/MS spectrum follows a binomial distribution:

$$P_{(X=k)} = \binom{N}{k} p^k (1-p)^{N-k} \quad \text{(Eq. 1)}$$

where ***p*** is the probability of matching a fragment ion in a sampling event, and ***N*** and ***k*** represent the theoretical and observed site-determining fragment ions from the MS/MS spectrum.

For each modified peptide, the site confidence (SC) score for a PTM at position ***i*** is calculated as:

$$SC_i = 1 - \sum_j P_j \quad \text{(Eq. 2)}$$

where $\boldsymbol{P_j}$ is the false positive (FP) probability that a PTM is located at position ***i*** but not at position ***j*** in the same peptide.

To calculate the SC score, the MS/MS spectra were preprocessed to create a list of observed fragment ions that contained the 6 most intense fragment ions per 100 *m/z* units. Masses of theoretical ions for each identified peptide were obtained from MS-Product (http://prospector.ucsf.edu/prospector/). For each identified peptide with a PTM at position ***i***, the

alternative forms of modifications include: 1) the same PTM at other possible sites, 2) a different PTM with similar mass shift (<0.02 Da, identical to the OMSSA tolerance of precursor ions) in this peptide. For each alternative PTM at position ***j, N*** (total number of site-determining ions), ***k*** (number of observed fragment ions that matched the theoretical ion using a mass tolerance of 0.5 Da) and ***p*** (=0.06) were used to calculate the $\boldsymbol{P}_j$. Then the SC score was determined by 1 minus the sum of $\boldsymbol{P}_j$ of all alternative forms.

RESULTS

Overview of the ISPTM approach

The ISPTM work flow is summarized in Figure 1. Scripts written in Python were used to perform tasks including spectra refining, filtering the spectra of unmodified peptides, setting the pool of PTMs and the number of variable modifications in each iterative search, generating the commands for OMSSA searches, collecting and filtering the identification results, and annotating the site confidence of PTMs. The python scripts are fully automatic in each step, minimizing the user's intervention. The outputs of ISPTM follows the same format of standard OMSSA csv outputs, with a new column indicating the SC score for each PTM site. The Python scripts and instructions have been deposit on Google Code (https://code.google.com/p/isptm-python/). The ISPTM analyses of the synthetic peptides and the NM data were both finished in less than 48 hrs.

ISPTM Analysis of the Synthetic Modified Peptide Mixtures with Known Modifications

In the synthetic modified peptide mixtures, peptide "NGDTASPKEYTAGR" with 3 different methylated forms of lysine (methylation, dimethylation and trimethylation) were identified in a single conventional search allowing all 13 modifications as variable modifications. In ISPTM, an iterative database search was applied and matches were found when the mono-, di- and tri-methyl modifications were tested, respectively (Supplemental Figure 1). We evaluated the performance of ISPTM using the synthetic modified peptide mixtures and compared it to the conventional all-in-one search. In total, 41 peptides were identified from 278 spectra by the conventional search of sample #1, while 45 peptides were identified from 358 spectra by the ISPTM method using the 13 modifications taken two-at-time (IS-13). Using a false discovery rate (FDR) < 0.1, only 13 peptides (out of 109 spectra) from the conventional search were acceptable, while 32 peptides (out of 239 spectra) were acceptable from the ISPTM method. A detailed comparison of conventional and ISPTM search results with different strategies for analysis of the synthetic peptide data is displayed in Supplemental Table 3.

The overall performance of the conventional search and ISPTM approaches with multiple strategies for sample#1 was compared using the receiver operating characteristic (ROC) plot (Figure 2A). The ROC plot demonstrated that an iterative search using the IS-13 strategy achieved the best discriminating power between the authentic and false positive identifications. The discriminating power was essentially the same if another 13 false modifications were included in the ISPTM search (IS-26). We further tested the performance of ISPTM on the synthetic peptides by using the 207 modifications in the OMSSA modifications pool in the search. In these analyses, variable modifications were used one-at-time (IS-Single), two-at-time (IS-Double) or three-at-time (IS-Triple). ROC analysis indicated that IS-13 and IS-26 strategies had higher discriminating power than the IS-Single, IS-Double, and IS-Triple strategies. Interestingly, performances of all these ISPTM strategies were essentially the same if ROC analysis was applied to identification results from sample #2 (Figure 2B). Overall, all the iterative searches showed superior discriminating power compared to the conventional all-in-one approach. The discriminating power of the IS-Single, IS-Double, and IS-Triple strategies increase for sample #2,

compared to sample #1, mainly because the peptides in sample #2 are more complex than in sample #1 and many natural modifications, such as oxidation of methionine and acetylation of the protein N-terminus are present.

We also noticed that the PSM score may change when different numbers of variable modifications were used. As shown in Figure 2C, for the same spectra with same identification results, the PSM score [represented by the $-\log_{10}$(E-value)] was plotted for the conventional and the IS-Double search results. The regression line with a slope ≈ 1 and an intercept of 1.29 indicates that the PSM score for the IS-Double search is slightly higher than the PSM score for the all-in-one search. This is the reason why the discriminating power decreased in the conventional all-in-one search. However, there is almost no difference in the PSM score when the IS-Double and IS-Triple search results were compared (slope = 1, intercept ≈ 0, Figure 2D).

Analysis of ISB data

We employed InsPecT, MODa, and our ISPTM approach to analyze the ISB data. By FDR < 0.01, InsPecT, MODa and ISPTM identified a total of 5,790, 11,233 and 9,556 spectra, respectively. Among these spectra, 1,639, 3,012 and 2442 were identified as modified peptides by InsPecT, MODa and ISPTM, respectively. All peptide/protein identifications are listed in Supplemental Table 4, with a PTM frequency matrix[11] was developed for the modified spectra for each program. A Venn gram shows the different coverage of identified peptides, as well as the modified peptides (peptides with identical sequence, modification site, and mass shift) by these approaches (Figure 3A and 3B). Overall, InsPect, MODa and ISPTM provide complementary identification coverage for the ISB dataset. MODa identified more peptides than InsPect and ISPTM. As shown in Table 1, the most frequent modifications identified by both InsPecT and MODa results were the sodium and potassium adducts. Other frequent PTMs identified by both programs include oxidation of methionine, carbamidomethylation of cysteine, and several biologically relevant PTMs, such as dehydration by beta-elimination of serine and threonine, and peptide N-terminal acetylation. In the ISPTM analysis, the most frequent modification was deamidation of asparagine. Consistent with the InsPecT and MODa results, oxidation of methionine, acetylation of the protein N-terminal, and beta-elimination of serine and threonine were identified. Interestingly, ISPTM also identified many peptides with cyclization of the N-terminal S-carbamidomethyl cysteine (Pyro-CamC) and N-terminal pyro-glutamic acid (Pyro-Glu).

ISPTM Analysis of the Global PTMs in Nuclear Matrix Samples

We evaluated the ISPTM approach using a pair of complex biological samples: nuclear matrix protein digests of mouse Pro-B cells before (Control) and after DNA damage (ADR-treated). Using the basic search for unmodified peptides, with E-value < 0.01, 9,315 and 5,527 MS/MS spectra were identified from the Control and ADR-treated samples, respectively. The number of false positive peptides, identified at the spectrum level, from the Control and ADR-treated datasets were 8 and 51 (both FDR < 0.01), respectively. Using the ISPTM approach with two variable modifications at a time, 28,595 modified peptide spectra were identified by Evalue < 0.1. By applying the FDR cutoff of 0.01, we identified 1,921 unique peptide sequences from 5,068 MS/MS spectra, of which 1,700 spectra were from the Control samples and 3,368 spectra were from ADR-treated samples. In the ISPTM results, 32.5% (625/1921) of the modified peptides were identified with an unmodified form in the basic search. At the protein level, 62.1% (907/1460) of the modified proteins were identified in the basic search. MODa was used to analyze the NM data as well. By FDR < 0.01, MODa identified 5,492 and 3,102 spectra of unmodified peptides as well as 1,857 and 4,437 spectra of modified peptides from the Control and ADR-treated samples, respectively. For the combined Control and ADR-treated nuclear matrix samples, ISPTM identified more

unique peptides and proteins than MODa (Figure 3C and 3D). However, MODa identified more modified peptides with either identical sequences or identical modifications than ISPTM (Figure 3E and 3F).

The frequencies of identification for modified peptides obtained by the ISPTM approach were shown in Supplemental Table 5. First, we applied an SC cutoff of 0.8 to remove the unconfident modifications. Then manual curation was performed for the results. For example, we found that OMSSA assigned ubiquitination, methylation and sumoylation to lysines and arginines at the C-terminus of some peptides. But a modified lysine or arginine is generally not recognized by trypsin for digestion and therefore could not appear at the C-terminus of a tryptic peptide. We also identified some spectra with O-GlcNAcylation to serine and threonine. Because O-GlcNAc is readily lost as an oxonium ion during collision-induced dissociation,[19] it renders the identification of O-GlcNAc modified peptides very difficult. Thus such assignments were removed from the identification results. As a result, we obtained 4,166 spectra identified for a total of 1,636 unique peptides (an entire list of identified peptides with PTMs is presented in Supplemental Table 6). The three most frequent modifications were carbamylation of lysine, carbamylation of the peptide N-terminal, and acetylation of protein N-terminal. For example, a peptide with carbamylation at lysine was shown in Supplemental Figure 2A. This modification is induced presumably by urea in the sample lysis buffer.[20] Another reagent commonly used during proteomics sample preparation is iodoacetamide (IAA). It is used to alkylate cysteines exposed by reduction of disulfide bonds. We found that histidine, glutamic acid, aspartic acid, and lysine were also alkylated by IAA.[21] A peptide with carboxyamidomethylation at histidine was shown in Supplemental Figure 2B. Pyro-Glu modification appears in 128 spectra (Supplemental Figure 2C), because N-terminal glutamine or glutamic acid residues are known to form Pyro-glu under aqueous conditions.[22] Another frequent chemical modification was Pyro-CamC at N-terminal cysteines, as shown in Supplemental Figure 2D. This modification has been reported to be caused by enzymatic digestion of proteins that have been S-alkylated by IAA.[23]

Deamidation of asparagine and glutamine was identified in 247 and 34 spectra in the NM samples (Supplemental Figures 3A and 3B). It has been reported that deamidation is involved in the “DNA damage-induced cellular response”.[24] Moreover, this modification can be caused chemically, especially during sample storage at higher temperature or when the asparagine or glutamine is followed by glycine.[25] Two additional PTMs, oxidation of methionine and acetylation of the protein N-terminal, which are often included in routine database searches, were identified in 183 and 249 spectra, respectively (Supplemental Figures 3C and 3D). ISPTM results indicated that the chemical modifications are quite abundant in the proteomics samples. It is important that they are included in routine database searches because their presence may affect the identification of other modifications. For instance, Figure 4A shows an annotated MS/MS spectrum of peptide “GVLKVFLENVIR” derived from histone H4 position 57~68. Previous studies have reported that the lysine residue at position 60 (H4K60) can be acetylated[26], ubiquitinylated[27], or formylated[28] physiologically. In our study, all three forms of modification on H4K60 were identified, but all peptides were also carbamylated on the N-terminal glycine (Figure 4B – 4D). Using the conventional database search approach, these spectra would not likely have been identified because carbamylation of the N-terminal glycine is not a common modification and therefore is not normally included as a variable modification.

The numbers of identified unique peptides corresponding to selected PTMs from the Control and ADR-treated NM samples are compared (Figure 5A). The number of peptides with deamidation of asparagine, which is the most frequent PTM, is similar in the two samples. The number of peptides with oxidized methionine decreased about 50% after ADR

treatment. Oxidative stress may cause oxidation of methionine *in vivo.*[29] However, this modification may also occur in vitro during the sample preparation. The number of peptides with dimethylated arginine increased in the ADR-treated sample. It has been reported that dimethylation of ribonucleoproteins (RNP) increases their ability to bind DNA and to promote gene transcription.[30] Table 2 lists a number of peptides that were found to carry dimethylated arginine, and their corresponding proteins. Dimethylated arginine has been reported previously for some of these proteins: Hnrnpa0, Hnrnpa1, Pabpc1, Ewsr1, Snrpb, Hnrnph1, and Hnrnpu.[30] A representative peptide with dimethylated arginine at Hnrnpa0 is shown in Supplemental Figure 4A. Neutral-loss of a monomethylamine (H_2N-CH_3) group indicated that this is a symmetric dimethylation site.[31] Other RNP proteins (Hnrnpa1) and Pabpc1 (Supplemental Figure 4B) were identified with dimethylation, indicating that this modification is essential for normal mRNA metabolism. Interestingly, dimethylation of arginine at Ewsr1 (Supplemental Figure 4C), Snrbp, Rbm33 (Supplemental Figure 4D), Hnrnph1, and Hnrnpu were only observed after ADR treatment. The function of dimethylation on these proteins remains unknown.

A glycyl (GG) modification on lysine is a degradation signal for ubiquitination.[32] Figure 5B shows the annotated MS/MS spectrum of “LIFAGKQLEDGR”, which is a signature tryptic peptide of the protein with K48 poly-ubiquitination.[33] Interestingly, the ubiquitination modification site on histone H4 (H4K60) was found to be the site of formylation after ADR treatment. As a secondary modification that results from oxidative DNA damage, formylation of lysine in histone proteins may interfere with the signaling functions and thus contribute to the pathophysiology of oxidative and nitrosative stress.[34] All the above data indicate that PTMs on many NM proteins, especially the core histones, were altered by treatment with ADR. Such modifications may change the activity and function of these proteins in response to ADR-induced DNA damage.

DISCUSSION

PTMs are extremely important for maintaining protein structure and function. We present in this paper a novel strategy named ISPTM to identify the proteins with complex patterns of PTMs from LC-MS/MS data. Compared to the conventional all-in-one search strategy, ISPTM can effectively control the search space by including a very limited number of PTMs in a search and has higher discriminating power for the true PTMs as we demonstrated in Figure 2A and 2B. In contrast, when a large number of PTMs needed to be tested in a sample, it will be increasingly difficult to use the conventional all-in-one search because of the exponential increase of search space and reduced PSM score to discriminate the true PTM identifications from the false identifications (Figure 2C). The unique feature of the ISPTM approach is that it performs an exhaustive search for hundreds of different modifications expected to be found in complex protein samples, including both naturally occurring and chemical modifications. Our data indicated that a large portion of peptides are chemically modified by carboxyamidomethylation, carbamylation and deamidation, but these chemical modifications are generally not considered in routine database searches. Importantly, our approach has demonstrated that identifying peptides with various (either chemical or biological) modifications in a sample can not only increase the spectral identification rate but also can increase the chance of identifying key protein regulators and their possible PTMs.

One limitation of ISPTM is that it is not designed to discover totally unknown modifications. All modified peptides are identified from a pre-defined pool of modifications. Nevertheless, the current UNIMOD database (www.unimod.org) contains more than a thousand modifications.[35] This pool could be employed in place of the OMMSA pool, which currently contains 207 modifications. However, instead of testing all PTMs in this

pool (i.e. Mascot error tolerant search), users can choose a small subset of interested PTMs to perform ISPTM. For a relatively simple mixture like the ISB data, we demonstrated that the performance of the ISPTM approach is equivalent to the blind searches engines InsPecT and MODa, when all possible modifications were tested.

When analyzing more complex proteome mixtures like the NM dataset here, ISPTM identified about 72% more spectra (14842/8594) of unmodified peptides, whereas it identified about 19% less (5068/6294) spectra of modified spectra, compared to the outputs of MODa. Here we interpret these results in several aspects: First, a restrictive engine such as OMSSA is better than unrestrictive engines in identifying unmodified peptides. Second, ISPTM separates the modified and unmodified peptides and applies FDR individually. By an FDR of 0.01 in the NM data, E-value cutoff of unmodified peptides was 0.01, while cutoff of modified peptides was 2.8E-6. This might be another reason that ISPTM identified fewer modified peptides than MODa. Third, both results contain false positive identifications even though an FDR cutoff has been applied. However, a fixed modification at a specific site in restrictive searches can minimize the artifacts. For instance, carbamylation (+43) only occurs at a peptide N-terminal and lysine, but we also observed a number of spectra were identified within other sites in MODa. Another advantage of the ISPTM approach over blind search or *de novo* methods is that the identification results are very easy to interpret, because all modifications are known and with clear site specificity. Finally, the large number of potential modifications provides the both algorithms with wide latitude for making assignments. Consequently, even peptides with strong scores can prove to be assigned incorrectly. Thus we strongly suggest that modification results be confirmed by manual sequencing and orthogonal approaches such as site-directed mutagenesis or MS/MS analysis of the synthetic peptide with a specific modification.[36]

In this paper we introduced an SC score method to access the site confidence of the identification results. Although both the SC score and a former A-score[37] methods are based on a cumulative binomial distribution model (measuring the likelihood of matching at least the number of matched site-determining ions by chance), we think that the SC score developed in this manuscript may have several advantages. First, A-score is an ambiguous score that only distinguishes the top two candidate sites, but SC score considers all possible candidate sites. Second, A-score is restricted on the same modification at different sites (i.e., phosphorylation on S/T/Y). However, if many PTMs with identical or close mass shifts are involved in a search like in the case of ISPTM search, it is difficult to determine exactly which PTMs are occurring and distinguish them. For instance, deamidation and citrullination have exactly the same mass-shift. And the mass-shift between acetylation and tri-methylation is very close, differing only by 0.036 Da. Such a small difference is undetectable in low-resolution MS but detectable in high-resolution MS. Our scripts calculate the mass shifts of all possible PTMs included in the search and the mass tolerance of all identified peptides automatically, providing the SC score without user intervention. Third, because the SC score is a confidence score, users can apply a certain cutoff (i.e., 0.8) to filter out the ambiguous PTM identification results. To summarize, the SC score developed here allows for more PTM assignments in a high throughput fashion. However, in the current design of ISPTM, it does not allow for assignment of novel PTMs unless this novel PTM is included in the ISPTM search.

The OMSSA search engine was chosen for testing our ISPTM approach in the current study because it is open-source and platform-independent.[9] However, in principle, ISPTM can be applied to other search engines, such as SEQUEST and Mascot. The computational resource required is an important concern for large scale PTM identification of complex proteome data. In this study, the cumulative CPU hours for analyzing the NM datasets by ISPTM was 4,535 hours, while the cumulative CPU hours for analyzing the NM datasets by MODa was

834 hours (Supplemental Table 7). Dependent on the number of cores/CPUs available for parallel computing, an ISPTM search of complex proteome datasets could be completed in a few hours. In this study, we used a 1,151 node Linux cluster for our analyses because we were analyzing up to 207 modifications. Such a large computing resource currently may not be available to all investigators. However, most studies would be expected to involve a smaller set of PTMs (less than 20). We have shown that the ISPTM approach has superior performance in testing 13 modifications compared to the all-in-one search. For those studies, a desktop PC would be suitable for iteratively testing one or two modifications at a time. Moreover, the issue of computing power should be addressed by the rapid development of modern computational technology such as supercomputers and cloud computing.[38] Indeed, any academic researcher in the United States already has access to a large computational resource via OSG (Open Science Grid, https://www.opensciencegrid.org) or, more broadly, XSEDE (Extreme Science and Engineering Discovery Enviroment, https://www.xsede.org/).

CONCLUSIONS

In summary, we have developed a novel approach, termed ISPTM, for the systematic identification of PTMs in proteome samples. The ISPTM approach enables conventional database search methods to be used for systematic PTM identification. The results obtained with ISPTM demonstrated that chemical modifications such as carboxyamidomethylation of histidine, glutamic acid, aspartic acid, and lysine and carbamylation of lysine are abundant when IAA and urea are used in sample preparation. With the increasing size of the PTM knowledge database, the ISPTM approach will bring the level of PTM identification from the era of limited identification and quantitation to the level of global PTM discovery for complex biological samples.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dr. Lawrence Schopfer for the editing of this manuscript. The mass spectrometry data were collected in the Mass Spectrometry and Proteomics Core Facility at the University of Nebraska Medical Center (UNMC) which is supported by the Nebraska Research Initiative. We thank Jim Keagy from Thermo Scientific for arranging demo experiments on the Q-Exactive mass spectrometer. We thank the Proteomics Standards Research Group (sPRG) from the Association of Biomolecular Resource Facilities for providing the synthetic modified peptide mixtures. This work was completed utilizing the Holland Computing Center of the University of Nebraska and the Open Science Grid. This work was financially supported by the Department of Pathology and Microbiology at UNMC and NEHHS LB606 (S.J.D), National Institutes of Health (NIH) grants AI076475 (Z.Z.). X.H. and H.P. were supported in part by scholarships from the Chinese Scholarship Council and M.L. was supported by a scholarship from the College of Medicine at UNMC.

Abbreviations

PTMs	Post-translational modifications
LC-MS/MS	liquid chromatography coupled with tandem mass spectrometry
PSM	peptide-spectrum-match
ISPTM	iterative search for peptide identification with PTMs
NM	nuclear matrix
ADR	Adriamycin
ISB	Institute for Systems Biology

DTT	dithiothreitol
IAA	iodoacetamide
IPI	international protein index
SC	site confidence
FP	false positive
ROC	receiver operating characteristic
FDR	false discovery rate
RNP	ribonucleoproteins
Pyro-CamC	S-carbamoylmethyl-L-cysteine
Pyro-Glu	Pyro-glutamic acid
O-GlcNAc	O-linked N-acetylglucosamine

References

1. Mann M, Jensen ON. Proteomic analysis of post-translational modifications. Nat Biotechnol. 2003; 21(3):255–61. [PubMed: 12610572]
2. Jensen ON. Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. Curr Opin Chem Biol. 2004; 8(1):33–41. [PubMed: 15036154]
3. Doyle HA, Mamula MJ. Posttranslational modifications of self-antigens. Ann N Y Acad Sci. 2005; 1050:1–9. [PubMed: 16014515]
4. Witze ES, Old WM, Resing KA, Ahn NG. Mapping protein post-translational modifications with mass spectrometry. Nat Methods. 2007; 4(10):798–806. [PubMed: 17901869]
5. Nielsen ML, Savitski MM, Zubarev RA. Extent of modifications in human proteome samples and their effect on dynamic range of analysis in shotgun proteomics. Mol Cell Proteomics. 2006; 5(12): 2384–91. [PubMed: 17015437]
6. Young NL, Plazas-Mayorca MD, Garcia BA. Systems-wide proteomic characterization of combinatorial post-translational modification patterns. Expert Rev Proteomics. 2010; 7(1):79–92. [PubMed: 20121478]
7. Eng JK, Mccormack AL, Yates JR. An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino-Acid-Sequences in a Protein Database. Journal of the American Society for Mass Spectrometry. 1994; 5(11):976–989.
8. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999; 20(18):3551–67. [PubMed: 10612281]
9. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH. Open mass spectrometry search algorithm. J Proteome Res. 2004; 3(5):958–64. [PubMed: 15473683]
10. Fu Y, Xiu LY, Jia W, Ye D, Sun RX, Qian XH, He SM. DeltAMT: a statistical algorithm for fast detection of protein modifications from LC-MS/MS data. Mol Cell Proteomics. 2011; 10(5):M110 000455.
11. Tsur D, Tanner S, Zandi E, Bafna V, Pevzner PA. Identification of post-translational modifications by blind search of mass spectra. Nat Biotechnol. 2005; 23(12):1562–7. [PubMed: 16311586]
12. Tanner S, Pevzner PA, Bafna V. Unrestrictive identification of post-translational modifications through peptide mass spectrometry. Nat Protoc. 2006; 1(1):67–72. [PubMed: 17406213]
13. Na S, Bandeira N, Paek E. Fast multi-blind modification search through tandem mass spectrometry. Mol Cell Proteomics. 2012; 11(4):M111 010199. [PubMed: 22186716]
14. Creasy DM, Cottrell JS. Error tolerant searching of uninterpreted tandem mass spectrometry data. Proteomics. 2002; 2(10):1426–34. [PubMed: 12422359]

15. MacCoss MJ, McDonald WH, Saraf A, Sadygov R, Clark JM, Tasto JJ, Gould KL, Wolters D, Washburn M, Weiss A, Clark JI, Yates JR. 3rd, Shotgun identification of protein modifications from protein complexes and lens tissue. Proc Natl Acad Sci U S A. 2002; 99(12):7900–5. [PubMed: 12060738]
16. Huang X, Tian C, Liu M, Wang Y, Tolmachev AV, Sharma S, Yu F, Fu K, Zheng J, Ding SJ. Quantitative proteomic analysis of mouse embryonic fibroblasts and induced pluripotent stem cells using 16O/18O labeling. J Proteome Res. 2012; 11(4):2091–102. [PubMed: 22375802]
17. Klimek J, Eddes JS, Hohmann L, Jackson J, Peterson A, Letarte S, Gafken PR, Katz JE, Mallick P, Lee H, Schmidt A, Ossola R, Eng JK, Aebersold R, Martin DB. 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. 2008; 7(1):96–103. [PubMed: 17711323]
18. Huang X, Tolmachev AV, Shen Y, Liu M, Huang L, Zhang Z, Anderson GA, Smith RD, Chan WC, Hinrichs SH, Fu K, Ding SJ. UNiquant, a program for quantitative proteomics analysis using stable isotope labeling. J Proteome Res. 2011; 10(3):1228–37. [PubMed: 21158445]
19. Greis KD, Hayes BK, Comer FI, Kirk M, Barnes S, Lowary TL, Hart GW. Selective detection and site-analysis of O-GlcNAc-modified glycopeptides by beta-elimination and tandem electrospray mass spectrometry. Anal Biochem. 1996; 234(1):38–49. [PubMed: 8742080]
20. Lin MF, Williams C, Murray MV, Conn G, Ropp PA. Ion chromatographic quantification of cyanate in urea solutions: estimation of the efficiency of cyanate scavengers for use in recombinant protein manufacturing. J Chromatogr B Analyt Technol Biomed Life Sci. 2004; 803(2):353–62.
21. Zachara NE, Hart GW. Cell signaling, the essential role of O-GlcNAc! Biochim Biophys Acta. 2006; 1761(5–6):599–617. [PubMed: 16781888]
22. Saito S, Yano K, Sharma S, McMahon HE, Shimasaki S. Characterization of the post-translational modification of recombinant human BMP-15 mature protein. Protein Sci. 2008; 17(2):362–70. [PubMed: 18227435]
23. Geoghegan KF, Hoth LR, Tan DH, Borzilleri KA, Withka JM, Boyd JG. Cyclization of N-terminal S-carbamoylmethylcysteine causing loss of 17 Da from peptides and extra peaks in peptide maps. J Proteome Res. 2002; 1(2):181–7. [PubMed: 12643538]
24. Li C, Thompson CB. Cancer. DNA damage, deamidation, and death. Science. 2002; 298(5597): 1346–7. [PubMed: 12434041]
25. Kameoka D, Ueda T, Imoto T. A method for the detection of asparagine deamidation and aspartate isomerization of proteins by MALDI/TOF-mass spectrometry using endoproteinase Asp-N. J Biochem. 2003; 134(1):129–35. [PubMed: 12944379]
26. Choudhary C, Kumar C, Gnad F, Nielsen ML, Rehman M, Walther TC, Olsen JV, Mann M. Lysine acetylation targets protein complexes and co-regulates major cellular functions. Science. 2009; 325(5942):834–40. [PubMed: 19608861]
27. Danielsen JM, Sylvestersen KB, Bekker-Jensen S, Szklarczyk D, Poulsen JW, Horn H, Jensen LJ, Mailand N, Nielsen ML. Mass spectrometric analysis of lysine ubiquitylation reveals promiscuity at site level. Mol Cell Proteomics. 2011; 10(3):M110 003590. [PubMed: 21139048]
28. Wisniewski JR, Zougman A, Mann M. Nepsilon-formylation of lysine is a widespread post-translational modification of nuclear proteins occurring at residues involved in regulation of chromatin function. Nucleic Acids Res. 2008; 36(2):570–7. [PubMed: 18056081]
29. Ghesquiere B, Jonckheere V, Colaert N, Van Durme J, Timmerman E, Goethals M, Schymkowitz J, Rousseau F, Vandekerckhove J, Gevaert K. Redox proteomics of protein-bound methionine oxidation. Mol Cell Proteomics. 2011; 10(5):M110 006866. [PubMed: 21406390]
30. Ong SE, Mittler G, Mann M. Identifying and quantifying in vivo methylation sites by heavy methyl SILAC. Nat Methods. 2004; 1(2):119–26. [PubMed: 15782174]
31. Brame CJ, Moran MF, McBroom-Cerajewski LD. A mass spectrometry based method for distinguishing between symmetrically and asymmetrically dimethylated arginine residues. Rapid Commun Mass Spectrom. 2004; 18(8):877–81. [PubMed: 15095356]
32. Xu G, Paige JS, Jaffrey SR. Global analysis of lysine ubiquitination by ubiquitin remnant immunoaffinity profiling. Nat Biotechnol. 2010; 28(8):868–73. [PubMed: 20639865]

33. Wang D, Cotter RJ. Approach for determining protein ubiquitination sites by MALDI-TOF mass spectrometry. Anal Chem. 2005; 77(5):1458–66. [PubMed: 15732931]

34. Jiang T, Zhou X, Taghizadeh K, Dong M, Dedon PC. N-formylation of lysine in histone proteins as a secondary modification arising from oxidative DNA damage. Proc Natl Acad Sci U S A. 2007; 104(1):60–5. [PubMed: 17190813]

35. Zachara NE, Hart GW, Cole RN, Gao Y. Detection and analysis of proteins modified by O-linked N-acetylglucosamine. Curr Protoc Mol Biol. 2002; Chapter 17(Unit 17):6. [PubMed: 18265305]

36. Zhang J, Chen Y, Zhang Z, Xing G, Wysocka J, Zhao Y. MS/MS/MS reveals false positive identification of histone serine methylation. J Proteome Res. 2010; 9(1):585–94. [PubMed: 19877717]

37. Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol. 2006; 24(10): 1285–92. [PubMed: 16964243]

38. Halligan BD, Geiger JF, Vallejos AK, Greene AS, Twigger SN. Low cost, scalable proteomics data analysis using Amazon's cloud computing services and open source search algorithms. J Proteome Res. 2009; 8(6):3148–53. [PubMed: 19358578]

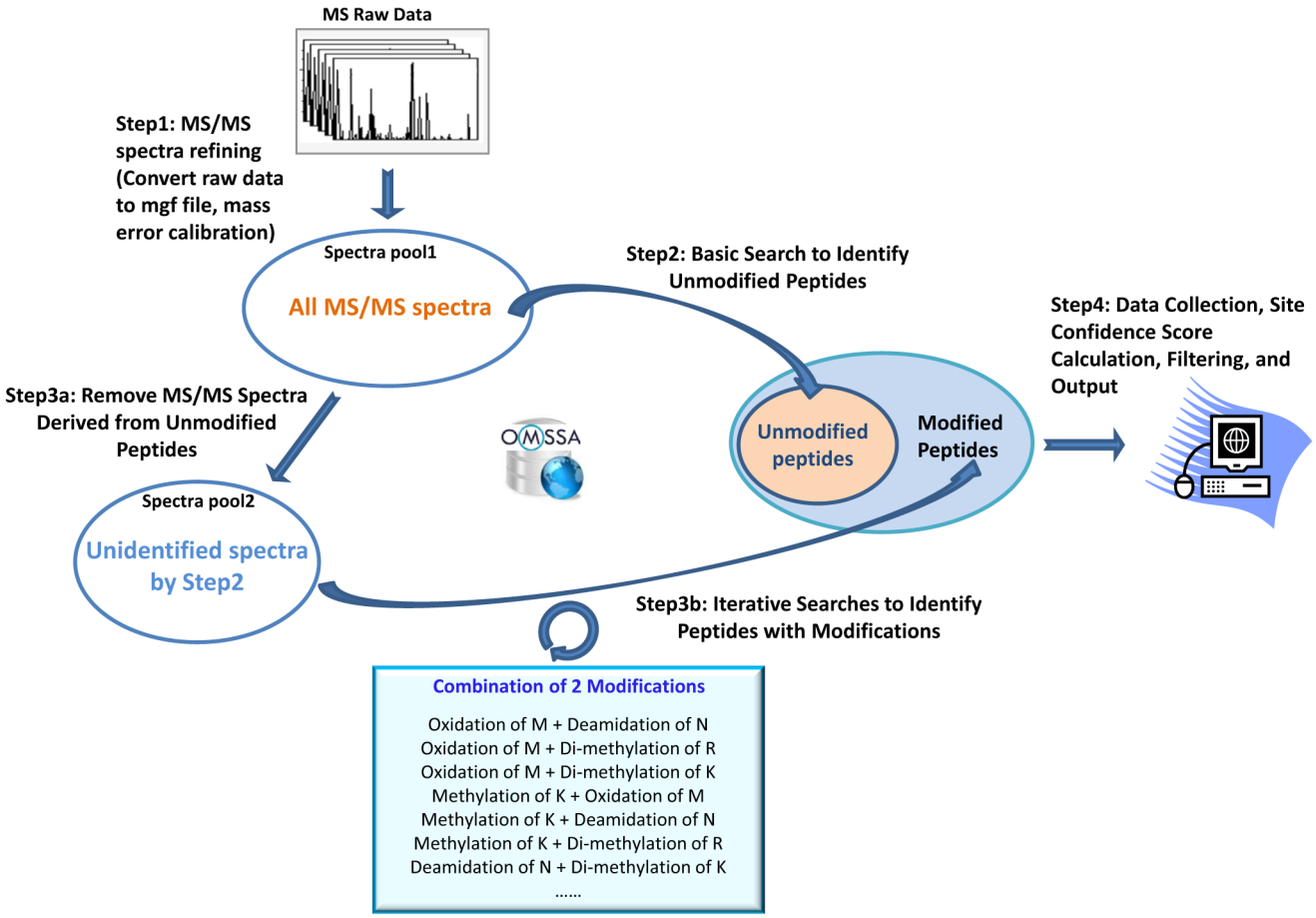

Figure 1.

Work flow of the ISPTM approach. Four steps are involved: 1) MS/MS spectra refining, 2) basic search, 3) iterative searches, and 4) data collection, filtering, and output. All steps are automated by Python scripts. The OMSSA search engine is used for all database searches.

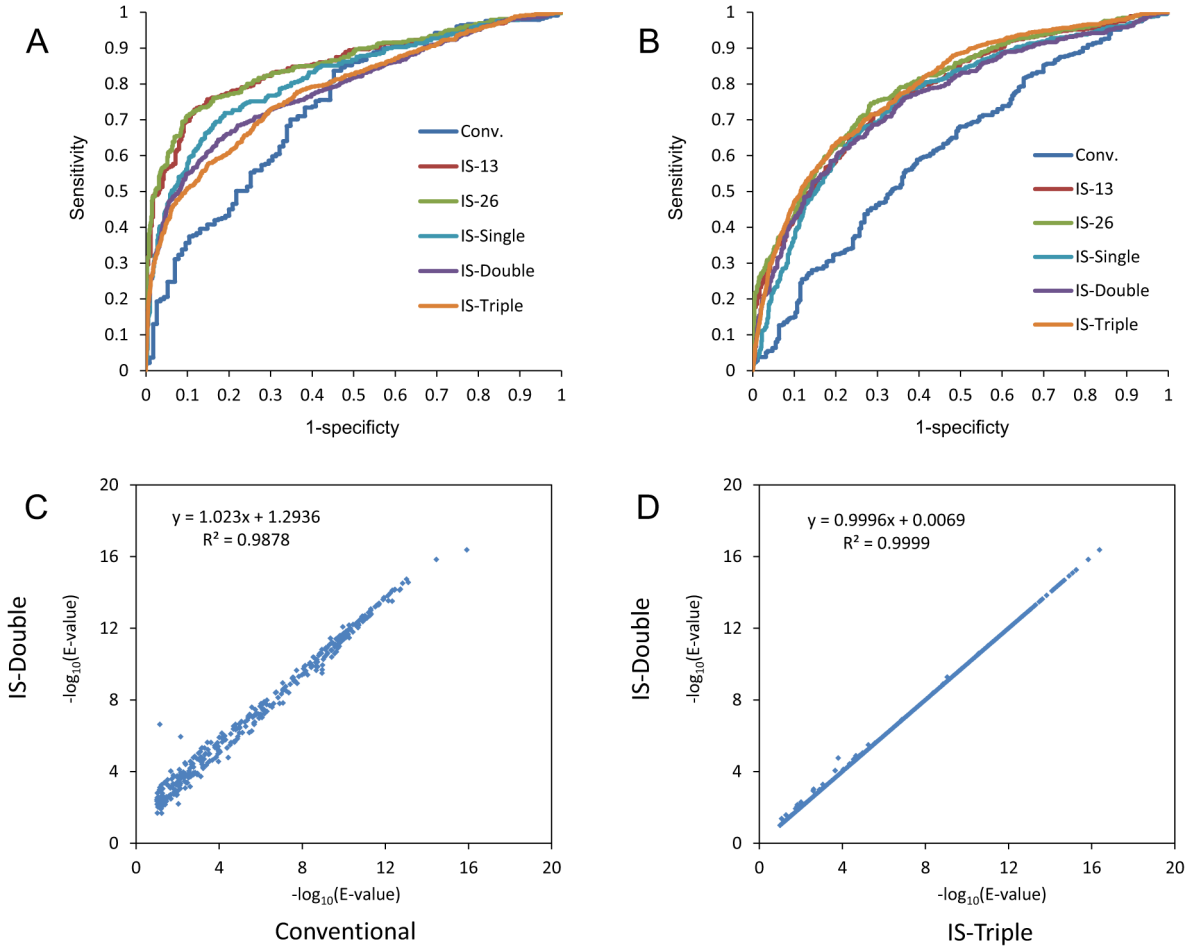

Figure 2.

Comparison of the synthetic peptides data identified by conventional search and ISPTM. (A–B) Receiver operating characteristic (ROC) plots illustrate the discriminating power of different search strategies for (A) sample #1 and (B) sample #2. Search strategies include the conventional (Conv.) all-in-one search using 13 variable modifications, and variations on ISPTM strategy: iteratively testing 13 authentic modifications used two-at-time (IS-13), testing 13 authentic modifications plus 13 false modifications used two-at-time (IS-26), testing all 207 modifications provided by OMSSA used one-at-time (IS-Single), testing all 207 modifications provided by OMSSA used two-at-time (IS-Double), and testing all 207 modifications provided by OMSSA used three-at-time (IS-Triple). (C–D) OMSSA E-value comparison of the same identification results using different search strategies. For the same spectrum with same OMSSA identification results in Sample #1, the PSM score by $-\log_{10}$(E-value) from (C) the conventional all-in-one search and IS-Double search strategies, and (D) from the IS-Double and IS-Triple search strategies were plotted. Linear regression was applied for each plot and the regression equation and R^2 are indicated in each plot.

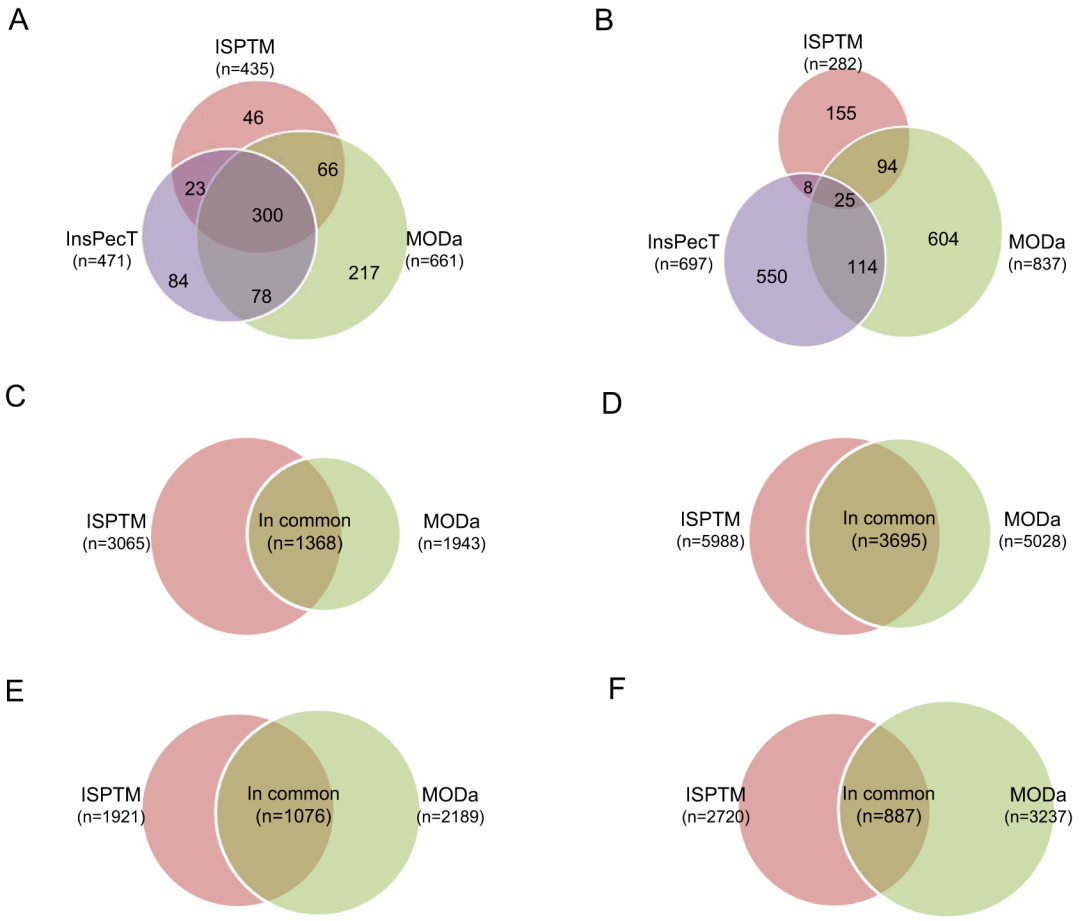

Figure 3.

Comparison of the ISB data analyzed by ISPTM, MODa and InsPecT approaches, as well as the complex nuclear matrix data analyzed by ISPTM and MODa. (A–B) Different coverages of the (A) unique peptides (peptides with identical AA sequence but different modifications) and the (B) modified peptides (peptides with identical AA sequence and identical modification) identified by three programs for the ISB data. (C–F) Different coverages of the (C) unique proteins, the (D) unique peptides, the (E) unique modified peptides (modified peptides with identical AA sequence but different modifications), and the (F) modified peptides (modified peptides with identical AA sequence and identical modification) by ISPTM and MODa for the nuclear matrix data.

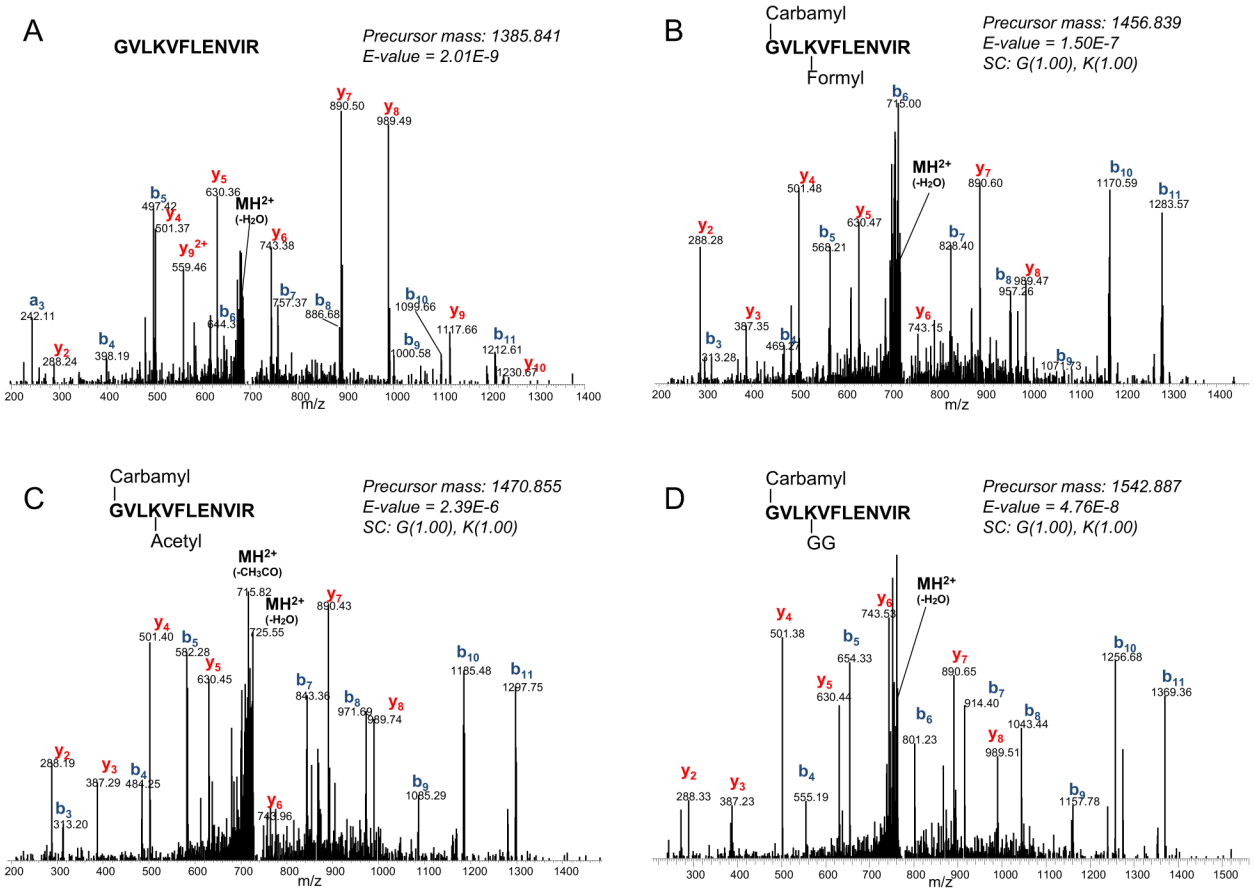

Figure 4.

The MS/MS spectra of peptide “GVLKVFLENVIR” derived from Histone H4 (IPI00407339) with a variety of PTMs on lysine 60 (H4K60): (A) Unmodified peptide, (B) formylated lysine, (C) acetylated lysine, and (D) ubiquitinylated lysine. All of these peptides were carbamylated on the N-terminus. Mass and charge of the precursor, OMSSA E-value, and the modification site confidence score are indicated in each MS/MS spectrum.

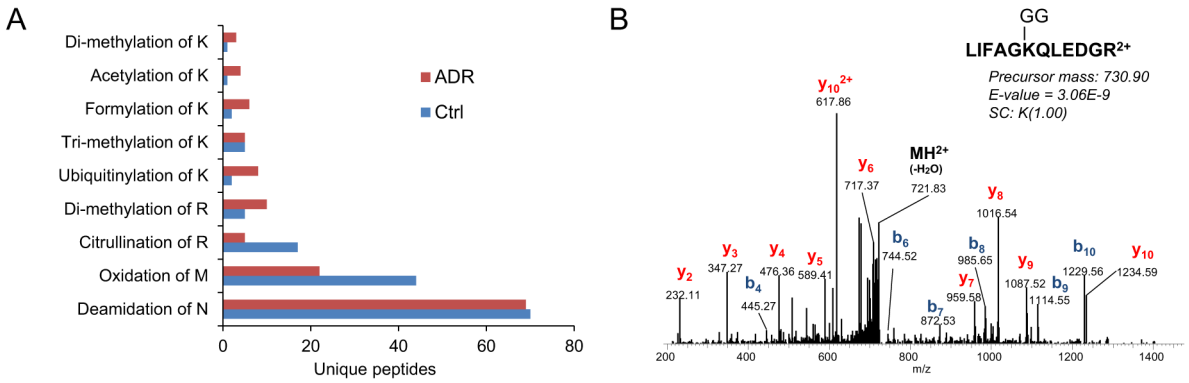

Figure 5.

Changes in PTM on the nuclear matrix proteins from mouse pro-B cells after DNA damage, analyzed by ISPTM. (A) A bar graph showing the number of unique peptides found for a representative set of PTMs that were identified in the Control (Ctrl) and ADR treated samples. (B) The MS/MS spectrum of peptide “LIFAGK(GG)QLEDGR” which was derived from a protein with K48 poly-ubiquitination, that appeared after DNA damage. Mass and charge of the precursor, OMSSA E-value, and the modification site confidence score are indicated in each MS/MS spectrum.

Table 1

The frequent modifications of ISB data analyzed by InsPecT, MODa, and ISPTM.

Residues	Mass Shift	Spectra	Annotations
Modifications by InsPecT:			
Nonspecific	22	294	Sodium
Nonspecific	38	151	Potassium
M	16	92	Oxidation
C	209	85	Carbamidomethylation by DTT
N or protein N-terminus	−17	83	Amonia loss
Protein N-terminus	42	44	Acetylation
S	42	44	Acetylation
M	15	31	Unknown
A	26	29	A->P substitution
N-terminus	128	25	Unknown
Modifications by MODa:			
Nonspecific	22	558	Sodium
N	1	514	Deamidation
Nonspecific	38	279	Potassium
M	16	273	Oxidation
C	152	163	Carbamidomethylation
N-terminus	−17	83	Amonia loss
R	−43	51	Arg->Leu/Ile substitution
A, T	26	65	A: ->P substitution; T: unknown
D, E	−1	62	Amidation
S, T	−18	25	Beta elimination
Modifications by ISPTM:			
N	1	1070	Deamidation
M	16	537	Oxidation
N-terminus	42	340	Acetylation of protein N-terminal
N-terminal Q	−17	150	Pyro-glu modification of N-terminal Q
C	−17	91	Pyro-CamC
W	16	57	Oxidation
Q	1	48	Deamidation
S, T	−18	46	Beta elimination
D	14	17	Methylation
P	16	14	Hydroxylation

Table 2

Proteins that were dimethylated on arginine identified from the control and Adriamycin-treated NM proteins

Protein ID	Gene Name	Protein Name	[a] Peptide	Site	[b] −log (E-value)	ADR #SC[c]	Ctrl #SC	[d] Combinatory PTMs	Ref
IPI00109813	Hnrnpa0	Heterogeneous nuclear ribonucleoprotein A0	SNSGPYR*GGYGGGYGGGSF	[h] 7	6.89	1	4		Ref [30]
IPI00817004	Hnrnpa1	Heterogeneous nuclear ribonucleoprotein A1.	SGSGNFGGGR*GGGFGGNDNFGR	10	7.50	4	3	Hydroxylation of N:5 (1), deamidation of N: 17 (1)	Ref [30]
IPI00331552	Pabpc1	Pabpc1 protein	VANTSTQTMGPR*PAAAAAAATPAVR	[i] 12	10.66	1	4	Oxidation of M:9 (1)	[e] UniProt
IPI00742310	Ewsr1	RNA-binding protein EWS	R*GGPGGPPGPLMEQMGGR	[j] 1	11.36	5	0	Oxidation of M:12 (1)	UniProt
IPI00875791	Dhx9	ATP-dependent RNA helicase A Small nuclear	R*GYGGGYFGQGR*GGGGGGGY VPLAGAAGGPGIGR*AAGR	1 & 12	11.36	2	1		
IPI00114052	Snrpb	ribonucleoprotein-associated protein B	*GIPAGVPMPQAPAGLAGPVR	14 & 18	6.29	1	0		UniProt
IPI00759858	Rbm33	RNA-binding protein 33	DPFFLGVSGEPR*FPSHLFLEQR	[k] 12	9.20	1	0		
IPI00224729	Hnrnph1	Heterogeneous nuclear ribonucleoprotein H1	R*GAYGGGYGGYDDYNGYNDGYGFGSDR	1	17.74	1	0	Deamidation of N:15 (1)	UniProt
IPI00458583	Hnrnpu	Heterogeneous nuclear ribonucleoprotein U	R*GNMPQR*GGGGGSGGIGYPYPR	1 & 7	5.55	1	0		Ref [30]

[a] Peptides with dimethylation site labeled with an asterisk "*".

[b] E-value reflects the quality of the peptide spectrum matching. An average $-\log_{10}$(E-value) was calculated for peptides with multiple identified spectra.

[c] SC stands for Spectra count which indicates the number of identified peptides in the control (Ctrl) or Adriamycin (ADR) treated groups.

[d] If the identified peptide carried other modifications, the modification name, site and spectra count (in parenthesis) are indicated.

[e] Label "UniProt" indicates this site was reported in the UniProt database (www.uniprot.org).

[h] Supplemental Figure 5A

[i] Supplemental Figure 5B

[j] Supplemental Figure 5C

[k] Supplemental Figure 5D